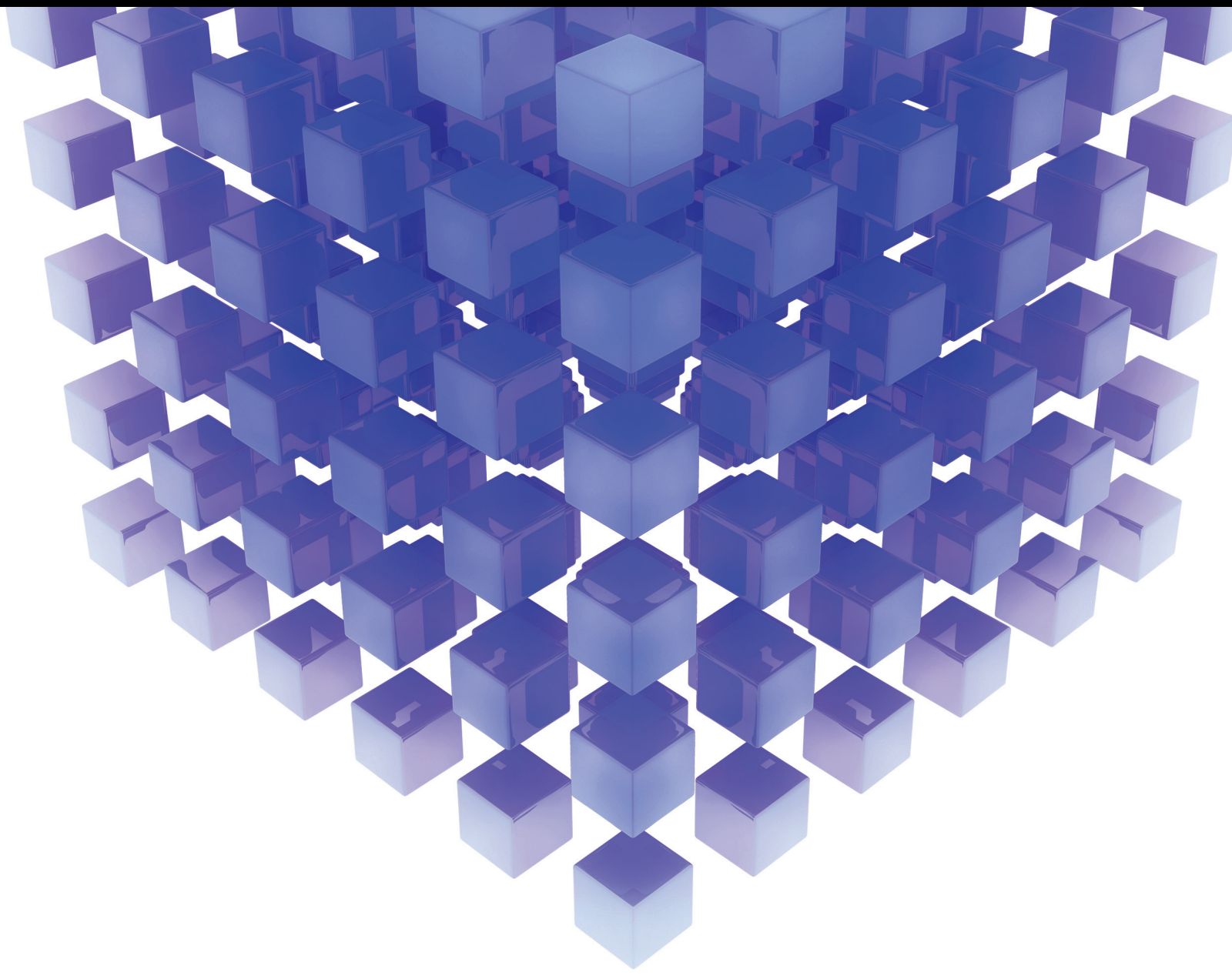
Mathematical Problems in Engineering

Advanced Modeling and Services Based Mathematics for Ubiquitous Computing

Guest Editors: Jong Hyuk Park, Hong Shen, Jian-nong Cao, Fatos Xhafa, and Young-Sik Jeong



Advanced Modeling and Services Based Mathematics for Ubiquitous Computing

Mathematical Problems in Engineering

Advanced Modeling and Services Based Mathematics for Ubiquitous Computing

Guest Editors: Jong Hyuk Park, Hong Shen, Jian-nong Cao, Fatos Xhafa, and Young-Sik Jeong

Editorial Board

Contents

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 745472, 3 pages
http://dx.doi.org/10.1155/2015/745472

Editorial

Advanced Modeling and Services Based Mathematics for Ubiquitous Computing

Jong Hyuk Park,[1] Hong Shen,[2] Jian-nong Cao,[3] Fatos Xhafa,[4] and Young-Sik Jeong[5]

[1]*Department of Computer Science & Engineering, Seoul National University of Science & Technology, Seoul 139-743, Republic of Korea*
[2]*School of Computer Science, University of Adelaide, Adelaide, SA 5000, Australia*
[3]*Department of Computing, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong*
[4]*Department of Computer Science, Technical University of Catalonia, 08034 Barcelona, Spain*
[5]*Department of Multimedia Engineering, Dongguk University, Seoul 100-715, Republic of Korea*

Correspondence should be addressed to Young-Sik Jeong; ysjeong.dk@gmail.com

Received 8 June 2015; Accepted 8 June 2015

1. Introduction

Recent advancements on Ubiquitous Computing have been a great challenge to computer science and engineering. Ubiquitous Computing systems manage huge numbers of heterogeneous mobile devices which continuously connect real-world objects, and most data are automatically generated through wireless communication environments [1–3]. Ubiquitous Computing frameworks might help support the interaction between "humans and objects" and allow for more complex structures like intelligent computing and application development. Many Ubiquitous Computing frameworks seem to focus on real time data logging solutions which offer some basis to work with many "humans and objects." Future developments might lead to specific software development environments to create the software and to work with the hardware used in the Ubiquitous Computing [1–5].

This special issue aims to provide an advanced theory and application for helping researchers to conduct new research and reviewing articles that present latest and practical findings that can contribute to the future evolutions of mathematics in Ubiquitous Computing applications.

Among many manuscripts we have received, only high quality manuscripts were finally selected for this special issue. Each selected manuscript was blindly reviewed by at least three reviewers consisting of guest editors and external reviewers. We present a brief overview of each manuscript in the following.

2. Related Works

Recent advancements in advanced modeling and services based mathematics for Ubiquitous Computing have created new research topics including (1) Mathematical and numerical modeling for Ubiquitous Computing, (2) optimization methods, mathematics modeling, and services for Ubiquitous Computing, (3) numerical analysis for security and emergencies for Ubiquitous Computing, (4) methods for improving efficiency or accuracy of M2M applications, (5) vehicle autodiagnosis for Ubiquitous Computing, (6) computational models of communication mechanisms for Ubiquitous Computing, (7) adaptive and dynamic algorithms for Ubiquitous Computing, (8) applied cryptography and security issues for Ubiquitous Computing, (9) advanced modeling and services for IoT (Internet of Things) applications, and (10) ubiquitous sensor networks and RFID for Ubiquitous Computing.

In these several topics of this special issue, some articles proposed the following findings.

N. Javaid et al. proposed the routing model overhead produced by three reactive protocols; AODV, DSR, and DYMO. They choose the three routing protocols because these are widely used in literature. Their main focus was to measure

routing overhead for LL and NL feedback mechanisms. To analyze the link sensing mechanisms of AODV, DSR, and DYMO, they conducted simulations in NS-2. The overhead was measured for nodes' different mobility and densities.

M. Aazam and E.-N. Huh proposed a resource management model, keeping in view different types of services, different customer types, customer characteristic, pricing, and refunding. The presented model was implemented and evaluated using CloudSim 3.0.3 toolkit. Their results and discussion validated the model and its efficiency.

S. Lee et al. proposed a NUI-specific SDK, called "Gesture SDK," for the development of NUI-based applications. Gesture SDK provides a gesture generator with which developers can directly define gestures. Further, a "Gesture Recognition Component" was provided, which enables defined gestures to be recognized by applications. They generated gestures using the proposed SDK and developed a "Smart Interior," NUI-based application using the Gesture Recognition Component. The results indicated that the recognition rate of the generated gestures was 96% on average.

S. Liu and X. Meng constructed a Region-Based Location Graph (RLG), which can combine with user short-ranged mobility formed by daily activity and long-distance mobility formed by social network tie. They can sequentially recommend local business information and long-distance business information to users. Moreover, it can combine user-based collaborative filtering with item-based collaborative filtering and successfully generate recommendation for cold start users. Consequently it can alleviate cold start problem which traditional recommender systems often suffer from. The experiments on real dataset confirmed the effectiveness of the proposed method compared to other ones.

H. Kim et al. implemented a mobile diagnosing system that provides user-centered interfaces for more precisely estimating and diagnosing engine conditions through communications with the self-developed ECU only for industrial CRDI engine use. For the implemented system, a new protocol was designed and applied based on OBD-II standard to receive engine data values of the developed ECU.

I. Y. Jung et al. proposed a security framework for health information management of the self-organizing software platform (SoSp). The proposed framework was designed to ensure easy detection of identification information for typical users. In addition, it provides powerful protection for the user's health information.

H. Kim et al. proposed a new framework for generating 3D city models that satisfy both the visual and physical requirements for ground-oriented virtual reality applications. To ensure its usability, the framework must be cost-effective and allow for automated creation. To achieve these goals, they leveraged a mobile mapping system that automatically collects high-resolution images and supplements sensor information such as the position and direction of the captured images. To resolve problems stemming from sensor noise and occlusions, they developed a fusion technique to incorporate digital map data. This paper described the major processes of the overall framework and the proposed techniques for each step and presented experimental results from a comparison with an existing 3D city model.

G. Yi and N. Y. Yen designed a tool for visualizing social network data from the famous social networking website Facebook. It also proposed a new interaction and navigation technique that uses Kinect to explore and interact with social networks [6]. The tool as well as the new method of interaction should help users in interacting and exploring their social networks. They would also help learners experience better learning interaction.

A. Mahmud et al. proposed a medium access control protocol that can work in the absence of a CCC and reduce the possible overhead to a greater extent. In their proposed protocol, CR users took advantage of similar spectrum availability in their neighborhood for resource utilization. They also proposed a contention-based spectrum allocation mechanism that works in a distributed manner over different available channels. Simulation results showed that this approach can reduce broadcast overhead significantly while maintaining connectivity success similar to its counterparts.

J. Sim et al. examined the influence of dataset characteristics and patterns of missing data on the performance of classification algorithms using various datasets. The moderating effects of different imputation methods, classification algorithms, and data characteristics on performance were also analyzed. The results were important because they could suggest which imputation method or classification algorithm to use depending on the data conditions. The goal was to improve the performance, accuracy, and time required for Ubiquitous Computing.

K.-C. Lin et al. proposed a feature selection model combining the modified AFSA (MAFSA) with SVM. MAFSA was used to simulate the mechanism underlying the endocrine system in order to create a different search space for every individual fish in order to enhance the efficiency with which optimal solutions are derived.

G. Yi conducted clustering under the assumption that the functional classifier inside the cluster had similar functions and utilized the features extracted from the inside of the cluster as the learning data. When finding protein whose function is unknown, the model that predicts GO (or the controlled vocabulary) was defined through the learning and learned data documents of those proteins whose function was already defined. This was the existing functional prediction, which is the method to harmonize appropriately those frequently used methods such as sequence similarity, protein-interaction, and context-free ones; thus, it could increase the prediction probability of GO.

J.-R. Chang and B. Chang developed a model to investigate the tourists' preference. Ten attributes of tourist destinations were used in this study. Fuzzy set theory was adopted as the main analysis method to find the tourists' preference. In their research, 248 types of data were used. Besides the evaluations for the factors, the overall evaluations (namely, satisfied, neutral, and dissatisfied) for every tourism destination were also inquired. After screening, 201 types of these data were usable. Among these 201 types of data, 141 (70.15%) were classified into "satisfied" with the tourism destination, 49 (24.38%) were "neutral," and 11 (5.47%) were

"dissatisfied." Eight rules were obtained with the method of fuzzy preprocess. Regarding the condition attributes, three of the original 10 attributes were found influential, namely, level of prices, living costs, information, and tourist services as well as tourist safety of the tourism destinations. These study results showed that top management of tourism destinations should put resources in these fields first, in order to allow limited resources to perform to their maximum effectiveness.

Y. Zhang et al. proposed a new efficient and provable secure digital rights management authentication scheme using smart card based on elliptic curve cryptography. To demonstrate the scheme is provable secure, they introduced a security model AFP05 and analyzed the scheme in this model. In the following, they gave the proof that the proposed scheme was secure in the AFP05 model. As known to all, one-way hash function is more efficient than the operation of scalar multiplication and pairings. Moreover, the pairing operation costs much more than the scalar multiplication operation. The effort of evaluating one pairing operation is approximately three times the effort of evaluating one scalar multiplication operation. So, they cut down some pairings operation of point on elliptic curve and used hash function instead to increase the scheme's efficiency.

J. Uddin et al. proposed a method for the reliable fault detection and classification of induction motors using two-dimensional (2D) texture features and a multiclass support vector machine (MCSVM). The proposed model first converts time-domain vibration signals to 2D gray images and then utilizes the global neighborhood structure (GNS) map to extract texture features of the converted gray images. GNS maps were calculated by averaging the local neighborhood structure (LNS) maps of central pixels. The principle component analysis (PCA) is then used to select the most significant feature dimensions.

H. Cho and B. Kim suggested Improved Cooperative Collision Warning System (ICCWS) that considers the curvature of the road and is based on intervehicle communication. To predict the radius of curvature of the road, the Arc Relative Distance (ARD), the real relative distance to a preceding vehicle on a curved road has been used. The risk of collision with the preceding vehicle was decided by calculating an index of the risk of collision on a curved road using the computed ARD. The effect of ICCWS proposed through this simulation has been reviewed, and the improvement in performance in following a preceding vehicle has been analyzed quantitatively via comparative analysis with the conventional forward collision warning system.

W.-S. Lee et al. proposed a subsurface scattering-based object rendering technique that was optimized for smartphone games. They employed a subsurface scattering method that is utilized for a real time smartphone game. Their example game was designed to validate how the proposed method can be operated seamlessly in real time. Finally, they showed the comparison results between bidirectional reflectance distribution function, bidirectional scattering distribution function, and their proposed subsurface scattering method on a smartphone game.

Finally S. Nam and I. Kim presented an efficient spatial reasoning algorithm working on a mixture of directional and topological relations between spatial entities and then explained the implementation of a spatial reasoner based on the proposed algorithm. Their algorithm not only has the checking function for path-consistency within each directional or topological relation set, but also provides the checking function for cross-consistency between them. This paper also presented an application system developed to demonstrate the applicability of the spatial reasoner and then introduced the results of the experiment carried out to evaluate the performance of the spatial reasoner.

Acknowledgment

We would like to thank all authors for their contributions to this special issue. We also extend our thanks to the external reviewers for their excellent help in reviewing the manuscripts.

Jong Hyuk Park
Hong Shen
Jian-nong Cao
Fatos Xhafa
Young-Sik Jeong

References

[1] M. Friedewald and O. Raabe, "Ubiquitous computing: an overview of technology impacts," *Telematics and Informatics*, vol. 28, no. 2, pp. 55–65, 2011.

[2] Y.-S. Jeong, N. Chilamkurti, and L. J. García Villalba, "Advanced technologies and communication solutions for internet of things," *International Journal of Distributed Sensor Networks*, vol. 2014, Article ID 896760, 3 pages, 2014.

[3] Gartner, *Gartner's Hype Cycle Special Report for 2011*, Gartner, 2012, http://www.gartner.com/technology/research/hype-cycles/.

[4] H. Ning and Z. Wang, "Future internet of things architecture: like mankind neural system or social organization framework?" *IEEE Communications Letters*, vol. 15, no. 4, pp. 461–463, 2011.

[5] Y.-S. Jeong and J. H. Park, "High availability and efficient energy consumption for cloud computing service with grid infrastructure," *Computers and Electrical Engineering*, vol. 39, no. 1, pp. 15–23, 2013.

[6] R. Francese, I. Passero, and G. Tortora, "Wiimote and Kinect: gestural user interfaces add a natural third dimension to HCI," in *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12)*, pp. 116–123, May 2012.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 604108, 9 pages
http://dx.doi.org/10.1155/2015/604108

Research Article

Feature Selection and Parameter Optimization of Support Vector Machines Based on Modified Artificial Fish Swarm Algorithms

Kuan-Cheng Lin,[1] Sih-Yang Chen,[1] and Jason C. Hung[2]

[1] *Department of Management Information Systems, National Chung Hsing University, Taichung 40227, Taiwan*
[2] *Department of Information Technology, Overseas Chinese University, Taichung 40721, Taiwan*

Correspondence should be addressed to Kuan-Cheng Lin; kuanchenglin@gmail.com

Received 21 August 2014; Accepted 22 September 2014

Academic Editor: Jong-Hyuk Park

Rapid advances in information and communication technology have made ubiquitous computing and the Internet of Things popular and practicable. These applications create enormous volumes of data, which are available for analysis and classification as an aid to decision-making. Among the classification methods used to deal with big data, feature selection has proven particularly effective. One common approach involves searching through a subset of the features that are the most relevant to the topic or represent the most accurate description of the dataset. Unfortunately, searching through this kind of subset is a combinatorial problem that can be very time consuming. Meaheuristic algorithms are commonly used to facilitate the selection of features. The artificial fish swarm algorithm (AFSA) employs the intelligence underlying fish swarming behavior as a means to overcome optimization of combinatorial problems. AFSA has proven highly successful in a diversity of applications; however, there remain shortcomings, such as the likelihood of falling into a local optimum and a lack of multiplicity. This study proposes a modified AFSA (MAFSA) to improve feature selection and parameter optimization for support vector machine classifiers. Experiment results demonstrate the superiority of MAFSA in classification accuracy using subsets with fewer features for given UCI datasets, compared to the original FASA.

1. Introduction

Advances in information and communications technology have led to a rapid increase in the data processing and computing power of handheld devices. This has made it possible to obtain information anytime and anywhere, ushering in the era of ubiquitous computing (ubicomp) and the Internet of Things (IoTs). Applications, such as photo sharing and social networking, create enormous volumes of digital data, which is available to aid in decision-making. Feature selection is particularly effective in the classification of big data in fields such as data mining, pattern recognition [1], bioinformation [2], arrhythmia classification [3], and numerous others.

Supervised learning methods, such as decision trees [4], support vector machines (SVM) [5–7], and neural networks [8–11], are used to classify data into appropriate categories. The classification of data requires the classification of data we already know and then divide them into training data and testing data. After building a classification model using the training data, the testing data is used to evaluate the model according to classification accuracy. SVM is based on the statistical theories proposed by Chervonenks [12] and the principle of minimizing structural risk. SVM is commonly used to solve classification or regression problems by finding the optimal hyperplane. SVM provides excellent classification accuracy using a small training set and is easy to implement. This study adopted SVM as a classifier in conjunction with the wrapper method of feature selection.

Feature selection is used to filter out large amounts of unnecessary data, much of which is irrelevant, redundant, and/or noisy. Irrelevant features are unrelated to a given goal and redundant features represent the same information,

despite containing different types of data. Noisy features contain wrong or missing data. Sifting through this unnecessary data would result in enormous computing costs or even skew the results. Feature selection is used to identify the features that are essential to a given task.

Two methods are commonly used for feature selection: filter and wrapper. The filter approach is based on assigning weights to every feature, such as distance or dependability, and then combining the features with the highest weights in order to obtain an optimal subset. The wrapper involves collocating using metaheuristic algorithm and then assembling an optimized subset of features by eliminating or combining features and calculating a fitness value for each feature subset on the basis of classification accuracy. The filter approach tends to be quicker but suffers from lower classification accuracy. The need for the classifier to conduct training extends the processing of the wrapper approach; however, the classification accuracy is far higher.

One common wrapper method involves compiling a subset of optimal features of the highest relevance with the most accurate description of the characteristics in the dataset. This can be very time consuming due to the fact that any increase in the number of features exponentially expands the number of combinations in the feature subset, which is known as the curse of dimensionality [13]. This study used a metaheuristic algorithm to obtain good-enough or near-optimal feature subsets within a reasonable amount of time. By reducing much of the unnecessary data, feature selection processes can enhance classification accuracy and reduce processing time.

Metaheuristic algorithms are widely used to solve problems of optimization, such as schedule management [14, 15], function optimization [16], and intrusion detection [17]. Metaheuristic algorithms combine random search functions with empirical rules, and many of these methods have been inspired by mechanisms found in nature, such as genetic algorithms (GA) [18–20], based on gene mutation and mating, and particle swarm optimization (PSO) [21], based on the movements of flocks of birds. In 2002, the artificial fish swarm algorithm (AFSA) [22] was proposed to solve problems of optimality by simulating the movement of schools of fish and the intelligence underlying these behaviors. Numerous studies have demonstrated the efficacy of AFSA [14, 17, 23]; however, a number of shortcomings must still be addressed. In [24], various defects were pointed out, such as the likelihood of falling into a local optimum and a lack of multiplicity.

This study proposed a feature selection model combining the modified AFSA (MAFSA) with SVM. MAFSA is used to simulate the mechanism underlying the endocrine system in order to create a different search space for every individual fish in order to enhance the efficiency with which optimal solutions are derived.

Section 2 introduces SVM and the MAFSA. Section 3 outlines the proposed method based on a combination of these principles. Section 4 describes our experiments and results and in Section 5, we draw conclusions and describe our future work.

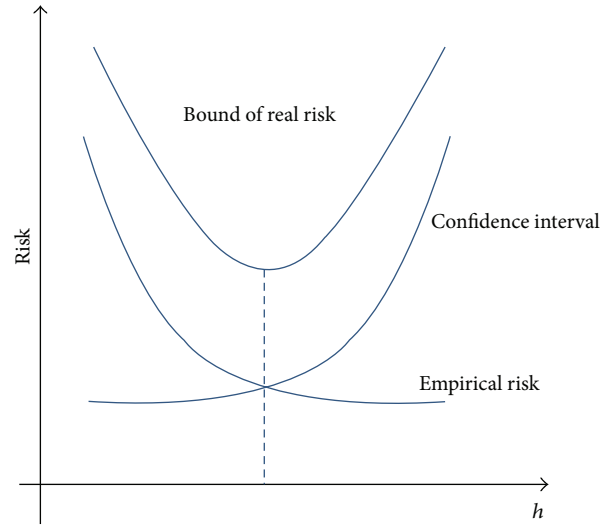


Figure 1: Connection between confidence interval and empirical risk.

2. Background

2.1. Support Vector Machine. Since its introduction by Vapnik in 1995, [5], SVM has become a very popular classifier due to its ease of use and high classification accuracy even when using small training sets. This method of supervised learning is based on the Vapnik-Chervonenkis (VC) dimension and structural risk minimization theory [12, 25]. The VC dimension is used for function sets with only two values: 0 and 1. The function set has N samples and regardless of how the position of the function samples is changed, the dimension can be separated from the samples. For example, with h as the maximum number of samples, an increase in h will lead to an increase in N. As shown in (1), the upper bound of generalization error (R) is the sum of training error (R_{Emp}) and confidence interval (CI):

$$R \leq R_{\text{Emp}} + \Phi\left(\frac{h}{N}\right), \tag{1}$$

where R represents the generalization error (also called testing error), R_{Emp} represents training error, and Φ represents the CI. An increase in N leads to a reduction in the CI and an increase in the VC dimension leads to an increase in h. When the VC dimension increases, the differences between testing error and training error also enlarge. Thus, reducing the complexity of the classification model and alleviating testing error require that we minimize training error as well as VC dimensions.

The value for N in the samples is influenced by VC dimension, such that an increase in empirical risk reduced the CI. Conversely, a reduction in empirical risk increases the CI without affecting the total risk. Thus, we must consider empirical risk as well as confidence interval and the tradeoff between them in order to minimize the total risk. Figure 1 [12] illustrates the connection between the CI and empirical risk.

The primary function in SVM involves finding the optimal hyperplane and using it for the classification of data. As

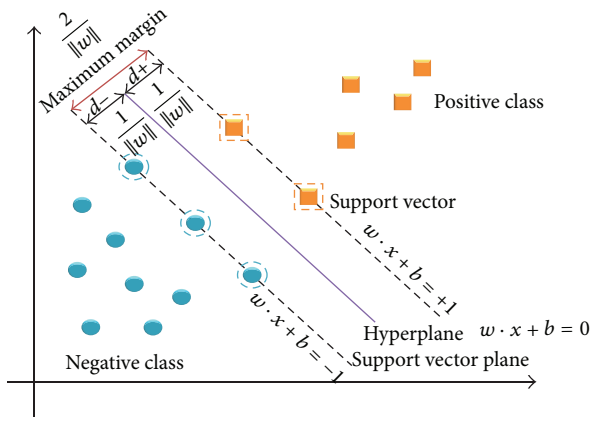


Figure 2: Optimal hyperplane.

shown in Figure 2, the optimal has maximal margin to those two classes. In Figure 2, the round and square points in the maximal margin line are the support vectors. This illustrates why the optimal hyperplane (with maximal margin) is able to achieve the highest classification accuracy.

The optimal hyperplane is able to classify only two classes; however, in real world situations, classification problems nearly always involve more than this. Thus, we need to use kernel function (Φ) in order to map the data into a higher VC dimension plane. Three common kernel functions can be used for different situations: radial basis functions (RBFs), polynomials, and sigmoids, as shown in formulas (2), (3), and (4), respectively. The RBF kernel function provides the best performance and versatility of these three methods; therefore, we adopted RBFs as the SVM kernel function.

RBF kernel is

$$\Phi\left(x_i - x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|\right). \tag{2}$$

Polynomial kernel is

$$\Phi\left(x_i - x_j\right) = \left(1 + x_i \cdot x_j\right). \tag{3}$$

Sigmoid kernel is

$$\Phi\left(x_i - x_j\right) = \tanh\left(kx_i \cdot x_j - \delta\right). \tag{4}$$

In [25], the authors introduce the soft margin method to process the mislabel examples located within the margin. The soft margin suggests the slack variables which measure the misclassification degree of the samples. And, the penalty parameter, C, is used to weight the misclassification degree. It was shown that setting the right values of penalty parameter and RBF kernel parameter for SVM could greatly enhance the effectiveness of classification in [26].

2.2. Artificial Fish Swarm Algorithm (AFSA). AFSA is a meta-heuristic algorithm combining the concept of random search and empirical rules. AFSA solves optimization problems by simulating the movement of schools of fish and the intelligence underlying these behaviors. There are three types

Table 1: Representation of feature sets.

C	G	F_1	$\cdots F_i \cdots$	F_n

Table 2: Parameters of AFSA.

Parameter name	Formula
Distance (F_i, F_j)	$\sum_{k=1}^{k} \left\|F_i(k) - F_j(k)\right\|$
Vision (F_i)	$\frac{\sum_{i=1}^{N}\sum_{k=1}^{N} \text{Distance}(F_i, F_k)}{\text{Total number of Fishes}}$
Neighbor (F_i)	$\{F_k \mid 0 < \text{distance}(F_i, F_k) \le \text{vision}\}$
Center (F_i)	$F_{\text{center}}(i) = \begin{cases} 0, & \sum_{k=1}^{k} F_k(i) < \frac{k}{2} \\ 1, & \sum_{k=1}^{k} F_k(i) \ge \frac{k}{2} \end{cases}$
Crowd degree (F_i)	$\frac{\text{Neighbors of } F_i}{\text{Total number of Fishes}}$

of AFSA: follow, swarm, and prey. AFSA repeatedly executes these three functions for every individual fish in order to find an optimal solution. Every fish represents a solution and a fitness number represents the value of each solution. Solving the problem of feature selection requires conversion into a form that can be dealt with by AFSA. Table 1 presents the form of feature sets for each fish.

In Table 1, C and G represent the parameters of RBF kernel functions in SVM, and F_1 to F_n represent the features. A feature value of 0 means that it is not selected, while a value of 1 means that the feature is selected.

The steps involved in feature selection for AFSA are presented as follows.

(1) Initialization. Randomly initialize the value of the feature subsets of each fish.

(2) Evaluate Fitness. Use the fitness function to evaluate the fitness value of the feature subset of every fish.

(3) Optimization Steps of Fish Swarm. Search for an optimal solution by executing the three search steps (follow, swarm, and prey) for each fish.

(4) Output the Optimal Feature Subset. If the terminal conditions of algorithm are satisfied, then stop the algorithm and output the optimal feature subset. Otherwise, initiate the algorithm to proceed through another iteration starting at Step (2).

The three optimal search steps of AFSA (follow, swarm, and prey) are presented in the following and the parameters are listed in Table 2.

(i) Follow. Executing follow for F_i would involve comparing its fitness value with that of the best fish in its vicinity (among its neighbors). If the best fitness value among the neighboring fish is better than that of the fish in question and the crowd degree of the neighbor does not exceed the maximum, then the feature subset and parameters are replaced with those of that best neighbor. As long as the replacement of subset and

parameters is successful, then the algorithm executes follow for next fish; otherwise, swarm is executed for F_i.

(ii) Swarm. In the event that the follow function fails for F_i, then the algorithm would proceed to swarm. At this point, the fitness value of F_i would be compared with that of the center of the neighboring fish, as shown in Table 2. If the fitness value of the center is better than that of the fish in question and the crowd degree of the center does not exceed the maximum crowd degree, then the feature subset and parameters are replaced with those of the center. As long as the replacement of subset and parameters is successful, then the algorithm executes follow for next fish; otherwise, prey is executed for F_i.

(iii) Prey. In the event that the swarm function fails for F_i, then the algorithm would proceed to prey. In this step, the algorithm makes random changes to the features of F_i in order to create new random fish. The maximum number of changes never exceeds the vision of the fish in question. If the fitness value of the random fish exceeds that of F_i, then feature subset and parameters are replaced with those of the random fish. As long as the replacement of subset and parameters is successful, then the algorithm executes follow for next fish; otherwise, the algorithm would continue searching for random fish until reaching the stipulated maximum number of tries.

The parameters of AFSA are defined as follows.

(i) Distance. The distance between F_i and F_j was calculated according to the formula presented in Table 2. For example, if both fish have the same dimension k and if the first feature of both fish was 0, then the distance between F_i and F_j would remain the same. In contrast, if the first feature of F_i was 0 and the first feature of F_j was 1, then the distance between the two fish would be increased by 1. The total distance between the two fish is equal to the sum of the differences of all k features.

(ii) Vision. The vision of every fish was calculated using the formula in Table 2. Vision also determines the maximum number of random feature changes that will be implemented when initiating *prey*.

(iii) Neighbor. The neighbors of F_i were determined using the formula in Table 2. Any fish F_j with a distance exceeding 0 but not exceeding the vision of F_i is deemed a neighbor of F_i.

(iv) Center. The center of F_i was calculated using the formula in Table 2. The center can be viewed as an individual fish. If, among more than half of the fish in the neighborhood of F_i, the first feature is 0, then the value of the center is set to 0. If, among more than half of these fish, the first feature is not 0, then the value of the center is set to 1.

(v) Crowd Degree. The crowd degree of F_i was calculated using the formula in Table 2. This parameter represents the density of fish in the vicinity of F_i.

(vi) Maximum Crowd Degree. In the execution of steps *follow* and *swarm*, we sought to prevent the agglomeration of all fish at the same point by designating that if the crowd degree of F_i exceeded the maximum, then no other fish would be permitted to approach this location. In other words, no other fish would be able to replace its feature subset using with that of F_i.

(vii) Maximum Number of Attempts. This is the maximum number of times that *prey* could be executed.

3. Modified Artificial Fish Swarm Algorithm

Researchers have revealed a number of shortcomings in the function of AFSA, such as the likelihood of falling into a local optimum and the lack of multiplicity. This study developed a modified artificial fish swarm algorithm with two fundamental changes: dynamic vision and improvement of the ability to search for the best fish swarm.

3.1. Dynamic Vision. The vision parameter plays a crucial role in the performance of AFSA, by determining the number of neighboring fish with which the target fish will interact, which largely determines the success of steps *follow* and *swarm*. Setting the vision parameter higher increases the likelihood of finding fish with higher fitness levels; however, this can lead to the centralization of the fish swarm to a particular location. This also tends to limit the diversity of species, which can easily fall into a local optimum. Setting the vision parameter lower reduces the number of neighboring fish (and thus the likelihood of finding a fish with higher fitness values) and causes the school to scatter while increasing the diversity of species. This increases the search space as well as the likelihood of finding an optimal solution; however, the time required to reach convergence is extended. Finding a reasonable balance with regard to the assignment of the vision parameter can be troublesome. To overcome this difficulty, this study developed a mechanism referred to as dynamic vision, in which each fish is assigned visions parameters suitable to its conditions. For example, individual fish with a lower fitness value require greater vision in order to find a solution quickly. Conversely, fish with a higher fitness value require a smaller vision parameter in order to enhance local searches.

We employed the endocrine-based formula, as outlined in [26], in order to provide dynamic vision. Individual fish obtain their own vision parameter values according to their fitness values. For example, if the fitness value is above the average, then the vision parameter is decreased, and vice versa. The endocrine-based formula is presented as follows:

$$\mathrm{EM}(i) = f_1\left(\frac{f_{\max} - f_i}{f_{\max} - f_{\mathrm{avg}}}\right) \cdot \left[\frac{\pi}{2} + f_2\left(f_i - \frac{f_{i-1} + f_{i+1}}{2}\right)\right], \tag{5}$$

$$\mathrm{Vision}(i) = \mathrm{Vision}(\mathrm{static}) \cdot \mathrm{EM}(i) \cdot \mathrm{CV}, \tag{6}$$

where EM(i) represents the endocrine of fish F_i, $f_{\max}$ represents the maximum fitness value of the fish in the school, and f_{avg} represents the average fitness value in the school. f_{i-1} represents the fitness values of fish f_{i-1}; f_{i+1} represents the fitness value of fish f_{i+1}. In order to adjust the range of

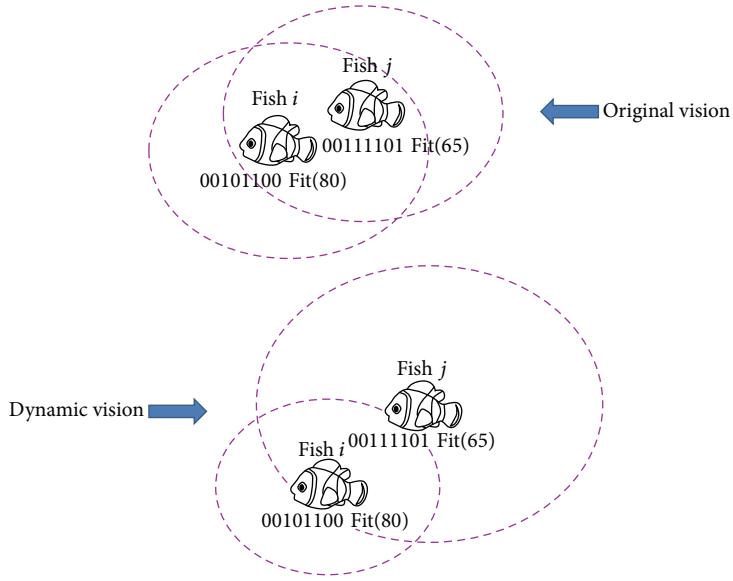


Figure 3: Original vision and dynamic vision.

the endocrine system, $f_1(x) = \text{atan}(x)$, $f_2(x) = \text{atan}(-x)$. In formula (6), vision(i) represents the original vision parameter value calculated as the average distance between each fish and CV is the adjustment constant. Figure 3 presents an example of dynamical vision. In the original vision parameter, fish i and j would have had the same vision, regardless of their fitness values. Following adjustment using the endocrine-based formula, the fitness value of fish i is improved to 80, which decreases its vision in order to enhance its ability to perform local searches. In contrast, the fitness value of fish j dropped to 65, which enhanced the vision of fish j, thereby enhancing global search capacity in order to obtain solutions more quickly.

3.2. Searching for the Best Fish Swarm. This study employed a simple method of searching for the best fish swarm to enhance local search ability and prevent falling into a local optimum. After each iteration has finished, the algorithm copies the fish with the best fitness value into a fish swarm and then executes a simple local search to enhance the possibility of finding an optimal solution. Four parameters were added to facilitate this change: best fish number (BFN), best fish search number (BSN), best fish mutation rate (BMR), and minimum change number (MCN).

BFN represents the number of fish copied. For example, setting BFN = 5 would cause the five best fish to be copied to the best fish swarm. BSN represents the local search number for each fish in the best fish swarm. Setting BSN = 10 would cause the algorithm to execute local ten searches for each fish in the best fish swarm. BMR represents the mutation rate of local searches and the MCN represents the minimum number of feature changes in each local search. For example, suppose that feature number was set to 8, BMR was 0.1, and MCN was set to 1, $8 * 0.1 = 0.8$. In this case, 0.8 is less than the value of MCN (1); therefore, 1 random change would be executed in every local search in the best fish swarm. Figure 5 presents an example of a local search intended to assemble a swarm of the best fish in the vicinity of a given fish. After five local searches among this collection of high scoring fish, the fish that had previously been identified as best fish 1 is replaced by the fish with the best feature subset. Figure 4 presents a flowchart of the MAFSA.

Table 3: UCI dataset.

Number	Dataset	Number of classes	Number of features	Number of instances
1	Bupa	2	6	345
2	Pima	2	8	768
3	Glass	6	9	214
4	Vowel	11	10	528
5	Heart	2	13	270
6	Australian	2	14	690
7	Vehicle	4	18	846
8	Robot	4	24	5456
9	German	2	24	1000
10	Sonar	2	60	208

4. Experiment Results

4.1. Experiment Environment. To compare the performance of MAFSA and AFSA, we used datasets commonly used in machine learning, the UCI dataset [27]. Ten datasets with different numbers of records, different features, different classes, and from different fields are presented in Table 3. The parameters of MAFAS and AFSA are presented in Table 4.

As shown in Table 4, AFSA does not have the BFN, BSN, BMR, or MCN parameters, which are used only for assembling a swarm of the best fish when using MAFSA. C and G are the parameters of SVM with C representing

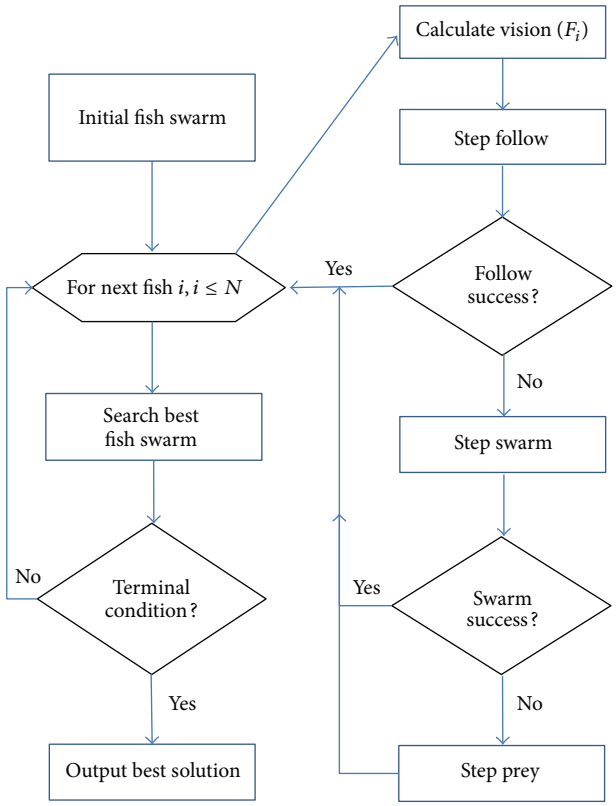


Figure 4: Flowchart of modified artificial fish swarm algorithm (MAFSA).

Table 4: Parameters of AFSA and MAFSA.

Parameters	AFSA	MAFSA
Fish number	30	30
Maximum try number of prey	20	20
Maximum crowd degree	0.5	0.5
CV	N/A	0.5
BFN	N/A	5
BSN	N/A	20
BMR	N/A	0.15
MCN	N/A	1
C	[0.01, 1024]	[0.01, 1024]
G	[0.00001, 8]	[0.00001, 8]

the penalty parameter and G representing the parameters of the RBF kernel function. This used libsvm [28] as a classifier. In [26], it was shown that setting the right parameters for SVM could greatly enhance the effectiveness of classification. As mentioned in Section 2.2, the proposed algorithm determines the optimal values for these two parameters. To ensure the reliability of the experiment, we used tenfold cross-validation.

Figure 6 presents a flowchart of the proposed model used for feature selection using MAFSA and SVM. To achieve tenfold cross-validation, the dataset was divided into two parts, training data and testing data. Various feature subsets were created along with throughout the process of MAFSA. After deleting the features and corresponding data which were not selected, the training data was input into SVM to build a classification model. SVM then outputs a value representing the accuracy of classification for each feature subset in the form of a fitness value. MAFSA would stop and output the optimal feature subset only after the terminal conditions were satisfied. In this case, the terminal condition of each fold was a lack of changes in the optimal features subset for a period of 1 hour.

4.2. Experiment 1: Feature Selection without Parameter Optimization. To compare the effectiveness of AFSA and MAFSA, we examined three experiment results: (1) the dataset without features, (2) the dataset after feature selection using AFSA and SVM, and (3) the dataset after feature selection using MAFSA and SVM. Experiment 1 did not involve search for the optimal SVM parameters C and G. In other words, Experiment 1 used the default values of parameters C and G found in libsvm. The results are presented in Table 5.

In Table 5, the first group is the dataset without feature selection, which did not provide satisfactory results with regard to classification accuracy or the number of selected features. Furthermore, because this group used data without feature selection, the number of selected features was the same as the in the original. The second group was the result for a dataset that underwent feature selection using AFSA, and the third group was the result after feature selection using MAFSA. The classification accuracy of Groups 2 and 3 were far better than those of Group 1, and the number of selected features was far less than that of Group 1. MAFSA provided the best classification accuracy in all ten of the datasets; however, in five of the datasets, the results were on par with those obtained using AFSA. MAFSA resulted in fewer selected features in eight of the ten datasets.

In five of the datasets with fewer features, MAFSA was unable to exceed the classification accuracy of AFSA. Nonetheless, MAFSA still produced fewer features. Experiment 1 did not involve SVM parameters C or G; therefore, AFSA and MAFSA were able to search nearly all of the possible feature subsets. In some folds of the datasets, more than one of the feature subsets provided the same best classification accuracy. For this reason, MAFSA resulted in fewer selected features but maintained the same classification accuracy, thereby demonstrating the superior search ability of MAFSA.

4.3. Experiment 2: With Feature Selection and Parameter Optimization. Experiment 2 compares the performance of AFSA and MAFSA in which SVM parameters C and G were also considered. In other words, AFSA and MAFSA searched not only for the optimal feature subset but also the optimal parameters of C and G. The results are presented in Table 6.

In Table 6, MAFSA is shown to have higher classification accuracy in eight of the ten datasets as well as fewer selected features in six of the ten datasets. After determining SVM parameters C and G, the classification accuracy in all ten of

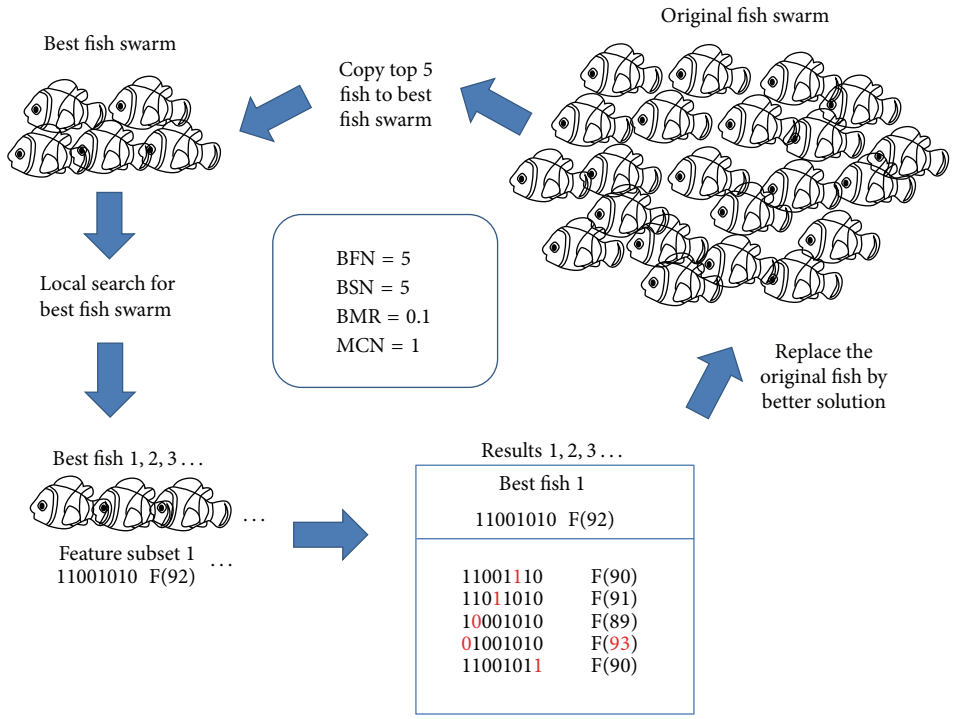


Figure 5: Process of assembling swarm of best fish.

Table 5: Result of Experiment 1.

Dataset	Number of original features	Group 1 No feature selection		Group 2 Feature selection by AFSA		Group 3 Feature selection by MAFSA	
		Number of selected features	Average accuracy rate (%)	Number of selected features	Average accuracy rate (%)	Number of selected features	Average accuracy rate (%)
Bupa	6	6	58.85	3.3	59.73**	3.2*	59.73**
Pima	8	8	77.21	5*	80.47**	5.6	80.47**
Glass	9	9	48.11	4.8	64.89**	4.4*	64.89**
Vowel	10	10	71.91	9	77.57	8.5*	77.67**
Heart	13	13	81.85	7.2	94.07**	7.1*	94.07**
Australian	14	14	85.79	8	88.26**	7.7*	88.26**
Vehicle	18	18	71.98	12.4*	75.41	12.4*	75.53**
Robot	24	24	85.90	17*	86.73	17.6	87.48**
German	24	24	74.90	14.7	82.90	12.4*	83.50**
Sonar	60	60	73.19	29.7	93.33	27.2*	94.28**

*It indicates the fewest number of selected features.
**It indicates the highest classification accuracy.

the datasets improved significantly. MAFSA was shown to have higher classification accuracy in most of the datasets; however, it little differentiated AFSA from MAFSA. We therefore used statistical analysis to verify the improvements obtained by using MAFSA. Data analysis was performed using SPSS. We employed the Friedman test, which is a statistical analysis method used to test the performance of algorithms. The Friedman statistics test is used for nonparametric statistical verification of whether k related samples differ significantly. We use the classification accuracy of AFSA and MAFSA in the ten datasets in Table 6 to verify the superior performance of MAFSA. The results indicated a significance value of 0.02, which is less than 0.05; that is, the test provides 95% CI that MAFSA provides an improvement in classification accuracy over the results obtained using AFSA.

5. Conclusion and Future Works

This research proposed a modified version of the artificial fish swarm algorithm in conjunction with support vector machine for feature selection. MAFSA differs in its use of

Table 6: Result of Experiment 2.

Dataset	Number of original features	Feature selection by AFSA		Feature selection by MAFSA	
		Number of selected features	Average accuracy rate (%)	Number of selected features	Average accuracy rate (%)
Bupa	6	4.1	84.94	3.9*	85.22**
Pima	8	4.8	83.72	4.2*	83.85**
Glass	9	4.9*	89.69	5.1	91.58**
Vowel	10	7.8	100**	7*	100**
Heart	13	6.1*	97.03	6.2	97.77**
Australian	14	7.1	93.62**	5.7*	93.33
Vehicle	18	11.2	91.01	11*	92.08**
Robot	24	7.6*	96.57	8.3	97.25**
German	24	14.2	83.7	13.7*	84.6**
Sonar	60	27.2*	99.04	29.1	100**

*It indicates the fewest number of selected features.
**It indicates the highest classification accuracy.

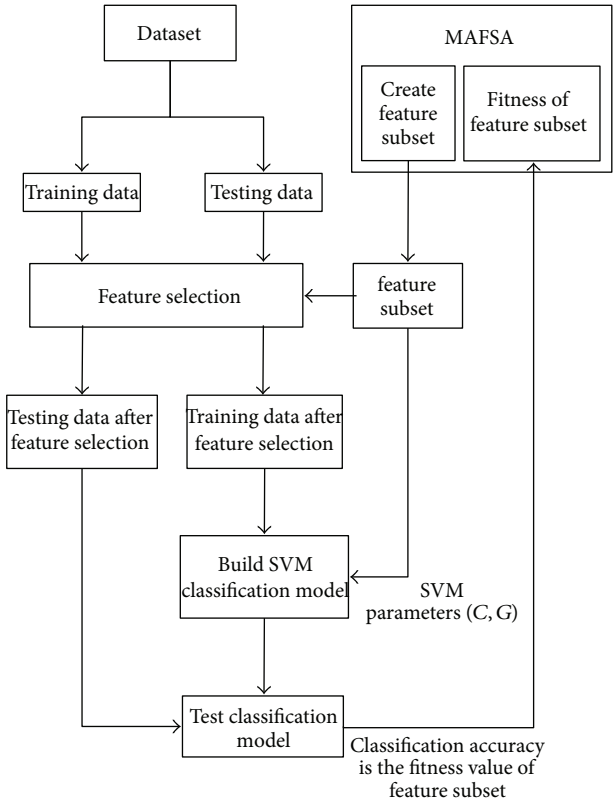


Figure 6: Flowchart of feature selection of MAFSA and SVM.

an endocrine-based formula to provide dynamic vision to assign an appropriate search space for each fish. We also included a mechanism by which a swarm of the best fish can be assembled to enhance local search ability. Experiments demonstrate that MAFSA is superior to AFSA with regard to classification accuracy as well as the number of selected features.

Nonetheless, MAFSA still has room for improvement. For example, all of the parameters of MAFSA could be selected dynamically to further enhance adaptability to different datasets. We designed binary-coded algorithms to deal with the encoding of feature selection. We expect to apply MAFSA to the optimization problem of continuum. We expect that MAFSA could be used to solve real world optimization problems in a wide range of applications.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] J. K. Kishore, L. M. Patnaik, V. Mani, and V. K. Agrawal, "Application of genetic programming for multicategory pattern classification," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 3, pp. 242–258, 2000.

[2] R. Stevens, C. Goble, P. Baker, and A. Brass, "A classification of tasks in bioinformatics," *Bioinformatics*, vol. 17, no. 2, pp. 180–188, 2001.

[3] E. Namsrai, T. Munkhdalai, M. Li, J.-H. Shin, O.-E. Namsrai, and K. H. Ryu, "A feature selection-based ensemble method for arrhythmia classification," *Journal of Information Processing Systems*, vol. 9, no. 1, pp. 31–40, 2013.

[4] J. R. Quinlan, *C4.5: Programs for Machine Learning*, The Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann, San Mateo, Calif, USA, 1993.

[5] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000.

[6] Y. S. Hwang, J. B. Kwon, J. C. Moon, and S. Je Cho, "Classifying malicious web pages by using an adaptive support vector machine," *Journal of Information Processing Systems*, vol. 9, no. 3, pp. 395–404, 2013.

[7] J. Uddin, R. Islam, and J. Kim, "Texture feature extraction techniques for fault diagnosis of induction motors," *Journal of Convergence*, vol. 5, no. 2, pp. 15–20, 2014.

[8] G. P. Zhang, "Neural networks for classification: a survey," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 30, no. 4, pp. 451–462, 2000.

[9] M. Malkawi and O. Murad, "Artificial neuro fuzzy logic system for detecting human emotions," *Human-Centric Computing and Information Sciences*, vol. 3, article 3, 2013.

[10] A. K. Gopalakrishna, "A subjective job scheduler based on a backpropagation neural network," *Human-Centric Computing and Information Sciences*, vol. 3, article 17, 2013.

[11] H. Lee, H. Kim, and J. Seo, "An integrated neural network model for domain action determination in goal-oriented dialogues," *Journal of Information Processing Systems*, vol. 9, no. 2, pp. 259–270, 2013.

[12] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[13] R. Jensen and Q. Shen, *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches*, IEEE Press Series on Computational Intelligence, Wiley, Oxford, UK, 2008.

[14] S. Farzi, "Efficient job scheduling in grid computing with modified artificial fish swarm algorithm," *International Journal of Computer Theory and Engineering*, vol. 1, no. 1, pp. 13–18, 2009.

[15] O. Mirzaei and M.-R. Akbarzadeh-T, "A novel learning algorithm based on a multi-agent structure for solving multi-mode resource-constrained project scheduling problem," *Journal of Convergence*, vol. 4, no. 1, pp. 47–52, 2014.

[16] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[17] T. Liu, A.-L. Qi, Y.-B. Hou, and X.-T. Chang, "Feature optimization based on artificial fish-swarm algorithm in intrusion detections," in *Proceedings of the International Conference on Networks Security, Wireless Communications and Trusted Computing (NSWCTC '09)*, vol. 1, pp. 542–545, Wuhan, China, April 2009.

[18] J. H. Holland, *Adaptation in natural and artificial systems*, University of Michigan Press, Ann Arbor, Mich, USA, 1975.

[19] H. T. T. Binh and S. H. Ngo, "All capacities modular cost survivable network design problem using genetic algorithm with completely connection encoding," *Human-Centric Computing and Information Sciences*, vol. 4, article 13, 2014.

[20] H. T. T. Binh, "Multi-objective genetic algorithm for solving the multilayer survivable optical network design problem," *Journal of Convergence*, vol. 5, no. 1, pp. 20–25, 2014.

[21] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, pp. 1942–1948, IEEE, Perth, Australia, December 1995.

[22] X. L. Li, Z. J. Shao, and J. X. Qian, "Optimizing method based on autonomous animats: fish-swarm Algorithm," *System Engineering Theory and Practice*, vol. 22, no. 11, pp. 32–38, 2002.

[23] J.-P. Wang and M.-J. Hu, "A solution for TSP based on artificial fish algorithm," in *Proceedings of the International Conference on Computational Intelligence and Natural Computing (CINC '09)*, pp. 26–29, Wuhan, China, June 2009.

[24] C. Wang, T. Xue, and L. Sun, "A survey of Artificial Fish-swarm Algorithm and its Improve Technique," 2014, http://www.paper.edu.cn/html/releasepaper/2008/03/581/.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[26] K.-C. Lin, S.-H. Hsu, and J. C. Hung, "Adaptive SVM-based classification systems based on the improved endocrine-based PSO algorithm," *Lecture Notes in Computer Science*, vol. 7669, pp. 543–552, 2012.

[27] UC Irvine Machine Learning Repository, 2014, http://archive.ics.uci.edu/ml/.

[28] C. C. Chang and C. J. Lin, *LIBSVM: A Library for Support Vector Machines*, 2014, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 643842, 10 pages
http://dx.doi.org/10.1155/2015/643842

Research Article

The Development of a Tourism Attraction Model by Using Fuzzy Theory

Jieh-Ren Chang and Betty Chang

National Ilan University, No. 1, Section 1, Shen-Lung Road, Yilan City 26047, Taiwan

Correspondence should be addressed to Jieh-Ren Chang; jrchang@niu.edu.tw

Received 28 September 2014; Accepted 9 March 2015

Academic Editor: Jong-Hyuk Park

The purpose of this study is to develop a model to investigate the tourists' preference. Ten attributes of tourist destinations were used in this study. Fuzzy set theory was adopted as the main analysis method to find the tourists' preference. In this study, 248 pieces of data were used. Besides the evaluations for the factors, the overall evaluations (namely, satisfied, neutral, and dissatisfied) for every tourism destination were also inquired. After screening, 201 pieces of these data could be used. In these 201 pieces of data, 141 were classified into "satisfied" with the tourism destination, accounting for 70.15%, and 49 were "neutral," accounting for 24.38%, while 11 were "dissatisfied," accounting for 5.47%. Eight rules were obtained with the method of fuzzy preprocess. Regarding the condition attributes, three of the original ten attributes were found influential, namely, level of prices, living costs, information and tourist services, and tourist safety of the tourism destinations. From the results of this study, it is shown that top management of tourism destinations should put resources in these fields first, in order to allow limited resources to perform to maximum effectiveness.

1. Introduction

With the rapid economic and social development, the increase in GDP every year, and people's growing concern toward recreations, the tourism industry has been developing vigorously. In many countries, the tourism industry is a main industry that deserves our policy attention, and obviously it has become a global socioeconomic phenomenon [1]. A successful tourism industry can enhance regional economic development, as well as becoming a source of rich foreign exchange income [2]. The tourism industry is one of the main industries that determine the world's long-term economic growth [3].

The tourism industry has a far-reaching influence on many aspects such as social and economic development, culture, city development, and revival; in particular, it has the greatest influences on economics [4]. The output value of the tourism industry accounts for US$ 6 trillion of the global economy which is 9% of global GDP [5]. UNWTO (2011) [4] further estimated that, by 2020, the number of global tourists will reach 1.6 billion and 2 million people; global tourism earnings will reach as much as two trillion US dollars. The data indicates that tourism can bring in an enormous amount of economic benefits [6].

However, in modern society, no tourism industry can escape from international competition due to globalization. In this situation, how to increase international competitiveness of the tourism industry has become one of the greatest concerns.

This study investigated into tourist destinations. A tourist destination (such as city or region) is no longer viewed as a place that features unique natural landscape, culture, or art; instead, it is seen as a compound product that satisfies the tourists' need [1]. Now, many countries are actively developing their own tourist destinations' international competitiveness [2]. However, how to enhance tourist destinations' attractiveness to tourists relies on more than a single factor; it requires an overall plan to increase the tourist destinations' competitiveness in the international market [7].

The goal of this study is to establish a model for managing tourist destinations. The management of all tourism destinations should focus on enhancing their attractiveness and quality, as well as effectively using the limited resources in the current environment [8]. Therefore, this study explores

various tourist destinations from the perspective of tourists. In addition, how these tourism destinations attract tourists and the tourists' evaluations are also included in this study [9].

Fuzzy model is similar to the thinking model of human beings [10, 11]. This study therefore uses fuzzy model to analyze the preference rules of tourist destinations. Hereby, this research aims to develop a model for investigation of tourist destinations management. It adopts fuzzy set theory as the main analysis method for tourism industry to find the tourists' preference. In the second part of this study, it does literature exploration on the competitiveness and attractiveness of tourism destination. The third section focuses on introduction of fuzzy set theory and fuzzy rules extraction algorithm [12]. The fourth section gives a possible explanation for the results. Finally, the authors draw a conclusion and the suggestions for future research in the last section.

2. Review and Discussion of the Literature

Tourism refers to people's temporary movement from their residence or working environment to a destination, and all related facility or services in the destination which are to be provided to tourists are covered in the tourism industry [13].

Gunn [14] reported that the so-called destination refers to local residents' "location"; on the other hand, it is a "playground" for tourists from other areas. A better explanation for a playground is a tourist destination, as it can be a specific tourist attraction, a town, a certain region in a country, an entire country, or even a bigger area on the planet [15]. Cracolici and Nijkamp [1] believed a tourist destination is a supplier that satisfies tourists physically and mentally; two parts are included: "structural" and "nonstructural." "Structural" refers to the natural landscape and cultural resources in a tourist destination, while "nonstructural" means human resources, perceptions, and so forth. Accordingly, if an area plans to develop tourism, the key point basically lies in how to present a destination that attracts tourists [16] and how to become more appealing and competitive than any other areas' destinations.

The critical factor model of tourist destinations' competitiveness established by Cracolici and Nijkamp [1] based on the concept of Crouch and Ritchie [8] encompasses physiography, culture and history, market ties, activities, events, and the tourism superstructure. Ten attractions of tourist destinations were compiled and used as the attributes in this study, as follows: (F1) reception and sympathy of local residents, (F2) artistic and cultural cities, (F3) landscape, environment, and nature, (F4) hotels and other accommodation, (F5) typical foods, (F6) cultural events (concerts, art exhibitions, festivals, etc.), (F7) level of prices, living costs, (F8) quality and variety of products in the shops, (F9) information and tourist services, and (F10) tourist safety.

3. Establishment of Fuzzy Decision Rules

3.1. Introduction of Fuzzy Concept. Fuzzy theory has been widely studied and successfully applied in various fields, which has got remarkable achievements so far. The fuzzy set defined by Professor Zadeh is represented by characteristic function $\mu_A(x)$ in mathematics, in which the value of membership function is the degree of element x belonging to a fuzzy set A. Therefore, the function matches the elements in the universal set to another set that is between 1 and 0:

$$\mu_A : X \longrightarrow [0, 1], \tag{1}$$

where $x \in X$, X indicates the universal set that is defined for the specific problem, while $[0, 1]$ refers to the range of real numbers between 0 and 1. Accordingly, this study will apply the two operating factors in the deduction of if-then fuzzy rules and membership function.

The vague linguistics between "yes" and "no" could be all represented by membership function values, which is the basic concept of fuzzy set theory. It aims to illustrate fuzzy phenomenon by clear and strict mathematic methods.

In this study, the tourism management of towns with cultural heritage is investigated. Ten attributes about tourism management of these towns were used for the study. In addition, fuzzy set theory is utilized to obtain the rules of tourist preference. During the fuzzy deduction, we collect various data from complicated environments and apply them in fuzzy deduction rules and membership functions to make the final decisions.

For the tourism management of cultural heritage towns, ten properties are investigated. Besides, the fuzzy set theory is also used to obtain the rules of tourists' preference. To sum up, the fuzzy system theory is scientific, advanced, and practical and can also provide correct guidance to our work. A new learning method for automatically deriving fuzzy rules and membership functions from a given set of training instances is proposed here as the knowledge acquisition facility [17]. Notation and definitions are introduced below.

Data preprocess and fuzzy rule establishment are included in the fuzzy learning algorithm. A set of training instances are collected from the environment. Our task here is to generate automatically reasonable membership functions and appropriate decision rules from these training data, so that they can represent important features of the data set. In order to avoid the disturbance of ineffective information, all the data should be preprocessed in advance [18].

The support set of fuzzy set D is a crisp set; it includes all the elements in the universe set U, but the membership value in D must be greater than 0 as follows:

$$\text{supp}(D) = \{x \in U \mid \mu_D(x) > 0\}. \tag{2}$$

The center of a fuzzy set is defined as if the membership values which correspond to fuzzy set D from every element in supp(D) are finite (basically 1 is supposed to be the maximum value). In this situation, the position of the maximum value or the medium point of the maximum value is defined as the center of the fuzzy set as shown in Figure 1. The typical center of a fuzzy set is shown in Figure 1.

Fuzzy set includes all the points in the set U. Concerning set D, when the membership value is equal to 0.5, it is the vaguest point.

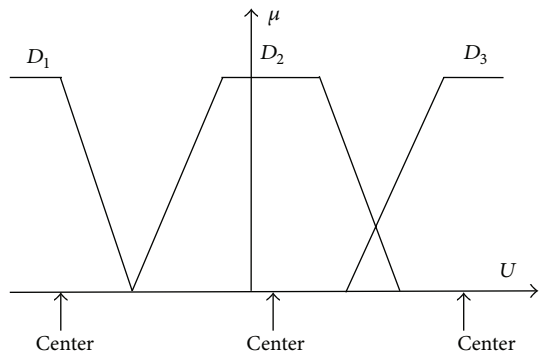


Figure 1: Typical centers in fuzzy set.

In order to obtain the support set higher than a certain level, α-cut is used to extract the support set and α-cut of fuzzy set D is a definite set D_α as follows:

$$D_\alpha = \{x \in U \mid \mu_D(x) \geq \alpha\}. \tag{3}$$

Fuzzy proposition includes two types, namely, atomic fuzzy proposition and compound fuzzy proposition. An atomic fuzzy proposition is a single fuzzy proposition as follows:

$$q_1 \text{ is } a_1, \tag{4}$$

where q is a linguistic variable and a_1 is the linguistic value of q_1.

A compound fuzzy proposition is using conjunctions such as "and," "or," and "not" to joint atomic fuzzy propositions to make fuzzy intersection set, fuzzy union set, and fuzzy compensate set. For example, q_1 stands for "information and tourist services," q_2 stands for "level of prices, living costs," a_1 and a_2 stand for linguistic values "very good" and "barely acceptable," and then the compound fuzzy proposition will be as follows:

$$q_1 \text{ is } a_1 \text{ and } q_2 \text{ is } a_2. \tag{5}$$

Fuzzy rules are made of "if-then" and fuzzy propositions as shown in rule r:

$$\begin{aligned} r&: \text{If } q_1 \text{ is } a_1 \text{ and } q_2 \text{ is } a_2 \\ &\text{Then } y \text{ is } b_1. \end{aligned} \tag{6}$$

In an "if fuzzy proposition," the questionnaire analysis is set as a condition attribute and, in a "then fuzzy proposition," the questionnaire analysis is set as a decision attribute. When linguistic variable q_1 is a_1 and q_2 is a_2, linguistic variable y will be b_1; therefore with fuzzy rules, the linguistic causal relationship can be inferred. All the fuzzy rules can be put together to make a fuzzy rule database and this database includes various corresponding fuzzy rules.

Fuzzy inferences mean making inferences with all the rules in fuzzy rule database. There are three types in fuzzy inferences, namely, type 1, type 2, and type 3, which stand for singleton, linguistic, and linear inference rules. In this study, linguistic inference rules were used, and the method proposed by Tsukamoto was applied.

3.2. Deleting Ineffective Data. In order to avoid the interruption from ineffective data, preprocessing is necessary before data analysis [19]. There are many different methods that can be used for preprocessing. However, one preprocessing method may not be suitable for all of the fields. In this study, a novel preprocessing method of screening ineffective data for questionnaires was proposed. Here we define the effective data as honest data and ineffective data as dishonest data. Some attributes and data might be deleted to let decision-makers obtain precise and useful data in questionnaire analysis process. In this process, it is supposed that data from some respondents can be neglected. This type can be considered as a form of majority verdict which can obtain the main consensus from the majority of the questionnaire respondents. Concerning the data analysis in this study, the answers from questionnaires responded by tourists were used for data analysis. The effective data are defined as responses from the majority of tourists. The ineffective data, on the other side, include dishonest data and data from respondents with special preference.

3.3. Establishing Questionnaire Rules. The method of deleting ineffective data will be reported in this part. First of all, the authors assumed that most people have similar perception. Therefore, concerning a specific tourism destination, it is supposed that the scoring toward a specific attribute from the questionnaire respondents would be aggregated in a range. In the space of condition attribute, every decision attribute forms a block space and has its own center; those data with bias might be far from the center and more likely to be ineffective data. In addition, in the space of condition attribute, the intersection with different decision attribute might be small or empty; this assumption is to make sure that the classification of decision attribute is identifiable.

With establishing fuzzy rules, the authors can screen ineffective data with the method of fuzzy inference. Concerning the content of the questionnaire, there are n subquestion items in each of the questions, and these n subquestion items stand for condition attribute items as follows:

$$Q = \{q_1, q_2, \ldots, q_n\}, \tag{7}$$

where $n \in N$ and N stands for the set of positive integers.

The overall evaluation a respondent made is the decision attribute y in a fuzzy rule. Supposing that a respondent answered a specific question item q_p, the set of linguistic values is as follows:

$$a^p = \left\{a_1^p, a_2^p, \ldots, a_{j_p}^p\right\}, \tag{8}$$

where $1 \leq p \leq n$ and $p \in N$, j_p is the number of the linguistic values of a specific condition attribute, and $j_p \in N$.

After answering all the subquestions, the respondent must select a linguistic value from set B as the overall evaluation, where set B is a set of linguistic values as follows:

$$B = \{b_1, b_2, \ldots, b_i\}, \tag{9}$$

where i is the number of decision attribute linguistic values and $i \in N$.

The data of the answers from respondents were transferred into fuzzy rules. For example, when the linguistic value of the decision attribute inference is b_h, the first fuzzy rule will be as follows:

$$r_1^h\text{: } q_1 \text{ is } a_{p_1}^1 \text{ and } q_2 \text{ is } a_{p_2}^2 \text{ and } \cdots \text{ and } q_n \text{ is } a_{p_n}^3 \longrightarrow y \text{ is } b_h, \tag{10}$$

where $1 \le h \le i$ and $h \in N$.

Then all the fuzzy rules would be put together in fuzzy rule database as follows:

$$R = \left\{r^1, r^2, \ldots, r^i\right\}. \tag{11}$$

The linguistic values of decision attribute in fuzzy rule of R could be classified into i categories and every category would correspond to the linguistic values in set B as follows:

$$r^h = \left\{r_1^h, r_2^h, \ldots, r_{k_h}^h\right\}, \tag{12}$$

where r^h stands for the fuzzy rule classification of b_h and the number of rules is k_h.

The previous part reported the principles of fuzzy rules for multiple condition attribute to single decision attribute. It is found from rule classification that the distribution space of b_h corresponds to set Q in (7) as follows:

$$F_h = \left\{a^{-1}(b_h), a^{-2}(b_h), \ldots, a^{-n}(b_h)\right\}, \tag{13}$$

where F_h is the linguistic value distribution space of b_h and $a^{-1}(b_h)$ stands for the distribution situation of a^1, which is corresponded from b_h.

4. Results and Discussion

4.1. Overview of the Research Data. In this study, 248 data used were retrieved. Most of the respondents are the office workers and young persons in Taiwan. In these 248 data, 201 of the tourist sites the respondents mentioned include the sites in northern parts, central parts, southern parts, and eastern parts of Taiwan. And the other 47 ones are international tourist sites out of Taiwan. In these data, tourists' evaluations for each of the factors about the tourism destinations were included. Besides the evaluations for the factors, the overall evaluations (namely, satisfied, neutral, and dissatisfied) for every tourism destination were also inquired. After screening, 201 of these data could be used. In these 201 data, 141 were classified into "satisfied" with the tourism destination, accounting for 70.15%, and 49 were "neutral," accounting for 24.38%, while 11 were "dissatisfied," accounting for 5.47%. The numbers and percentages of data classified into each category were shown in Table 1. The evaluation of the attribute "level of prices, living costs," has three fuzzy linguistic terms of levels ("good," "barely acceptable," and "poor.") On the other hand, the attribute "tourist safety" has four fuzzy linguistic terms of levels ("very good," "good," "poor," and "very poor.") The levels of these two attributes were shown in Table 2. Through the method of fuzzy preprocess, 8 rules were obtained. These fuzzy rules were shown in Table 3. Concerning the condition attributes, two of the original ten attributes were found influential, namely, level of prices, living costs (F7), and tourist safety (F10) of the tourism destinations.

Table 1: The numbers and percentages of overall evaluation.

	Satisfied	Neutral	Dissatisfied	Total
Numbers of data classified into each category	141	49	11	201
Percentage	70.15%	24.38%	5.47%	100.00%

4.2. Fuzzy Rules Analysis. The results of the fuzzy rules analysis were shown in Table 3. According to fuzzy mathematics, only two (F7, level of prices, living costs, and F10, tourist safety) of the 10 attributes were strongly influential attributes. From these rules, the following results can be obtained.

(1) From Rule 2 and Rule 3, when F7 (level of prices, living costs) received "good," the overall evaluations would be "satisfied" if F10 (tourist safety) received "good" or "very good."

(2) From Rule 1 and Rule 3, when F10 (tourist safety) received "very good," the overall evaluations would be "satisfied" even if F7 (level of prices, living costs) received "barely acceptable."

(3) From Rule 4 and Rule 5, when F7 (level of prices, living costs) received "barely acceptable," the overall evaluations would be neutral, if F10 (tourist safety) received the level of "good" or "poor."

(4) From Rule 6 and Rule 7, when F7 (level of prices, living costs) received "poor," the overall evaluations would be dissatisfied, if F10 (tourist safety) received the level of "poor" or "very poor."

(5) From Rule 6 and Rule 8, if F10 (tourist safety) received "very poor," the overall evaluations would be dissatisfied, no matter F7 (level of prices, living costs) received "barely acceptable" or "poor."

(6) Comparing Rule 1 and Rule 4, F7 (level of prices, living costs) received "barely acceptable" in both rules and at this time F10 (tourist safety) would be a key for the overall evaluations. F10 (tourist safety) received "very good" in Rule 1 and the overall evaluations were "satisfied," while, in Rule 4, the overall evaluations were "neutral" as F10 (tourist safety) received "poor."

(7) While comparing Rule 2 and Rule 5, F10 (tourist safety) received "good" in both of these rules. F7 (level of prices, living costs) would be a key for the overall evaluations in this situation. In Rule 2 F10 (tourist safety) received "good" and the overall evaluations were "satisfied"; in Rule 5, however, the overall evaluations were "neutral" as F7 (level of prices, living costs) received "barely acceptable."

Table 2: Levels of attributes.

Attributes	Numbers of levels	Fuzzy linguistic terms of levels (form high level to low level)
F7: level of prices, living costs	3 levels	"Good," "barely acceptable," and "poor."
F10: tourist safety	4 levels	"Very good," "good," "poor," and "very poor."

Table 3: The 8 rules derived from fuzzy analysis.

	F7 level of prices, living costs	F10 tourist safety	Evaluation
Rule 1	Barely acceptable	Very good	Satisfied
Rule 2	Good	Good	Satisfied
Rule 3	Good	Very good	Satisfied
Rule 4	Barely acceptable	Poor	Neutral
Rule 5	Barely acceptable	Good	Neutral
Rule 6	Poor	Very poor	Dissatisfied
Rule 7	Poor	Poor	Dissatisfied
Rule 8	Barely acceptable	Very poor	Dissatisfied

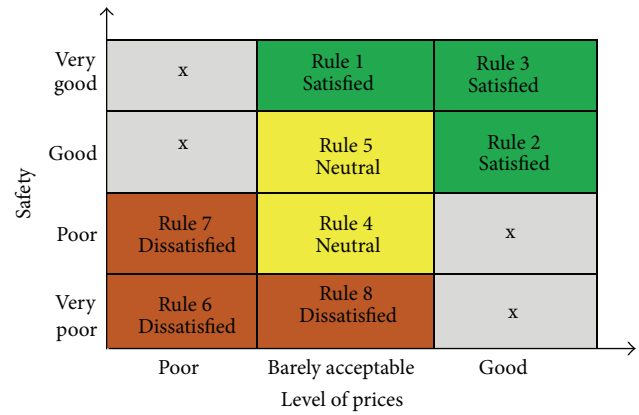


Figure 2: Fuzzy rule base (for all tourists).

(8) Comparing Rule 4 and Rule 7, as F10 (tourist safety) received "poor" in both rules, F7 (level of prices, living costs) would be a key for the overall evaluations. For example, F7 (level of prices, living costs) received "barely acceptable" in Rule 4 and the overall evaluations were "neutral," while, in Rule 7, F7 (level of prices, living costs) received "poor" and the overall evaluations turned to "dissatisfied" consequently.

(9) Comparing Rule 5 and Rule 8, when F7 (level of prices, living costs) received "barely acceptable," F10 (tourist safety) played a crucial role for deciding the overall evaluations. In other words, if F10 (tourist safety) received "good," the overall evaluations would be "neutral." On the other hand, if F10 (tourist safety) received "very poor," the overall evaluations would be "dissatisfied."

(10) From the comparison of Rule 1, Rule 4, Rule 5, and Rule 8, it was found that F7 (level of prices, living costs) received "barely acceptable" in each of the rules. In Rule 1, for example, the overall evaluations were "satisfied" since F10 (tourist safety) received "very good." The overall evaluations of Rule 4 and Rule 5 were "neutral" on the other hand as F10 (tourist safety) received either "good" or "poor." In Rule 8, however, the overall evaluations were "dissatisfied" when F10 (tourist safety) received "very poor."

In this section, the rules in Table 3 were represented as in Figure 2. In Figure 2, the upper right corner (areas of Rule 1, Rule 2, Rule 3, and Rule 5) shows that when the attribute "tourist safety" was evaluated as "good" or "very good," the attribute "level of prices, living costs" was also evaluated as "good" or "barely acceptable," and the overall evaluations were satisfied or neutral. The reason might be that most tourists already had sufficient information about the level of local living costs before they made decision for their destinations. The tourist might therefore think the level of price is agreeable. On the other hand, the lower left corner (areas of Rule 6, Rule 7, Rule 8, and Rule 4) shows that when the attribute "tourist safety" was evaluated as "poor" or "very poor," the attribute "level of prices, living costs" was evaluated as "poor" or "barely acceptable," and the overall evaluations were dissatisfied or neutral. It is believed that the poor safety might impair tourists' confidence. To sum up, tourist safety is the attribute the tourists care about the most.

In Figure 2, "x" stands for no rules in that exact area. According to Figure 2, no rules were found in the upper left corner; these areas stand for destinations with high safety and high price. The reason for no rules here might be that most respondents are office workers and young persons, they made very different evaluations about these destinations, and therefore no consistent rules could be produced. Besides, there are no rules either in the lower right corner. This lower right corner area stands for tourist destinations with poor safety. Since tourist safety was the attribute the tourists care about the most, very few tourists would select these destinations.

4.3. Comparison of the Results from Tourists of Different Ages. Tourism is getting more and more popular in the 21st century. However, tourists of different ages might have various demands and different preference regarding tourism destinations. In order to investigate the tourist preference of different ages, the authors divided the data of tourists into two groups: one group is of tourists above 30 years old and the other group is of tourists of 30 years old and below.

4.3.1. Results from Tourists of 30 Years Old and Below. In the group of tourists of 30 years old and below, there are 139 pieces of data collected from these tourists. After programming

Table 4: Levels of attributes (tourists of 30 years old and below).

Attributes	Numbers of levels	Fuzzy linguistic terms of levels (form high level to low level)
F7: level of prices, living costs	3 levels	"Good," "barely acceptable," and "poor."
F9: information and tourist services	3 levels	"Good," "barely acceptable," and "poor."
F10: tourist safety	4 levels	"Very good," "good," "poor," and "very poor."

Table 5: The 13 rules derived from fuzzy analysis (tourists of 30 years old and below).

	F7 level of prices, living costs	F9 information and tourist services	F10 tourist safety	Evaluation
Rule 1	Barely acceptable	Barely acceptable	Very good	Satisfied
Rule 2	Barely acceptable	Good	Good	Satisfied
Rule 3	Barely acceptable	Good	Very good	Satisfied
Rule 4	Good	Barely acceptable	Good	Satisfied
Rule 5	Good	Barely acceptable	Very good	Satisfied
Rule 6	Good	Good	Good	Satisfied
Rule 7	Good	Good	Very good	Satisfied
Rule 8	Barely acceptable	Barely acceptable	Poor	Neutral
Rule 9	Barely acceptable	Barely acceptable	Good	Neutral
Rule 10	Poor	Poor	Very poor	Dissatisfied
Rule 11	Poor	Barely acceptable	Very poor	Dissatisfied
Rule 12	Barely acceptable	Poor	Very poor	Dissatisfied
Rule 13	Barely acceptable	Poor	Poor	Dissatisfied

with fuzzy set theory, three of the attributes were found to be crucial, namely, "level of prices, living costs" (F7), "information and tourist services" (F9), and "tourist safety" (F10). The evaluations of both of the attributes "level of prices, living costs" and "information and tourist services" were divided into three fuzzy linguistic terms of levels "good," "barely acceptable," and "poor" while the evaluation of the attribute "tourist safety" could be divided into four fuzzy linguistic terms of levels "very good," "good," "poor," and "very poor" as shown in Table 4. Thirteen fuzzy rules were derived from fuzzy computing as shown in Table 5.

According to the fuzzy rules obtained from the data of tourists of 30 years old and below, the following results can be obtained.

(1) Comparing Rule 4 and Rule 9, F9 (information and tourist services) received "barely acceptable" and F10 (tourist safety) received "good" in both rules; at this time F7 (level of prices, living costs) would play a crucial role in deciding the overall evaluations. For example, when F7 received "good" in Rule 4, the overall evaluation would be "satisfied" while, in Rule 9, F7 received "barely acceptable" and the overall evaluation was then "neutral."

(2) Comparing Rule 4 and Rule 9, F9 (information and tourist services) received "barely acceptable" and F10 (tourist safety) received "good" in both rules; at this time F7 (level of prices, living costs) would play a crucial role in deciding the overall evaluations. For example, when F7 received "good" in Rule 4, the overall evaluation would be "satisfied" while, in Rule 9, F7 received "barely acceptable" and the overall evaluation was then "neutral."

(3) Comparing Rule 1, Rule 8, and Rule 9, both of F7 (level of prices, living costs) and F9 (information and tourist services) received "barely acceptable" in each of the rules. In this situation, F10 (tourist safety) would be a key for the overall evaluations. In Rule 1, F10 (tourist safety) received "very good" and the overall evaluation was "satisfied," while, in Rule 8, F10 (tourist safety) received "poor"; and in Rule 9, F10 (tourist safety) received "good" and the overall evaluations of both of Rule 8 and Rule 9 were "neutral."

(4) Comparing Rule 8 and Rule 13, F7 (level of prices, living costs) received "barely acceptable" and F10 (tourist safety) received "poor" in both rules; at this time F9 (information and tourist services) would play an influential role in deciding the overall evaluations. For example, when F9 received "barely acceptable" in Rule 8, the overall evaluation would be "neutral," while, in Rule 13, F9 received "poor" and the overall evaluation was then "dissatisfied."

According to the results of fuzzy analysis, for tourists of 30 years old and below, three (F7, level of prices, living costs, F9, information and tourist services, and F10, tourist safety) of the 10 attributes were strongly influential attributes. Compared with the results in the previous section, there was

Table 6: Levels of attributes (tourists above 30 years old).

Attributes	Numbers of levels	Fuzzy linguistic terms of levels (form high level to low level)
F7: level of prices, living costs	4 levels	"Very good," "good," "poor," and "very poor."
F9: information and tourist services	4 levels	"Very good," "good," "poor," and "very poor."
F10: tourist safety	4 levels	"Very good," "good," "poor," and "very poor."

an extra influential attribute, namely, information and tourist services (F9). In order to analyze the relationship among these three attributes, 3 figures based on three different levels (good, barely acceptable, and poor) of information and tourist services were generated.

Figure 3(a) shows the rule base of tourists of 30 years old and below when information and tourist services of the destinations are good. Only four rules were generated in the upper right corner of Figure 3(a). These 4 rules are all evaluated as "satisfied" with very good or good in safety and good or barely acceptable in living cost. On the other hand, there were no rules created in other areas in Figure 3(a). In the condition of sufficient information, tourists would try their best to avoid going to destinations with poor safety or poor level of prices. Similar to the condition in Figure 2, no rules were found in the upper left corner and the lower right corner.

Figure 3(b) shows the rule base of tourists of 30 years old and below when information and tourist services of the destinations are barely acceptable. Comparing Figure 3(b) with Figure 3(a), Rule 9 in Figure 3(b) is in the same position as Rule 2 in Figure 3(a). However, the overall evaluation of Rule 9 in Figure 3(b) is neutral and that of Rule 2 in Figure 3(a) is satisfied; the authors therefore inferred that good information and tourist services of the destinations may promote the image of a tourist site.

Figure 3(c) shows the rule base of tourists of 30 years old and below when information and tourist services of the destinations are poor. Comparing Figure 3(c) with Figure 3(b), Rule 12 in Figure 3(c) is in the same position as Rule 8 in Figure 3(b). Nevertheless, the overall evaluation of Rule 12 in Figure 3(c) is dissatisfied and that of Rule 8 in Figure 3(b) is neutral; it is therefore inferred that poor information and tourist services of a tourist site may degrade the overall evaluation of a destination. On the other hand, there were no rules generated in other areas in Figure 3(c). Actually, very few people know destinations with poor information. Besides, it is supposed that a tourist site with good safety and living cost condition will soon be popular in this Internet era, and then those cases will be transferred into the section of sufficient information such as the cases in Figures 3(a) and 3(b).

4.3.2. Results from Tourists above 30 Years Old. In the group of tourists above 30 years old, there are pieces of 34 data collected from these tourists. After programming with fuzzy set theory, three of the attributes were found to be crucial, namely, "level of prices, living costs" (F7), "information and tourist services" (F9), and "tourist safety" (F10). The evaluation of all the attributes "level of prices, living costs," "information and tourist services," and "tourist safety" was shown as four fuzzy linguistic terms of levels ("very good," "good," "poor," and "very poor") as shown in Table 6. Fourteen fuzzy rules were derived from fuzzy computing as shown in Table 7.

According to the fuzzy rules obtained from the data of tourists above 30 years old, the following results can be obtained.

(1) Comparing Rule 8 and Rule 14, F7 (level of prices, living costs) received "good" and F10 (tourist safety) received "very poor" in both rules; at this time F9 (information and tourist services) would play a crucial role in deciding the overall evaluations. For example, when F9 received "good" in Rule 8, the overall evaluation would be "neutral," while, in Rule 14, F9 received "very poor" and the overall evaluation was then "dissatisfied."

(2) Comparing Rule 4 and Rule 11, F9 (information and tourist services) received "very good" and F10 (tourist safety) received "good" in both rules; at this time F7 (level of prices, living costs) would be a key for the overall evaluations. In Rule 4, F7 (level of prices, living costs) received "very good" and the overall evaluation was "satisfied," while, in Rule 11, F7 (level of prices, living costs) received "good" and the overall evaluation of Rule 11 was then "neutral."

According to the results of fuzzy analysis, for tourists above 30 years old, three (F7, level of prices, living costs, F9, information and tourist services, and F10, tourist safety) of the 10 attributes were strongly influential attributes. Besides, there are four levels in each of the three attributes as shown in Table 6. In order to analyze the relationship among these three attributes, 4 figures based on four different levels (very good, good, poor, and very poor) of information and tourist services were generated.

Figure 4(a) shows the rule base of tourists above 30 years old when information and tourist services of the destinations are very good. Seven rules were generated: four rules of satisfied were in the upper right corner of Figure 4(a) and the other three rules are of neutral. Comparing Figure 4(a) with Figure 4(b), Rule 7 in Figure 4(a) is in the same position as Rule 13 in Figure 4(b). However, the overall evaluation of Rule 7 in Figure 4(a) is neutral and that of Rule 13 in Figure 4(b) is dissatisfied; it is therefore inferred that better information

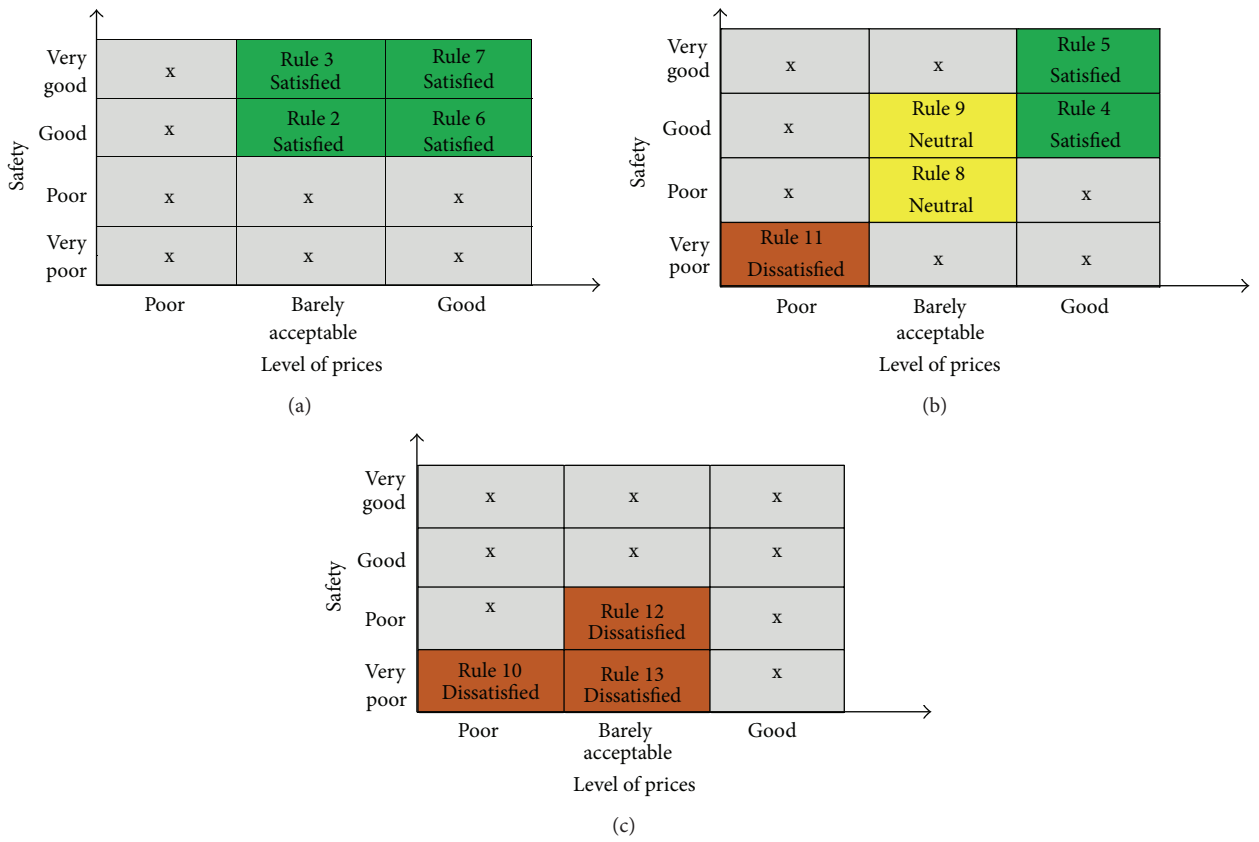


FIGURE 3: (a) Tourists of 30 years old and below/information and tourist services are good. (b) Tourists of 30 years old and below/information and tourist services are barely acceptable. (c) Tourists of 30 years old and below/information and tourist services are poor.

TABLE 7: The 14 rules derived from fuzzy analysis (tourists above 30 years old).

	F7 level of prices, living costs	F9 information and tourist services	F10 tourist safety	Evaluation
Rule 1	Poor	Very good	Very good	Satisfied
Rule 2	Good	Very good	Very good	Satisfied
Rule 3	Very good	Poor	Very good	Satisfied
Rule 4	Very good	Very good	Good	Satisfied
Rule 5	Very good	Very good	Very good	Satisfied
Rule 6	Very poor	Very good	Poor	Neutral
Rule 7	Poor	Very good	Poor	Neutral
Rule 8	Good	Good	Very poor	Neutral
Rule 9	Good	Good	Poor	Neutral
Rule 10	Good	Good	Good	Neutral
Rule 11	Good	Very good	Good	Neutral
Rule 12	Very good	Good	Poor	Neutral
Rule 13	Poor	Good	Poor	Dissatisfied
Rule 14	Good	Very poor	Very poor	Dissatisfied

and tourist services of a tourist site may promote the overall evaluation of a destination.

It is found that very few rules are in Figures 4(c) and 4(d). Similarly, there are few rules found in Figure 3(c) (three rules). The authors inferred that destinations with poor information and tourist services have fewer tourists. There is only one rule especially in each of Figures 4(c) and 4(d) because of lack of data from tourists above 30 years old. It is therefore concluded that tourists in this group (tourists above 30 years old) seldom travel to destinations with poor

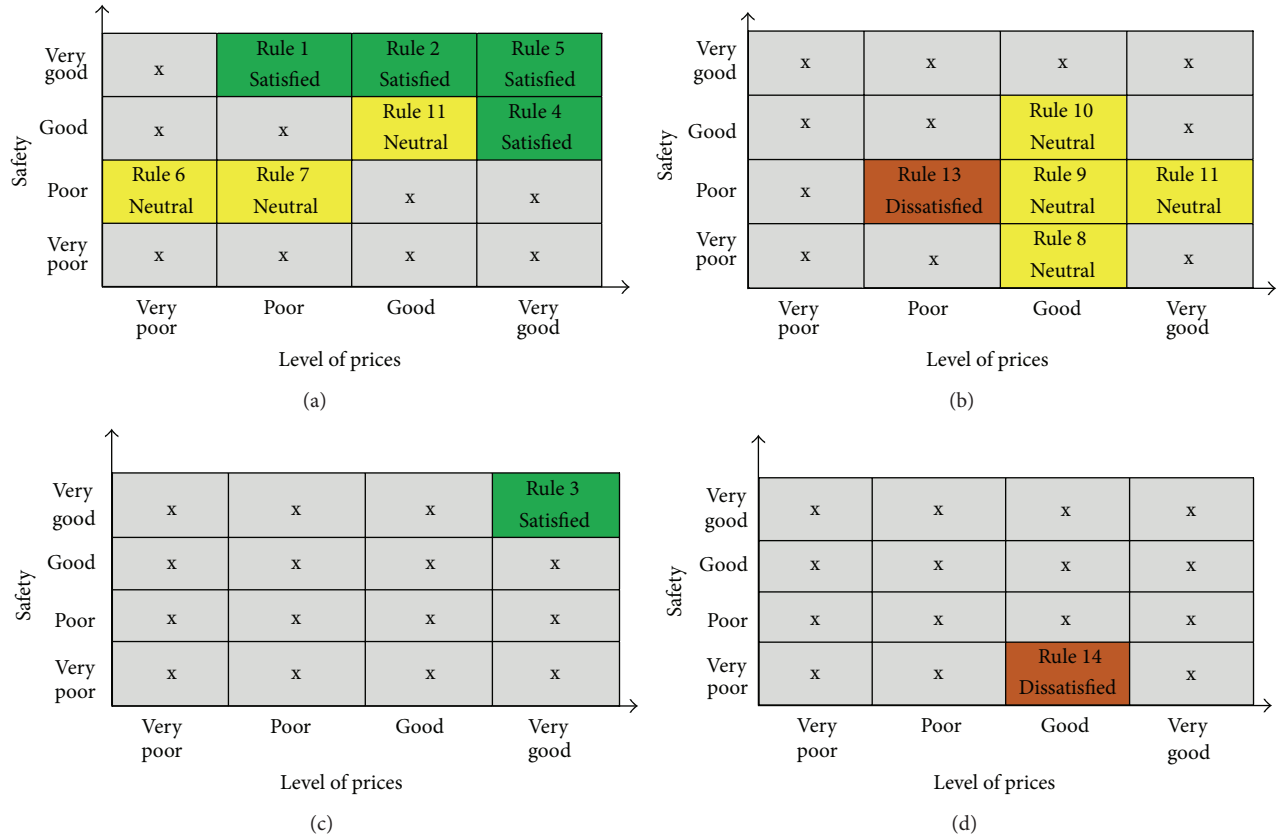


Figure 4: (a) Tourists above 30 years old/information and tourist services are very good. (b) Tourists above 30 years old/information and tourist services are good. (c) Tourists above 30 years old/information and tourist services are poor. (d) Tourists above 30 years old/information and tourist services are very poor.

information and tourist services. In other words, tourists above 30 years old need good information and tourist services when they select destinations for tour.

5. Conclusion

In this study, F7 (level of prices, living costs) and F10 (tourist safety) were found influential factors through fuzzy algorithm analysis [20]. From this research, a fuzzy rule database of tourism destinations is established to provide a fuzzy system inference decision-making model. This decision-making rule model can be provided to the tourism managers as a reference to establish tourism management. Tourism planners can use the ten attributes as a reference.

However, the budgets of some tourism destinations are often limited. This research simplified the ten constituent elements into two; in other words, two key attributes were found. While the budgets are limited, the tourism destinations could use the resource in the most crucial attributes to create comparatively large benefit.

From the rule analysis, it can be speculated that when tourists visit a tourism destination, they value "level of prices, living costs" (F7) and "tourist safety" (F10) of this area.

In order to investigate the tourist preference of different ages, the authors divided the data of tourists into two groups: one group is of tourists above 30 years old and the other group is of tourists of 30 years old and below. It was found that tourists of different ages showed their different preferences in three fields, namely, "level of prices, living costs" (F7), "information and tourist services" (F9), and "tourist safety" (F10). In other words, if the tourism industry would satisfy tourists' demands and preferences, especially for tourists of different ages, they have to focus on information and tourist services as well.

On the basis of the results of this study, it is shown that top management of tourism destinations should put resources in these fields first, in order to allow limited resources to perform to maximum effectiveness for the positive evaluations by tourists.

Lastly, this study still has parts that can be further researched or improved. In terms of the fuzzy linguistics, attribute F7 (level of prices, living costs) is of 3 levels, while attribute F10 (tourist safety) is of 4 levels, and 8 rules were produced. If other attributes such as tourists' age or gender are further considered, more focused rules will be obtained, which will assist in providing management of tourism destinations with more precise reference rules. At the same time, this can help decision-makers to make future development plans for tourism destinations that they manage, so as to cater to the preferences of different groups.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] M. F. Cracolici and P. Nijkamp, "The attractiveness and competitiveness of tourist destinations: a study of Southern Italian regions," *Tourism Management*, vol. 30, no. 3, pp. 336–344, 2009.

[2] W.-W. Wu, "Beyond Travel & Tourism competitiveness ranking using DEA, GST, ANN and Borda count," *Expert Systems with Applications*, vol. 38, no. 10, pp. 12974–12982, 2011.

[3] A. Kyriakidis, H. Hancock, S. Oaten, and R. Bashir, "Capturing the visitor economy: a framework for success," in *The Travel & Tourism Competitiveness Report 2009*, J. Blanke and T. Chiesa, Eds., pp. 65–77, World Economic Forum, Geneva, Switzerland, 2009.

[4] World Tourism Organization UNWTO, *UNWTO Global Summit on City Tourism 2011*, 2011, http://www.unwto.org/.

[5] World Travel Tourism Council, *Travel and Tourism 2011*, 2011, http://www.wttc.org/.

[6] I. Hwon, "Mining consumer attitude and behavior," *Journal of Convergence*, vol. 4, no. 2, pp. 29–35, 2013.

[7] G. Peng, K. Zeng, and X. Yang, "A hybrid computational intelligence approach for the VRP problem," *Journal of Convergence*, vol. 4, no. 2, pp. 1–4, 2013.

[8] G. I. Crouch and J. R. B. Ritchie, "Tourism, competitiveness, and societal prosperity," *Journal of Business Research*, vol. 44, no. 3, pp. 137–152, 1999.

[9] J. C. Augusto, V. Callaghan, D. Cook, A. Kameas, and I. Satoh, "Intelligent environments: a manifesto," *Human-Centric Computing and Information Sciences*, vol. 3, no. 1, p. 12, 2013.

[10] C. T. Lin and C. S. G. Lee, *Neural Fuzzy Systems*, Prentice-Hall, Singapore, 1999.

[11] M. Malkawi and O. Murad, "Artificial neuro fuzzy logic system for detecting human emotions," *Human-Centric Computing and Information Sciences*, vol. 3, article 3, 2013.

[12] O. P. Verma, V. Jain, and R. Gumber, "Simple fuzzy rule based edge detection," *Journal of Information Processing Systems*, vol. 9, no. 4, pp. 575–591, 2013.

[13] A. Matheison and G. Wall, *Tourism: Economic, Physical and Social Impacts*, Longman, New York, NY, USA, 1982.

[14] C. A. Gunn, *Tourism Planning*, 1988.

[15] C. R. Goeldner and J. R. B. Ritchie, *Tourism: Principles, Practices, Philosophies*, John Wiley & Sons, Hoboken, NJ, USA, 2006.

[16] L. Dwyer, P. Forsyth, and P. Rao, "The price competitiveness of travel and tourism: a comparison of 19 destinations," *Tourism Management*, vol. 21, no. 1, pp. 9–22, 2000.

[17] M. Lee, "Design of an intelligent system for autonomous groundwater management," *Journal of Convergence*, vol. 5, no. 1, pp. 26–31, 2014.

[18] E. Namsrai, T. Munkhdalai, M. Li, J.-H. Shin, O.-E. Namsrai, and K. H. Ryu, "A feature selection-based ensemble method for arrhythmia classification," *Journal of Information Processing Systems*, vol. 9, no. 1, pp. 31–40, 2013.

[19] M. Brahami, B. Atmani, and N. Matta, "Dynamic knowledge mapping guided by data mining: application on Healthcare," *Journal of Information Processing Systems*, vol. 9, no. 1, pp. 1–30, 2013.

[20] H. Binh and S. Ngo, "All capacities modular cost survivable network design problem using genetic algorithm with completely connection encoding," *Human-Centric Computing and Information Sciences*, vol. 4, no. 1, article 13, 2014.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 271374, 6 pages
http://dx.doi.org/10.1155/2015/271374

Research Article

A Study on Development of Engine Fault Diagnostic System

Hwa-seon Kim, Seong-jin Jang, and Jong-wook Jang

Department of Computer Engineering, Dong Eui University, 176 Eomgwangno Busan-jin-gu, Busan 614, Republic of Korea

Correspondence should be addressed to Hwa-seon Kim; doeunrain@naver.com

Received 1 July 2014; Accepted 27 October 2014

Academic Editor: Jong-Hyuk Park

This study implemented a mobile diagnosing system that provides user-centered interfaces for more precisely estimating and diagnosing engine conditions through communications with the self-developed ECU only for industrial CRDI engine use. For the implemented system, a new protocol was designed and applied based on OBD-II standard to receive engine data values of the developed ECU. The designed protocol consists of a message structure to request data transmission from a smartphone to ECU and a response message structure for ECU to send data to a smartphone. It transmits 31 pieces of engine condition information simultaneously and sends the trouble diagnostic code. Because the diagnostic system enables real-time communication through modules, the engine condition information can be checked at any time. Thus, because when troubles take place on the engine, users can check them right away, quick response and resolution are possible, and stable system management can be expected.

1. Introduction

ECU in CRDI system enables an engine to operate under optimal conditions by analyzing sensor information. The program and data parts of this ECU can be changed only by the manufacturer. Therefore, with respect to a diagnostic device for engines, it is not easy to use or understand the content without an expert. In order to solve these problems, this study independently developed ECU dedicated for an industrial CRDI engine. This study suggested an appropriate protocol for the developed ECU with the application of OBD-II standard in order to collect data values of the developed ECU and developed user-centered diagnostic devices for a mobile by receiving an input of data through communication after application.

The diagnostic devices provide user-centered diagnostic services and prevents accidents caused due to the engine malfunction by providing real-time communications with the use of wired system and Bluetooth module as a wireless system [1, 2] to transmit and receive engine fault diagnosis signals and sensor output signals and air pollution such as excessive gas exhaust and emission of incomplete combustion gas by controlling to operate an engine under the optimal conditions through the knocking diagnosis. Therefore, it is expected to contribute to eco industry which has received attention recently.

2. Development Engine Fault Diagnosis System

This study developed a mobile diagnostic system based on OBD-II for the industrial CRDI engine. Figure 1 shows that this system is able to verify engine information and existence of malfunction therein by Bluetooth communication between the ECUs using OBD-II protocol.

With creation of the mobile application for engine diagnostic system, an administrator may confirm automotive information in real time without other devices. Drivers may always check the status of engine with smartphone which they always carry and may promptly respond to the malfunction in engine if any occurs.

For this experiment, an automotive simulator with which communication test could be conducted in the same way as an actual car was developed and tested.

This study developed a mobile diagnostic system based on OBD-II for the industrial CRDI engine. This system is able to verify engine information and existence of malfunction therein by Bluetooth communication between the ECUs using OBD-II protocol.

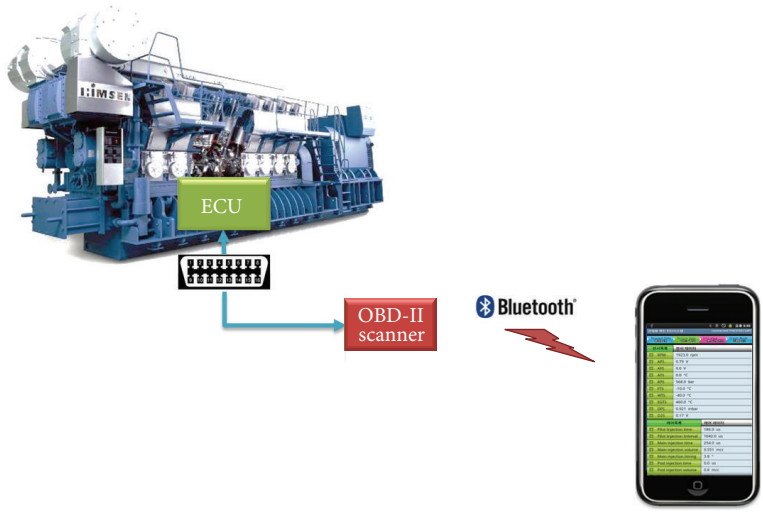


Figure 1: Diagnosis schematic diagram.

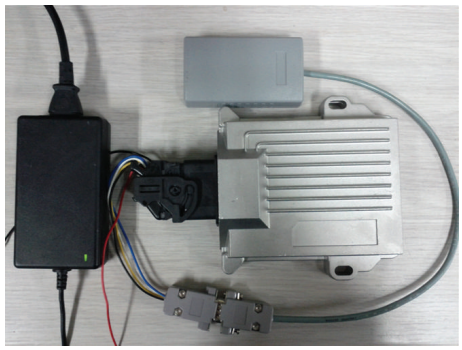


Figure 2: OBD-II simulator.

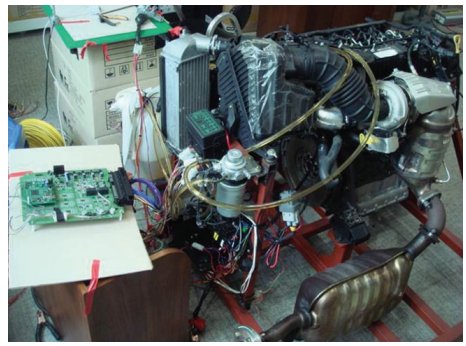


Figure 3: Apply the actual engine test.

With creation of the mobile application for engine diagnostic system, an administrator may confirm automotive information in real time without other devices. Drivers may always check the status of engine with smartphone which they always carry and may promptly respond to the malfunction in engine if any occurs.

This study developed a mobile diagnostic system based on OBD-II for the industrial CRDI engine. This system is able to verify engine information and existence of malfunction therein by Bluetooth communication between the ECUs using OBD-II protocol.

For this experiment, an automotive simulator with which communication test could be conducted in the same way as an actual car was developed and tested. Figure 2 shows the OBD-II simulator which was personally manufactured. This consists of dongles for Bluetooth communication and OBD-II connector. Figure 3 is the screen which the developed ECU is equipped in the actual engine.

2.1. Design of Bluetooth OBD-II Protocol Structure. The OBD-II is a standard that visualizes the information on main system of vehicles or on failure transmitted from sensors attached to a vehicle to ECU from a center console or external device by using the serial communication function [3].

Table 1: OBD-II message format.

Priority/type (1 byte)	Target address (1 byte)	Source address (1 byte)	Data byte (7 byte)	Checksum

The OBD-II message format consists of 1-byte priority, target address, source address header, 7-byte data, and checksum and is basically used as a protocol for SAE-J1850 and ISO [4–6]. The CAN OBD message format consists of ID bits (11 or 29), DLC, 7 data bytes, and checksum (CRC-15 processing method) [7–9]. Table 1 shows OBD-II message format [10–12].

The developed OBD-II protocol has been manufactured based on the existing OBD-II standard. This differs from the existing OBD-II standard protocol structure. In case of the OBD-II protocol standard, automotive information of only one PID which was requested and can be read and responded. However, the developed industrial automotive OBD-II protocol read all the automotive information and transmits such information at once.

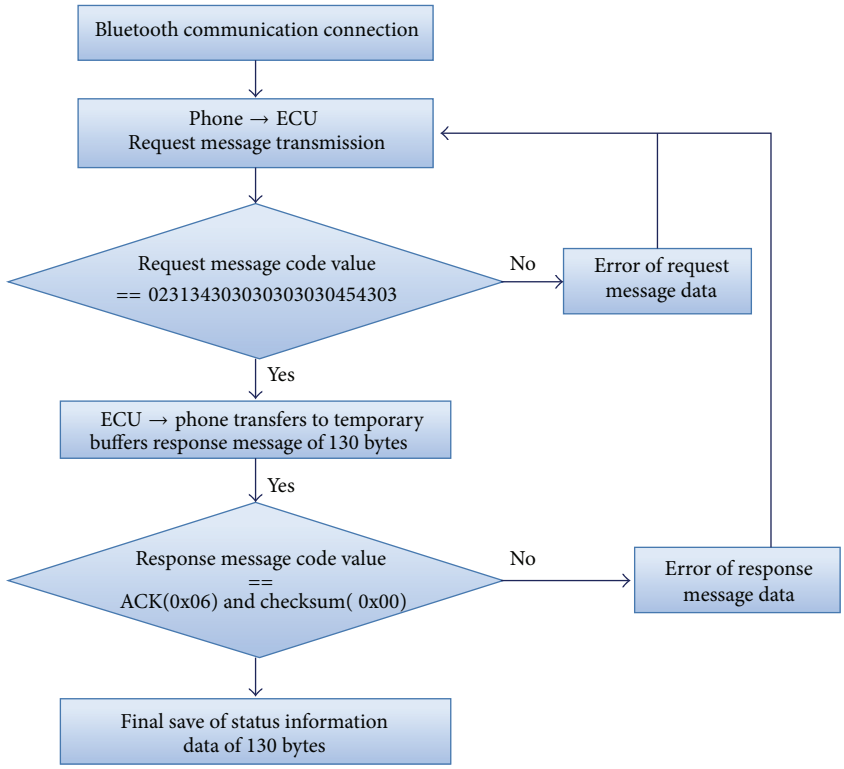


Figure 4: Flowchart of engine status information collection algorithm.

Table 2: Proposed OBD-II protocol request message structure.

Command STX	Command ID	Info.	Opt 1	Opt 2	Checksum	Command ETX

Table 3: Proposed OBD-II protocol response message structure.

Data STX	Data 1	⋯	Data 31	Checksum	Data ETX

2.1.1. OBD-II Protocol Structure for Obtaining Engine Status Information. The OBD-II message may be obtained from the automotive ECU by using automotive diagnostic tools. Figures 4 and 5 are proposed OBD-II protocol for this study. As seen in Table 2, the message consists of header, data, and checksum and saves 12 bytes of data in total and uses HEX codes.

In case of the proposed OBD-II protocol, it is designed to read the whole sensor information of vehicle by a response message at once when automotive information is requested to ECU. ECU provides 31 types of sensor information which actual service centers practically use. Table 3 shows the proposed OBD-II protocol response message structure.

Table 4 is detailed code of OBD-II protocol status information request message.

2.1.2. OBD-II Protocol Structure for Obtaining Engine Trouble Code. There is a function to inform drivers that there is malfunction in the electronic control engine by lighting up the malfunction indicator lamp (MIL) and to set diagnostic trouble code (DTC) according to the details of malfunction and to automatically record such codes in the RAM of ECU if there is malfunction in electronic control engine or in exhaust gas related parts [13, 14]. This function was originally to set the OBD in order to easily verify the location to be inspected if automotive malfunction occurs but, thanks to speedy development of computers, it came to play a role of conducting ready-test (monitoring of exhaust gas equipment) as well as making freeze frame (function to record DTC on ECU) when malfunction occurs in input and output of ECU (computer) [15–17].

Table 4: Detailed code of OBD-II protocol status information request message.

Content	BYTE information	HEX code
Command STX	0x02	0x02
Command ID	0x14	0x31
		0x34
Info.	0x00	0x30
		0x30
Opt 1	0x00	0x30
		0x30
Opt 2	0x00	0x30
		0x30
Checksum	0xEC	0x45
		0x43
Command ETX	0x03	0x03

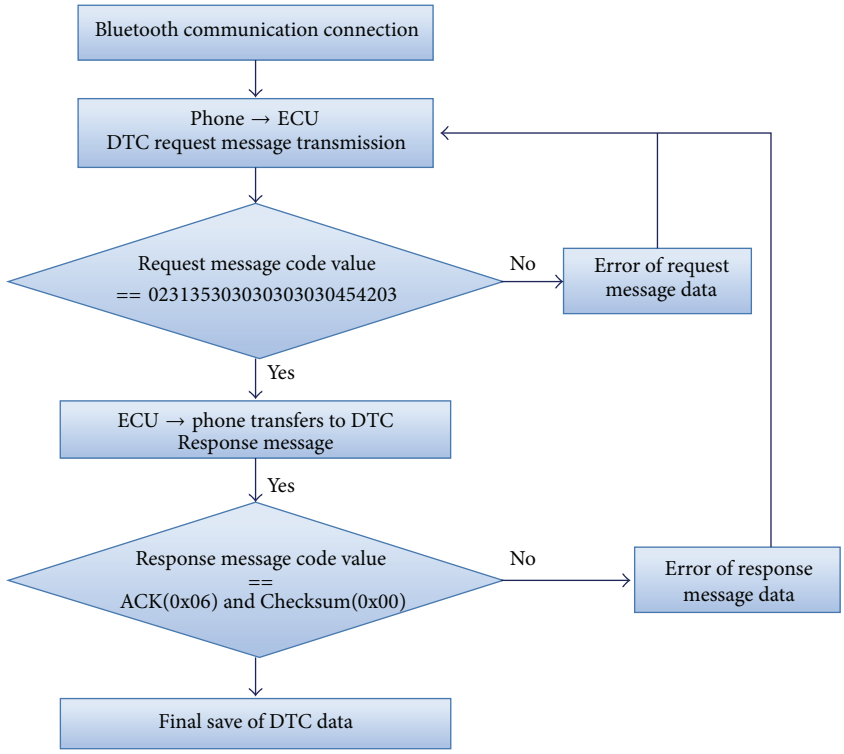


Figure 5: Flowchart of diagnostic trouble code collection algorithm.

Table 5: ECU DTC response message structure.

Command STX	Command ID	MODE	DTC code	Checksum	Command ETX

Therefore, a self-diagnosis function is the priority to be inspected when malfunction occurs in the car equipped with electronic control engine.

If disconnection or short circuit occurs in a sensor or actuator, ECU makes a comparison with preset voltage value and judges existence of malfunction and memorizes the preset DTC on the RAM. Such DTC information is memorized on the RAM of ECU (computer), so unless power provided for ECU is cut, such information is not deleted.

Table 5 shows the structure of DTC response message of ECU and Table 6 shows detailed code of OBD-II protocol DTC response message.

Table 6: Detailed code of OBD-II protocol DTC reponse message.

Content	BYTE information	HEX code
Command STX	0x02	0x02
Command ID	0x15	0x31
		0x35
Info.	0x00	0x30
		0x30
Opt 1	0x00	0x30
		0x30
Opt 2	0x00	0x30
		0x30
Checksum	0xEB	0x45
		0x42
Command ETX	0x03	0x03

2.2. ECU Information Collection Algorithm

2.2.1. Algorithm for Obtaining Engine Status Information. In order to collect the information of ECU, data are transmitted through the process as described in Figure 4.

First of all, if Bluetooth communication is connected, a data request message is transmitted to ECU. If the input request message is identical to 023134303030303030454303, ECU transmits OBD-II response message of 130 bytes to temporary buffers. All the data between STX and ETX are converted into HEX codes and sent. The calculation of checksum is of longitudinal redundancy check (LRC) and the lower byte of the sum of data between STX and ETX and checksum should be zero. If the response message is ACK(0x06) and the checksum value is 0x00, 31 automotive data of 130 bytes are finally saved.

2.2.2. Algorithm for Obtaining Engine Diagnostic Code. In order to collect DTC of ECU, such codes are transmitted through the process as described in Figure 5.

Collection process of ECU diagnostic trouble codes is similar to the automotive information collection algorithm as explained above. First of all, if Bluetooth communication is connected, data request message is sent to ECU. If the input

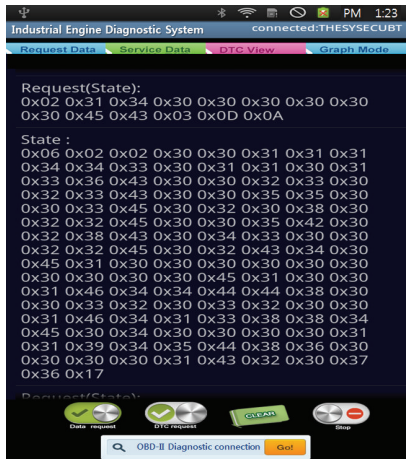


Figure 6: Status information communication between devices.

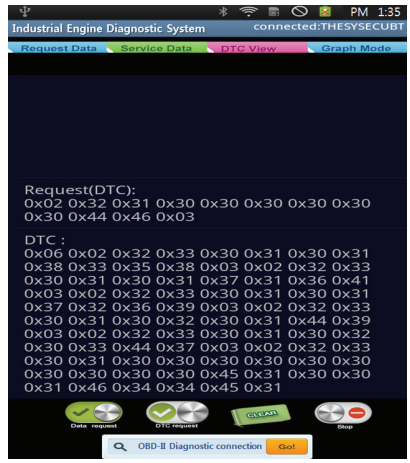


Figure 8: DTC information transmission between devices.

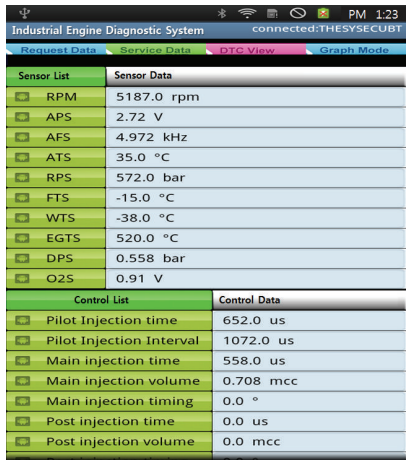


Figure 7: Status information.

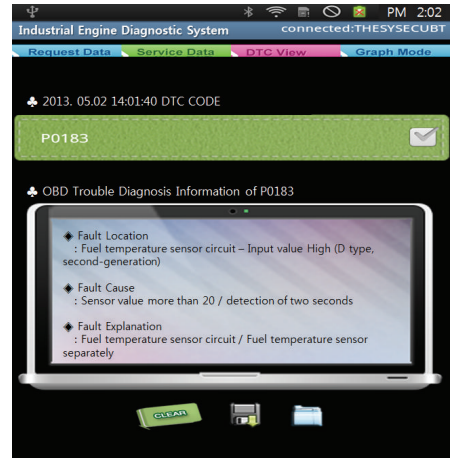


Figure 9: DTC information.

request message is identical to 023135303030303030454203, ECU transmits OBD-II response message of 14 bytes to temporary buffers. Then, if response message is ACK(0x06) and the checksum is 0x00, automotive information of 14 bytes becomes finally saved.

3. Bluetooth Mobile Application Software for OBD-II Protocol Diagnosis

The following shows a screen used for communication between devices. If clicking the button named "Data Request" in order for application to request the connected device for automotive status information, the connected device transmits the status information to the application. Figure 6 shows status information communication between devices. If phone orders ECU to read status information through communication protocol, ECU sends out a data response. In order to disconnect the communication between devices for a while, if clicking the "Stop" button, such intercommunication becomes suspended. If clicking the "Clear" button, the screen showing transmission of codes becomes wiped out. In order to reconnect the communication, if clicking the "Request" button, the communication between devices restarts. The communication between devices is made with HEX codes, so it is difficult for a user to verify the information instinctively. In order to make a user promptly understand the status information, such information was made through parsing process so that it may be shown on the screen as in Figure 7.

Figure 8 is a screen showing DTC information transmission between devices. When malfunction occurs in ECU, the concerned trouble codes are searched and if the data which are identical to codes saved in DB exist, information related to such codes are notified. If DTC is found, such found DTC is shown on the screen as in Figure 9. If clicking DTC, the concerned DTC information is shown on the screen.

4. Conclusion

In this study, with OBD-II protocol, a smart phone engine diagnostic system using Bluetooth communication was developed. In this study, instead of handling information that can be controlled only by manufacturing companies, it was made possible to select necessary information only and take control at first hand.

It is unnecessary to passively receive and deal with all the data including even needless one, so the administrator may handle information satisfying his needs only. Therefore, with this system it was made possible that information of engine condition may be identified in real time and that if engine has malfunction, by notifying diagnostic trouble codes and information, the user and administrator may promptly respond to such malfunction. In other words, the diagnostic devices provide user-centered diagnostic services and prevents accidents caused due to the engine malfunction by providing real-time communications with the use of wired system and Bluetooth module as a wireless system to transmit and receive engine fault diagnosis signals and sensor output signals and air pollution such as excessive exhaust gas emission and emission of incomplete combustion gas by controlling to operate an engine under the optimal conditions through the knocking diagnosis. Therefore, it is expected to contribute to eco industry which has received attention recently.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the Brain Busan 21 Project in 2014 and Dong-eui University Grant (no. 2014AA324).

References

[1] A. Sinha and D. K. Lobiyal, "Performance evaluation of data aggregation for cluster-based wireless sensor network," *Human-Centric Computing and Information Sciences*, vol. 3, article 13, 2013.

[2] H.-R. Lee, K.-Y. Chung, and K.-S. Jhang, "A study of wireless sensor network routing protocols for maintenance access hatch condition surveillance," *Journal of Information Processing Systems*, vol. 9, no. 2, pp. 237–246, 2013.

[3] HK-e car, http://www.hke-car.com/.

[4] D. J. Oliver, "Implementing the J1850 protocol," http://smartdata.usbid.com/datasheets/usbid/2000/2000-q4/j1850_wp.pdf.

[5] CAN bus, http://en.wikipedia.org/wiki/CAN_bus.

[6] Bosch, "CAN Specification Version 2.0," Bosch, 1991.

[7] Open source project using OBD-II, http://www.opendiag.org/.

[8] OBD-II, http://www.obdii.com/.

[9] OBD-II PIDs, http://en.wikipedia.org/wiki/OBD-II_PIDs.

[10] ISO-14230, Road vehicless-Diagnostics systems-Keyword Protocol 2000.

[11] ISO-15765, Road vehicles-Diagnostics on Controller Area Network.

[12] M. Yoon, Y.-K. Kim, and J.-W. Chang, "An energy-efficient routing protocol using message success rate in wireless sensor networks," *Journal of Convergence*, vol. 4, no. 1, pp. 15–22, 2013.

[13] J. Kin, S.-C. Chen, Y.-T. Shih, and S.-H. Chen, "A study on remote on-line diagnostic system for vehicles by integrating the technology of OBBD, GPS, and 3G," *World Academy of Science, Engineering and Technology*, vol. 56, 2009.

[14] B. Lee, "OBD-II(exhaust)," Kyung Young sa, 2005.

[15] On-Board Diagnostics, http://en.wikipedia.org/wiki/OBD-II#OBD-II.

[16] A. I. Santini, *OBD-II: Function, Monitors and Diagnostic Techniques*, Cengage Learning, 2010.

[17] ISO 14230-4, Road Vehicles-Diagnostic Systems, https://law.resource.org/pub/us/cfr/ibr/004/iso.14230-4.2000.pdf.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 538613, 14 pages
http://dx.doi.org/10.1155/2015/538613

Research Article

Missing Values and Optimal Selection of an Imputation Method and Classification Algorithm to Improve the Accuracy of Ubiquitous Computing Applications

Jaemun Sim,[1] Jonathan Sangyun Lee,[2] and Ohbyung Kwon[2]

[1] *SKKU Business School, Sungkyunkwan University, Seoul 110734, Republic of Korea*
[2] *School of Management, Kyung Hee University, Seoul 130701, Republic of Korea*

Correspondence should be addressed to Ohbyung Kwon; obkwon@khu.ac.kr

Received 18 June 2014; Revised 29 September 2014; Accepted 11 October 2014

Academic Editor: Jong-Hyuk Park

In a ubiquitous environment, high-accuracy data analysis is essential because it affects real-world decision-making. However, in the real world, user-related data from information systems are often missing due to users' concerns about privacy or lack of obligation to provide complete data. This data incompleteness can impair the accuracy of data analysis using classification algorithms, which can degrade the value of the data. Many studies have attempted to overcome these data incompleteness issues and to improve the quality of data analysis using classification algorithms. The performance of classification algorithms may be affected by the characteristics and patterns of the missing data, such as the ratio of missing data to complete data. We perform a concrete causal analysis of differences in performance of classification algorithms based on various factors. The characteristics of missing values, datasets, and imputation methods are examined. We also propose imputation and classification algorithms appropriate to different datasets and circumstances.

1. Introduction

Ubiquitous computing has been the central focus of research and development in many studies; it is considered to be the third wave in the evolution of computer technology [1]. In ubiquitous computing, data must be collected and analyzed accurately in real time. For this process to be successful, data must be well organized and uncorrupted. Data preprocessing is an essential but time- and effort-consuming step in the process of data mining. Several preprocessing methods have been developed to overcome data inconsistencies [2].

Data incompleteness due to missing values is very common in datasets collected in real settings [3]; it presents a challenge in the data preprocessing phase. Data is often missing when user input is required. For example, in human-centric computing, systems often require user profile data for the purpose of personalization [4]. In the case of Twitter, text data is used for sentiment analysis in order to analyze user behaviors and attitudes [5]. As a final example, in ubiquitous commerce, customer data has been used to personalize services for users [6]. Values may be missing when users are reluctant to provide their personal data due to privacy concerns or lack of motivation. This is especially true for optional data requested by the system.

Missing values can also be present in sensor data. Sensor data is usually in quantitative form. Sensors provide physical information regarding temperature, sound, or trajectory. Sensor technology has advanced over the years; it is an essential source of data for ubiquitous computing and is used for situation awareness and circumstantial decision-making. For example, human interaction sensors read and react to current situations [7]. Analysis of image files for face recognition and object detection using sensors is widely used in ubiquitous computing [8]. However, incorrect data and missing values are possible even using advanced sensor technology due to mechanical and network errors. Missing values can interfere with decision-making and personalization, which can ultimately lead to user dissatisfaction. In many cases, the impact of missing data is costly to users of data analysis methods such as classification algorithms.

Data incompleteness may have negative effects on data preprocessing and decision-making accuracy. Extra time and effort are required to compensate for missing data. Using uncertain or null data results in fatal errors in the classification algorithm, and deleting all records that contain missing data (i.e., using the listwise deletion method) reduces the sample size, which might decrease statistical power and introduce potential bias to the estimation [9]. Finally, unless the researcher can be sure that the data values are missing completely at random (MCAR), then the conclusions resulting from a complete-case analysis are most likely to be biased.

In order to overcome issues related to data incompleteness, many researchers have suggested methods of supplementing or compensating for missing data. The missing data imputation method is the most frequently used statistical method developed to deal with missing data problems. It is defined as "a procedure that replaces the missing values in a dataset by some plausible values" [3]. Missing values occur when no data is stored for a given variable in the current observation.

Many studies have attempted to validate the missing data imputation method of supplementing or compensating for missing data by testing it with different types of data; other studies have attempted to develop the method further. Studies have also compared the performance of various imputation methods based on benchmark data. For example, Kang investigated the ratio of missing to complete data in various datasets and compared the average accuracy of several imputation methods, such as MNR, k-NN, CART, ANN, and LLR [10]. The results demonstrated that k-NN performed best on datasets with less than 10% of data missing and LLR performed best on those with more than 10% of data missing.

However, after multiple tests using complete datasets, not much difference in performance was observed, and some datasets were linearly inferior. In Kang's study [10], many datasets with equivalent conditions yielded different results. Thus, the fit between the dataset characteristics and the imputation method must also be considered. Previous studies have compared imputation methods by varying the ratio of missing to complete data or evaluating performance differences between complete and incomplete datasets. However, the reasons for these different results between datasets under equivalent conditions remain unexplained. Various factors may affect the performance of classification algorithms. For example, the interrelationship or fitness between the dataset, imputation method, and characteristics of the missing values may be important to the success or failure of the analytical process.

The purpose of this study is to examine the influence of dataset characteristics and patterns of missing data on the performance of classification algorithms using various datasets. The moderating effects of different imputation methods, classification algorithms, and data characteristics on performance are also analyzed. The results are important because they can suggest which imputation method or classification algorithm to use depending on the data conditions. The goal is to improve the performance, accuracy, and time required for ubiquitous computing.

2. Treating Datasets Containing Missing Data

Missing information is an unavoidable aspect of data analysis. For example, responses may be missing to items on survey instruments intended to measure cognitive and affective factors. Various imputation methods have been developed and used for treatment of datasets containing missing data. Some popular methods are listed below.

(1) Listwise Deletion. Listwise deletion (LD) involves the removal of all individuals with incomplete responses for any items. However, LD reduces the effective sample size (sometimes greatly, resulting in large amounts of missing data), which can, in turn, reduce statistical power for hypothesis testing to unacceptably low levels. LD assumes that the data are MCAR (i.e., their omission is unrelated to all measured variables). When the MCAR assumption is violated, as is often the case in real research settings, the resulting estimates will be biased.

(2) Zero Imputation. When data are omitted as incorrect, the zero imputation method is used, in which missing responses are assigned an incorrect value (or zero in the case of dichotomously scored items).

(3) Mean Imputation. In this method, the mean of all values within the same attribute is calculated and then imputed in the missing data cells. The method works only if the attribute examined is not nominal.

(4) Multiple Imputations. Multiple imputations can incorporate information from all variables in a dataset to derive imputed values for those that are missing. This method has been shown to be an effective tool in a variety of scenarios involving missing data [11], including incomplete item responses [12].

(5) Regression Imputation. The linear regression function is calculated from the values within the same attribute and then used as the dependent variable. The other attributes (except the decision attribute) are then used as independent variables. Then the estimated dependent variable is imputed in the missing data cells. This method works only if all considered attributes are not nominal.

(6) Stochastic Regression Imputation. Stochastic regression imputation involves a two-step process in which the distribution of relative frequencies for each response category for each member of the sample is first obtained from the observed data.

In this paper, the details of the seven imputation methods used herein are as follows.

(i) Listwise Deletion. All instances are deleted that contain more than one missing cell in their attributes.

(ii) Mean Imputation. The missing values from each attribute (column or feature) are replaced with the mean of all known values of that attribute. That is, let X_i^j be the jth missing attribute of the ith instance, which is imputed by

$$X_i^j = \sum_{k \in I(\text{complete})} \frac{X_k^j}{n_{|I(\text{complete})|}}, \quad (1)$$

where I(complete) is a set of indices that are not missing in X_i and $n_{|I(\text{complete})|}$ is the total number of instances where the jth attribute is not missing.

(iii) Group Mean Imputation. The process for this method is the same as that for mean imputation. However, the missing values are replaced with the group (or class) mean of all known values of that attribute. Each group represents a target class from among the instances (recorded) that have missing values. Let $X_{m,i}^j$ be the jth missing attribute of the ith instance of the mth class, which is imputed by

$$X_{m,i}^j = \sum_{k \in I(m\text{th class incomplete})} \frac{X_{m,k}^j}{n_{|I(m\text{th class incomplete})|}}, \quad (2)$$

where I(mth class incomplete) is a set of indices that are not missing in $X_{m,i}^j$ and $n_{|I(m\text{th class incomplete})|}$ is the total number of instances where the jth attribute of the mth class is not missing.

(iv) Predictive Mean Imputation. In this method, the functional relationship between multiple input variables and single or multiple target variables of the given data is represented in the form of a linear equation. This method sets attributes that have missing values as dependent variables and other attributes as independent variables in order to allow prediction of missing values by creating a regression model using those variables. For a regression target y_i, the MLR equation with d predictors and n training instances can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_d x_{id} \quad \text{for } i = 1, \ldots, n. \quad (3)$$

This can be rewritten in matrix form such that $y = X\beta$, and the coefficient β can be obtained explicitly by taking a derivative of the squared error function as follows:

$$\begin{aligned} \min E(\beta) &= \frac{1}{2}(y - X\beta)^T (y - X\beta), \\ \frac{\partial E(\beta)}{\partial \beta} &= X^T X\beta - X^T y = 0, \\ \beta &= \left(X^T X\right)^{-1} \cdot X^T y. \end{aligned} \quad (4)$$

(v) Hot-Deck. This method is the same in principle as case-based reasoning. In order for attributes that contain missing values to be utilized, values must be found from among the most similar instances of nonmissing values and used to replace the missing values. Therefore, each missing value is replaced with the value of an attribute with the most similar instance as follows:

$$X_i^j = X_k^j, \quad k = \arg\min_P \sqrt{\sum_{j \in I(\text{complete})} \text{Std}_j \left(X_i^j - X_P^j\right)^2}, \quad (5)$$

where Std_j is the standard deviation of the jth attribute which is not missing.

(vi) k-NN. Attributes are found via a search among nonmissing attributes using the 3-NN method. Missing values are imputed based on the values of the attributes of the k most similar instances as follows:

$$X_i^j = \sum_{P \in k\text{-NN}(X_i)} k\left(X_i^{I(\text{complete})}, X_P^{I(\text{complete})}\right) \cdot X_P^j, \quad (6)$$

where k-NN(X_i) is the index set of the kth nearest neighbors of X_i based on the nonmissing attributes and $k(X_i, X_j)$ is a kernel function that is proportional to the similarity between the two instances X_i and X_j ($k = 4$).

(vii) k-Means Clustering. Attributes are found through formation of k-clusters from nonmissing data, after which missing values are imputed. The entire dataset is partitioned into k clusters by maximizing the homogeneity within each cluster and the heterogeneity between clusters as follows:

$$\arg\min_{C^{h(\text{complete})}} \sum_{i=1}^{k} \sum_{X_j^{I(\text{complete})} \in C_i^{h(\text{complete})}} \left\| X_j^{I(\text{complete})} - C_i^{I(\text{complete})} \right\|^2, \quad (7)$$

where $C_i^{I(\text{complete})}$ is the centroid of $C_i^{I(\text{complete})}$ and $C^{I(\text{complete})}$ is the union of all clusters ($C^{I(\text{complete})} = C_1^{I(\text{complete})} \cup \cdots \cup C_k^{I(\text{complete})}$). For a missing value X_i^j, the mean value of the attribute for the instances in the same cluster with $X_i^{I(\text{complete})}$ is imputed thus as follows:

$$\begin{aligned} X_i^j &= \frac{1}{\left|C_k^{I(\text{complete})}\right|} \cdot \sum_{X_P^{I(\text{complete})} \in C_k^{I(\text{complete})}} X_P^j \\ \text{s.t. } k &= \arg\min_i \left| X_j^{I(\text{complete})} - C_i^{I(\text{complete})} \right|. \end{aligned} \quad (8)$$

3. Model

In this paper, we hypothesize an association between the performance of classification algorithms and the characteristics of missing data and datasets. Moreover, we assume that the chosen imputation method moderates the causality between these factors. Figure 1 illustrates the posited relationships.

3.1. Missing Data Characteristics. Table 1 describes the characteristics of missing data and how to calculate them. The pattern of missing data characteristics may be univariate, monotone, or arbitrary [11]. A univariate pattern of missing data occurs when missing values are observed for a single variable only; all other data are complete for all variables.

Table 1: The characteristics of missing data.

Variables	Meaning	Calculation
Missing data ratio	The number of missing values in the entire dataset as compared to the number of nonmissing values	The number of empty data cells/total cells
Patterns of missing data	Univariate Monotone Arbitrary	Ratio of missing to complete values for an existing feature compared to the values for all features
Horizontal scatteredness	Distribution of missing values within each data record	Determine the number of missing cells in each record and calculate the standard deviation
Vertical scatteredness	Distribution of missing values for each attribute	Determine the number of missing cells in each feature and calculate the standard deviation
Missing data spread	Larger standard deviations indicate stronger effects of missing data	Determine the weighted average of the standard deviations of features with missing data (weight: the ratio of missing to complete data for each feature)

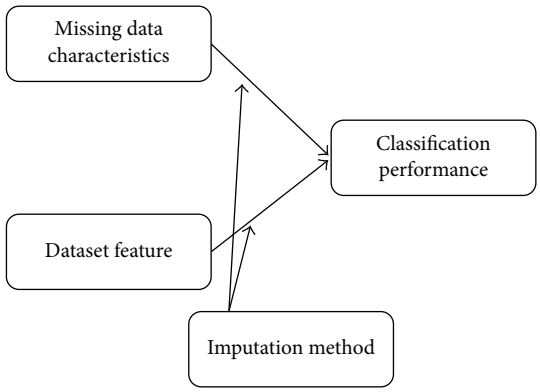


Figure 1: Research model.

A monotone pattern occurs if variables can be arranged such that all $Y_{j+1}, \ldots, Y_k$ are missing for cases where Y_j is missing. Another characteristic, missing data spread, is important because larger standard deviations for missing values within an existing feature indicate that the missing data has greater influence on the results of the analysis (Figure 2).

3.2. Dataset Features. Table 2 lists the features of datasets. Based on the research of Kwon and Sim [15], in which characteristics of datasets that influence classification algorithms were identified, we considered the following statistically significant features in this study: missing values, the number of cases, the number of attributes, and the degree of class imbalance. However, the discussion of missing values is omitted here because it has already been analyzed in detail by Kwon and Sim [15].

3.3. Imputation Methods. Table 3 lists the imputation methods used in this study. Since datasets with categorical decision attributes are included, imputation methods that do not accommodate categorical attributes (e.g., regression imputation) are excluded from this paper.

Table 2: Dataset features.

Variables	Description
Number of cases	Number of records in the dataset
Number of attributes	Number of features characteristic of the dataset
Degree of class imbalance	Ratio

3.4. Classification Algorithms. Many studies have compared classification algorithms in various areas. For example, the decision tree is known as the best algorithm for arrhythmia classification [16]. In Table 4, six types of representative classification algorithms for supervised learning are described: C4.5, SVM (support vector machine), Bayesian network, logistic classifier, k-nearest neighbor classifier, and regression.

4. Method

We conducted a performance evaluation of the imputation methods and classification algorithms described in the previous section using actual datasets taken from the UCI dataset archive. To ensure the accuracy of each method in cases with no missing values, datasets with missing values were not included. Among the selected datasets, six (Iris, Wine, Glass, Liver Disorder, Ionosphere, and Statlog Shuttle) were included for comparison with the results of Kang [10]. These datasets are popular and frequently utilized benchmarks in the literature, which makes them useful for demonstrating the superiority of the proposed idea.

Table 5 provides the names of the datasets, the numbers of cases, and the descriptions of features and classes. The numbers in parentheses in the last two columns represent the number of features and classes for the decision attributes. For example, in dataset Iris, "Numeric (4)" indicates that there are four numeric attributes, and "Categorical (3)" means that there are three classes in the decision attribute.

Since UCI datasets have no missing data, target values in each dataset were randomly omitted [10]. Based on

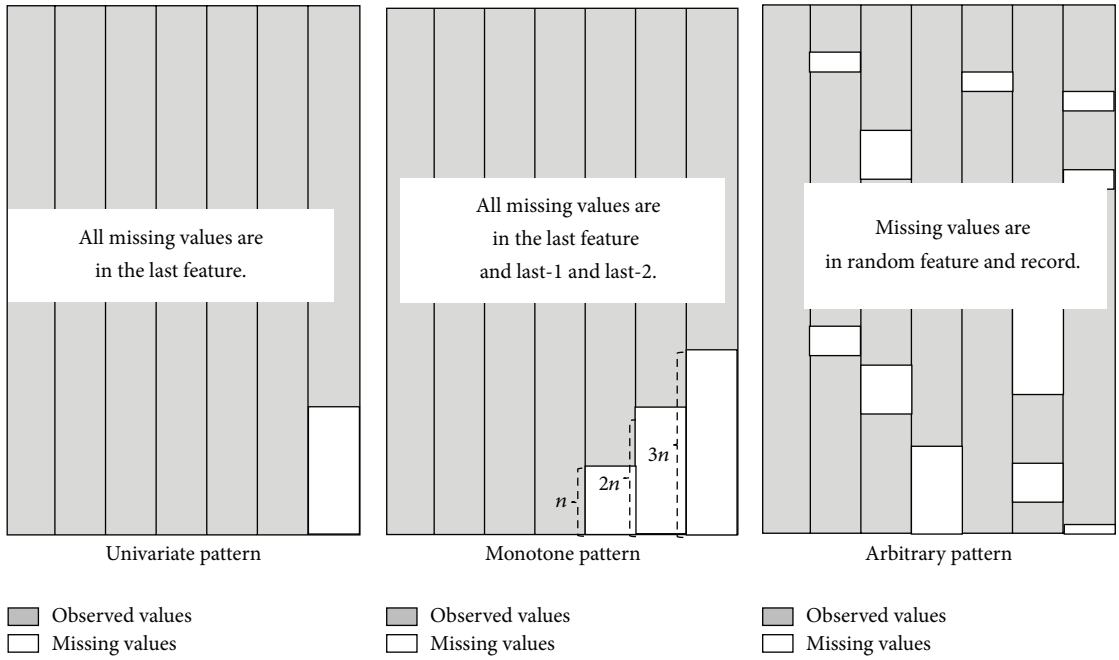


Figure 2: Missing data patterns.

Table 3: Imputation methods.

Imputation methods	Description
Listwise deletion	Perhaps the most basic traditional technique for dealing with missing data. Cases with missing values are discarded, restricting the analyses to cases for which complete data are available.
Mean imputation	Involves replacing missing data with the overall mean for the observed data.
Group mean imputation	A missing value is replaced by the mean of a subset of the data, based on other observed variable(s) in the data.
Predictive mean imputation	Also called regression imputation. Predictive mean imputation involves imputing a missing value using an ordinary least-squares regression method to estimate missing data.
Hot-deck	Most similar records are imputed to missing values.
k-NN	The attribute value of k is imputed to the most similar instance from nonmissing data.
k-means clustering	k numbers of sets are created that are homogeneous on the inside and heterogeneous on the outside.

Table 4: Classification algorithms.

Algorithms	Description
C4.5	Estimates the known data using learning rules. C4.5 gradually expands the conditions of the algorithm, splitting the upper node into subnodes using a divide-and-conquer method until it comes to the end node.
SVM	Classifies the unknown class by finding the optimal hyperplane with the maximum margin that reduces the estimation error.
Bayesian network	A probability network with a high posterior probability given the instances. Such a network can provide insight into probabilistic dependencies among the variables in the training dataset.
Logistic classifier	Takes the functional form of logistic CDF (cumulative distribution function). This function relates the probability of some event to attribute variables through regression coefficients and alpha and beta parameters, which are estimated from training data [13].
k-nearest neighbor classifier	Simple instance-based learner that uses the class of the nearest k training instances for the class of the test instances.
Regression	The class is binarized, and one regression model is built for each class value [14].

Table 5: Datasets used in the experiments.

Dataset	Number of cases	Features	Decision attributes
Iris	150	Numeric (4)	Categorical (3)
Wine	178	Numeric (13)	Categorical (3)
Glass	214	Numeric (9)	Categorical (7)
Liver disorder	345	Numeric (6)	Categorical (2)
Ionosphere	351	Numeric (34)	Categorical (2)
Statlog Shuttle	57,999	Numeric (7)	Categorical (7)

the list of missing data characteristics, three datasets with three different missing data ratios (5%, 10%, and 15%) and three sets representing each of the missing data patterns (univariate, monotone, and arbitrary) were created for a total of nine variations for each dataset. In total, 54 datasets were imputed for each imputation method, as 6 datasets were available. We repeated the experiment for each dataset 1000 times in order to minimize errors and bias. Thus, 5,400 datasets were imputed in total for our experiment. All imputation methods were implemented using packages written in Java. In order to measure the performance of each imputation method, we applied imputed datasets to the six classification algorithms listed in Table 4.

There are various indicators to measure performance, such as accuracy, relative accuracy, MAE (mean absolute error), and RMSE (root mean square error). However, RMSE is one of the most representative and widely used performance indicators in the imputation research. Therefore, we also adopted RMSE as the performance indicator in this study. The performance of the selected classification algorithms was evaluated using SPSS 17.0.

RMSE measures the difference between predicted and observed values. The term "relative prediction accuracy" refers to the relative ratio of accuracy, which is equivalent to 1 when there are no missing data [10]. The no-missing-data condition was used as a baseline of performance. As the next step, we generated a missing dataset from the original no-missing-dataset and then applied an imputation method to replace the null data. Then a classification algorithm was conducted to estimate the results of the imputed dataset. With all combinations of imputation methods and classification algorithms, a multiple regression analysis was conducted using the following equation to understand the input factors, the characteristics of missing data, and those of the datasets, in order to determine how the selected classification algorithms affected performance:

$$y_p = \sum_{\forall j \in M} \beta_{pj} x_j + \sum_{\forall k \in D} \chi_{pk} z_k + \varepsilon_p. \quad (9)$$

In this equation, x is the value of the characteristics of the missing data (M), z is the value of each dataset's characteristics in the set of dataset (D), and y is a performance parameter. Note that M = {missing data ratio, patterns of missing data, horizontal scatteredness, vertical scatteredness, missing data spread} and D = {number of cases, number of attributes, degree of class imbalance}. In addition, $p = 1$ indicates relative prediction accuracy, $p = 2$ represents RMSE, and $p = 3$ means elapsed time. We performed the experiment using the Weka library source software (release 3.6) to determine the reliability of the implementation of the algorithms [17]. We did not use the Weka GUI tool but developed a Weka library-based performance evaluation program in order to conduct the automatized experiment repeatedly.

5. Results

In total, 32,400 datasets (3 missing ratios × 3 imputation patterns × 6 imputation methods × 100 trials) were imputed for each of the 6 classifiers. Thus, in total, we tested 226,800 datasets (32,400 imputed dataset × 7 classifier methods). The results were divided by those for each dataset, classification algorithm, and imputation method for comparison in terms of performance.

5.1. Datasets. Figure 3 shows the performance of each imputation method for the six different datasets. On the x-axis, three missing ratios represent the characteristics of missing data, and on the y-axis, performance is indicated using the RMSE. All results of three different variations of the missing data patterns and tested classification algorithms were merged for each imputation method.

For Iris data (Figure 3(a)), the mean imputation method yielded the worst results and the group mean imputation method the best results.

For Glass Identification data (Figure 3(b)), hot-deck imputation was the least effective method and predictive mean imputation was the best.

For Liver Disorder data (Figure 3(c)), k-NN was the least effective, and once again, the predictive mean imputation method yielded the best results.

For Ionosphere data (Figure 3(d)), hot-deck was the worst and k-NN the best.

For Wine data (Figure 3(e)), hot-deck was once again the least effective method, and predictive mean imputation the best.

For Statlog data (Figure 3(f)), unlike the other datasets, the results varied based on the missing data ratio. However, predictive mean imputation was still the best method overall and hot-deck the worst.

Figure 3 illustrates that the predictive mean imputation method yielded the best results overall and hot-deck imputation the worst. However, no imputation method was generally superior in all cases with any given dataset. For example, the k-NN method yielded the best performance for the Ionosphere dataset, but for the Liver Disorders dataset, its performance was lowest. In another example, the group mean imputation method performed best for the Iris and Wine datasets, but its performance was only average for other datasets. Therefore, the results were inconsistent, and determining the best imputation method is impossible. Thus, the imputation method cannot be used as an accurate predictor of performance. Rather, the performance must be influenced by other factors, such as the interaction between the characteristics of the dataset in terms of missing data and the chosen imputation method.

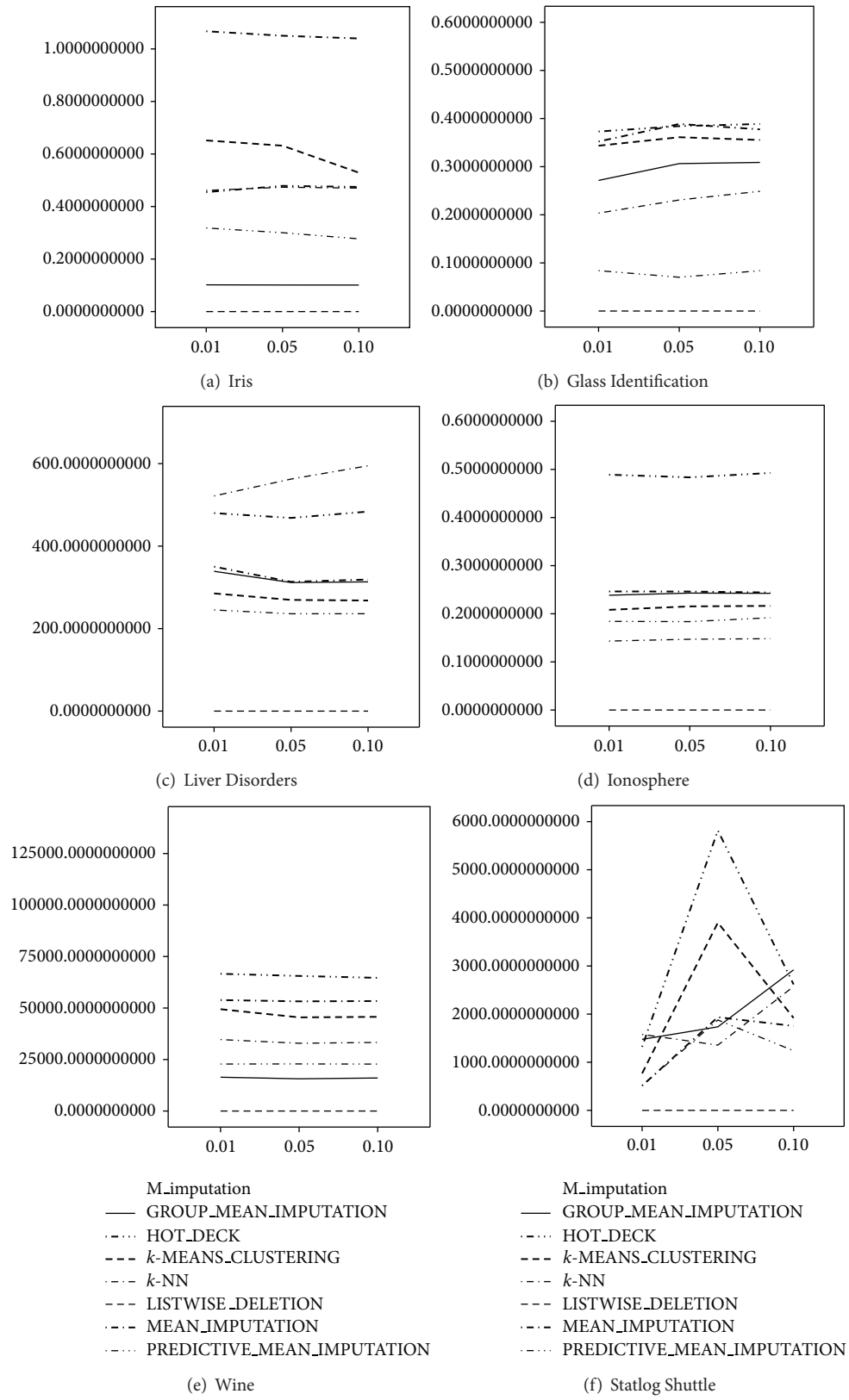


Figure 3: Comparison of performances of imputation methods for each dataset.

Table 6: Factors influencing accuracy (RMSE) for each algorithm (standard beta coefficient): mean imputation.

Data characteristic	trees.J48	BayesNet	SMO	Regression	Logistic	IBk
N_attributes	−.076**	−.075**	−.178**	−.072**	.115**	.007
N_cases	−.079**	−.049**	.012	−.017	−.032	−.048**
C_imbalance	.117**	.239**	.264**	.525**	.163**	.198**
R_missing	.051*	.078**	.040	.080**	.076**	.068**
SE_HS	.249**	.285**	.186**	.277**	.335**	.245**
SE_VS	−.009	−.013	−.006	−.013	−.016	−.010
Spread	−.382**	−.430**	−.261**	−.436**	−.452**	−.363**
P_missing_dum1	−.049	−.038	−.038	−.037	−.045	−.038
P_missing_dum2	−.002	.014	.002	.011	.001	.011

Note 1: N_attributes: number of attributes, N_cases: number of cases, C_imbalance: degree of class imbalance, R_missing: missing data ratio, SE_HS: horizontal scatteredness, SE_VS: vertical scatteredness, spread: missing data spread, and missing patterns: univariate (P_missing_dum1 = 1, P_missing_dum2 = 0), monotone (P_missing_dum1 = 0, P_missing_dum2 = 1), and arbitrary (P_missing_dum1 = 1, P_missing_dum2 = 1)
Note 2: RMSE indicates error; therefore, lower values are better.
Note 3: $^{*}P < 0.05$, $^{**}P < 0.01$.

Table 7: Factors influencing accuracy (RMSE) for each algorithm (standard beta coefficient): group mean imputation.

Data characteristic	trees.J48	BayesNet	SMO	Regression	Logistic	IBk
N_attributes	−.068**	−.072**	−.179**	−.068**	.115**	.010
N_cases	−.082**	−.050**	.011	−.018	−.034*	−.047**
C_imbalance	.115**	.228**	.260**	.517**	.156**	.197**
R_missing	.050**	.085**	.043	.084**	.095**	.066**
SE_HS	.230**	.268**	.178**	.273**	.300**	.248**
SE_VS	−.008	−.012	−.006	−.013	−.013	−.010
Spread	−.296**	−.439**	−.264**	−.443**	−.476**	−.382**
P_missing_dum1	−.043	−.032	−.034	−.035	−.035	−.041
P_missing_dum2	.002	.024	.004	.016	.021	.013

Note 1: N_attributes: number of attributes, N_cases: number of cases, C_imbalance: degree of class imbalance, R_missing: missing data ratio, SE_HS: horizontal scatteredness, SE_VS: vertical scatteredness, spread: missing data spread, and missing patterns: univariate (P_missing_dum1 = 1, P_missing_dum2 = 0), monotone (P_missing_dum1 = 0, P_missing_dum2 = 1), and arbitrary (P_missing_dum1 = 1, P_missing_dum2 = 1)
Note 2: RMSE indicates error; therefore, lower values are better.
Note 3: $^{*}P < 0.05$, $^{**}P < 0.01$.

5.2. Classification Algorithm. Figure 4 shows the performance of the classification algorithms by imputation method and ratio of missing data. As shown in the figure, the performance of each imputation method was similar and did not vary depending on the ratio of missing data, except for listwise deletion. For listwise deletion, as the ratio of missing to complete data increased, the performance deteriorated. In the listwise deletion method, all records are deleted that contain missing data; therefore, the number of deleted records increases as the ratio of missing data increases. The low performance of this method can be explained based on this fact.

The differences in performance between imputation methods were minor. The figure displays these differences by classification algorithm. Using the Bayesian network and logistic classifier methods significantly improved performance compared to other classifiers. However, the relationships among missing data, imputation methods, and classifiers remained to be explained. Thus, a regression analysis was conducted.

In Figure 4, the results suggest the following rules.

(i) IF the missing rate increases AND IBK is used, THEN use the GROUP_MEAN_IMPUTATION method.

(ii) IF the missing rate increases AND the logistic classifier method is used, THEN use the HOT_DECK method.

(iii) IF the missing rate increases AND the regression method is used, THEN use the GROUP_MEAN_IMPUTATION method.

(iv) IF the missing rate increases AND the BayesNet method is used, THEN use the GROUP_MEAN_IMPUTATION method.

(v) IF the missing rate increases AND the trees.J48 method is used, THEN use the k-NN method.

5.3. Regression. The results of the regression analysis are presented in Tables 6, 7, 8, 9, 10, and 11. The analysis was conducted using 900 datasets (3 missing ratios × 3 missing

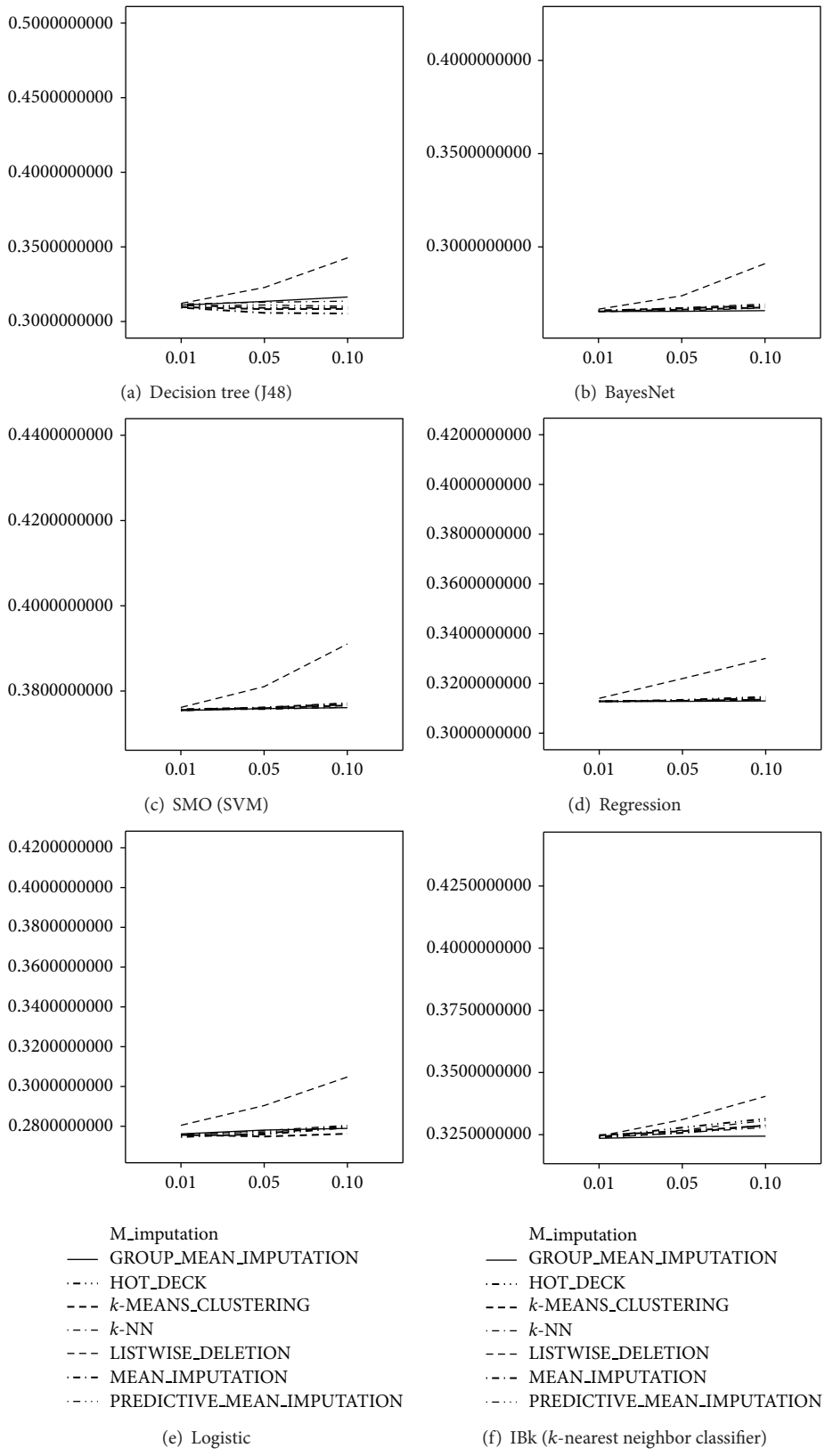


Figure 4: Comparison of classifiers in terms of classification performance.

Table 8: Factors influencing accuracy (RMSE) for each algorithm (standard beta coefficient): Predictive_Mean_Imputation.

Data characteristic	trees.J48	BayesNet	SMO	Regression	Logistic	IBk
N_attributes	−.076**	−.076**	−.178**	−.063**	.123**	.016
N_cases	−.084**	−.049**	.012	−.017	−.034*	−.047**
C_imbalance	.117**	.242**	.263**	.523**	.153**	.198**
R_missing	.050*	.079**	.043	.085**	.080**	.068**
SE_HS	.223**	.279**	.182**	.268**	.322**	.242**
SE_VS	−.008	−.013	−.006	−.013	−.015	−.009
Spread	−.328**	−.432**	−.262**	−.434**	−.465**	−.361**
P_missing_dum1	−.042	−.035	−.034	−.028	−.044	−.036
P_missing_dum2	.008	.012	.004	.018	.007	.011

Note 1: N_attributes: number of attributes, N_cases: number of cases, C_imbalance: degree of class imbalance, R_missing: missing data ratio, SE_HS: horizontal scatteredness, SE_VS: vertical scatteredness, spread: missing data spread, and missing patterns: univariate (P_missing_dum1 = 1, P_missing_dum2 = 0), monotone (P_missing_dum1 = 0, P_missing_dum2 = 1), and arbitrary (P_missing_dum1 = 1, P_missing_dum2 = 1)
Note 2: RMSE indicates error; therefore, lower values are better.
Note 3: $^{*}P < 0.05$, $^{**}P < 0.01$.

Table 9: Factors influencing accuracy (RMSE) for each algorithm (standard beta coefficient): Hot_deck.

Data characteristic	trees.J48	BayesNet	SMO	Regression	Logistic	IBk
N_attributes	−.080**	−.073**	−.176**	−.071**	.115**	.007
N_cases	−.081**	−.049**	.012	−.018	−.034*	−.047**
C_imbalance	.135**	.237**	.261**	.524**	.133**	.211**
R_missing	.062**	.083**	.044	.084**	.075**	.070**
SE_HS	.225**	.275**	.183**	.271**	.313**	.254**
SE_VS	−.009	−.013	−.006	−.013	−.014	−.010
Spread	−.365**	−.428**	−.265**	−.427**	−.441**	−.361**
P_missing_dum1	−.035	−.037	−.034	−.033	−.048	−.038
P_missing_dum2	.012	.015	.004	.012	−.004	.009

Note 1: N_attributes: number of attributes, N_cases: number of cases, C_imbalance: degree of class imbalance, R_missing: missing data ratio, SE_HS: horizontal scatteredness, SE_VS: vertical scatteredness, spread: missing data spread, and missing patterns: univariate (P_missing_dum1 = 1, P_missing_dum2 = 0), monotone (P_missing_dum1 = 0, P_missing_dum2 = 1), and arbitrary (P_missing_dum1 = 1, P_missing_dum2 = 1)
Note 2: RMSE indicates error; therefore, lower values are better.
Note 3: $^{*}P < 0.05$, $^{**}P < 0.01$.

Table 10: Factors influencing accuracy (RMSE) for each algorithm (standard beta coefficient): k-NN.

Data characteristic	trees.J48	BayesNet	SMO	Regression	Logistic	IBk
N_attributes	−.085**	−.079**	−.181**	−.068**	.122**	.006
N_cases	−.083**	−.049**	.011	−.018	−.034*	−.047**
C_imbalance	.143**	.249**	.260**	.521**	.152**	.211**
R_missing	.054*	.078**	.041	.085**	.075**	.071**
SE_HS	.234**	.290**	.182**	.269**	.328**	.255**
SE_VS	−.010	−.013	−.006	−.013	−.014	−.011
Spread	−.332**	−.427**	−.264**	−.431**	−.450**	−.369**
P_missing_dum1	−.038	−.041	−.035	−.029	−.057	−.035
P_missing_dum2	.003	.008	.005	.017	.000	.011

Note 1: N_attributes: number of attributes, N_cases: number of cases, C_imbalance: degree of class imbalance, R_missing: missing data ratio, SE_HS: horizontal scatteredness, SE_VS: vertical scatteredness, spread: missing data spread, and missing patterns: univariate (P_missing_dum1 = 1, P_missing_dum2 = 0), monotone (P_missing_dum1 = 0, P_missing_dum2 = 1), and arbitrary (P_missing_dum1 = 1, P_missing_dum2 = 1)
Note 2: RMSE indicates error; therefore, lower values are better.
Note 3: $^{*}P < 0.05$, $^{**}P < 0.01$.

Table 11: Factors influencing accuracy (RMSE) for each algorithm (standard beta coefficient): *k*-MEANS_CLUSTERING.

Data characteristic	trees.J48	BayesNet	SMO	Regression	Logistic	IBk
N_attributes	−.080**	−.078**	−.181**	−.068**	.117**	.009
N_cases	−.079**	−.049**	.012	−.017	−.033	−.047**
C_imbalance	.136**	.240**	.263**	.524**	.145**	.206**
R_missing	.057*	.079**	.041	.084**	.079**	.057*
SE_HS	.236**	.289**	.183**	.271**	.315**	.264**
SE_VS	−.009	−.013	−.006	−.013	−.014	−.011
Spread	−.362**	−.439**	−.262**	−.440**	−.474**	−.363**
P_missing_dum1	−.037	−.042	−.036	−.032	−.038	−.046
P_missing_dum2	.002	.013	.001	.014	.009	.004

Note 1: N_attributes: number of attributes, N_cases: number of cases, C_imbalance: degree of class imbalance, R_missing: missing data ratio, SE_HS: horizontal scatteredness, SE_VS: vertical scatteredness, spread: missing data spread, and missing patterns: univariate (P_missing_dum1 = 1, P_missing_dum2 = 0), monotone (P_missing_dum1 = 0, P_missing_dum2 = 1), and arbitrary (P_missing_dum1 = 1, P_missing_dum2 = 1)
Note 2: RMSE indicates error; therefore, lower values are better.
Note 3: $^{*}P < 0.05$, $^{**}P < 0.01$.

patterns × 100 trials). Each dataset was generated randomly to meet the preconditions. We conducted the performance evaluation by randomly assigning each dataset to test/training sets at a 3 : 7 ratio. The regression analysis included the characteristics of the datasets and the patterns of the missing values as independent variables. Control variables, such as the type of classifier and imputation method, were also included. The effects of the various characteristics of the data and missing values on classifier performance (RMSE) were analyzed. Three types of missing ratios were treated as two dummy variables (P_missing_dum1, 2: 00, 01, 10). Tables 6–11 illustrate the results of the regression analysis of the various imputation methods. The results suggest the following rules regardless of which imputation method is selected:

(i) IF N_attributes increases, THEN use SMO.

(ii) IF N_cases increases, THEN use trees.J48.

(iii) IF C_imbalance increases, THEN use trees.J48.

(iv) IF R_missing increases, THEN use SMO.

(v) IF SE_HS increases, THEN use SMO.

(vi) IF Spread increases, THEN use Logistic.

Figure 5 displays the coefficient pattern of the decision tree classifier for each imputation method. Dataset characteristics are illustrated on the x-axis and the regression coefficients for each imputation method on the y-axis. For all imputation methods except listwise deletion, the classifiers' coefficient patterns seemed similar. However, significant differences were found in the coefficient patterns using other algorithms. For example, for all imputation methods, a higher beta coefficient of the number of attributes (N_attributes) was observed for the logistics algorithm than for any other algorithm. Thus, the logistics algorithm exhibited the lowest performance (highest RMSE) in terms of the number of attributes. In terms of the number of cases (N_cases), SMO performed the worst. When the data were imbalanced, the regression method was the least effective one. For the missing ratio, the regression method showed the lowest performance except in comparison to listwise deletion and mean imputation. For the horizontal scattered standard error (SE_HS), SMO had the lowest performance. For missing data spread, the logistic classifier method had the lowest performance.

Moreover, for each single factor (e.g., spread), even if the results for two algorithms were the same, their performance differed depending on which imputation method was applied. For example, for the decision tree (J48) algorithm, the mean imputation method had the most negative effect on classification performance for horizontal scattered standard error (SE_HS) and spread, while the listwise deletion and group mean imputation methods had the least negative effect.

The similar coefficient patterns shown in Figure 5 indicate that the differences in impact of each imputation method on performance were insignificant. In order to determine the impact of the classifiers, more tests were needed. Figure 6 illustrates the coefficient patterns when the ratio of missing to complete data is 90%. Under these circumstances, the distinction between imputation methods according to dataset characteristics is significant. For example, very high or very low beta coefficients may be observed for most dataset characteristics except the number of instances and class imbalance.

Figure 7 shows the RMSE based on the ratio of missing data for each imputation method. As the ratio increases, the performance drops (RMSE increases); this is not an unexpected result. However, as the ratio of missing to complete data increases, the differences in performance between imputation methods become significant. These results imply that the characteristics of the dataset and missing values affect the performance of the classifier algorithms. Furthermore, the patterns of these effects differ depending on the imputation methods and classifiers used.

Lastly, we estimate the accuracy (RMSE) of each method by conducting a multiple regression analysis. As shown in Table 12, the results confirmed a significant association between the characteristics of the missing data and the method of imputation with the performance of each classification in terms of RMSE. In total, 226,800 datasets (3

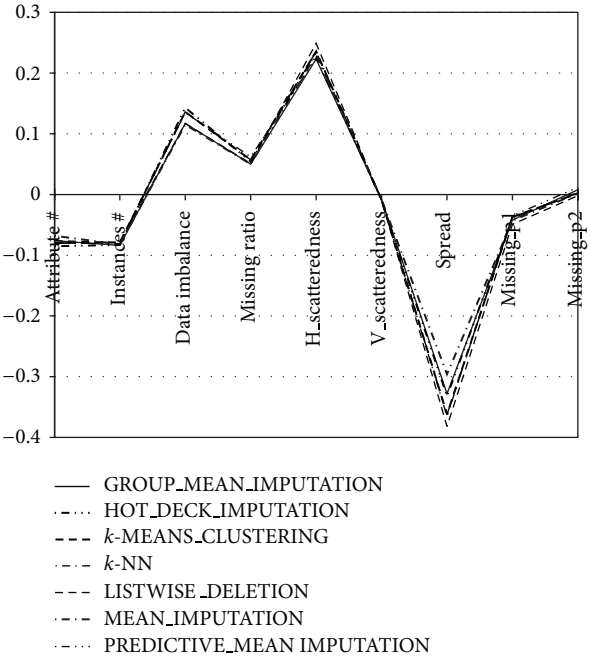


Figure 5: Coefficient pattern of the decision tree algorithm (RMSE).

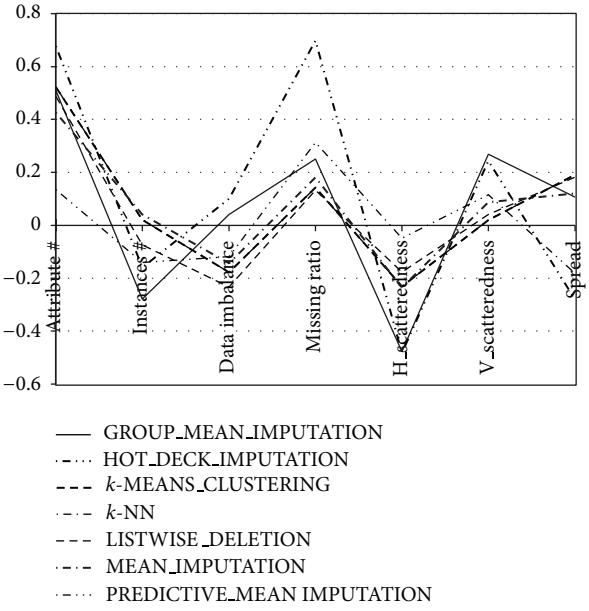


Figure 6: Coefficient pattern of the decision tree algorithm based on a 90% missing ratio (RMSE).

missing ratios × 3 missing patterns × 100 trials × 6 imputation methods × 7 classification methods) were analyzed. The results have at least two implications. First, we can predict the classification accuracy for an unknown dataset with missing data only if the data characteristics can be obtained. Second, we can establish general rules for selection of the optimal combination of a classification algorithm and imputation algorithm.

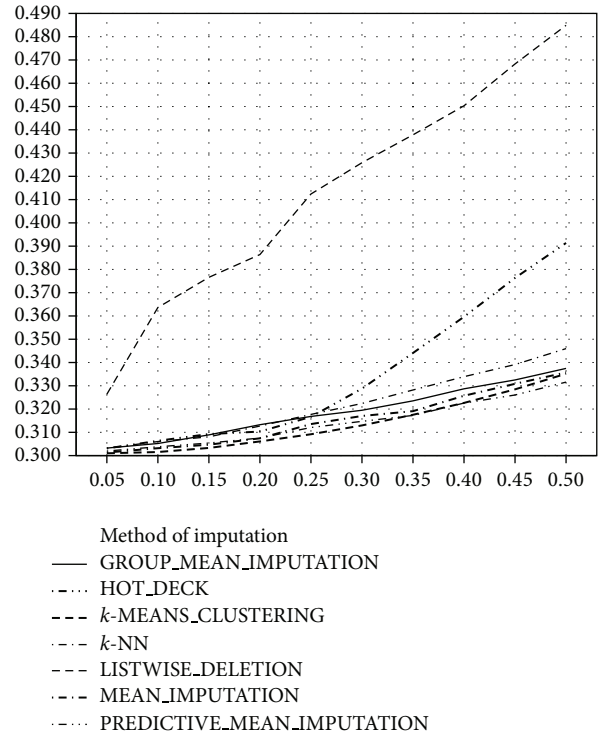


Figure 7: RMSE by ratio of missing data.

Table 12: Factors influencing accuracy (RMSE) of classifier algorithms.

Data characteristic	*B*	Data characteristic	*B*
(constant)	.060**	M_imputation_dum1	.012**
R_missing	.083**	M_imputation_dum2	−.001*
SE_HS	−.005**	M_imputation_dum4	000
SE_VS	.000**	M_imputation_dum5	000
Spread	.017**	M_imputation_dum6	.001**
N_attributes	−.008**	M_imputation_dum7	−.001*
C_imbalance	−.003**	P_missing_dum1	−.006**
N_cases	.002**	P_missing_dum3	.000

Note 1: Dummy variables related to imputation methods: LISTWISE DELETION (M_imputation_dum1 = 1, others = 0), MEAN_IMPUTATION (M_imputation_dum2 = 1, others = 0), GROUP_MEAN_IMPUTATION (M_imputation_dum3 = 1, others = 0), PREDICTIVE_MEAN_IMPUTATION (M_imputation_dum4 = 1, others = 0), HOT_DECK (M_imputation_dum5 = 1, others = 0), *k*-NN (M_imputation_dum6 = 1, others = 0), and *k*-MEANS_CLUSTERING (M_imputation_dum7 = 1, others = 0). Missing patterns: univariate (P_missing_dum1 = 1, P_missing_dum2 = 0, P_missing_dum3 = 0), monotone (P_missing_dum1 = 0, P_missing_dum2 = 1, P_missing_dum3 = 0), and arbitrary (P_missing_dum1 = 1, P_missing_dum2 = 1, P_missing_dum3 = 1). *B*: standard beta coefficient.
Note 2: $^{*}P < 0.1$, $^{**}P < 0.05$.

6. Conclusion

So far, the prior research does not fully inform us of the fitness among datasets, imputation methods, and classification algorithms. Therefore, this study ultimately aims to establish a rule set which guides the classification/recommender system developers to select the best classification algorithm based

on the datasets and imputation method. To the best of our knowledge, ours is the first study in which the performance of classification algorithms with multiple dimensions (datasets, imputation data, and imputation methods) is discussed. Prior research examines only one dimension [15]. In addition, as shown in Figure 3, since the performance of each method differs according to the dataset, the results of prior studies on imputation methods or classification algorithms depend on the datasets on which they are based.

In this paper, factors affecting the performance of classification algorithms were identified as follows: characteristics of missing values, dataset features, and imputation methods. Using benchmark data and thousands of variations, we found that several factors were significantly associated with the performance of classification algorithms. First, as expected, the results show that the missing data ratio and spread are negatively associated with the performance of the classification algorithms. Second and as a new finding to our best knowledge, we observed that the number of missing cells in each record (SE_HS) was more sensitive in affecting the classification performance than the number of missing cells in each feature (SE_VS). Further, we found it interesting that the number of features negatively affects the performance of the logistic algorithm, while other factors do not.

A disadvantage of logistic regression is its lack of flexibility. The assumption of a linear dependency between predictor variables and the log-odds ratio results in a linear decision boundary in the instance space, which is not valid in many applications. Hence, in the case of data imputation, the logistic algorithm must be avoided. Next, in response to concerns about class imbalance, which has been discussed in data mining research [18, 19], we found that the degree of class imbalance was the most significant data feature to decrease the predicted performance of classification algorithms. In particular, SMO was second to none in predicting SE_HS in any imputation situation; that is, if a dataset has a high number of records in which the number of missing cells is large, then SMO is the best classification algorithm to apply.

The results of this study suggest that optimal selection of the imputation method according to the characteristics of the dataset (especially the patterns of missing values and choice of classification algorithm) improves the accuracy of ubiquitous computing applications. Also, a set of optimal combinations may be derived using the estimated results. Moreover, we established a set of general rules based on the results of this study. These rules allow us to choose a temporally optimal combination of classification algorithm and imputation method, thus increasing the agility of ubiquitous computing applications.

Ubiquitous environments include a variety of forms of sensor data from limited service conditions such as location, time, and status, combining various different kinds of sensors. Using the rules deducted in this study, it is possible to select the optimal combination of imputation method and classification algorithm for environments in which data changes dynamically. For practitioners, these rules for selection of the optimal pair of imputation method and classification algorithm may be developed for each situation depending on the characteristics of datasets and their missing values. This set of rules will be useful for users and developers of intelligent systems (recommenders, mobile applications, agent systems, etc.) to choose the imputation method and classification algorithm according to context while maintaining high prediction performance.

In future studies, the predicted performance of various methods can be tested with actual datasets. Although, in prior research on classification algorithms, multiple benchmark datasets from the UCI laboratory have been used to demonstrate the generality of the proposed method, performance evaluations in real settings would strengthen the significance of the results. Further, for brevity, we used a single performance metric, RMSE, in this study. For example, FP rate, as well as TP rate, is very crucial when it comes to investigating the effect of class imbalance, which is considered in this paper as an independent variable. Although the performance results would be very similar when using other metrics such as misclassification cost and total number of errors [20], more valuable findings may be generated from a study including these other metrics.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Strategic R&D Program for Industrial Technology (10041659) and funded by the Ministry of Trade, Industry, and Energy (MOTIE).

References

[1] J. Augusto, V. Callaghan, D. Cook, A. Kameas, and I. Satoh, "Intelligent environments: a manifesto," *Human-Centric Computing and Information Sciences*, vol. 3, no. 12, pp. 1–18, 2013.

[2] R. Y. Toledo, Y. C. Mota, and M. G. Borroto, "A regularity-based preprocessing method for collaborative recommender systems," *Journal of Information Processing Systems*, vol. 9, no. 3, pp. 435–460, 2013.

[3] G. Batista and M. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 519–533, 2003.

[4] R. Shtykh and Q. Jin, "A human-centric integrated approach to web information search and sharing," *Human-Centric Computing and Information Sciences*, vol. 1, no. 1, pp. 1–37, 2011.

[5] H. Ihm, "Mining consumer attitude and behavior," *Journal of Convergence*, vol. 4, no. 2, pp. 29–35, 2013.

[6] Y. Cho and S. Moon, "Weighted mining frequent pattern based customers RFM score for personalized u-commerce recommendation system," *Journal of Convergence*, vol. 4, no. 4, pp. 36–40, 2013.

[7] N. Howard and E. Cambria, "Intention awareness: improving upon situation awareness in human-centric environments," *Human-Centric Computing and Information Sciences*, vol. 3, no. 9, pp. 1–17, 2013.

[8] L. Liew, B. Lee, Y. Wang, and W. Cheah, "Aerial images rectification using non-parametric approach," *Journal of Convergence*, vol. 4, no. 2, pp. 15–21, 2013.

[9] K. J. Nishanth and V. Ravi, "A computational intelligence based online data imputation method: an application for banking," *Journal of Information Processing Systems*, vol. 9, no. 4, pp. 633–650, 2013.

[10] P. Kang, "Locally linear reconstruction based missing value imputation for supervised learning," *Neurocomputing*, vol. 118, pp. 65–78, 2013.

[11] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art," *Psychological Methods*, vol. 7, no. 2, pp. 147–177, 2002.

[12] H. Finch, "Estimation of item response theory parameters in the presence of missing data," *Journal of Educational Measurement*, vol. 45, no. 3, pp. 225–245, 2008.

[13] S. J. Press and S. Wilson, "Choosing between logistic regression and discriminant analysis," *Journal of the American Statistical Association*, vol. 73, no. 364, pp. 699–705, 1978.

[14] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. H. Witten, "Using model trees for classification," *Machine Learning*, vol. 32, no. 1, pp. 63–76, 1998.

[15] O. Kwon and J. M. Sim, "Effects of data set features on the performances of classification algorithms," *Expert Systems with Applications*, vol. 40, no. 5, pp. 1847–1857, 2013.

[16] E. Namsrai, T. Munkhdalai, M. Li, J.-H. Shin, O.-E. Namsrai, and K. H. Ryu, "A feature selection-based ensemble method for arrhythmia classification," *Journal of Information Processing Systems*, vol. 9, no. 1, pp. 31–40, 2013.

[17] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2nd edition, 2005.

[18] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man and Cybernetics C: Applications and Reviews*, vol. 42, no. 4, pp. 463–484, 2012.

[19] Q. Yang and X. Wu, "10 challenging problems in data mining research," *International Journal of Information Technology & Decision Making*, vol. 5, no. 4, pp. 597–604, 2006.

[20] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 63–77, 2006.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 537078, 9 pages
http://dx.doi.org/10.1155/2015/537078

Research Article

A Fully Distributed Resource Allocation Mechanism for CRNs without Using a Common Control Channel

Adil Mahmud, Youngdoo Lee, and Insoo Koo

School of Electrical and Computer Engineering, University of Ulsan, Ulsan 680-749, Republic of Korea

Correspondence should be addressed to Insoo Koo; iskoo@ulsan.ac.kr

Received 18 June 2014; Accepted 11 October 2014

Academic Editor: Jong-Hyuk Park

Spatial and temporal variations in the available spectrum for cognitive devices add additional complexity in designing a medium access strategy for cognitive radio (CR) networks. To ease this complication, a common control channel (CCC) is often used as a common platform to exchange control messages, so the CR users can make decisions concerning resource allocation. Much of the literature in this area opts for forming groups and negotiating a CCC from the available channels rather than considering a dedicated common control channel. However, if a primary user appears on the CCC, all the CR users have to vacate that channel and negotiate another CCC, and thus overhead continues to grow with an agile primary network. We propose a medium access control protocol that can work in the absence of a CCC and reduce the possible overhead to a greater extent. In our proposed protocol, CR users take advantage of similar spectrum availability in their neighborhood for resource utilization. We also propose a contention-based spectrum allocation mechanism that works in a distributed manner over different available channels. Simulation results show that this approach can reduce broadcast overhead significantly while maintaining connectivity success similar to its counterparts.

1. Introduction

Immense market demand for wireless communications technologies, including video streaming and other applications, precipitates spectrum scarcity. But it has been observed that the static nature of users limits the ability to dynamically use the whole spectrum at any given time. Statistics show that, in some cases, spectrum usage even falls below 5–10% [1]. In such situations, new users are deprived of spectrum resources, which is absolutely unreasonable.

The ideas of cognitive radio (CR) and dynamic spectrum access bring a promising opportunity to solve the spectrum scarcity problem [2]. A CR device, also called a secondary user (SU) or simply a node, has a cognitive capability with which it obtains information about the networks surrounding it and can opportunistically access unused spectrum without interfering with licensed users, also known as primary users (PUs) [2, 3]. This opportunistic spectrum utilization is commonly known as spectrum sharing. Spectrum sharing among SUs is a major issue to focus on in a CR network (CRN). There are many solutions to medium access control (MAC) for traditional sensor and ad hoc networks. Some of those are centralized protocols [4–8] and some are decentralized [9]. But none of those protocols are directly applicable to a CRN because of different spectrum etiquette. Nevertheless, the advancement in MAC mechanisms for a CRN is worth mentioning. A number of proposals with a variety of unique ideas have already emerged so far for CRNs. Those proposals comprise both centralized [10, 11] and decentralized approaches [12, 13], which may be further categorized on the basis of common control channel (CCC) usage. Though the centralized network approach using a dedicated common control channel (DCCC) simplifies the task of spectrum management, assignment, and utilization, it has been proven impractical in a CRN.

Keeping the above-mentioned constraints in mind, several research articles have been proposed, primarily aimed at reducing coordination overhead while maintaining spectrum management efficiency at the level of centralized solutions. One group of such proposals considers choosing a CCC from

among the available channels in a neighborhood in designing distributed MAC protocols for CRN [12, 14]. Despite the remarkable success over their centralized counterparts, as well as over the protocols using a DCCC, the protocols that consider a CCC or coordination channel have drawbacks in CR application systems. For example, a CCC can be a viable center of attack for an intruder. Moreover, when a PU appears on the CCC, all the nodes using that channel as a CCC must leave it and negotiate for establishing a new CCC. As a result, this approach is vulnerable and, therefore, inappropriate in a CRN when PU activity is high. To solve the problems mentioned above, some researchers opted to design MAC protocols without using any sort of CCC. For example, Timalsina et al. [15] proposed a protocol that minimizes wait time and delay but uses two radios. As a result, their solution increases hardware cost.

In addition, a distributed broadcast protocol for a multihop CR ad hoc network was proposed by Song and Xie [16] in which a channel rendezvous technique employing broadcasts instead of using a CCC was adopted.

In this paper, we propose a resource allocation mechanism that does not require a CCC. For this work, we were motivated by the fact that neighboring CR users generally experience similar spectrum availability. Zhao et al. [14] also showed the degree of correlation in spectrum availability among neighbors in their experiments. In their research, a coordinated scheme was developed where CR nodes in a small neighborhood select the most commonly available channel among them as a coordination channel, and the transmission pairs utilize that channel to negotiate data transmission channels. However, a coordinated scheme still requires additional control overhead to establish a coordination channel. Instead, in this paper, we propose a unique contention-based method, followed by each CR user, to reduce the extra overhead caused by the coordination channel selection process. Another advantage of our approach is that contention for occupying a channel is distributed over the available channels in the vicinity. However, in such a unique contention-based method, the prime issue will be how to allocate resources (i.e., channels) to users. Therefore, we also propose a channel acquisition scheme to address this issue where the basic concept of carrier sense multiple access (CSMA) is adopted as a tool.

The rest of the paper is arranged as follows. In Section 2, we present assumptions and the system model. The proposed protocol is described in detail in Section 3. Simulation results are given in Section 4. Finally the paper is concluded in Section 5.

2. Assumptions and the System Model

In this section, we first mention the assumptions we make in designing the protocol, and, then, we describe the system model.

2.1. Assumptions. In this paper, we refer to the licensed network as the primary network and the unlicensed network as the secondary network. We assume there are N uncorrelated channels in the primary network. Let the set of channels

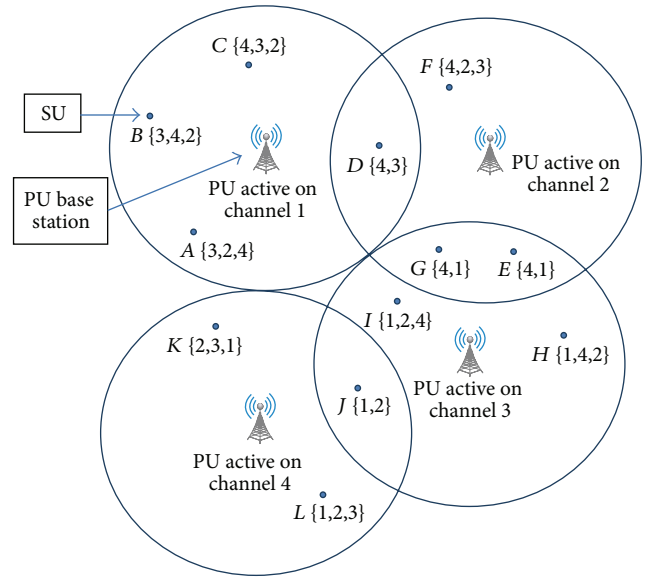


Figure 1: An example of CR network topology.

for the primary network be C ($C = \{1, 2, \ldots, N\}$). Each SU is equipped with a single radio having cognition capability and holding predefined characteristics such as transmission range, transmit power, and modulation. We also assume that SUs can sense the channels with high reliability. How they sense channels with high reliability is beyond the scope of this work. In our proposed model, each of the SUs can sense and utilize a number of channels available at its position. No two SUs within transmission range of each other can use the same channel. The SUs can be static or quasistatic, which means channel availability does not change frequently unless PUs abruptly occupy the channels.

2.2. System Model. Figure 1 shows a simple exemplary CR network suitable for explaining the proposed scheme. There are four primary network base stations serving their users (PUs) on different channels (channels 1–4). At the same time, several SUs, denoted as A, B, C, and so forth, reside inside the primary network. According to the regulations of the opportunistic spectrum access policy, no SU can use a channel such that the SU interferes with any of the PUs.

In this paper, SUs operate in overlay mode, which means that an SU will use a channel if there is no PU activity on that particular channel. The available channels for SUs are indicated within the curly brackets beside each of the nodes in Figure 1. For example, the available channel set for SU A is denoted as $\{3, 2, 4\}$. Generally, the available channel set of an SU is expressed as follows:

$$A_i = \{c \mid c \in C \text{ and } d_{i,c} > t_b\}, \quad (1)$$

where A_i is the available channel set of node i, $d_{i,c}$ is the distance of node i from a PU or a primary network base station that is currently using channel c, and t_b is the radial distance covered by a primary base station. For convenience, the coverage area of a node or a base station is assumed to be

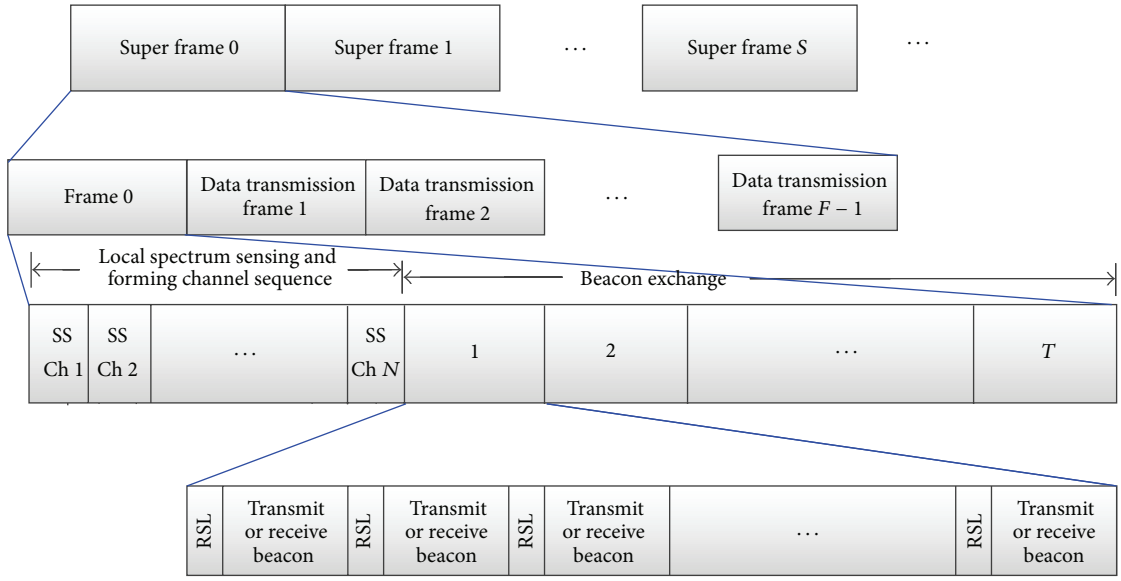


Figure 2: The frame structure of the proposed scheme where each super frame is composed of frame 0 (used for spectrum sensing and discovering neighbors) and frames 1 to $F - 1$ (used for data transfer).

circular. We further assume that two SUs can communicate only if they are within transmission range of each other. We denote the transmission range of an SU by the symbol t_s, which is the radius of the circular area it can cover. Moreover, we consider $t_b > t_s$, which is a common assumption in the literature.

3. General Frame Structure

In the proposed framework, time is divided into super frames; each super frame consists of F frames. This sort of structure is used in the IEEE 802.22 wireless regional area network (WRAN) MAC framework [12]. Though the exterior construction is similar to that of the WRAN structure, we organize the frames in a different way to fit the proposed protocol as follows: frame 0 is used for discovering the available channels of neighbors. For this, the first part of frame 0 is used for spectrum sensing and the rest of it is used for beacon exchange. A number of papers consider node information or beacons to be exchanged among all the neighbors but do not give any explanation about the way to exchange this information [14, 17, 18]. Therefore, in Section 3.2, we propose a beacon exchange method. The other frames of a super frame, that is, frames 1 to $F - 1$, are called *data transmission frames* in this paper. These frames are used for transmitting data. Because nodes should be allocated conflict-free channels to communicate with fairness, we further divide these frames into a number of parts to achieve these goals. We will explain data transmission frames in detail in Section 3.3.

3.1. Spectrum Sensing. At the beginning of frame 0, each node sequentially senses all N channels to check channel availability. For decision-making, the nodes can use any detection method, such as energy detection, which is what we propose in this work. The spectrum decision is obtained according to the following well-known equation:

$$x(t) = \begin{cases} n(t), & H_0, \\ hs(t) + n(t), & H_1, \end{cases} \tag{2}$$

where h denotes gain in the channel between PU and SU, $s(t)$ denotes the signal transmitted by the PU, and $n(t)$ denotes additive noise for the SU.

After sensing all channels, a node creates a channel sequence for an available channel set based on the received normalized signal energy according to the following equation:

$$E_c^i = \frac{\left(\sum_{j=1}^{Ns} \left|x_c(j)\right|^2\right)/N_s}{\max\left\{\left|x_c(j)\right|^2\right\}}, \tag{3}$$

where E_c^i is the normalized energy received on the cth channel by the ith node, $x_c(j)$ is the jth sample of the received PU signal on the cth channel, and N_s is the number of samples received during each sensing interval. The lower the received energy of a channel, the higher its position in the channel sequence. If required, a node hops from one channel to another following this channel sequence. Hence, we can also define the channel sequence as the hopping sequence of a node. Next, nodes need to know their neighboring nodes' information, specifically their channel sequences. This is done by beacon exchange, which is described next.

3.2. Beacon Exchange. Nodes are required to discover their neighbors as well as to know each other's channel sequences. To do this, the second part of frame 0 is divided into T slots, as shown in Figure 2. During the first slot, a node tunes itself to its first channel, which is also called the *best channel* in this

paper, according to its channel sequence. Then in the second slot, it moves to the second channel according to its channel sequence and so on. In each of the slots, a node resides on a particular channel and broadcast beacon, which contains its ID, channel sequence, its neighboring nodes' IDs, and their channel sequences.

Two nodes become direct neighbors (DNs) of each other if they exchange beacons. A neighbor, for which node ID and channel sequence are obtained from one of the DNs, is categorized as a neighbor-of-a-neighbor (NoN) node. A node keeps separate records for DNs and NoNs. An NoN becomes a DN if the nodes directly exchange beacons at some stage.

In a particular slot, in order to avoid collision among the nodes on a channel, each node randomly selects a *random short listening* (RSL) time and listens on the channel if a node has already started beaconing. The node with the shortest RSL time broadcasts its beacon, while the other nodes on the channel receive it. After a node successfully broadcasts its beacon in a slot, it goes into a listening state and waits to receive other nodes' beacons during the rest of the time in that slot. The remaining nodes on that channel follow the same mechanism mentioned above. However, if the number of available channels of a node is less than T, it repeats broadcasting beacons on its available channels in the order of its channel sequence. Since the number of slots should be limited to meet the frame size constraint, it is important to formulate an efficient rendezvous method so that most, if not all, of the nodes within transmission range of a node can exchange beacons. However, in this study, we assume that all the neighbors within transmission range are discovered, either as a DN or as an NoN.

To illustrate the beacon exchange process, we take an example where there are three channels, $c1$, $c2$, and $c3$, and five nodes within transmission range of each other: $n1\{c1, c2\}$, $n2\{c1, c3\}$, $n3\{c1\}$, $n4\{c2, c3\}$, and $n5\{c1, c2\}$. Here, channel numbers enclosed in curly brackets after each node represent the channel sequence of a node. During the first subslot, we have nodes $n1$, $n2$, $n3$, and $n5$ on channel $c1$ whereas node $n4$ is on channel $c2$. Node $n4$ is alone on channel $c2$, so it broadcasts its beacon after waiting for its RSL time but receives no neighbor beacon. On the other hand, all four nodes on channel $c1$ wait for their respective RSL times. Let $n1$ be the winner in the contention to transmit first, which means it has the shortest RSL time. So $n1$ broadcasts its beacon while other nodes receive it. After broadcasting its beacon, a node goes into a listening state to receive the beacons of other nodes. The other nodes ($n2, n3, n5$) exchange their beacons in the same way. During the second subslot, nodes $n1$ and $n5$ move to channel $c2$, nodes $n2$ and $n4$ move to channel $c3$, and node $n3$ keeps repeating its broadcast on channel $c1$. It is important to note here that node $n4$ can make a direct neighborhood only with node $n2$. But other nodes will learn about $n4$ through $n2$. With the help of our proposed data communications mechanism described in Section 3.3, the other nodes will be able to communicate with node $n4$ using $n2$ as a relay. In Section 3.3, we address the process of accessing channels and the overall data communications mechanism among nodes.

3.3. Data Communication. Henceforth, we call a node that has data to send the *sender* and a node that is supposed to receive data the *receiver*. As mentioned earlier, frame 1 to frame $F-1$ of a super frame will be used for data transmission, so we call these frames data transmission frames. A data transmission frame is divided into three segments. Moreover, the way of transmitting in a data transmission frame differs based on neighborhood status between sender and receiver (i.e., DN or NoN) and, similarly, on channel availability. In Section 3.3.1, we discuss the transmission frame structure, and, in Section 3.3.2, we discuss transmission cases.

3.3.1. Transmission Frame Structure. A transmission frame is divided into three parts: contention, transmission, and reporting. Figure 3(a) depicts such a frame structure. In the following parts (1), (2), and (3), we discuss details for each part.

(1) Contention. The contention period is divided into M slots. At the start of each slot, nodes perform fast sensing (FS), a technique similar to the one used in a WRAN [6]. With this approach, a node can recheck in a particular frame for the activity of PU on the channel where it resides. This provides extra protection against interference with primary users. FS plays a vital role in a primary network where PUs are very active.

We assume that not all nodes have data to send in a particular frame. There are some nodes that have data to send, while some nodes do not. Among the nodes that have no data to send, some are receiving nodes, whereas the other nodes may be used as relay nodes. A node that has no data to send always prefers to tune itself to the best channel among its available channels in any contention time slot. Here, an available channel means that the channel of concern is occupied by neither PU nor SU in that time slot. So, if a channel is occupied by a transmission pair, all the nodes having no data to send move to the next channel in their respective channel sequences. On the other hand, a sender follows its intended receiver's available channel sequence. If a node finds there is no active PU on a particular channel, the node waits for the RSL time to avoid collision with other contenders.

If the sender wins contention to use the channel (i.e., the sender has the shortest RSL time), it sends a transmission request to send (TRTS) message to the receiver, while the receiver responds with transmission clear to send (TCTS), forming a transmission pair. At this stage, the transmission pair starts data transmission without waiting for the actual transmission time to begin while the other contenders on that channel jump to next channels according to their respective receivers' channel lists. Such a scenario is depicted in Figure 3(b). The benefit of starting transmission right after the occupation of a channel is twofold. First, the transmission pair gets more time for data transmission, which increases total throughput. Second, if there is no activity until the actual transmission time starts, other nodes might wrongfully determine that the channel is vacant if they somehow do not receive, or miss, the TRTS-TCTS messages.

It is also possible that a sender is unable to occupy a common channel between itself and its intended receiver because

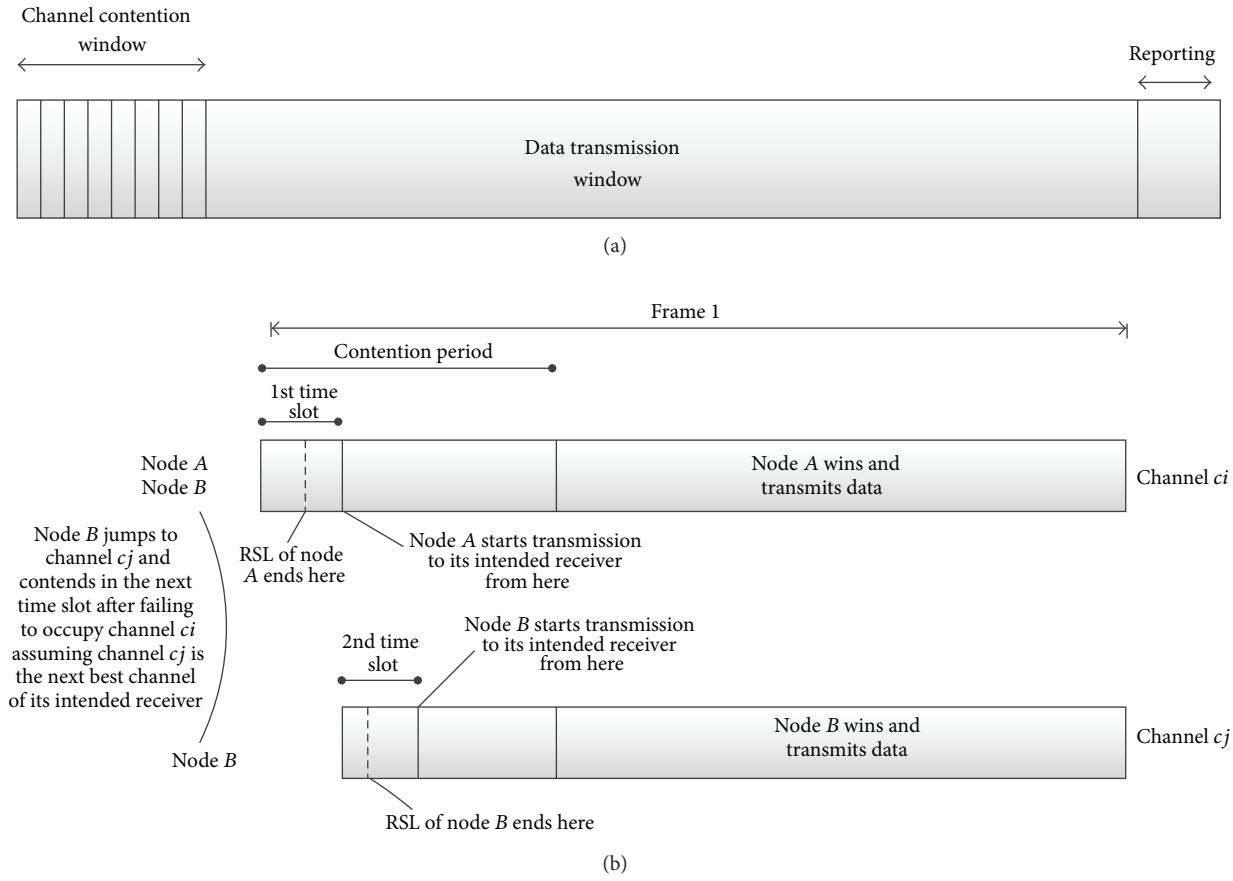


Figure 3: (a) Structure of data transmission frame and (b) the proposed contention-based channel occupation method.

of repeated failure during the contention period. We call such a node an *unsuccessful sender*. The unsuccessful sender returns to its intended receiver's best channel after the end of the contention period.

(2) Transmission. The data transmission slot starts just after the designated contention period ends, although the winning transmission pairs commence communications before the start of the transmission slot, as described in Section 3.3.1(1). The start of the transmission frame signifies that the contention period ends, and, henceforth, no node is able to contend for any channel for transmission until the end of the frame. Senders that could not win a channel, or could not find a relay node, will quietly wait on their intended receivers' best channels until the end of the transmission period.

(3) Reporting. During the reporting slot, prospective senders for the next frame will tune themselves to their receiver's best channel. An unsuccessful sender will broadcast an "unsuccessful" message to notify nodes about its failure in the current frame so that it gets priority during the contention period of the next frame. This reporting mechanism will help reduce delays. However, the mechanism for giving priority to a node over others is not discussed in this study and is left for future research.

3.3.2. Transmission Cases. As we mentioned before, there are three possible transmission cases, based on channel availability and neighborhood status between sender and receiver. These cases are discussed in the next three parts (1), (2), and (3). A flow chart of all the cases is depicted in Figure 4.

(1) Case 1: Receiver Is a DN, and Sender Has Receiver's Best Channel. When a node has data to send to one of its DNs and it has the best channel for that neighbor, the sender first moves to the best channel of that receiver during the first time slot of the contention period. Because there may be several nodes that want to send data in the same time slot and on the same channel, the sender waits for the RSL time to avoid collision and to check whether any other node has already occupied the channel. If no other node occupies the channel within the elapsed RSL time, the node sends TRTS and waits for TCTS from its intended receiver. Once the receiver acknowledges TCTS, communication begins between them. Overhearing these TRTS-TCTS messages, the other nodes leave that channel. However, if the sender fails to win during contention, following the channel sequence of its intended receiver, it switches itself to the next channel, if that channel is available to it. If the channel is not available, the sender moves to the best channel of its receiver and waits until the end of the transmission period. When the reporting time starts,

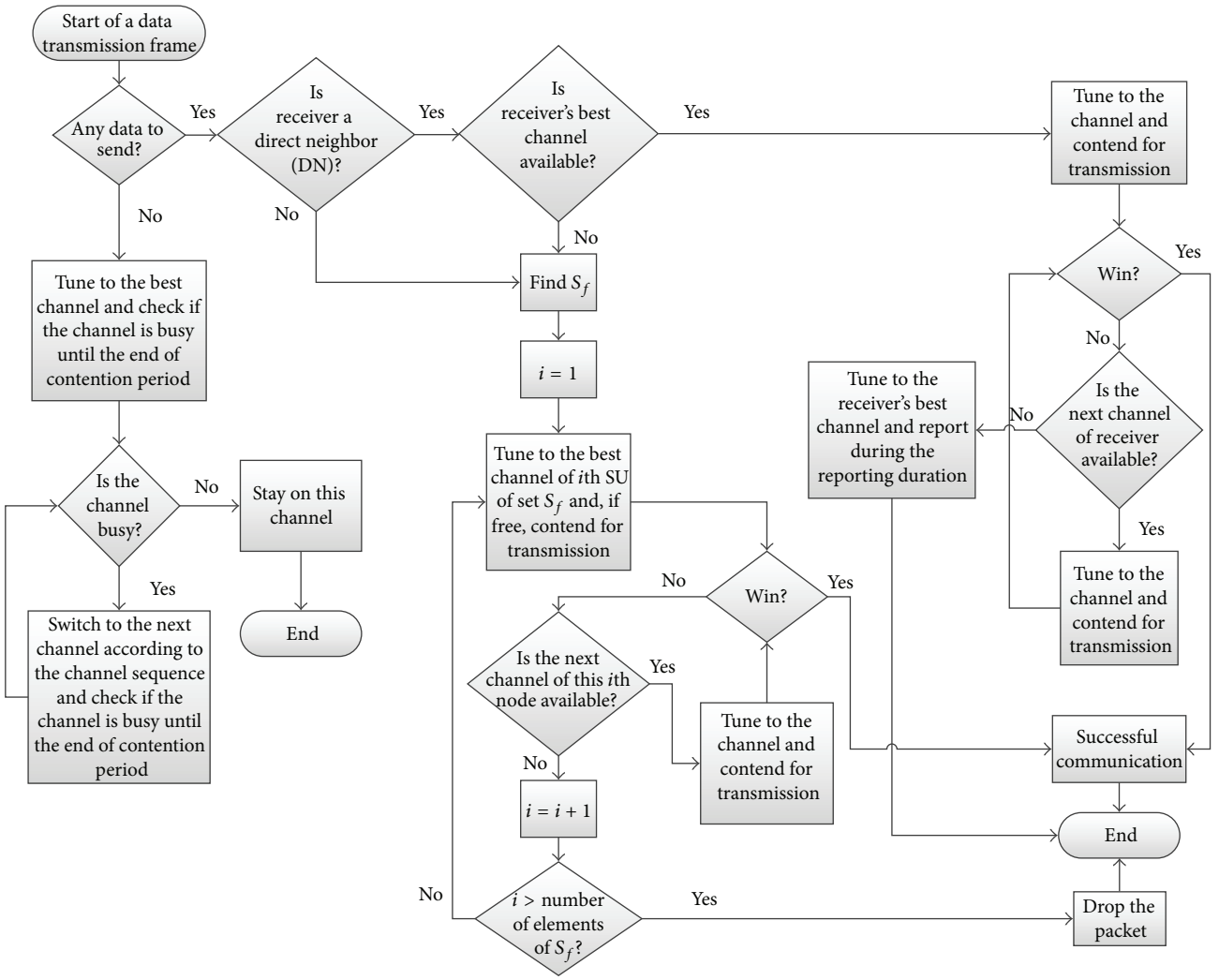


Figure 4: Flow chart of data transmission mechanism in the proposed scheme.

the sender broadcasts an unsuccessful message to get priority in the next frame, as mentioned in Section 3.3.1(3).

(2) Case 2: Receiver Is a DN, but Sender Does Not Have Receiver's Best Channel. This situation can occur when a node does not have the best channel of a neighbor but meets with it on any other channel during the beacon exchange period. Since the sender does not have the best channel of the receiver, it cannot directly communicate with it. In this situation, the sender relies on any DN that has the best channel of the receiver. Therefore, the sender first finds a set of neighbors, S_r, with the best channel of the receiver in their channel sequences. S_r is defined as follows:

$$S_r = \{k \mid \text{The node } k \text{ has the best channel of receiver in its channel list}\}. \quad (4)$$

Next, the sender will find another set of neighbors, S_f, such that $S_f \subset S_r$ and the best channels of all of the neighbors in S_f are available on the sender's available channel list. Formally, the definition of S_f is as follows:

$$S_f = \{m \mid \text{The node } m \text{ is a neighbor of the receiver and sender has the best channel of node } m\}. \quad (5)$$

Finally, the sender sends data to one of the neighbors in S_f according to the method described in Section 3.3.2(1). The neighbor, also called the relay node, then sends the data to the intended receiver in the next frame, which also follows the method described in Section 3.3.2(1).

(3) Case 3: Receiver Is One of the NoNs. A neighbor of a neighbor can be within transmission range of a sender, or it can be outside the range. But this information is not known to the sender. In such a situation, the sender first moves to the best channel of the intended receiver and proceeds with a mechanism similar to that described in Section 3.3.2(1). If the sender does not receive TCTS from the receiver,

Table 1: Nodes and their channels at the beginning of a frame.

Channel number	Nodes on the channel
$c1$	$n1, n2, n3, n4, n5$
$c2$	$n6, n7, n8, n9$
$c3$	$n10, n11, n12$
$c4$	$n13, n14, n15, n16$

Table 2: Sender-receiver pairs, along with their channel sequences.

Sender	Receiver
$n3$ $\{c2, c1, c4\}$	$n1$ $\{c1, c4\}$
$n4$ $\{c4, c1\}$	$n14$ $\{c4, c1\}$
$n7$ $\{c1, c4, c2\}$	$n15$ $\{c4, c2, c1\}$
$n12$ $\{c3, c4, c2\}$	$n8$ $\{c2, c3\}$
$n16$ $\{c3, c1, c4\}$	$n2$ $\{c1, c3\}$

it assumes that the receiver is outside of its transmission range. In that case, the sender proceeds with the method described in Section 3.3.2(2).

(4) An Illustrative Example. To illustrate the mechanism, let us consider an example where 16 nodes (numbered as $n1$–$n16$) look for opportunities to communicate within a primary network consisting of four channels ($c1$, $c2$, $c3$, and $c4$). Table 1 shows which nodes are residing on which channels at the beginning of a data transmission frame. In Table 2, we consider some sender-receiver pairs. The channel sequences for each of the sender and receiver nodes are provided within curly brackets. According to the protocol, nodes record these channel sequences from the beacon exchange step described in Section 3.2.

According to the mechanism, the nodes that do not have data to send initially stay tuned to their best channels according to their channel sequences. Consequently, receivers $n1$ and $n2$ should be on channel $c1$, receiver $n8$ should be on channel $c2$, and receivers $n14$ and $n15$ should be on channel $c4$. At the beginning of a transmission frame, senders $n3$ and $n16$ move to channel $c1$, senders $n4$ and $n7$ move to channel $c4$, and sender $n12$ moves to channel $c2$.

During the contention period, let us assume that senders $n3$ and $n4$ win use of channels $c1$ and $c4$, respectively, whereas node $n12$ is the only sender on channel $c2$. Soon after winning contention, the sender-receiver pairs exchange TRTS-TCTS messages and transmit their data. Meanwhile, node $n2$ moves to channel $c3$, which is still free (i.e., not yet occupied by an SU). Because a sender knows the next channel on which it should search for the receiver, node $n16$ moves to channel $c3$ and communicates with the receiver. On the other hand, node $n15$ (and, consequently, node $n7$, because it wants to send data to node $n15$) moves to channel $c2$ and then to channel $c1$ but finds no channel free. In this case, node $n7$ returns to channel $c4$ (as it is the best channel of its intended receiver) and broadcasts an unsuccessful message during the reporting time slot.

4. Performance Analysis

In this section, we demonstrate and compare the performance of our proposed scheme. To measure performance, we specifically focus on the number of broadcast packets and the connectivity among the nodes. Reduction of broadcast traffic is the main goal of this study, and to achieve this goal, we propose a protocol that does not rely on a common control channel. But without a common control channel, it is difficult to establish connectivity among the nodes and, thus, the resource allocation task becomes complicated. We compare our proposed scheme with the one proposed by Zhao et al. [14] on the basis of the above-mentioned performance metrics, that is, number of broadcast packets and connectivity among nodes. The reason behind choosing the protocol proposed by Zhao et al. [14] is that this protocol is also designed based on spectrum homogeneity in the neighborhood.

4.1. Simulation Setup. We use MATLAB to simulate performance and randomly deploy four primary network base stations in a 100-by-100 area. Each base station serves the PUs on one of the four channels, and each channel is considered busy all the time, because we assume that primary users are always communicating with their nearest base stations. So, an SU can only use those channels on which no PU is active inside its sensing range. Within the coverage area of the base stations, we deploy 100 secondary users, but the number of SUs we consider for measuring performance is chosen at random. More specifically, each time we run the simulation, we randomly consider a group of nodes that are at most two hops away from each other. Because the positions of the SUs change each time, their available channel sets (as well as their channel sequences) also change. We argue that, with these sorts of settings, the results offer better credibility. The values for t_b and t_s are set to 30 and 20, respectively.

4.2. Comparison of the Number of Broadcast Packets. As mentioned before, we compare our proposed protocol with the one proposed by Zhao et al. [14]. Both protocols use the beacon broadcast to discover neighbors and exchange available channel sets. But, unlike the proposal by Zhao et al. [14], we avoid forming coordination channels. Rather, with the available information of the nodes, we propose a contention-based resource allocation mechanism that greatly reduces the number of broadcast packets. This is evident from Figure 5, where we can see that the number of broadcast packets is always higher under the Zhao et al. protocol [14], compared to our proposed scheme.

4.3. Comparison of Connectivity. Connectivity is a major issue for a resource allocation mechanism, especially in the absence of any sort of coordination channel or CCC. A CCC or coordination channel provides a common platform for all the nodes to exchange information and negotiate a channel. Since no CCC or coordination channel is considered here, it is important to observe the nodes' rates of success when connecting with each other. To show connectivity performance, we considered two cases: (i) connectivity with nodes within

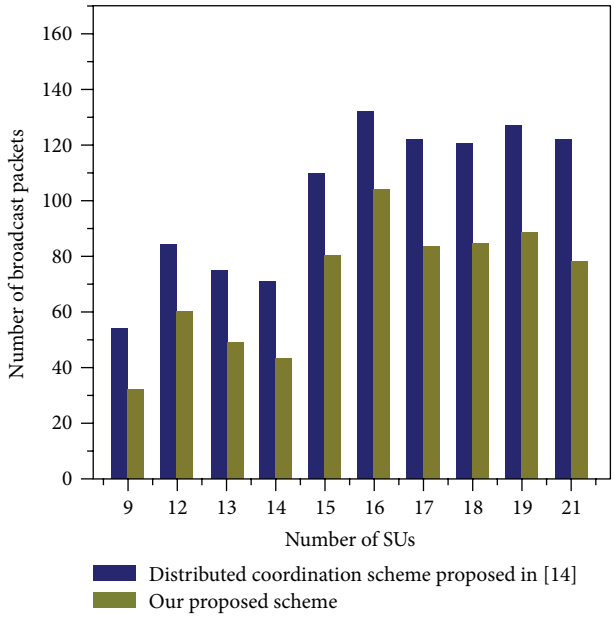


Figure 5: Number of broadcast packets when four PU channels are considered, and the number of SUs is chosen at random.

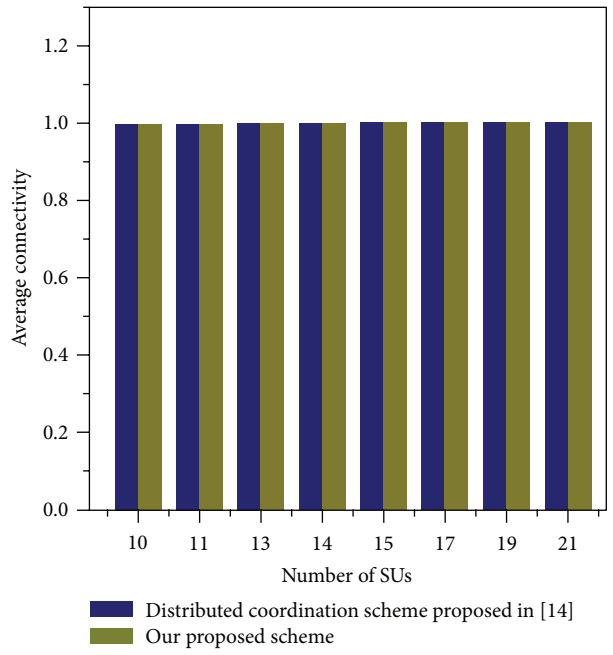


Figure 6: Average connectivity, according to the number of SUs, when four PU channels are considered. All receivers are within transmission range of senders. The number of SUs randomly changes between 10 and 21.

transmission range and (ii) connectivity when nodes are two hops away. The simulation result for Case (i) is shown in Figure 6. We can see that the performance is exactly equal under both protocols. Moreover, connectivity is 100% in both cases, which means that, with the current settings, any node

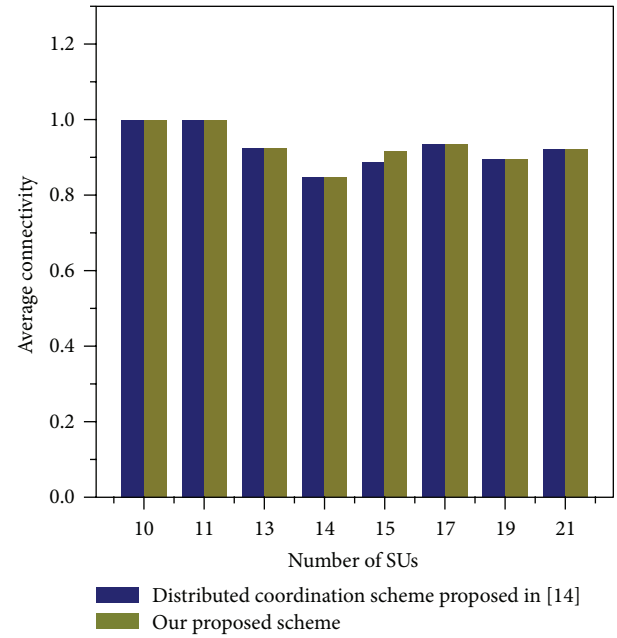


Figure 7: Average connectivity, according to the number of SUs for a multihop (two-hop) scenario when four PU channels are considered. The number of SUs randomly changes between 10 and 21.

can connect to any other node within its transmission range. The reason is that nodes within transmission range have very similar spectrum availability. In Figure 7, we can see the average connectivity in Case (ii). It is clear that connectivity is almost the same for both schemes, except for some minimal differences. This signifies that, with two-hop nodes, there can be some packets lost due to differences in spectrum availability.

5. Conclusion

A fully distributed mechanism for spectrum sharing and access is proposed in this paper. While a common strategy is to choose a CCC, of any form, to solve these problems, we propose a strategy that does not need a CCC. Rather, it works in a distributed fashion over available channels. This removes the need for extra coordination overhead caused by negotiation for a CCC. In a CRN, where channel availability differs temporally and spatially and possibly quite frequently, our proposed protocol prevents the network from being overwhelmed by coordination overhead. In simulations, we compared broadcast performance with a protocol that uses a coordination channel [14]. Results show that our proposed method reduces broadcasts significantly while keeping connectivity in the network similar to that of the reference work.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the KRF funded by the MEST (NRF-2013R1A2A2A05004535) and the Ministry of Education (2013R1A1A2063779).

References

[1] M. McHenry, "Spectrum white space measurements," in *Proceedings of the New America Foundation Broadband Forum*, June 2003.

[2] J. Mitola III and G. Q. Maguire Jr., "Cognitive radio: making software radios more personal," *IEEE Personal Communications*, vol. 6, no. 4, pp. 13–18, 1999.

[3] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, 2005.

[4] M. Yoon, Y.-K. Kim, and J.-W. Chang, "An energy-efficient routing protocol using message success rate in wireless sensor networks," *Journal of Convergence*, vol. 4, no. 1, pp. 15–22, 2013.

[5] A. Kusdaryono, K. O. Lee, and Y. Lee, "A clustering protocol with mode selection for wireless sensor network," *Journal of Information Processing System*, vol. 7, no. 1, pp. 29–42, 2011.

[6] Y. Yang, F. Zhen, T.-S. Lee, and M.-S. Park, "TASL: a traffic-adapted sleep/listening MAC protocol for wireless sensor network," *Journal of Information Processing Systems*, vol. 2, no. 1, pp. 39–43, 2006.

[7] S. Kumar, M. Sarkar, S. Gurajala, and J. D. Matyjas, "MMMP: a MAC protocol to ensure QoS for multimedia traffic over multihop ad hoc networks," *Journal of Information Processing Systems*, vol. 4, no. 2, pp. 41–52, 2008.

[8] A. Sinha and D. K. Lobiyal, "Performance evaluation of data aggregation for cluster-based wireless sensor network," *Human-Centric Computing and Information Sciences*, vol. 3, no. 1, pp. 1–17, 2013.

[9] S. P. Algur and N. P. Kumar, "Novel user centric, game theory based bandwidth allocation mechanism in WiMAX," *Human-Centric Computing and Information Sciences*, vol. 3, no. 1, article 20, 2013.

[10] C. Cordeiro, K. Challapali, D. Birru, and N. Sai Shankar, "IEEE 802.22: the first worldwide wireless standard based on cognitive radios," in *Processding of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, pp. 328–337, Baltimore, Md, USA, November 2005.

[11] S.-Y. Lien, C.-C. Tseng, and K.-C. Chen, "Carrier sensing based multiple access protocols for cognitive radio networks," in *Proceedings of the IEEE International Conference on Communications (ICC '08)*, pp. 3208–3214, Beijing, China, May 2008.

[12] Y. Lee and I. Koo, "A distributed MAC protocol using virtual control channels for CRSNs," *Wireless Personal Communications*, vol. 71, no. 2, pp. 1021–1048, 2013.

[13] J. Jia, Q. Zhang, and X. Shen, "HC-MAC: a hardware-constrained cognitive MAC for efficient spectrum management," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 106–117, 2008.

[14] J. Zhao, H. Zheng, and G.-H. Yang, "Distributed coordination in dynamic spectrum allocation networks," in *Proceedings of the 1st IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '05)*, pp. 259–268, Baltimore, Md, USA, November 2005.

[15] S. K. Timalsina, S. Moh, I. Chung, and M. Kang, "A concurrent access MAC protocol for cognitive radio ad hoc networks without common control channel," *Eurasip Journal on Advances in Signal Processing*, vol. 2013, no. 1, article 69, 2013.

[16] Y. Song and J. Xie, "A distributed broadcast protocol in multihop cognitive radio ad hoc networks without a common control channel," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '12)*, pp. 2273–2281, Orlando, Fla, USA, March 2012.

[17] J. Liu and S.-H. Chung, "An efficient load balancing scheme for multi-gateways in Wireless Mesh Networks," *Journal of Information Processing Systems*, vol. 9, no. 3, pp. 365–378, 2013.

[18] H.-R. Lee, K.-Y. Chung, and K.-S. Jhang, "A study of wireless sensor network routing protocols for maintenance access hatch condition surveillance," *Journal of Information Processing Systems*, vol. 9, no. 2, pp. 237–246, 2013.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 105245, 14 pages
http://dx.doi.org/10.1155/2015/105245

Research Article

Modeling Routing Overhead of Reactive Protocols at Link Layer and Network Layer in Wireless Multihop Networks

N. Javaid,[1,2] Z. A. Khan,[3] U. Qasim,[4] M. Jamil,[5] M. Ishfaq,[6] and T. A. Alghamdi[7]

[1]*CAST, COMSATS Institute of Information Technology, Islamabad, Pakistan*
[2]*EE Dept., COMSATS Institute of Information Technology, Islamabad, Pakistan*
[3]*Internetworking Program, FE, Dalhousie University, Halifax, Canada*
[4]*University of Alberta, Alberta, Canada*
[5]*RAI Department, SMME, National University of Sciences & Technology (NUST), Islamabad, Pakistan*
[6]*King Abdulaziz University, Rabigh, Saudi Arabia*
[7]*Umm Al-Qura University, Makkah, Saudi Arabia*

Correspondence should be addressed to N. Javaid; nadeemjavaid@comsats.edu.pk

Received 23 March 2014; Revised 26 August 2014; Accepted 8 September 2014

Academic Editor: Young-Sik Jeong

To keep information recent between two nodes, two types of link sensing feed-back mechanisms are used: link layer (LL) and network layer (NL). In this paper, we model and evaluate these link sensing mechanisms in three widely used reactive routing protocols: ad hoc on-demand distance vector (AODV), dynamic source routing (DSR), and dynamic MANET on-demand (DYMO). Total cost paid by a routing protocol is the sum of cost paid in the form of energy consumed (in terms of packet reception/transmission) and time spent (in terms of processing route information). Routing operations are divided into two phases: route discovery (RD) and route maintenance (RM). These protocols majorly focus on broadcast cost optimization performed by expanding ring search (ERS) algorithm to control blind flooding. Hence, our model relates link sensing mechanisms in RD and RM for the selected routing protocols to compute consumed energy and processing time. The proposed framework is evaluated via NS-2, where the selected protocols are tested with different nodes' mobilities and densities.

1. Introduction

Recent research mainly focusses on wireless multihop networks (WMhNs) due to increased use of wireless devices all around. A mobile node, in WMhNs, acts as a transmitter and as a router (relay node) for nodes not in the direct transmission range of each other. Applications of these networks range from a small room to large areas like a battlefield or a natural disaster. Performance of WMhNs majorly depends on the routing protocols operating these networks.

On the basis of route calculation, routing protocols are divided into two major categories: reactive and proactive. Protocols from the former category calculate routes when data request arrives, thus, also called "on-demand" protocols. Examples of these protocols are ad hoc on-demand distance vector (AODV) [1, 2], dynamic source routing (DSR) [3, 4], dynamic MANET on-demand (DYMO) [5, 6], and so forth, whereas protocols belonging to the later category periodically perform routing table calculation independently from data request arrival. These protocols include destination-sequence distant vector (DSDV) [7], fish-eye state routing (FSR) [8, 9], optimized link state routing (OLSR) [10, 11], and so forth.

Reactive protocols are well suited for highly dynamic networks, whereas proactive ones are designed for large networks with low mobility. Reactive protocols exchange lot of control (routing) information to accurately route data within or outside the network. There are two main phases in which these protocols calculate routing information: route discovery (RD) and route maintenance (RM). After computation and establishment of a route for the requested destination during RD phase, RM phase starts if a link breaks. The first step is to perform periodic links' sensing

in active route (which are established during RD for data transmission). Links are sensed by the routing protocols either from link layer (LL) or from network layer (NL). After detecting a link breakage while sensing, the second step (RM phase) repairs the link. The control messages generated by routing protocols and the time which is spent during RD and RM, collectively generate routing overhead.

In this paper, we model routing overhead produced by three reactive protocols: AODV, DSR, and DYMO. We choose the selected three routing protocols because these are widely used in literature. Our main focus is to measure routing overhead for LL and NL feed-back mechanisms. To analyze the link sensing mechanisms of AODV, DSR, and DYMO, we conduct simulations in NS-2. The overhead is measured for nodes' different mobilities and densities.

2. Related Work and Motivation

2.1. Related Work. Routing protocols play an important role for performance optimization in wireless networks. Many protocols, including bioinspired routing [12, 13], security based routing [14, 15], and balanced load routing [16, 17], have been proposed for wireless networks. However, we focus on the modeling of overhead for reactive routing protocols at NL and LL. In [18], authors focus on quality routing link metrics for wireless multihop networks. Authors in [19] address performance evaluation of two on-demand routing protocols in WMhNs, DSR, and AODV. They simulate these protocols for 20 and 40 number of sources with speeds of 2 m/s and 6 m/s. Simulation results show that AODV protocol is more suitable as compared to DSR for wireless transmission with rapid change of network topology.

In [20], authors compare AODV and DYMO, using packet-level simulations for different speeds (1 m/s, 9 m/s, and 15 m/s) in NS-2. They select throughput, routing overhead, and average packet size of the routing packets as performance metrics. From simulation results, they conclude that AODV achieves higher throughput than DYMO. This is because the distance vector information of AODV consumes less bandwidth than source routing of DYMO.

The control overhead of ad hoc routing protocols is surveyed in [21]. The authors classify reactive and proactive protocols as "hello protocol" and "flooding protocol," respectively. They conclude through simulations that "hello protocols" generate more control overheads than "flooding protocols" in mobile scenarios. Hence, "hello protocols" are more suitable for static scenarios and "flooding protocols" for mobile scenarios.

The failure or unability of a routing protocol to identify all disjoint paths between a pair of nodes is called path diminution. In [22], this phenomenon is studied. The paper states that diminution of path becomes unavoidable whenever a protocol becomes aware of multiple disjoint paths while discovering a single route. In order to mitigate path diminution, the paper discusses various schemes. However, as per conclusion none of the discussed schemes guarantee to discover all disjoint paths between a pair of nodes.

Saleem et al. in [23] analytically model and compare the routing control overhead. The flooding mechanism of reactive protocols is compared with relay node technique of proactive protocols.

Packet drop and link failure significantly degrade network performance. As the root cause of these problems is congestion so [24] presents congestion and link failure aware data delivery mechanism. This work jointly considers control approaches, congestion detection, and buffering while increasing reliability of delivering data. Simulation results validate better performance of the proposed mechanism as compared to the selected existing ones in terms of end to end delay, throughput, and reliability.

Saleem et al. [25] propose a framework for flooding cost in routing protocols. They evaluate the framework using two key performance metrics: routing overhead and route optimality for DSDV, DSR, AODV-LL, and gossiping.

A reactive traffic-aware routing strategy for urban vehicular environments has been proposed in [26]. The beauty of this strategy is the avoidance of dead ends and unnecessary routes. In this work, dynamic paths are created on the basis of prior global knowledge of the traffic for each vehicle. Moreover, this prior global knowledge is used by decision making nodes while taking critical decisions.

As mobility and PHY/MAC layers implementation affect the routing protocols in ad hoc networks so [27] presents simulation based analysis of the two reactive routing protocols: DSR and AODV. The simulations are carried with modified IEEE 802.11a (PHY/MAC layers) along with modified mobility models (freeway, traffic sign, and stop sign). From OPNET based simulation results, they conclude that modified version of AODV performs better than DSR in terms of delay, throughput, and number of retransmission attempts.

2.2. Motivation. In our previous work [28], we model the routing overhead incurred by AODV, DSR, and DYMO in terms of energy and time costs for the generated control packets. We presented a framework for RD and RM of the protocols. These protocols mainly focus on the optimization performed by expanding ring search (ERS) algorithm to minimize the overhead generated due to blind flooding. The proposed framework is evaluated via NS-2 to compare performance of the chosen routing protocols.

A comparative evaluation of AODV and DYMO is also presented in [29]. Both the protocols are compared through simulations using LL feed-back mechanism.

Inspired from [28, 29], we enhance the framework of [29] for LL and NL link sensing mechanisms. Moreover, to validate the proposed framework, we simulate AODV, DSR, and DYMO with two link sensing mechanisms: LL and NL. Simulations are performed against varying network sizes and nodes' mobilities.

3. Modeling the Cost Paid by Reactive Protocols

Flooding is a process used by routing protocols to exchange routing information throughout the network. In this process, each node acts as router and each node broadcasts route information to all of its neighbors. This process is repeated until destination is reached. Blind flooding is a simple approach in which each node rebroadcasts the packet, whenever it receives packet for the first time.

In [25], an approximate per packet cost paid by a protocol for RD using blind flooding is presented as

$$C_p = \begin{cases} P_s d_{\text{avg}}, & \text{if } h = 1, \\ P_s d_{\text{avg}} + d_{\text{avg}} \sum_{i=1}^{h-1} (P_s)^{i+1} \prod_{j=1}^{i} d_f[j], & \text{if } h > 1, \end{cases} \quad (1)$$

where d_{avg} denotes the average degree or the average number of immediate one-hop neighbors of a node. A node is isolated, if it has $d = 0$. In (1), h is the number of hops in the network, $d_f[j]$ is the expected forward degree of a node at jth hop, and P_s is the broadcast probability [30].

ERS [31], adopted by AODV, DSR, and DYMO, is one of the optimization techniques that have been proposed to control the routing overhead. As ERS is adopted by the three reactive routing protocols so we focus on its working. This technique sets search diameters, based on time-to-live (TTL) value in RD phase, to limit broadcast overhead. In ERS, these search diameters are known as rings. In order to calculate the packet cost of a ring "R_i," h is replaced by TTL_VALUE of that ring. Let $C_E(R_i)$ be the cost of any ring R_i. The selected protocols set TTL_VALUE in route REQuest (RREQ) message. In ERS, RD can either be successful or not; successful RD is stopped after receiving route REPly (RREP) message, whereas unsuccessful RD results in dissemination up to maximum TTL_VALUE and rediscovery attempts. An R_i which generates RREP(s) is called R_{rrep} and ring up to maximum TTL limit is known as $R_{\text{max\_limit}}$ resulting in either successful or unsuccessful RD. Therefore, $C_E(R_i)$ is computed from our previous work in [28] as

$$C_E(R_i) = \begin{cases} P_r d_{\text{avg}}, & \text{if TTL}(R_i) = 1, \\ P_r d_{\text{avg}} + d_{\text{avg}} \sum_{\text{TTL}=1}^{\text{TTL}(R_i)-1} (P_r)^{\text{TTL}+1} \prod_{j=1}^{\text{TTL}} d_f[j], & \\ & \text{if TTL}(R_i) > 1. \end{cases} \quad (2)$$

In ERS, RD requires adjustments in TTL values to find destination. If source node fails to find destination, then TTL is incremented. In ERS, gradual growth of broadcast ring takes place to reduce the chances of flooding throughout the network which results in different rings for different broadcast levels. The cumulative routing cost for expanding rings, during RD process "$C_{E\text{-RD}}$," is computed as

$$C_{E\text{-RD}} = \begin{cases} \sum_{R_i=1}^{R_{\text{max\_limit}}} (C_E(R_i)) \times (\text{RREQ}), & \\ & \text{if no RREP received,} \\ C_{E\text{-}R_{\text{rrep}}} & \text{if TTL}(R_{\text{rrep}}) = 1, \\ \sum_{R_i=1}^{R_{\text{rrep}}} (C_{E\text{-}R_i})_{R_i}, & \text{otherwise,} \\ \{R_{\text{rrep}} = 1, 2, 3, \ldots, \text{max\_limit}\}. & \end{cases} \quad (3)$$

On encountering a dynamic topology due to varying number of nodes and mobility rates, a routing protocol "p" pays some cost in the form of per packets consumed energy "C_E^p" and in the form of per packet time spent "C_T^p". In [28], authors have computed this cost as

$$C_{\text{total}}^p = C_E^p \times C_T^p. \quad (4)$$

3.1. Cost of Energy Consumption. Each reactive protocol performs two routing operations: RD and RM. Therefore, we define the cost for energy consumption, during RD and RM processes, C_E^p, respectively

$$C_E^p = C_{E\text{-RD}}^p + C_{E\text{-RM}}^p. \quad (5)$$

C_E^p is different for each reactive protocol due to different routing strategies. For example, in DSR, multiple routes in route cache (RC) reduce the routing overhead with the help of gratuitous RREPs (grat. RREPs) and packet salvaging (PSing), whereas, in AODV, route length is shortened by grat. RREPs to reduce the cost of RD process and successful Local Link Repair (LLR) process diminishes route rediscoveries.

Receiver of the route REQuest (RREQ) generates a route REPly (RREP) to the originator, if it either is itself the destination or it contains an active route (AR) to the destination (also known as gratuitous RREPs (grat. RREPs)). Grat. RREPs are generated if the node generating the RREP is not the destination itself but is a substitute node along the path from originator to destination.

3.1.1. Energy Consumed during RD. AODV, DSR, and DYMO use ERS mechanism for RD via broadcasting the RREQ messages from the source node. Successful RD results in unicast of RREP message to the originator node. Depending upon the generating node, the RREP message is of two types: dest. RREP and grat. RREP. An RREP which is generated from destination node is known as dest. RREP which is used by all the three reactive protocols. A source node can receive RREP from the nodes that contain alternate (short) route to the desired destination. These replies are only supported in AODV and DSR and are known as grat. RREPs.

```
(1) if R = active_mode then
(2)   for all l ∈ AR do
(3)     start LSM
(4)     if upstream node detects Link Breakage (LB) then
(5)       start LLR
(6)       if successful repair through LLR then
(7)         repairing node sends data to repaired route
(8)       else
(9)         broadcast RERR message
(10)        receiver of RERR deletes faulty route from RT
(11)        S starts route re-discovery based on RREQ_RETRIES
(12)      end if
(13)    else
(14)      repeat
(15)    end if
(16)  end for
(17) else
(18)   no action is performed for link checking
(19) end if
```

Algorithm 1: Route Maintenance in AODV.

The cost paid for RREQ packets and for RREPs, produced during RD, is computed as

$$C_{E\text{-RD}} = \begin{cases} \sum_{R_i=1}^{R_{\max\_limit}} \left(C_E\left(R_i\right)\right) \times (\text{RREQ}) \\ \qquad \text{if no RREP received} \\ C_E R_{\text{rrep}} \times (\text{RREQ}) \\ \qquad \text{if TTL}\left(R_{\text{rrep}}\right) = 1 + \sum_{n=1}^{n_{\text{rrep}}} (\text{RREP})_n \\ \sum_{R_i=1}^{R_{\text{rrep}}} \left(C_E\left(R_i\right)\right) \times (\text{RREQ}) + \sum_{n=1}^{n_{\text{rrep}}} (\text{RREP})_n \\ \qquad \text{otherwise,} \end{cases} \tag{6}$$

where n_{rrep} notation is used for number of nodes which unicast RREP to the sender. The ERS mechanism of AODV and DYMO is shown in Figure 1(a) with TTL_VALUES and waiting_time using LL feed-back mechanism, whereas Figure 1(b) shows NL feed-back mechanism.

The generation of RREP(s) in AODV and DSR (refer Figure 2) is also due to valid routes in routing table (RT) or RC. Therefore, R_i for DSR and AODV is less than DYMO due to the absence of grat. RREPs in DYMO. R_{rrep} can be $1, 2, 3, \ldots,$ max_limit, depending upon the hop-distance from source to destination.

3.1.2. Energy Consumed during RM. After the establishment of a successful route, during RD process, the next process is to perform link state monitoring (LSM). During this process, links of active routes (ARs) are sensed, as we have mentioned earlier that this sensing can be performed at NL and LL.

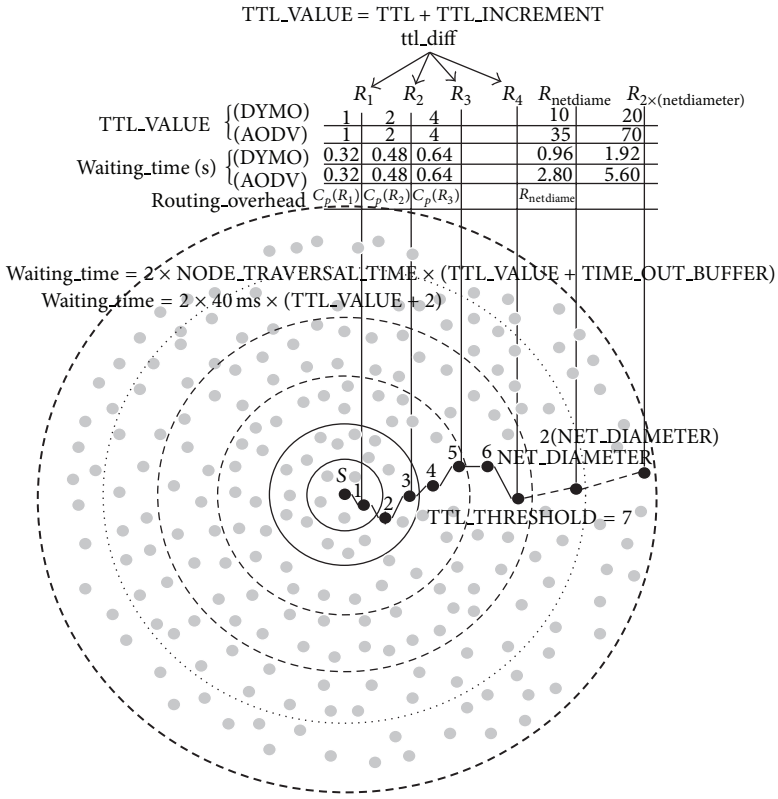
During RM process, different protocols pay different link monitoring costs and also use different supplementary maintenance strategies in case of link breakage. Therefore, this cost metric depends upon the respective routing protocols; $C_{E\text{-LLR}}^{\text{AODV}}$ for AODV and $C_{E\text{-PS}}^{\text{DSR}}$ for DSR, whereas DYMO does not use any supplementary mechanism. RM process for AODV, DSR, and DYMO is given in Algorithms 1, 2, and 3, respectively.

After the detection of route failure due to link breakage, there are three different scenarios for reactive protocols describing the route repair mechanism. The most simple mechanism describes that RD reinitiation process takes place under limited retries for route rediscovery process: RREQ_RETRIES = 3 in DYMO, RREQ_TRIES = 2 in AODV, and MaxMaintRexmt in DSR = 2 retransmissions.

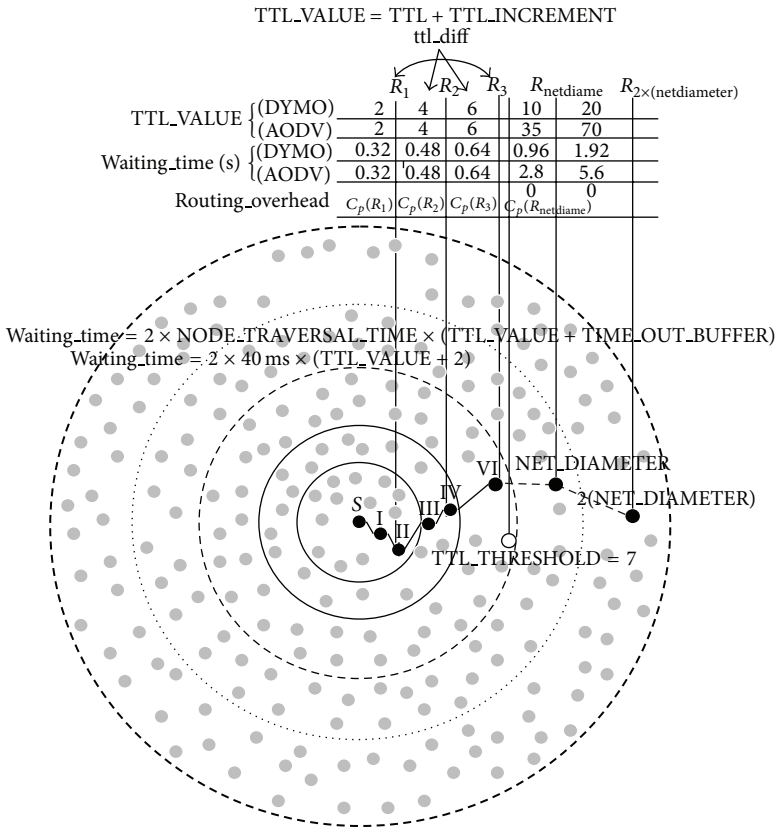
In AODV, after unsuccessful LLR and in DYMO ultimately after detecting link breakage, RERR messages are broadcasted by the node which detects any link breakage. If LLR becomes successful, then in a dense network, it saves the energy which is consumed for route rediscovery, otherwise reinitiating RD process after performing LLR strategy increases the energy consumption. DSR's PS technique reduces both the energy and time costs paid by the reactive protocols by diminishing the route rediscovery. In case of successful PS, RERR messages are broadcasted to neighbors for the deletion of useless routes, whereas absence of alternate route(s) in RC leads to failure of PS. In this situation, RERR messages are piggy-backed in the next RREQ messages during route rediscovery process. There are some approaches for detecting link breakages. In this paper, we focuss on two approaches for the detection of link breakage. In the first approach, link sensing is performed at link layer, which uses LL feed-back to check link's status. This checking is performed 100 times in every second. The principle behind notifying link breakage depends on failure of link level feedback.

In Algorithms 1, 2, and 3, we use "l" for link, "S" for source node, and "R" for route.

If a node in AR receives eight consecutive failures, then it notifies broken link status. The second approach uses sending and receiving of beacon messages on NL (i.e.,

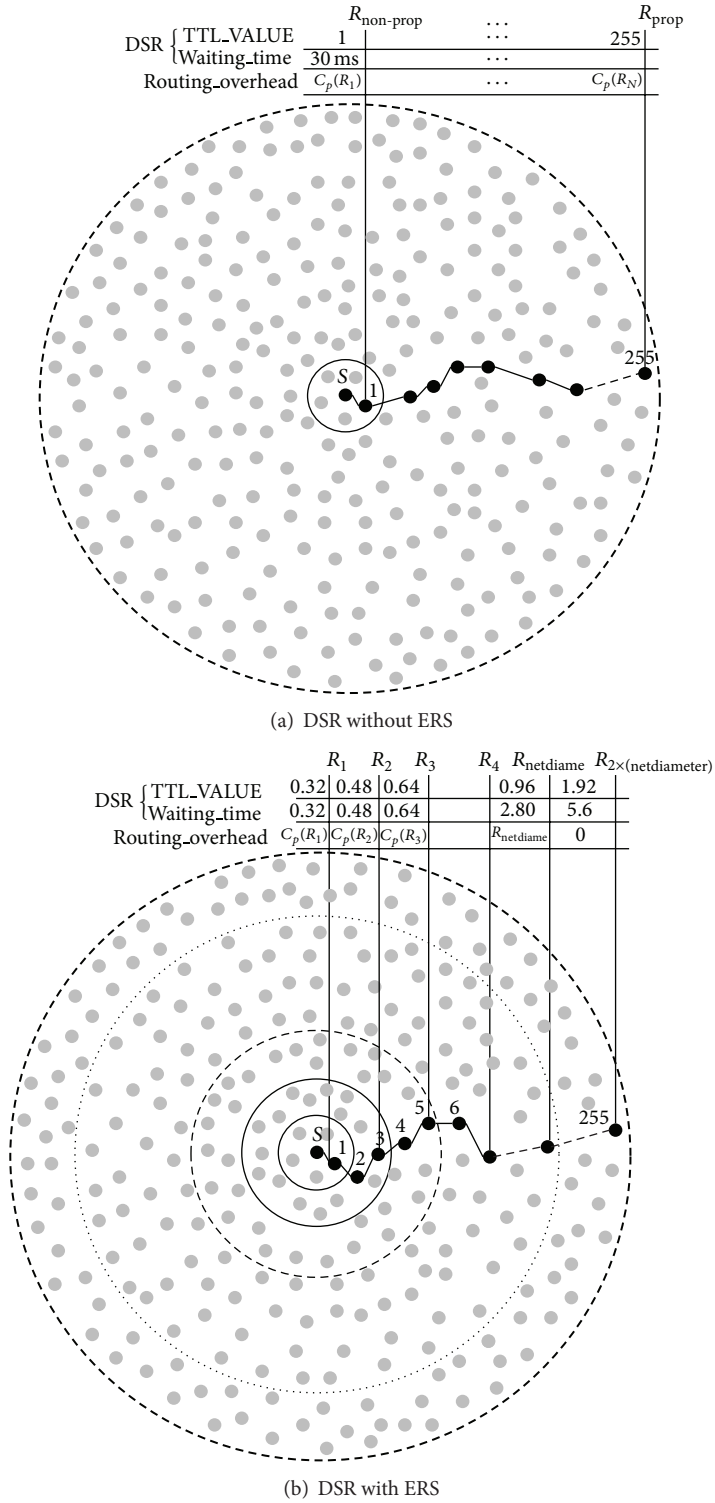


(a) AODV/DYMO RD using LL feed-back



(b) AODV/DYMO RD using NL feed-back

Figure 1: AODV and DYMO: waiting_time and TTL_VALUE for RD.



(a) DSR without ERS

(b) DSR with ERS

Figure 2: DSR: waiting_time and TTL_VALUE for RD.

HELLO messages and ACKnowledgement ACK) to check connectivity of ARs. If ACK is not received, after a specific number of tries, then link is notified as broken. The example of this approach is the use of HELLO message in AODV. HELLO messages are sent after every HELLO_INTERVAL and if ACK is not received for ALLOWED_HELLO_LOSS value, then link is considered as broken.

Following equations give RM cost for the three protocols with NL and LL feed-back mechanisms. Cost of LLR in AODV is given by (8), where R_{LLR} denotes the ring that

```
(1) for all l ∈ AR do
(2)    start LSM
(3)    if upstream node detects LB then
(4)       remaining nodes of path search alternative route in RC
(5)       if found alternative R then
(6)          send data to this R
(7)       else
(8)          S starts route re-discovery
(9)          S piggy-backs RERR with new RREQ
(10)      end if
(11)   else
(12)      upstream nodes receive LL feed-back from downstream nodes
(13)   end if
(14) end for
```

Algorithm 2: Route Maintenance in DSR.

```
(1) if R = active_mode then
(2)    for all l ∈ AR do
(3)       start LSM
(4)       if upstream node detects LB then
(5)          disseminate RERR message
(6)          receiver of RERR deletes faulty route from RT
(7)          S starts route re-discovery based on RREQ_TRIES
(8)       else
(9)          check links of AR
(10)      end if
(11)   end for
(12) else
(13)   do not perform any action
(14) end if
```

Algorithm 3: Route Maintenance in DYMO.

limits LLR activity. TTL_VALUE for R_{LLR} is calculated with LOCAL_ADD_TTL(= 2) and MIN_REPAIR_TTL (i.e., the last known hop-count to the destination). The per packet cost of LLR in AODV, $C_{E\text{-LLR}}^{\text{AODV}}$, depends upon the TTL_VALUE of R_{LLR}. Consider

$$C_{E\text{-RM}}^{\text{AODV}} = \begin{cases} C_{E\text{-LSM}} + C_{E\text{-LLR}}^{\text{AODV}} + \sum_{z=0}^{n} (\text{RERR})_z \\ \qquad \text{for NL feed-back} \\ C_{E\text{-LLR}}^{\text{AODV}} + \sum_{z=0}^{n} (\text{RERR})_z \\ \qquad \text{for LL feed-back} \end{cases} \tag{7}$$

$$\begin{aligned} &C_{E\text{-LLR}}^{\text{AODV}}(R_{\text{LLR}}) \\ &\quad = P_r d_{\text{avg}} + d_{\text{avg}} \sum_{\text{TTL}=1}^{\text{TTL}(R_{\text{LLR}})-1} (P_r)^{\text{TTL}+1} \prod_{j=1}^{\text{TTL}} d_f[j]. \end{aligned} \tag{8}$$

In large networks, successful LLR process is more useful because the chances of route rediscovery are reduced, which consumes more bandwidth. $\text{TTL}(R_{\text{LLR}})$ is computed as

$$\begin{aligned} &\max(\text{MIN\_REPAIR\_TTL}, 0.5 \times \#\text{hops}) \\ &\quad + \text{LOCAL\_ADD\_TTL}, \end{aligned} \tag{9}$$

where #hops represents the number of hops to the sender of the currently undeliverable data packet. TTL_VALUE for LLR is calculated from TTL $\geq$ MIN_REPAIR_TTL + LOCAL_ADD_TTL expression. Consider

$$C_{E\text{-RM}}^{\text{DSR}} = \begin{cases} C_{E\text{-LSM}} + C_{E\text{-PS}}^{\text{DSR}} + \sum_{z=0}^{n} (\text{RERR})_z \\ \qquad \text{for NL feed-back} \\ C_{E\text{-PS}}^{\text{DSR}} + \sum_{z=0}^{n} (\text{RERR})_z \\ \qquad \text{for LL feed-back} \end{cases}$$

$$C_{E\text{-PS}}^{\text{DSR}} = \sum_{k=n_{\text{BLB}}}^{n_{\text{PS}}} (\text{RREQ})_k$$

$$C_{E\text{-RM}}^{\text{DYMO}} = \begin{cases} C_{E\text{-LSM}} + \sum_{z=0}^{n} (\text{RERR})_z \\ \qquad \text{for NL feed-back} \\ \sum_{z=0}^{n} (\text{RERR})_z \\ \qquad \text{for LL feed-back,} \end{cases} \tag{10}$$

where $C_{E\text{-LSM}}$ is the LSM cost on NL, and is given below

$$C_{E\text{-LSM}} = \frac{\tau_{\text{AR}}}{\tau_{\text{LSM}}} \times N_{\text{AR}}, \tag{11}$$

where n_{BLB} is the node before link breakage and n_{PS} is any node from source to n_{BLB}. In wireless environment, route failures are due to frequent link breakages. We denote LSM cost at NL by $C_{E\text{-LSM}}$ and depending upon the link duration in AR, τ_{AR}, τ_{LSM} is the periodic interval for link sensing and N_{AR} represent the number of hops in an active route.

Broadcast needs to send z number of RERRs depending upon different situations for different protocols. In DYMO, link breakage causes the broadcasting of RERR messages. In AODV, when the probability of successful LLR becomes zero then it leads to the dissemination of RERRs. On the other hand, DSR piggybacks RERR messages along with next RREQs in case of route rediscovery process, while these RERR messages are broadcasted in case of PS success.

3.2. Cost of Time Consumption. The cost of end-to-end path computation time "C_T^p" in reactive protocols depends upon $C_{T\text{-RD}}^p$ and $C_{T\text{-RM}}^p$. Consider

$$C_T^p = C_{T\text{-RD}}^p + C_{T\text{-RM}}^p \tag{12}$$

with details in the following subsections.

3.2.1. Time Consumed during RD for DSR. Let τ denote constant time initially used for (NonpropRequestTimeout) [3] and its value is 30 ms. $R_{\text{max\_limit}}$, the maximum ring size, depends on the buffer time as well as the maximum allowed broadcasting during propagating RREQ (DiscoveryHopLimit = 255) [3]. Binary exponential back-off (BEB) is associated with each propagating ring (BEB mechanism doubles previous TTL_Value (refer Figure 1(b)) and waiting_time to reduce routing overhead). The expression for time consumed during RD by DSR "C_T^p" is given below:

$$C_{T\text{-RD}}^{\text{DSR}} = \begin{cases} \tau, & \text{if } R_{\text{rrep}} = 1, \\ \sum_{R_i=1}^{R_{\text{max\_limit}}} 2^{R_i - 1} \times \tau, & \text{if no RREP received,} \\ \sum_{R_i=1}^{R_{\text{rrep}}} 2^{R_i - 1} \times \tau, & \text{otherwise.} \end{cases} \tag{13}$$

3.2.2. Time Consumed during RD for AODV and DYMO. In both AODV and DYMO, TTL_VALUE (in IP header) is set to TTL_START (= 1 in case of link layer feedback otherwise = 2); then it is incremented by TTL_INCREMENT(= 2) up to TTL_THRESHOLD(= 7) [2]. When TTL_THRESHOLD is reached, TTL_VALUE is set to NET_DIAMETER (for AODV = 35 [2] and for DYMO = 10 [4]). For dissemination in the entire network, both TTL_START and TTL_INCREMENT are set to NET_DIAMETER. Moreover, maximum RREQ retries are 3 for DYMO [4] and maximum retries are 2 for AODV. The RREQ_TIME is set to 2 × NET_TRAVERSAL_TIME (whereas, NET_TRAVERSAL_TIME = 2 × NODE_TRAVERSAL_TIME × NET_DIAMETER). See complete detail of TTL_VALUE and waiting_time in Figure 1. Consider

$$C_{T\text{-RD}}^{\text{AODV,DYMO}} = \begin{cases} \sum_{R_i=1}^{R_{\text{max\_limit}}} \tau_1 \left(\text{TTL}\left(R_i\right) + \tau_2\right) \\ \qquad \text{if no RREP received} \\ \sum_{R_i=1}^{R_{\text{rrep}}} \tau_1 \left(\text{TTL}\left(R_i\right) + \tau_2\right) \\ \qquad \text{otherwise,} \end{cases} \tag{14}$$

where τ_1 = 2 × NODE_TRAVERSAL_TIME and τ_2 = TIME_OUT_BUFFER. There are two possibilities for AODV and DYMO: first case, when RD process becomes successful in threshold rings "$R_{\text{thereshold}}$," whereas in second case RD process needs to disseminate the request throughout the network, $R_{\text{netdiameter}}$. For these two rings, we define TTL($R_{\text{thereshold}}$) and TTL($R_{\text{netdiameter}}$). The earlier one represents TTL_VALUE in a ring that generates RREP(s) inside $R_{\text{thereshold}}$ with THRESHOLD and the later one shows TTL_VALUE for the entire network: TTL($R_{\text{netdiameter}}$) with NET_DIMETER.

3.2.3. Time Consumed during RM for AODV. AODV starts LLR process after noticing a link failure. $C_{T\text{-LLR}}$ gives the time cost of LLR which depends upon TTL_VALUE of the ring, R_{LLR}. In case of LLR failure, AODV disseminates RERR messages. $\tau_{\text{recv-RERR}}$ represents the time spent to reach RERR message from the node which detects link failure to the originator node. $C_{T\text{-re-RD}}^{\text{AODV}}$ (= $C_{T\text{-RD}}^{\text{AODV}}$) cost is to be paid to start route rediscovery based on the value RREQ_RETRIES(= 2). This cost is given as

$$C_{T\text{-RM}}^{\text{AODV}} = \begin{cases} \sum_{R_i=1}^{R_{\text{LLR}}} \tau_1 \left(\text{TTL}\left(R_i\right) + \tau_2\right) \\ \qquad \text{if LLR is successful} \\ \sum_{R_i=1}^{R_{\text{LLR}}} \tau_1 \left(\text{TTL}\left(R_i\right) + \tau_2\right) + \tau_{\text{recv-RERR}} \\ \qquad \text{if LLR fails} \\ \sum_{R_i=1}^{R_{\text{LLR}}} \tau_1 \left(\text{TTL}\left(R_i\right) + \tau_2\right) + \tau_{\text{recv-RERR}} + C_{T\text{-re-RD}} \\ \qquad \text{otherwise.} \end{cases} \tag{15}$$

3.2.4. Time Consumed during RM for DSR. After detecting a link failure, time τ_{PS} is utilized to check alternative routes in RC of intermediate nodes (from a node which detects link failure to a node having alternate route for this broken link, n_{PS}). In case of failure of PSing or in the case of presence of alternative route in RCing of the originator node, the cost for consumed time during RM for DSR is given as

$$C_{T\text{-RM}}^{\text{DSR}} = \begin{cases} \sum_{k=n_{\text{BLB}}}^{n_{\text{PS}}} \tau_k\,(\text{PS}), & \text{if PS is successful,} \\ \sum_{k=n_{\text{BLB}}}^{n_{\text{orig}}} \tau_k\,(\text{PS}) + C_{T\text{-re-RD}}^{\text{DSR}}, & \text{otherwise,} \end{cases} \tag{16}$$

where $C_{T\text{-re-RD}}^{\text{DSR}} = C_{T\text{-RD}}^{\text{DSR}}$, n_{BLB} is the node just before link breakage, and n_{orig} is the node which originates RD process.

3.2.5. Time Consumed during RM for DYMO. An RERR message is broadcasted by the node which detects link breakage. After $\tau_{\text{recv-RERR}}$ is consumed by the source node for receiving RERR message, source node initiates RD; $C_{T\text{-re-RD}}^{(\text{DYMO})}$ is based on RREQ_RETRIES(= 3). This cost is computed as

$$C_{T\text{-RM}}^{\text{DYMO}} = \begin{cases} \tau_{\text{recv-RERR}}, & \text{if RREQ\_TRIES expires,} \\ \tau_{\text{recv-RERR}} + C_{T\text{-re-RD}}^{(\text{DYMO})}, & \text{otherwise.} \end{cases} \tag{17}$$

4. Analytical Simulation Results Corresponding to the Designed Framework

We evaluate performance of our modeled framework in NS-2. For simulation setup, we have chosen continuous bit rate (CBR) traffic sources with packet size of 512 bytes. Nodes are dispersed in 1000 m × 1000 m of network square space allowing mobile nodes to move inside the network area. Links are provided with bandwidth of 2 Mbps to transmit on. We consider three performance metrics: packet delivery rate (PDR), end-to-end delay (E2ED), and normalized routing load (NRL) for our analysis. We simulate 50 nodes with variable pauses from 0 s to 900 s at 30 m/s for mobility analysis, whereas nodes with different densities, from 10 to 100, are simulated with 15 m/s and a constant pause of 2 s using random way point (RWP) mobility model. The random way point model is chosen due to the following reasons: (i) simplicity in implementation, and (ii) it meets all the required design considerations.

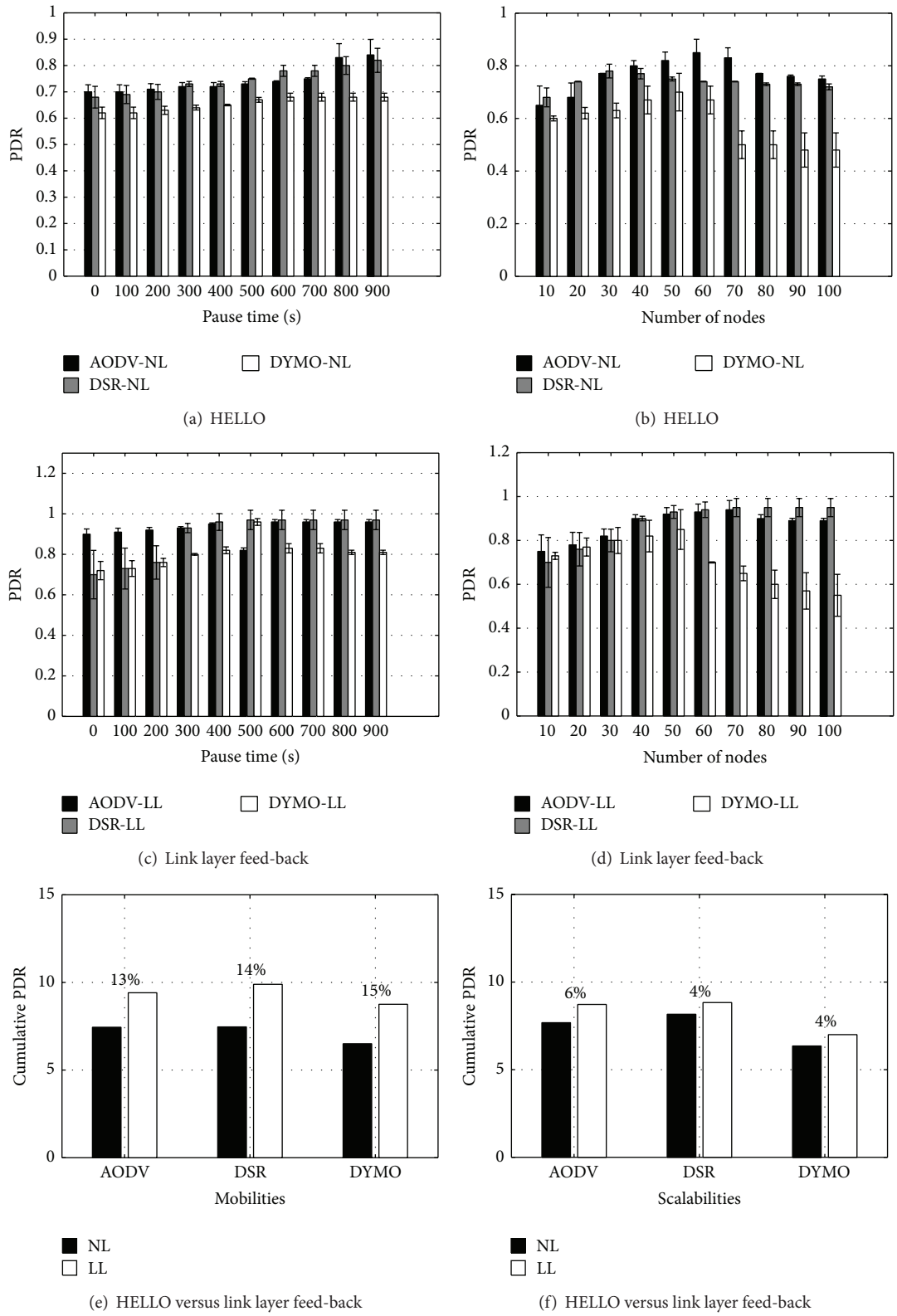
4.1. PDR. AODV attains highest PDR among all the selected protocols because during RD, timely-based route checking in the routing table provides correct route information and grat. RREPs are generated to reduce routing delay by shortening the routes, as depicted in Figures 3(a), 3(c), 3(b), and 3(d). Moreover, LLR mechanism helps to deliver more data packets in high node densities. AODV-LL achieves 5.2% and 6% more cumulative PDR as compared to AODV-NL (refer to Figures 3(e) and 3(f)). This is due to quick notification through LL feed-back mechanism which results in instant repairing, whereas, in DSR, RC contains stale routes due to very high *TAP_CACHE_SIZE* of 1024 bytes and high period for storage of routes in the RC, RouteCacheTimeout of 300 s. As there is no explicit mechanism to delete stale routes except RERR messages, so incorrect RCing in high nodes' mobilities generates faulty information and thus causes packets to be dropped. DSR-NL drops more data packets as compared to DSR-LL, as shown in Figures 3(e) and 3(f), because quick link sensing on LL provides more convergence as compared to NL feedback. In moderate and no-mobilities, DSR's throughput is the highest as compared to AODV and DYMO because RCing during RD phase and PSing for RM phase makes end-to-end path calculation quick, as portrayed in Figure 3.

DYMO does not implement any ancillary mechanism as grat. RREPs in AODV and DSR: LLR of AODV and RCing as well as PSing of DSR (refer Figure 3). Same as that of AODV-LL and DSR-LL, LL mechanism in DYMO-LL achieves more PDR as compared to DYMO-NL. DYMO-LL achieve 5% and 4% more cumulative PDR in different nodes' mobilities and densities, respectively, which can be seen in Figures 3(e) and 3(f). AODV outperforms all the selected protocols when network is denser. LLR, the distinguished feature of AODV, makes this protocol more suitable for high node densities due to reduction of routing latency.

4.2. E2ED. In all the selected nodes' mobilities as well as densities, DYMO attains lowest routing latency due to simple ERS mechanism and lack of checking routes in RC or in RT as depicted in Figure 4. In high mobilities, quick repair is needed after detecting link breakage for maintaining broken link(s) (path(s) reestablishment). To avoid these route-rediscoveries, AODV starts LLR process to quickly upkeep the broken link for achieving low routing latency. AODV, among the reactive protocols, attains highest delay as shown in Figures 4(a), 4(c), and 4(e). Because LLR for link breakages in routes sometimes results in increased path lengths, DYMO produces the lowest E2ED among the reactive protocols because it only uses ERS for route finding that causes less delay due to absence of PSing, RCing, and grat. RREPs. At higher mobilities, DSR-NL suffers the most, that is, highest E2ED (refer Figure 4(a)). The reasons include RCing and PSing failure in high dynamicity which introduce routing latencies. As DSR does not implement LLR, so its E2ED is less than AODV; however, during moderate and high nodes' mobilities, RC search frequently fails and results in increased delay.

Absence of grat. RREPs and any supplementary mechanism keeps the lowest E2ED of DYMO in all the node densities, as depicted in Figure 4(b). PS and grat. RREPs keep the delay low in medium and high traffic scenarios for DSR (while first checking the RC instead of simple ERS based RD process), augments the delay when population increases. Thus more delay of DSR is shown in Figures 4(b), 4(d), and 4(f), as compared to DYMO. AODV experiences the highest E2ED in all the node densities due to LLR process.

AODV-NL possesses less delay as compared to AODV-LL and same as that of DSR. In AODV and DSR, auxiliary



(a) HELLO (b) HELLO

(c) Link layer feed-back (d) Link layer feed-back

(e) HELLO versus link layer feed-back (f) HELLO versus link layer feed-back

Figure 3: PDR achieved by the selected routing protocols.

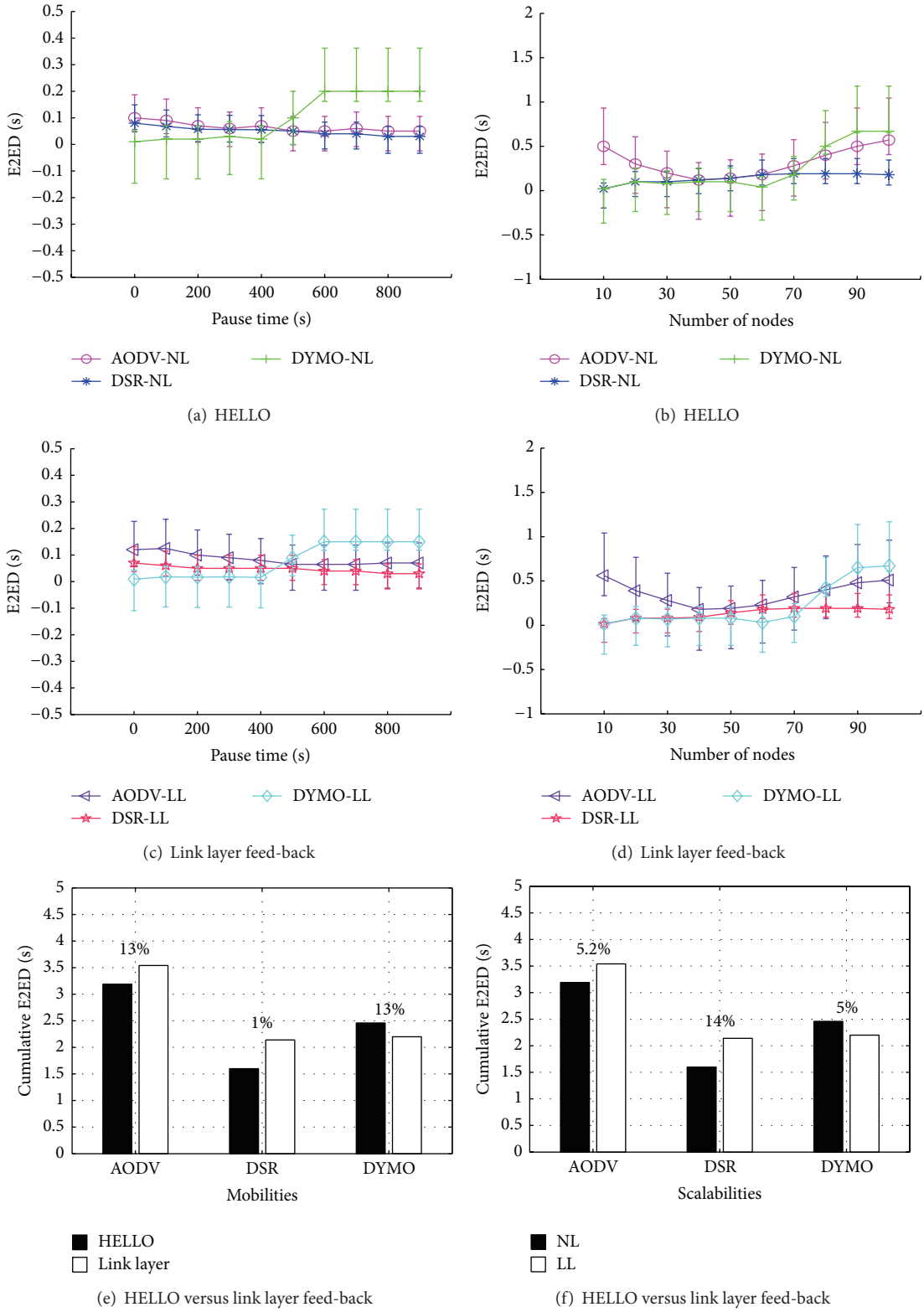


Figure 4: E2ED produced by the selected routing protocols.

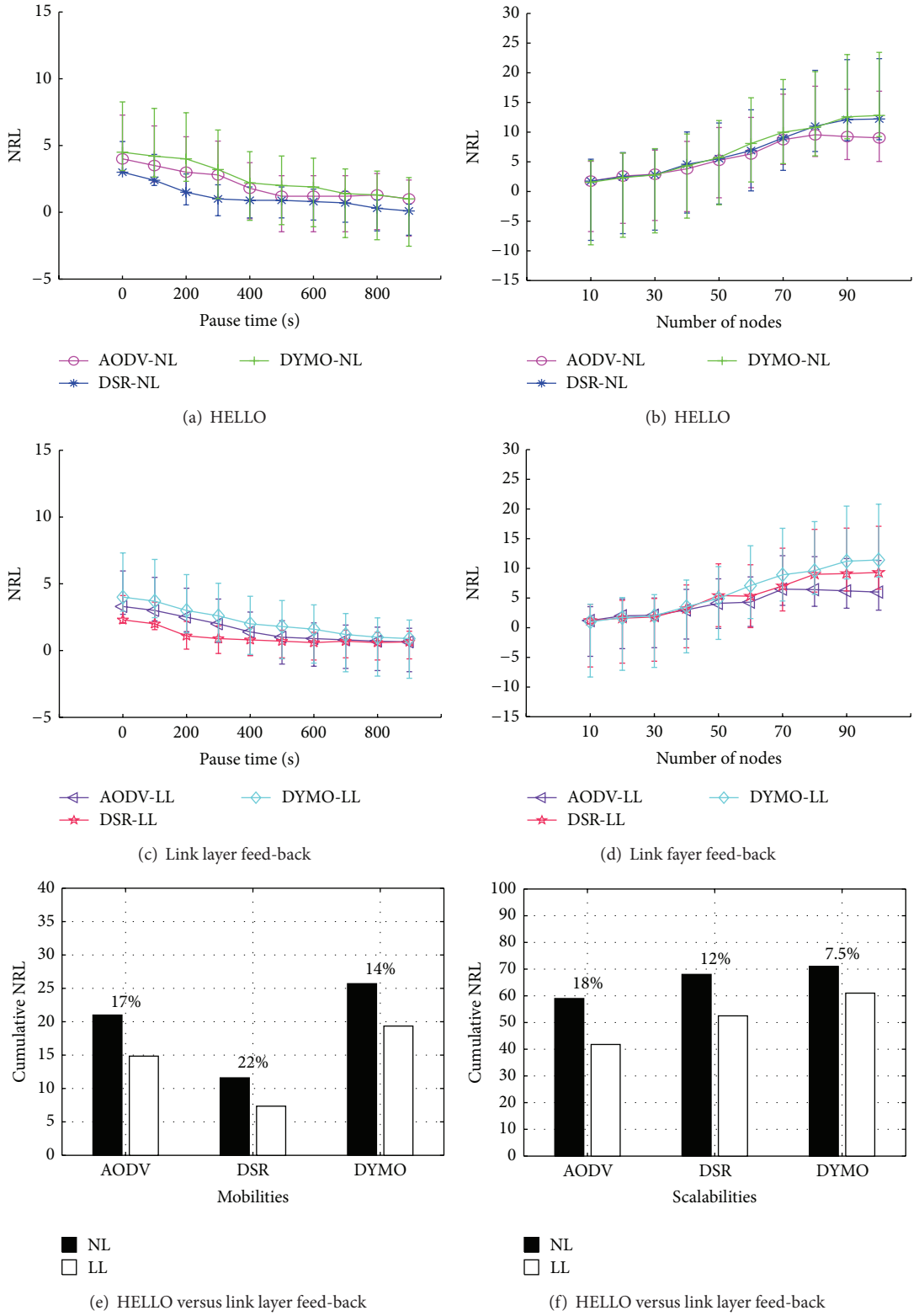


Figure 5: NRL produced by AODV, DSR, and DYMO protocols.

mechanisms during RM produce more routing latencies, because quick detection LB via LL feed-back mechanisms initiate these mechanisms (refer Figures 4(e) and 4(f)). As these mechanisms are absent in DYMO, thus, quick link failure detection in DYMO-LL results in less routing load as compared to DYMO-NL.

4.3. NRL. Due to absence of grat. RREPs, DYMO produces higher routing overhead among the selected reactive protocols, whereas in DSR route RCing and PSing due to promiscuous listening mode produce the lowest routing load in all the selected nodes' mobilities (refer Figure 5). Moreover, high rate of link breakages causes more route rediscoveries for path repairing. In DYMO, routing load increases with high nodes' mobilities because RERR messages are instantly broadcasted after detecting link breakage. Although, AODV uses grat. RREPs however due to the use of *HELLO* messages (to check the connectivity of the ARs in DYMO) and LLR, the DYMO causes more routing load than DSR, as shown in Figure 5(a).

One common noticeable behavior of the selected protocols is that, in high nodes' mobilities, routing overhead is higher as compared to moderate and low nodes' mobilities as shown in Figures 5(a), 5(c), and 5(e). Because, in response to link breakage, all the on-demand protocols disseminate RERR message to inform the route request generator about faulty links and thus prevent the use of invalid routes. As link breakage, in highly dynamic situations, is frequent, more RERR messages are generated resulting in high NRL.

In medium and high populations, routing load of DYMO is less than that of DSR and AODV (refer to Figures 5(b), 5(d), and 5(f)), whereas, in medium and high densities, AODV-NL attains the highest routing load. The *HELLO* messages (to check the connectivity of active routes), LLR and grat. RREPs increase the generation of routing packets. Each node that participates during RD (including intermediate nodes) of DSR, learns the routes to other nodes on the route due to source routing information in their RCs. During RD and RM phases, RCing and PSing processes are, respectively, used to get routes from RC of intermediate nodes. This approach is used to quickly access and to solve broken link issues by providing an alternative route. Thus, PSing and RCing mutually reduce the routing overhead in low node densities of 10, 20, and 30 nodes (refer to Figure 5(b)). However, in high densities, intermediate nodes, generating more grat. RREPs, augment routing overhead as shown in Figure 5(b). NL mechanism in all the three protocols augments routing packet cost as shown in Figure 5.

5. Conclusion

In wireless networks, routing protocols are responsible for efficient routing. In this paper, we select three reactive routing protocols: AODV, DSR, and DYMO. These protocols perform two phases for routing: RD and RM. During RM phase, LSM is more important to repair broken routes. We study two mechanisms of LSM: LL and NL feed-backs. A framework is also modelled for energy as well as time costs during RD and RM with LSM mechanism. For analytical comparison of these LSM mechanisms in the selected protocols, three performance metrics are chosen: PDR, E2ED, and NRL using NS-2. From analytical comparison, we deduce that LL mechanism is more suitable for LSM in reactive protocols as compared to NL feed-back mechanism.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] C. Perkins, E. Belding-Royer, and S. Das, *IETF RFC3561, "Ad hoc On-Demand Distance Vector (AODV) Routing"*, 2003, http://www.ietf.org/rfc/rfc3561.txt.

[2] S. Das, E. Belding-Royer, and C. Perkins, "Ad hoc on demand distance vector (AODV) routing," Mobile Ad-hoc Network (MANET) Working Group, IETF, 2003.

[3] D. Johnson, Y. Hu, and D. Maltz, "The dynamic source routing protocol (DSR) for mobile ad hoc networks for IPv4," IETF RFC, 2007.

[4] D. Johnson, D. Maltz, and J. Broch, "DSR: the dynamic source routing protocol for multi-hop wireless Ad Hoc networks," in *Ad Hoc Networking*, pp. 139–172, 2001.

[5] I. Chakeres and C. Perkins, "Dynamic MANET On-Demand (DYMO) Routing," IETF Draft Dymo-05, 2006.

[6] R. E. Thorup, *Implementing and evaluating the DYMO routing protocol [Ph.D. dissertation]*, Aarhus Universitet, Datalogisk Institute, 2007.

[7] C. Perkins and P. Bhagwat, "Highly dynamic Destination-Sequenced Distance-Vector routing (DSDV) for mobile computers," in *Proceedings of the Conference on Communications Architectures, Protocols and Applications (SIGCOMM '94)*, vol. 24, pp. 234–244, ACM SIGCOMM Computer Communication Review, 1994.

[8] M. Gerla, X. Hong, and G. Pei, "Fisheye State Routing Protocol (FSR) for Ad Hoc Networks," IETF Draft-01, 2000.

[9] G. Pei, M. Gerla, and T. W. Chen, "Fisheye state routing in mobile Ad Hoc networks," in *Proceedings of the IEEE International Conference on Communications (ICC '00)*, vol. 1, pp. 70–74, June 2000.

[10] T. Clausen and P. Jacquet, "Optimized link state routing protocol OLSR," Tech. Rep. IETF RFC-3626, 2003, http://www.ietf.org/rfc/rfc3626.txt.

[11] T. Clausen, P. Jacquet, C. Adjih et al., "Optimized link state routing protocol (OLSR)," 2003.

[12] S. Bitam and A. Mellouk, "Bee life-based multi constraints multicast routing optimization for vehicular ad hoc networks," *Journal of Network and Computer Applications*, vol. 36, no. 3, pp. 981–991, 2013.

[13] S. Bitam, A. Mellouk, and S. Zeadally, "HyBR: a hybrid bio-inspired bee swarm routing protocol for safety applications in Vehicular Ad hoc NETworks (VANETs)," *Journal of Systems Architecture*, vol. 59, no. 10, pp. 953–957, 2013.

[14] A. L. S. Orozco, J. G. Matesanz, L. J. G. Villalba, J. D. M. Diaz, and T.-H. Kim, "Security issues in mobile Ad Hoc networks," *International Journal of Distributed Sensor Networks*, vol. 2012, Article ID 818054, 6 pages, 2012.

[15] J. Ben Othman and L. Mokdad, "Enhancing data security in ad hoc networks based on multipath routing," *Journal of Parallel and Distributed Computing*, vol. 70, no. 3, pp. 309–316, 2010.

[16] V. Nazari Talooki, J. Rodriguez, and H. Marques, "Energy efficient and load balanced routing for wireless multihop network applications," *International Journal of Distributed Sensor Networks*, vol. 2014, Article ID 927659, 13 pages, 2014.

[17] A. Loutfi, M. Elkoutbi, J. Benothman, and A. Kobbane, "An energy aware algorithm for OLSR clustering," *Annales des Telecommunications/Annals of Telecommunications*, vol. 69, no. 3-4, pp. 201–207, 2014.

[18] N. Javaid, A. Bibi, A. Javaid, Z. A. Khan, K. Latif, and M. Ishfaq, "Investigating quality routing link metrics in wireless multi-hop networks," *Annals of Telecommunications*, vol. 69, no. 3-4, pp. 209–217, 2014.

[19] P. Jacquet and L. Viennot, "Overhead in mobile ad-hoc network protocols," Rapport de Recherche-Institu National de Recherche en Iformatique et en Automatique, 2000.

[20] I. Park and I. Pu, "Energy efficient expanding ring search," in *Proceedings of the Asia International Conference on Modelling and Simulation*, pp. 198–199, 2007.

[21] T. Lin, S. F. Midkiff, and J. S. Park, "A framework for wireless ad hoc routing protocols," in *Proceedings of the Wireless Communications and Networking Conference (WCNC '03)*, vol. 2, pp. 1162–1167, 2003.

[22] A. M. Abbas and B. N. Jain, "Path diminution in node-disjoint multipath routing for mobile ad hoc networks is unavoidable with single route discovery," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 5, no. 1, pp. 7–21, 2010.

[23] M. Saleem, S. A. Khayam, and M. Farooq, "On performance modeling of ad hoc routing protocols," *Eurasip Journal on Wireless Communications and Networking*, vol. 2010, Article ID 373759, 13 pages, 2010.

[24] M. Islam, A. Razzaque, M. R. Bosunia, A. Alamri, and M. M. Hassan, "Link failure and congestion-aware reliable data delivery mechanism for mobile ad hoc networks," *Annals of Telecommunications*, vol. 68, no. 9-10, pp. 539–551, 2013.

[25] M. Saleem, S. A. Khayam, and M. Farooq, "A formal performance modeling framework for bio-inspired ad hoc routing protocols," in *Proceedings of the 10th Annual Cnference on Genetic and Evolutionary Computation*, pp. 103–110, ACM, 2008.

[26] R. H. Khokhar, M. A. Ngadi, M. S. Latiff, and M. A. Amin, "Reactive traffic-aware routing strategy for urban vehicular environments," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 10, no. 3, pp. 149–163, 2012.

[27] M. A. Iqbal, F. Wang, X. Xu, S. M. Eljack, and A. H. Mohammad, "Reactive routing evaluation using modified 802.11a with realistic vehicular mobility," *Annales des Telecommunications*, vol. 66, no. 11-12, pp. 643–656, 2011.

[28] N. Javaid, A. Bibi, A. Javaid, and S. A. Malik, "Modeling routing overhead generated by wireless reactive routing protocols," in *Proceedings of the 17th Asia Pacific Conference on Communications (APCC '11)*, pp. 631–636, Sabah, Malaysia, October 2011.

[29] A. Ariza, A. Trivifio, E. Casilari, J.-C. Cano, C. T. Calafate, and P. Manzoni, "Assessing the impact of link layer feedback mechanisms on MANET routing protocols," in *Proceedings of the IEEE Symposium on Computers and Communications (ISCC '09)*, pp. 770–775, July 2009.

[30] J. Broch, D. A. Maltz, D. B. Johnson, Y. C. Hu, and J. Jetcheva, "A performance comparison of multi-hop wireless ad hoc network routing protocols," in *Proceedings of the 4th Annual ACM/IEEE International Conference on Mobile Computing and Networking*, pp. 85–97, ACM, 1998.

[31] R. Castañeda, S. R. Das, and M. K. Marina, "Query localization techniques for on-demand routing protocols in ad hoc networks," *Wireless Networks*, vol. 8, no. 2-3, pp. 137–151, 2002.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 647296, 7 pages
http://dx.doi.org/10.1155/2015/647296

Research Article

An Optimized Prediction Model Based on Feature Probability for Functional Identification of Large-Scale Ubiquitous Data

Gangman Yi

Department of Computer Science & Engineering, Gangneung-Wonju National University, Gangwon-do 220-711, Republic of Korea

Correspondence should be addressed to Gangman Yi; gangman@cs.gwnu.ac.kr

Received 18 August 2014; Accepted 9 September 2014

Academic Editor: Jong-Hyuk Park

Recently, there is a growing interest in the sequence analysis. In particular, the next generation sequencing (NGS) technique fragments the base sequence and analyzes the functions thereof. Its essential role is to arrange pieces of the base sequence together based on sequencing and to define the functions. The organization of unarranged piece of sequence is one of the active research areas; moreover, definition of gene function automatically is a popular research topic. The previous studies about the automatic gene function have mainly utilized the method that automatically defines protein functions by using the similarities of base sequence or the disclosed database and the protein interaction or context free method. This study aims to predict the category of protein whose function was not defined after learning automatically with GO by extracting the characteristics of protein inside the cluster. This study conducts clustering by using the protein interaction that is generated by the similarities of base sequence under the assumption that the proteins inside the cluster have similar function. The proposed method is to show an optimized result in accordance with the option after finding the option value that can give the outperformed prediction of GO, which classifies the functions based on the IPR and keywords inside the same cluster as the unique features.

1. Introduction

There is a growing demand to automatically predict the protein functions since there is a growing interest in the DNA analysis due to the development of technical equipment. Sequencing the sequence, which is composed of pieces, in the order is the basic technology of NGS. However, the studies to seek a method to predict automatically the protein functions or the so-called connected contig are being actively undertaken. The conventional method to predict the protein functions is to make groups as relevent functions mainly based on experiments. However, this method reached the limit in terms of time and efforts due to an increase in the quantity of data. In recent years, there have been a large amount of studies on the method to predict the protein functions in the most effective way by finding the relationship with the classifiers automatically through finding the features that can identify the functions.

The most basic protein classifier method is the method of defining the structure or functions through the protein family DB or finding the homology after determining the similarities between the sequences by using the sequence similarities program such as BLAST. One of the simple methods to define the protein functions automatically [1] is the method of conducting one-to-one mapping on the features and functions of protein. The most prominent example is InterPro2GO [2]. InterPro is protein domain database and GO is controlled vocabulary for gene annotation. GO is broadly subdivided into the three categories since it is the controlled vocabulary to annotate the functions of gene of various organisms and also it consisted of the hierarchical structure. InterPro2GO is the method that consisted of a simple manual mapping between InterPro term and GO term; this mapping table is made by checking the conserved common annotation and finding the specific level of GO term from the relevant family. This method is still possible to obtain InterPro by using InterProScan [3] even when one is not being able to know InterPro term just like a new sequence whose function is not defined. In addition, the study on a simple mapping work as to the GO annotation has been conducted in GOA [4, 5] projects.

Moreover, there have been some studies in which the accuracy as to the prediction was enhanced as compared with the conventional simple mapping method by defining the categories of features and protein functions more specifically with the machine learning method after extracting the DB based on the sequence similarities and the features based on the similarities [6–10]. The method for the protein prediction models using protein-interaction is being also utilized as a protein function classifier method [11, 12] in addition to the studies that defined the features and functional categories through the mathematical models or patterns. Enright et al. suggest the probability for the protein of each interaction using Bayesian [12]; thus, it suggested the probability of having functions unlike those conventional methods that would predict based on whether to "have relevant functions" or "have no relevant functions" [13, 14]. MCL [12] is the method of clustering by giving Option "*I*" in accordance with the mutual correlation between proteins by Markov model using the BLAST outcomes that is mainly used to measure the sequence similarity. Option "*I*" is the value to conduct the inflation and expansion in the Markov model; thus, it is possible to see the changes of elements inside the clustering depending on the value of *I*. Another function prediction method is the context-free method; it utilizes the context of several DBs of Uniprot or PubMed [15].

The research about sequence clustering algorithm is also one of the methods to attempt to group the relevant biological sequences. MCL [12] is a graph based unsupervised clustering by employing the Markov. BLASTClust [16] is an application of clustering the sequence by the single-linkage clustering. CD-HIT [17] is a fast method for clustering protein using the greedy algorithm. If a sequence is sufficiently similar to some clusters, that sequence is assigned to it; otherwise, make a new cluster. UBLAST [18] makes a cluster using USEARCH algorithm which searches high-scoring global alignment. Among these sequence algorithms, MCL provides unsupervised clustering option by the *I* value, which is able to look for the optimized performance by the sequence similarity; thus MCL is used as a clustering method at this suggested model.

The proposed model is to predict the classifier of that function by training the relationship with the features [19, 20] after clustering through protein-interaction. Protein-interaction makes a relational graph by using MCL tool based on the sequence similarities of protein, resulting in making a clustering for the protein. Protein inside the same cluster can be assumed to have similar functions. Yi et al. and Nhat et al. clustered by using the GO of gene and compared the results hereof with the other clustering algorithm [21, 22]. This study shows that there is a constant correlation between GO and clustering. In other words, protein inside the clustering can be explained to have the same GO since it performs a similar function. Based on this idea, that proteins in the same cluster can have similar GO, the suggested method seeks the related features. The features used in the proposed model are InterPro and keyword. They already proved the relationship between InterPro and GO in [23–25]. In conclusion, the purpose of this research is that the investigation of the relationship between selected feature such like IPR, keyword, and GO term within the cluster. The clusters make groups by the sequence similarity.

Table 1: Total data set.

	Protein	IPR	GO	Keyword
Number	5285	4862	4673	461

This study is managed as follows. It described the relationship diagram as to the model and test method that would describe and learn the properties of data used in Section 2 and also mainly stated the detailed description as to the learning model and test method. In Section 3, the most optimized option value of the proposed model is to be presented by comparing and analyzing the result values based on the model explained in Section 2. In Section 4, the conclusion and future direction are to be presented.

2. Method

2.1. Data Set. As for data, *Saccharomyces cerevisiae's* data was utilized at Uniprot. The features to be used are GO (gene ontology) and IPR in DR (database cross-reference) and keyword in the KW (keyword) line from FLAT format. The gene ontology (GO) is a controlled vocabulary of terms to describe protein functions. It consists of the three large categories that include "biological process," "molecular process," and "cellular component" and it is uniquely composed of the hierarchal structure. InterPro terms [26] are defined in the InterPro database, which is a curated protein domain database that acts as a central reference for several protein family and functional domain databases, including Prosite, Prints, Pfam, Prodom, SMART, TIGRFams, and PIR SuperFamily. We have previously shown that InterPro was an important source of features for identifying GO terms for proteins [25, 27]. Keyword performs the important role of reference for sequence as a predefined entry based on the functions, structure, or other categories. Among the extracted *Saccharomyces cerevisiae* data, those proteins not having GO were excluded and IEA (Inferred Electronic Annotation) out of the evidence codes of GO was not utilized. IEA is the property that was automatically extracted from the database, and usually it is not the feature that was revealed experimentally by manual. Thus, it is inappropriate for creating a model based on the calculating technique. To summarize, those proteins having at least one IPR or keyword while having more than one GO among the data of *Saccharomyces cerevisiae* were utilized as the data. As shown in Table 1, a total number of proteins used were 5,285. A total number of GOs owned by these proteins were 4,673, whereas a total number of IPRs and keywords owned by these proteins were 4862 and 461, respectively.

To verify the validity of the proposed model and find the optimized option by using the above data, the 10-fold cross validation technique was utilized. The 10-fold cross validation utilizes a certain part of data as the learning model and another part as the test. In other words, it is mainly utilized when testing the validity whether or not, the proposed method is correctly performed, or finding

the optimal parameter values of the proposed model. On average, one-fold consisted of approximately 520 proteins by dividing the proteins into 10 subsets randomly and only one subset out of those subsets is to be used as the testing data and the remaining 9 folds are to be used as the training data. All the protein data can obtain test results by the data learned by the data that excludes oneself one time through conducting the aforementioned task for each fold.

2.2. Training Method. Firstly, the training model executes BLAST in order to conduct the learning for the folds that consisted of the 9 subsets. BLAST is the program to determine the similarity between sequences; thus, it is the precedent phase process to execute Markov Cluster Algorithm (MCL), which is the program for creating a cluster. MCL algorithm is an unsupervised algorithm that can vary the cluster size promptly and variably based on graph. It is possible to create a matrix that can determine the similarity of sequence for each protein by using BLAST. The data to be expressed in the matrix at this point can become a real number or binary number such as *e*-value. The matrix becomes a node for each protein and can draw a graph by creating a weighted edge by the similarity of connection. Also, it can cluster with a certain threshold value in this graph. In the used training data, *e*-value option value was set at 0.01 when executing BLAST. *E*-value tends to have more similarity when being closer to 0. It represents all the cases that are shown when there is no option as to cut off; thus, the threshold value was set at 0.01 in order to reduce the weight as to those proteins whose similarity was small in terms of graph configuration by setting *e*-value.

Second, I value, which was the option value of MCL, was set at various values. I is the value to execute the inflation and expansion in the Markov model; thus, it is possible to obtain the optimized MCL results with the modified I value. The purpose is to find which parameter represents the optimal result condition by setting I with the four methods of 1.4, 2, 4, and 6. Figure 1 represents the number of proteins owned by each cluster ID when the default value of I is 2.0. As shown in the figure, there is a large quantity of proteins inside one cluster as for those clusters whose cluster ID number is small. However, those IDs with a higher number have a small quantity of proteins inside the cluster. In addition, it was found that it would be changed to a graph of long tail in accordance with I value.

As shown in Figure 1, the number of proteins in each cluster may be 2 or less. In other words, only the internal proteins belong to the corresponding cluster and at least one is to be used as the learning data and another one is to be used as the test data. Thus, those cases in which the number of proteins inside the cluster is less than one are to be excluded. This option is stipulated by Cutoff_2. As a result, 627 clusters are used as data. In addition, it is set in the cases of more than 5 (Cutoff_5) and 10 (Cutoff_10) to confirm the test results.

To express the common IPR, GO and keyword for each feature in each cluster in a formula, it can be derived from

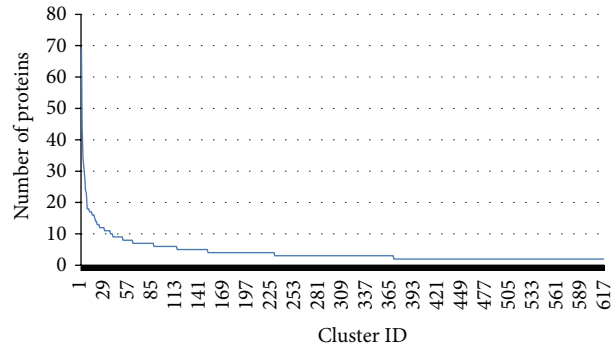


Figure 1: Number of proteins for each cluster at $I = 2.0$ (mcl).

the following formula. P_IPR_l means IPR owned by Protein P. The following can be derived by this formula,

$$\bigcap_{i=1}^{n} P_i\_\mathrm{IPR}_l. \tag{1}$$

Protein P_i contained in the cluster has n units and IPR contained in each protein has the quantity of l. l at this point may vary for each P_i in the cluster. GO and keyword can be expressed as shown in the following formula. m and n mean GO and keyword owned by each protein in the cluster and the quantity may vary depending on Protein P_i.

$$\bigcap_{i=1}^{n} P_i\_\mathrm{GO}_m, \qquad \bigcap_{i=1}^{n} P_i\_\mathrm{KW}_n. \tag{2}$$

The formula derived by the above two formulas is as follows. This formula stands for the common IPR, GO and keyword in each cluster, which is used as a training information.

$$C_{j=1}^{k}\left(\bigcap_{i=1}^{n} C_{jP_i\_\mathrm{IPR}_l}, \bigcap_{i=1}^{n} C_{jP_i\_\mathrm{GO}_m}, \bigcap_{i=1}^{n} C_{jP_i\_\mathrm{KW}_n}\right). \tag{3}$$

$C_{j=1}^{k}$ means a total number of Cluster ID and there is a total of k cluster IDs and the proteins in each cluster ID are to be expressed by C_{jP_i}. The common IPR, GO and keyword owned by the proteins of each cluster can be expressed by $C_{jP_i\_\mathrm{IPR}_l}$, $C_{jP_i\_\mathrm{GO}_m}$ and $C_{jP_i\_\mathrm{KW}_n}$, respectively.

In short, the flow chart of the procedure as to the testing is as shown in (a) in Figure 2. It is required to first execute BLAST using the learning data and set the *E*-value cutoff value at 0.01 at this point. It is then required to obtain the result composed of the 4 clustering results for each training data through the 4 Option I by using the BLAST results. At this point, it is required to examine whether the common IPR, GO, and keyword owned by all the proteins inside each cluster is as many as the number of clusters. The protein in which the common IPR and keyword exist while there is at least more than one common GO in each cluster is defined as the learning data. When the number of proteins inside the cluster is 2, 5, or 10, each of these cases is defined as Cutoff_2, Cutoff_5, and Cutoff_10.

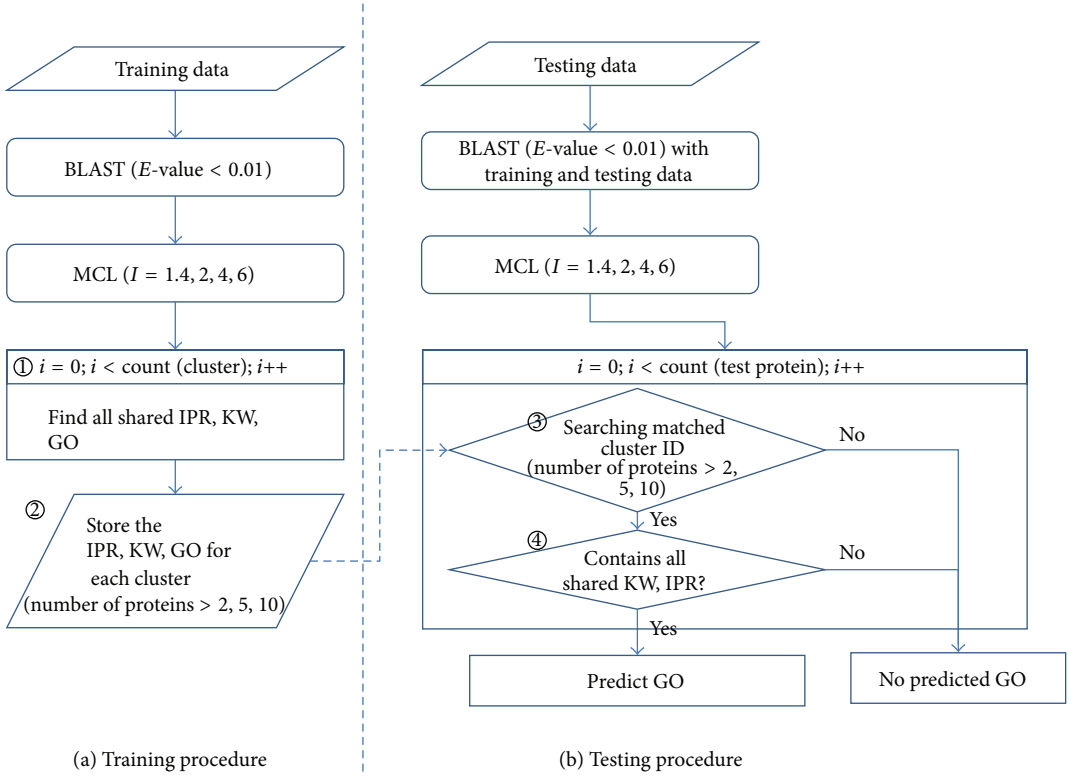


Figure 2: Flowchart for training and testing procedure.

Table 2: Percentage of number of proteins for each Option I at training result.

	$10 \leqq x$	$9 \leqq x < 10$	$8 \leqq x < 9$	$7 \leqq x < 8$	$6 \leqq x < 7$	$5 \leqq x < 6$	$4 \leqq x < 5$	$3 \leqq x < 4$	$2 \leqq x < 3$
I1.4	0.67	0.22	0.18	0.67	1.27	2.45	4.91	15.87	63.15
I2.0	0.47	0.25	0.29	0.90	1.35	2.58	5.36	18.36	69.43
I4.0	0.35	0.20	0.37	0.86	1.12	3.01	5.56	20.86	76.67
I6.0	0.27	0.16	0.49	0.76	1.06	2.99	6.03	21.49	77.75

As for Figure 3, the corresponding cluster has the three proteins (MCFS1_YEAST, MCFS2_YEAST, and YM60_YEAST) when I is 1.4 and the cluster ID is 292 and it has GO:00051792 jointly, which is expressed in blue. IPR and keyword also have "IPR000952," "IPR000073," "IPR012020," "hydrolase," "reference proteome," and "complete proteome" in common, which are represented in green and yellow, respectively. As for cluster ID 292, the number of proteins inside the cluster is 3; thus, it becomes the trained result that belongs to Cutoff_2. Nonetheless, it would not belong to Cutoff_5 and Cutoff_10, since the number of proteins inside the cluster is less than 5 or 10. Table 2 represents the percentage as to the mean number of proteins of 10-fold in accordance with each Option I when it is Cutoff_2 in the training result. For instance, the first-fold among the 10 folds when I is 1.4 has a total of 444 clusters and there are only 2 whose number of proteins inside the cluster are more than 10 while having at least more than one GO inside the cluster in common and having IPR and keyword as well. When converting it into percentage, it becomes $(2/444) * 100 = 0.4$ and 0.67 is its mean value as shown in Table 2 when obtaining the mean value of other 2~9-fold. In other words, it can be regarded as the mean probability distribution as to the number of proteins in accordance with Option I as the learning data of 10-fold. When I is 1.4, 2, 4, and 6, I has 63.15, 69.43, 76.67, and 77.75, respectively, when the number of proteins inside the cluster is 2. Mostly, the number of proteins inside the cluster tends to be small.

2.3. Testing Method. The flow chart of the procedure as to the testing is as shown in (b) in Figure 2. As a result of training model, each cluster ID has the common IPR, GO, and keyword (Figure 2-②). The test data first executes BLAST to test the sequence similarity just like the training method as shown in shown in Figure 2-(b). At this point, DB and query that executes BLAST utilize all the proteins used in the training as well as the test proteins. MCL conducts clustering based on the BLAST result; thus, the cluster ID in which the test data belongs can be obtained only when including the training data containing the test. It is required to find the cluster ID obtained from MCL for each tested protein

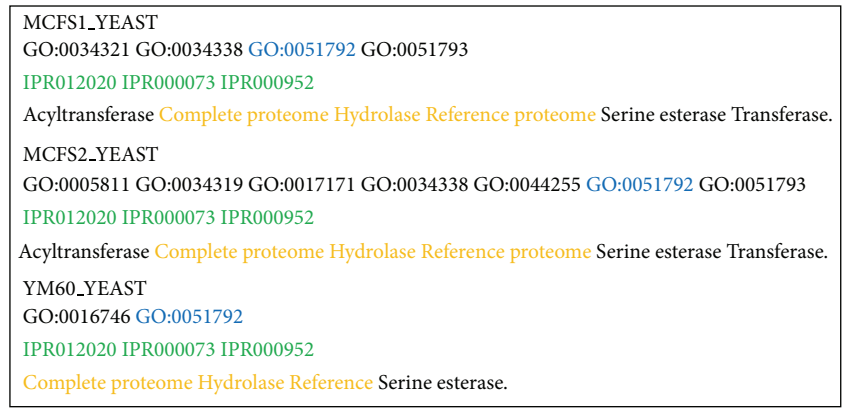


Figure 3: Example of cluster ID 292 at $I = 1.4$.

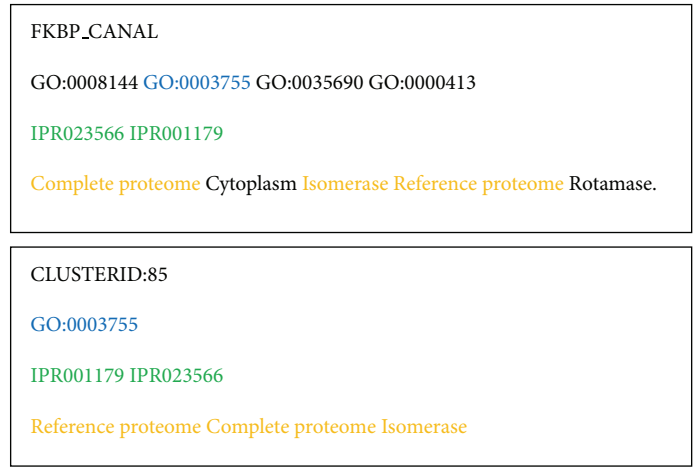


Figure 4: Example of tested protein.

(Figure 2-③) and confirm whether the corresponding cluster ID has the common IPR, GO, and keyword in the results obtained by the training model (Figure 2-④). At this point, the matching probability of IPR and keyword was set at more than 0.5. For instance, the test protein FKBP_CANAL has the 4 GOs, 2 IPRs, and 4 keywords (Figure 4). This protein belongs to cluster ID 85 and the proteins belonging to cluster ID 85 are shown to have the 5 common features such as "IPR001179," "IPR023566," "reference proteome," "complete proteome," and "isomerase" by referring to the data generated by the learning (Figure 2-②). The test protein FKBP_CANAL shows that the 5 features out of the 7 features of IPR and keyword are matched. The matching probability is more than 0.5; thus, GO:0003755 is to be set as the prediction GO of FKBP_CANAL. As for this protein, the accuracy is 0.25 since only one GO (GO:0003755) is predicted out of a total of 4 GOs (GO:0008144, GO:0003755, GO:0035690, and GO:0000413).

The formula thereof is as follows. If the protein to be test is Test_p, it will be possible to obtain a cluster ID to be obtained by the result of BLAST and MCL (j). If the ratio of number of common IPR ($C_{jP_i\_\text{IPR}_l} \cap \text{Test}_{p_\text{IPR}}$), KW ($C_{jP_{i\text{KW}_n}} \cap \text{Test}_{p\_\text{KW}}$) that is owned by the corresponding j cluster among IPR and KW owned by the proteins to be tested as compared with the number of lists of IPR ($\text{Test}_{p\_\text{IPR}}$) and KW ($\text{Test}_{p\_\text{KW}}$) of the proteins to be tested is more than 0.5, then the prediction will be conducted by GO ($C_{jP_{i\text{GO}_m}}$) of the proteins to test the common GO owned inside the cluster

$$\text{Predict}\left(j, \text{Test}_p\right) = C_{jP_{i\text{GO}_m}},$$
$$\text{if } \frac{\left(\left(C_{jP_{i\text{IPR}_l}} \cap \text{Test}_{p\_\text{IPR}}\right) \cup \left(C_{jP_{i\text{KW}_n}} \cap \textit{Test}_{p\_\text{KW}}\right)\right)}{\left(\text{Test}_{p\_\text{IPR}} \cup \text{Test}_{p\_\text{KW}}\right)} > 0.5. \tag{4}$$

3. Performance Evaluations

Figure 5 represents the number of proteins that had at least more than one GO by the proposed method in accordance with the Option I and cutoff value of MCL. At this point, the three cutoff options mean that the minimum numbers of proteins inside each cluster from the data generated by the training model were more than n. For example, "Cutoff_10" represents that the numbers of proteins in each cluster are more than 10. Absolutely, cluster should have more protein if the cutoff option is small. The more data can be utilized

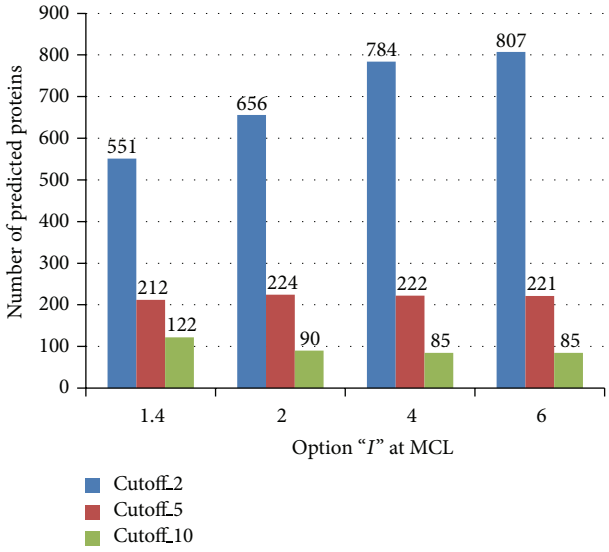


Figure 5: Number of proteins which has at least one GO by the test model.

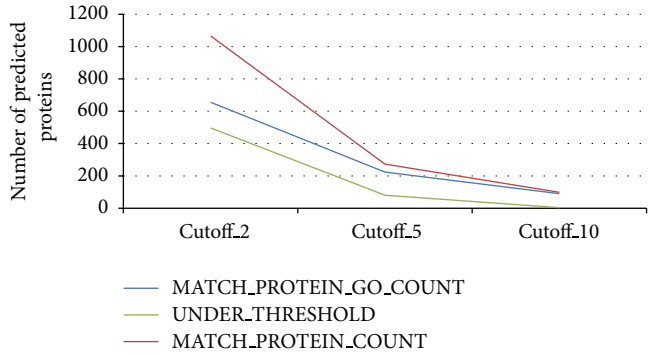


Figure 6: Tested number of proteins at $I = 2.0$.

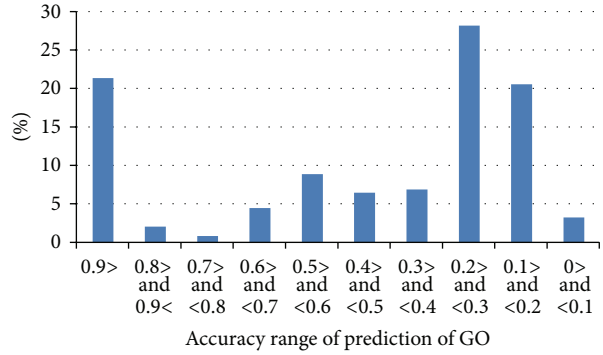


Figure 7: Accuracy of prediction of GO at $I = 2$ and Cutoff_5.

as a learning data, if the cutoff option is larger; thus, the number of predicted proteins would also be larger. However, the accuracy thereof would be lower with a smaller number of proteins learned inside the cluster. In the case of Cutoff_5, in other words, the date learned by at least more than 5 proteins, it will represent the largest number of predictions when I is 2. In addition, it will be shown that the number of predictions will slow down gradually from $I = 2$ option in the case of Cutoff_10.

MATCH_PROTEIN_COUNT, MATCH_PROTEIN_GO_COUNT, and UNDER_THRESHOLD when I was 2.0 based on the diagram of Figure 5 were investigated (Figure 6). MATCH_PROTEIN_COUNT is the protein having more than the predicted GOs as to the case where I is 2.0 by the entire learning data. Since all proteins are not predicted correctly, the numbers of proteins as to the case in which the accurately predicted GO were matched by comparing with the actual test data were defined as MATCH_PROTEIN_GO_COUNT. The gap of MATCH_PROTEIN_COUNT and MATCH_PROTEIN_GO_COUNT is incorrectly predicted GO. On the other hand, those proteins that could not have GO since IPR or keyword were matched with less than a certain threshold, which is defined as 0.5 in this experiment, were defined as UNDER_THRESHOLD. As for Cutoff_2, a large quantity of proteins had GO as expected when the number of proteins inside the learned cluster were at least more than 2. However, IPR and keyword were not matched since a majority was below the standard level; thus, there were many proteins that could not be predicted. As a result of Figures 5 and 6, the accuracy was enhanced with a higher cutoff value; however, the optimal condition was the case in which I was 2 and the cutoff was 5 since the number of predicted proteins was reduced. In other words, it was confirmed that they would be meaningful as the learning data only when the number of proteins used for learning inside the cluster was at least more than 5.

Figure 7 represents the prediction accuracy by diagram based on each protein when testing at each fold as to the case when the selected I is 2 and the cutoff is 5 as the optimal option. The number of GOs of the FKBP_CANA defined in Uniprot in the example of Figure 4 are GO:0008144, GO:0003755, GO:0035690, and GO:0000413. Of those, the accuracy was 0.25 by predicting GO:0003755. Using the number of GOs that was originally owned by UniProt and suggested method, it represents the prediction possibility of GO by the 10 fold cross validation and the horizontal axis means the percent within the scope of prediction accuracy. 0.9 represents the ratio between 0.9 and 1 and it is represented as a relatively high percentage since it is approximately 20 percent. A value of higher than 50 percent means the accuracy of 0.3 or higher. It is represented as the one to predict at least one GO out of the 3 GOs that have protein. It shows that the accuracy is high in the case where there is GO predicted through the suggested model.

4. Conclusions

This study conducted clustering under the assumption that the functional classifier inside the cluster had similar functions and utilized the features extracted inside the cluster as the learning data. When finding protein whose function is unknown, the model that predicts GO (or the controlled vocabulary) was defined through the learning and learned data documents of those proteins whose function was already

defined. This is the existing functional prediction, which is the method to harmonize appropriately those frequently used methods such as sequence similarity, protein-interaction, and context-free; thus, it could increase the prediction probability of GO.

However, in the case of used MCL cluster, the proteins in the cluster are often small in number with two or three higher serial number. When the number of proteins inside the cluster is small, they might not be able to perform the role of a cluster that would bind similar functions of protein since they are divided into pieces. The objective is to create a model to increase the GO predictability by using the features other than IPR and keywords as increasing the number of proteins inside the cluster by applying the other cluster methods in addition to MCL.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2063006).

References

[1] E. Elsayed, K. Eldahshan, and S. Tawfeek, "Automatic evaluation technique for certain types of open questions in semantic learning systems," *Human-Centric Computing and Information Sciences*, vol. 3, article 19, 2013.

[2] E. Camon, M. Magrane, D. Barrell et al., "The gene ontology annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro," *Genome Research*, vol. 13, no. 4, pp. 662–672, 2003.

[3] P. Jones, D. Binns, H.-Y. Chang et al., "InterProScan 5: genome-scale protein function classification," *Bioinformatics*, vol. 30, no. 9, pp. 1236–1240, 2014.

[4] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler, "The GOA database in 2009—an integrated Gene Ontology Annotation resource," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D396–D403, 2009.

[5] E. Camon, M. Magrane, D. Barrell et al., "The Gene Ontology Annotation (GOA) database: sharing knowledge in uniprot with gene oncology," *Nucleic Acids Research*, vol. 32, pp. D262–D266, 2004.

[6] J. Jung, G. Yi, S. A. Sukno, and M. R. Thon, "PoGO: prediction of gene ontology terms for fungal proteins," *BMC Bioinformatics*, vol. 11, article 215, 2010.

[7] S. Khan, G. Situ, K. Decker, and C. J. Schmidt, "GoFigure: automated gene ontology annotation," *Bioinformatics*, vol. 19, no. 18, pp. 2484–2485, 2003.

[8] D. M. A. Martin, M. Berriman, and G. J. Barton, "GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes," *BMC Bioinformatics*, vol. 5, article 178, 2004.

[9] A. Vinayagam, C. del Val, F. Schubert et al., "GOPET: a tool for automated predictions of gene ontology terms," *BMC Bioinformatics*, vol. 7, article 161, 2006.

[10] A. Vinayagam, R. König, J. Moormann et al., "Applying Support Vector Machines for gene ontology based gene function prediction," *BMC Bioinformatics*, vol. 5, article 116, 2004.

[11] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein-protein interaction data," *Journal of Computational Biology*, vol. 10, no. 6, pp. 947–960, 2003.

[12] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Research*, vol. 30, no. 7, pp. 1575–1584, 2002.

[13] K. Salim, B. Hafida, and R. S. Ahmed, "Probabilistic models for local patterns analysis," *The Journal of Information Processing Systems*, vol. 10, no. 1, pp. 145–161, 2014.

[14] K. J. Nishanth and V. Ravi, "A computational intelligence based online data imputation method: an application for banking," *Journal of Information Processing Systems*, vol. 9, no. 4, pp. 633–650, 2013.

[15] N. Daraselia, A. Yuryev, S. Egorov, I. Mazo, and I. Ispolatov, "Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks," *BMC Bioinformatics*, vol. 8, article 243, 2007.

[16] ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html.

[17] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[18] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 2010.

[19] A. P. James, B. Mathews, S. Sugathan, and D. K. Raveendran, "Discriminative histogram taxonomy features for snake species identification," *Human-Centric Computing and Information Sciences*, vol. 4, article 3, 2014.

[20] R. Pan, G. Xu, and P. Dolog, "Improving recommendations by the clustering of tag neighbours," *Journal of Convergence*, vol. 3, no. 1, pp. 13–20, 2012.

[21] G. Yi, S.-H. Sze, and M. R. Thon, "Identifying clusters of functionally related genes in genomes," *Bioinformatics*, vol. 23, no. 9, pp. 1053–1060, 2007.

[22] V. V. M. Nhat and N. H. Quoc, "A model of adaptive grouping scheduling in obs core nodes," *Jorunal of Convergence*, vol. 5, no. 1, pp. 9–13, 2014.

[23] S. Letovsky and S. Kasif, "Predicting protein function from protein/protein interaction data: a probabilistic approach," *Bioinformatics*, vol. 19, no. 1, pp. i197–i204, 2003.

[24] M. Deng, Z. Tu, F. Sun, and T. Chen, "Mapping gene ontology to proteins based on protein-protein interaction data," *Bioinformatics*, vol. 20, no. 6, pp. 895–902, 2004.

[25] J. Jung and M. R. Thon, "Automatic annotation of protein functional class from sparse and imbalanced data sets," in *Proceedings of the 1st International Workshop on Data Mining and Bioinformatics (VDMB '06)*, pp. 65–77, September 2006.

[26] N. Mulder and R. Apweiler, "InterPro and InterProScan: tools for protein sequence classification and comparison," *Methods in Molecular Biology*, vol. 396, pp. 59–70, 2007.

[27] J. Jung, *Automatic assignment of protein function with supervised classifiers [Doctoral thesis]*, 2008.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 345480, 9 pages
http://dx.doi.org/10.1155/2015/345480

Research Article

A Location-Based Business Information Recommendation Algorithm

Shudong Liu[1,2] and Xiangwu Meng[1,2]

[1] *Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2] *School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China*

Correspondence should be addressed to Shudong Liu; bupt.mymeng@gmail.com

Received 5 August 2014; Accepted 1 December 2014

Academic Editor: Jong-Hyuk Park

Recently, many researches on information (e.g., POI, ADs) recommendation based on location have been done in both research and industry. In this paper, we firstly construct a region-based location graph (RLG), in which region node respectively connects with user node and business information node, and then we propose a location-based recommendation algorithm based on RLG, which can combine with user short-ranged mobility formed by daily activity and long-distance mobility formed by social network ties and sequentially can recommend local business information and long-distance business information to users. Moreover, it can combine user-based collaborative filtering with item-based collaborative filtering, and it can alleviate cold start problem which traditional recommender systems often suffer from. Empirical studies from large-scale real-world data from Yelp demonstrate that our method outperforms other methods on the aspect of recommendation accuracy.

1. Introduction

Rapid technology development has brought an increasing number of mobile devices equipped with GPS capacities, such as laptops, PDAs, and mobile phones. It makes check-in behavior become a new life style of millions of users who share their locations, tips, and experience about points of interest (POI) with their friends in location-based social networks. Recently, how to provide timely and personalized information and sharing services based on users' location information is gradually attracting a lot of attention both from the industry and from the research community. It also forms a known and independent research area named the location-based services (LBS). In particular, personalized information recommendation is more important since it is beneficial for users to know new POIs or special promotions in marketplace and explore their city and for advertisers to launch advertisements to targeted users.

Recently, many researches on information (e.g., POI and ADs) recommendation based on location have been done in both research and industry [1–3]. Collaborative filtering (CF) is the mainstream of algorithm to solve this task. Both memory-based and model-based collaborative filtering methods have been proposed and investigated to learn users' preferences on the LBS from users' location check-in data [1, 4, 5]. However, previously proposed methods consider all check-ins in a whole and mobile users' basic laws governing human motion and dynamics are usually overlooked as well as rare researches on cold start problem which results from users' rarely rating on items in location-based recommender systems. As shown in [6], humans experience a combination of periodic movement that is geographically limited and seemingly random jumps that are correlated with their social networks. About 50% to 70% of all human movements are short-ranged and periodic both spatially and temporally and are not affected by the social network structure and about 10% to 30% of all human movements are long-distance and random and are usually influenced by social network ties. Hence, location-based recommendation should be sensitive to range of users' movement, and we will show how to

alleviate cold start problem in location-based recommender system using user's basic movement laws in this paper.

Hence in this paper, unlike the previous works, our goal is to provide users with information recommendation within the scope of users' movement in a very sparse rating system. The task is much harder than traditional location-based recommendation or prediction, because it recommends some interesting business information in the scope of user's daily movement. However, this task is more significant since it can provide various personalized favorite local pieces of information combined with long ranged travel information which is close to their friends' home. Thus, if we could divide users' movement region into two parts, the local part and remote part, then we can recommend their favorite business information in each part to them, but most of all, we should determine each user's scope of movement by exploiting their check-in log. In location-based services or social networks, the places users check in every time are often some points of interest or parkland they are visiting and we cannot obtain users' successive location trajectory from their check-in data, and more importantly, this results in a very sparse dataset in location-based recommender system.

Based on these two properties of check-in dataset and studies in [6], we focus on explicitly modeling users' local movements and long-distance travel preferences for recommendation in their check-in dataset. There are two challenges: (1) how to determine users' local movement region and remote movement region and (2) how to find users' favorite business information in each of their movement regions. To address these challenges, we propose a region-based location graph (RLG) and design new algorithms to make accurate top-N recommendation on RLG. The uniqueness of the proposed model is the introduction of users' movement region nodes which could help users find out their local neighbors and remote friends for collaborative information recommendation, including users' local movement region and remote movement region; furthermore, it captures users' local visit interest through user-local movement region connections and captures long-distance travel interest through user-remote movement region connections. As the two users' local movement regions are intersecting, we call the two users local neighbors and as the local movement region of a user and the remote movement region of the other user are intersecting, we call the two users remote friends.

To summarize, our main contributions are as follows.

(1) We construct a region-based location graph (RLG), in which region node connects with user node and business node, respectively.

(2) While the two regions of users' local movement are intersecting, we call them local neighbors and while the local movement region of a user and the remote movement region of the other user are intersecting, we call the two users remote friends. Based on RLG framework, we propose a novel location-based business information recommendation algorithm.

(3) We compare our approach with other methods on a real dataset and show the performance of our approach on alleviating cold start users problem in location-based recommender systems.

2. Related Work

In recent years, the technologies of mobile communication and mobile location have achieved great development, especially the usage of social network sites, which brings a new chance for social application of geospatial information. The willingness of users to share their current locations and experience originally facilitates the creation of location-based recommender systems (LBRS) based on users generated content and makes it receive much attention from the academic community and industry.

Currently, there are two lines of work to solve the task of location-based recommendation [5]. One line of research is conducted based on the GPS trajectory logs [7–11]. The GPS trajectory data usually consist of small number of users but dense records [12, 13]. Many collaborative filtering algorithms have been proposed and deemed location or POIs as item in traditional recommender systems, such as collective matrix factorization [8], tensor factorization [9], memory-based collaborative location model [10], and pattern recognition model [11]. The other line of work focuses on location-based social networks data, which is very sparse and large-scale [1, 4, 14]. Currently, geographical influences, for example, modeling the check-in probability to the distance of the whole check-in history by power-law distribution [1], modeling users' multicenter check-in behaviors via multicenter Gaussians [4], and mining user check-in behaviors [15], have been addressed and fused with traditional CF algorithms.

The crucial point of location-based recommender systems enables users to read or ask for recommendations in the vicinity of a specified location users are visiting or used to visit, so it is important that recommendations in location-based recommender systems must have strong binding to users' movement region. However, there are rarely previous researches on this issue and cold start problem because of the fact that very sparse data in location-based recommender systems is not yet well studied. As a user can only visit a limited number of locations, especially when a user travels to a new city, the user's locations matrix is very sparse, leading to a big challenge to traditional collaborative filtering-based location recommender systems. To this end, Bao et al. [16] proposed a location recommender system, which consists of two main parts: offline modeling and online recommendation. The offline modeling part models each individual's personal preference with a weighted category hierarchy and infers the expertise of each user in a city with respect to different category of locations according to their location histories using an iterative learning model. The online recommendation part selects candidate local experts in a geospatial range that matches the user's preferences using a preference-aware candidate selection algorithm and then infers a score of the candidate locations based on the opinions of the selected local experts. The significance of the recommender systems in location-based services and the promising solution motivate us to investigate further in this paper.

3. Data Model and Problem Definition

In this section, we briefly introduce the related data model and define users' favorite business information finding problem in their different movement regions.

3.1. Data Model. Unlike GPS trajectory data, users' check-in data are not continuous in both special and temporal dimensions in location-based social network. The places users check in are often some points of interest or parkland they are visiting. For example, when a user has dinner in a restaurant, he may share some information about this restaurant and his experience with his friends at a social network or review websites. So such check-in data indicate that the scope of user's daily motion can cover all businesses reviewed by him; in other words, the business information reviewed by a user is limited in the scope of his daily motion.

Suppose that $U = \{u_i \mid i \in [1, N]\}$ is a mobile user set, where N is the total number of mobile users. Each user has some essential attributes, such as gender, age, and occupation, which is denoted by the form of $\{u_{i,x} \mid x \in [1, X]\}$. For a user's reviewed and favorite businesses $I = \{i_j \mid j = 1, 2, \ldots\}$, which can be formed as the following triple ⟨user, business, rating⟩, each business i_j has some basic attributes, including location (e.g., longitude and latitude, denoted by $l_j = (l_{j,x}, l_{j,y})$) and service categories (e.g., restaurant, hotel, bar, and shopping mall, denoted by $s_j = \{s_{j,k} \mid k \in [1, K]\}$, $s_{j,k}$ is 1 or 0).

Our data are in the form of ⟨User, Business, Location⟩ triple which can be modeled by tripartite graph [17] or a tensor [18]. Although both tripartite graph and tensor treat location as a universal dimension shared by all users, as matter of fact, there is one-to-one correspondence between the business and location in users' check-in data and some users could never review lots of businesses which are out of their movement region in their daily lives. As argued in [6] most of the users' motions are composed of short-ranged daily travel between their homes and workplaces which is periodic both spatially and temporally and long-distance travel which is more influenced by social network ties. In a recommender system, the fixed correlation between business and location is typically not significant, while the movement region plays an important role in recommendation generating process, and the correlation between user and his movement regions is more relevant than that between user and location of business reviewed by them.

Therefore, according to the geographical position distribution of all businesses reviewed by each user, we divide a user's movement region into local movement region and remote movement region. Provided that (x_k, y_k) represents the geographical center of a user's movement region, $D_{\max}$ is the farthest distance between the center $o : (l_{x,o}, l_{y,o})$ and the position the user can reach in his daily life. If there exists a number D_{int} ($D_{\max} \le D_{\text{int}} < 0$), the percentage of businesses reviewed by a user in a circle region around the pointer of (x_k, y_k) with D_{int} as its radius can reach a fixed number ρ, and we call this circle region the user's local movement region R_{loc} and the other region his remote movement region R_{rem}. Namely,

$$\rho = \frac{\left|I'\right|}{|I|}, \tag{1}$$

where $I' \subset I, I' = \{i_j \mid d(l_j, o) \le D_{\text{int}},\ j = 1, 2, \ldots\}$, and $d(l_j, o)$ is the Euclidean distance between two points, and according to the conclusion in [6], we set ρ to 0.7. In addition, we call two users as local neighbors if their local movement regions are overlapping, denoted by N_{loc}, whereas we call them remote friends, if the remote movement region of a user and local movement region of the other user are overlapping, denoted by F_{rem}.

3.2. Problem Definition. Intuitively, location-based recommendation is trying to find potential favorite business information within users' entire motion range. With data model defined above, we formally define this problem as follows: there are some basic datasets of all users, including review log set $I = \{i_j \mid j = 1, 2, \ldots\}$, local neighbor set N_{loc} and remote friend set F_{rem}. For a specific query user, one recommendation method should return a ranked list of businesses which the user would like, and what is more, the ranking score in the process should consider both user's different movement region and social relationship.

4. RLG Construction

In this section, we treat the two movement regions of users as new nodes, which enable new linkages between users and the location of their reviewed business and construct a graph and name it region-based location graph (RLG), which contains three types of nodes: user node, movement region node, and business node. In this way, we can transform ⟨user, business, location⟩ into ⟨user, region, business⟩ by formation of users' two movement regions.

RLG is a bipartite graph $G(U, R, I, E, \omega)$, where U denotes the set of all user nodes, R is the set of users' movement region nodes, and I is the set of business nodes. $\omega : E \to R^*$ is a nonnegative weight function for all edges. Figure 1 is a simple example of RLG containing two user nodes, 4 region nodes, and 6 business nodes. It shows that user u_1 interacts with his two movement regions R_{11} and R_{12}; likewise, the user node u_2 interacts with his two movement region nodes R_{21} and R_{22}, Furthermore, region node R_{11} is also linked to business nodes i_1, i_2, and i_4 because user node u_1 reviewed business nodes i_1, i_2, and i_4 in his local movement region node R_{11}, and region node R_{12} is connected with business nodes i_3 because user node u_1 reviewed businesses node i_3 in his remote movement region node R_{12}. Similarly, region node R_{21} is connected with business nodes i_4 and i_5 and region node R_{22} is connected with business node i_6.

In RLG, each user node connects with two movement regions, and the two movement regions only connect with some business nodes which were reviewed by user. If two users coreviewed a business, then their two movement regions would be overlapping; for example, in Figure 1, user

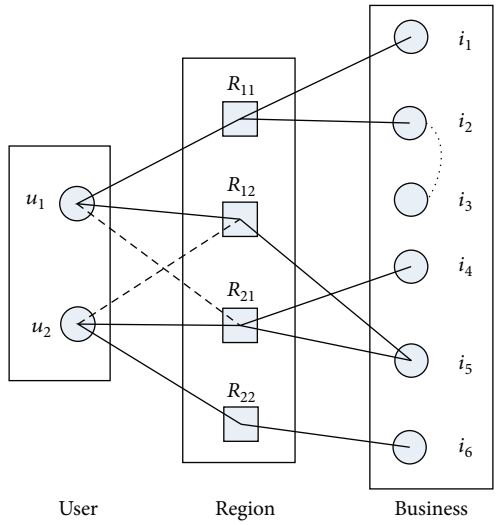


Figure 1: A simple example of RLG.

u_1 and user u_2 coreviewed, respectively, business i_4 in their local movement regions. This means that user u_2 is the local neighbor of user u_1 and the two regions are overlapping; that is to say, user u_1 could reach u_2's local movement region and user u_2 could reach u_1's local movement region, and they both would like some businesses in the region. Thus, based on the above empirical observations, we connect user node u_1 with region node R_{21} and connect user node u_2 with region node R_{11} in RLG (dotted lines), and if we start working from a user node (u_1), passing through a region node (R_{11}), we will find out his local neighbor (u_2) and we can reach an unknown business (i_5) in his local region; namely, $u_1 \rightarrow R_{11} \rightarrow u_2 \rightarrow R_{21} \rightarrow i_5$. In the same way, we can obtain another path $u_2 \rightarrow R_{21} \rightarrow u_1 \rightarrow R_{11} \rightarrow i_1(i_2)$ which connects user node u_2 with business node i_1 or i_2. Hence, in this way, we can help user search for favorite business information in his movement region from local neighbors or remote friends.

The edge weights of RLG between user node and region node are defined as

$$\omega\left(u_p, R_{q,k}\right) = \begin{cases} \omega_{*1}, & \text{if } p = q,\ k = 1, \\ \omega_{*2}, & \text{if } p = q,\ k = 2, \\ & \omega_{*2} = 1 - \omega_{*1},\ \omega_{*2} < \omega_{*1} \\ \operatorname{sim}\left(u_p, u_q\right)_{R_{p*} \cap R_{qk}}, & \text{if } p \neq q,\ u_p \in NB\left(u_q\right) \\ & \text{or } u_p \in RF\left(u_q\right) \\ 0, & \text{otherwis.} \end{cases} \tag{2}$$

Given an edge $e(u_p, R_{q,k})$, if $p = q$, $k = 1$, its weight will be ω_{*1} and if $p = q$, $k = 2$, its weight will be ω_{*2}, and, $\omega_{*1} > \omega_{*2}$, and since each user's local movement region or remote region is a part of his whole movement region and the probability of the user being active in local movement region is much greater than that in the remote region, so we let $\eta = \omega_{*1}/\omega_{*2}$ as a learning parameter in the following experiment, and in this way, we use different weights to model the influence of local mobility preferences and long-distance mobility preferences. $\operatorname{sim}(u_p, u_q)_{R_{p*} \cap R_{qk}}$ is the similarity between the two users who coreviewed lots of businesses. When one of the two users only reviewed one business, we use Jaccard coefficient (as shown in formula (6)) between the businesses reviewed by them as their similarity $\operatorname{sim}(u_p, u_q)_{R_{p*} \cap R_{qk}}$, while when the two users reviewed more than one business, we calculate $\operatorname{sim}(u_p, u_q)_{R_{p*} \cap R_{qk}}$ by using the Cosine similarity (as shown in formula (7)).

The edge weights of RLG between region node and business node are defined as

$$\omega\left(R_{qk}, i_j\right) = \begin{cases} \dfrac{r\left(u_p, i_j\right) - r_{\min}}{r_{\max} - r_{\min}}, & \text{if } \left(l_{jx}, l_{jy}\right) \in R_{qk},\ k = 1, 2 \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The definition means the higher the reviewing score is, the more the user likes the business in a movement region. We denote by $\Pr(G)$ the transition probability matrix of RLG:

$$\Pr(G) = \begin{pmatrix} \Pr(U \longrightarrow R), & 0 \\ 0, & \Pr(R \longrightarrow I) \end{pmatrix}, \tag{4}$$

where $\Pr(U \rightarrow R)$ is an $|U| \times |R|$ matrix representing the transition probability between user nodes and region nodes, as defined in (2) and $\Pr(R \rightarrow I)$ is an $|R| \times |I|$ matrix representing the transition probability between region nodes and business nodes, as defined in (3), and they are symmetric matrixes. We will choose random walk with restarting process to simulate location-based business recommendation process.

5. Making Recommendation on RLG

So far, several graph-based methods have been introduced into recommendation system [19–21] to model the interaction between users and items on a graph and to compute node similarity from a global perspective, instead of local pairwise computation of neighborhood [19]. They essentially transform the recommendation process into graph search problem in a graph. Random walk on graphs has shown that it has a rather good performance in graph-based recommendation systems. PageRank, one of typical random walk algorithms, has been widely used in search engines to rank items globally.

Now we describe the recommendation process as a graph search problem in RLG and use the example shown in Figure 1. Suppose that the system needs to recommend lots of businesses to an active user in one of his movement regions. We firstly determine the association between the user and each of businesses which have not been reviewed by this user. The businesses are then sorted according to the associations and top N businesses are chosen for recommendation.

In our model, the association between two nodes is determined by all paths connecting them. For the pair of a user node u_i and a business node i_j, we compute the association between them as the sum of weights of all

(1) Initialize $a_k(0) = 0$ for all $n_l \in N$, $a_i(0) = 1$
(2) for $t = 1, 2, 3, \ldots, T$, or until convergence do
(3) for each node $n_k \in N$ do
(4) $a_k(t) \leftarrow e_{ik}$
(5) for each node $n_l \in N$ do
(6) if $e_{kl} > 0$ or $t = T$
(7) $a_k(t) \leftarrow a_k(t) + \gamma e_{kl} a_k(t-1)$
(8) end for
(9) end for
(10) end for
(11) return $a_k(T)$, $a_k(T)$ is the association between an active u_i and a business via paths of the length T.

Algorithm 1

Table 1: The total number of users who reviewed N businesses.

$N \geq 400$	$300 \leq N < 400$	$200 \leq N < 300$	$100 \leq N < 200$	$50 \leq N < 100$	$30 \leq N < 50$	$10 \leq N < 30$	$5 \leq N < 10$	$N < 5$
3	8	33	152	418	534	2977	4808	34940

distinct paths that connect u_i and i_j. In this computation, we differentiate two types of paths-paths via local neighbor nodes and paths via remote friend nodes. The length of the two paths is 4. As we defined problem earlier, for a user u_i, if the larger score is on a business reviewed by most of his local neighbors, this user probably will review this business in his local movement region. We adapt modification [22] of PageRank algorithm to calculate the association between an active user and recommended business i_j. For the ease of the algorithms description, let N denote the set of nodes that can form paths from u_i to business nodes; let $n_k \in N$ denote a node regardless of this is a user node, region node, or business node; let e_{kl} denote the weight of the link between nodes $n_k \in N$ and $n_l \in N$; and let e_{il} denote the link weight between node u_i and $n_l \in N$; then, the matrix is formed from $\Pr(G)$. Furthermore, let $a_k(t)$ denote the association degree between u_i and node $n_k \in N$ when considering paths of length t ($t \leq T$). The algorithm for computing association between u_i and business nodes is shown in Algorithm 1.

Here, $\gamma \in [0, 1]$ is a parameter that downweighs longer paths. We fix γ with 0.5 in our experiments. The most time consuming part of this algorithm is from line 3 to line 9 which requires $O(N^2)$ computations over all e_{kl}. However, the matrix (e_{kl}) is very sparse with most elements that are equal to zero and are symmetric. This allows us to use sparse and triangular matrix representation for (e_{kl}), which can reduce the complexity to $O(C \cdot N)$, where C is the maximum number of nonzero elements for each row of matrix (e_{kl}).

6. Experimental Evaluation

6.1. Data Set. We use dataset from Yelp in our following experiment, and it is publicly available [23]. It is from a US city, Phoenix; each review has a location (being reviewed by users) that is associated with a unique pair of latitude and longitude coordinates. It contains 43873 users, 229907 reviews, and 11537 pieces of business information. About half of all users reviewed just only one business, and consequently the dataset is very sparse (99.9545% sparsity). The other pieces of information about the dataset are given in Table 1.

In the following section, we will discuss the location distribution of businesses reviewed by all users which can reveal all users' movement regions. We firstly randomly select ten users from the dataset, and the total number of their reviewed businesses, respectively, is 1, 1, 14, 22, 28, 49, 69, 91, 102, and 112, and the location distribution of businesses reviewed by ten users is as shown in Figure 2. The data from Figure 2 indicates that almost all businesses reviewed by each user are located in a certain region, and most of regions are overlapping, and obvious zoning appeared in the whole region. We can get an observation which is the same as the observation in [6]. To further verify the above-mentioned conclusion, we do some statistical analyses on the percentage of businesses reviewed by each user in three circle regions, whose center is user's movement center $o : (l_{x,o}, l_{y,o})$ and radii are, respectively, $D_{\max}/3$, $D_{\max}/2$, and $2D_{\max}/3$ (as shown in Figures 3, 4, and 5), and we, respectively, call them $D_{\max}/3$ region, $D_{\max}/2$ region, and $2D_{\max}/3$ region.

We can see from the three figures that the larger the number of businesses reviewed by a user is, the higher the proportion of businesses in the central zone is. It also demonstrates that most mobility of all users is restricted in a local region by their daily activities. Furthermore, the percentage of all users who reviewed more than half of businesses in $D_{\max}/2$ region is, respectively, 48.95%, 86.03%, and 99.48% in the three figures. Therefore, we call $D_{\max}/2$ region the local movement region of each user and the other is the remote region.

6.2. Evaluation Metric and Compared Methods. We use Hit ratio [24] as the metric for the top-N recommendation. Our dataset is split into training part and testing part: for each user, the latest businesses he reviewed are selected as test data and other businesses are selected as training data. When a recommendation is made, we always generate a list of N ($N = 10$) businesses named $R(u, R)$ for every user in his

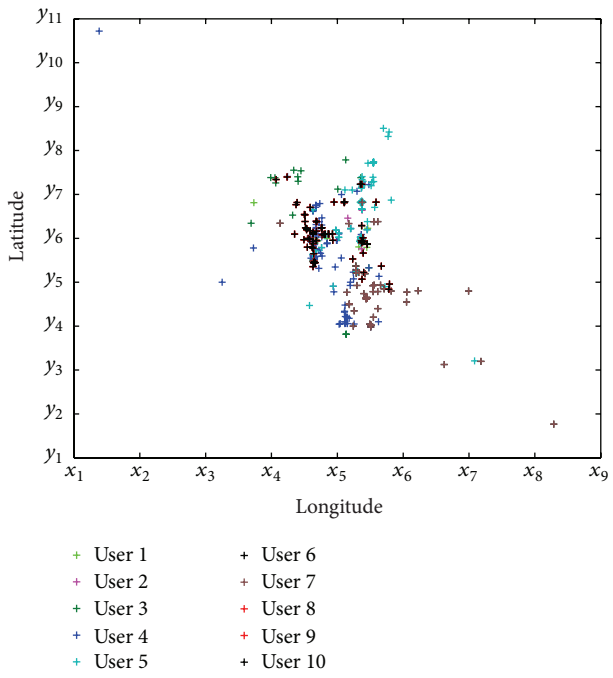


Figure 2: The location distribution of businesses reviewed by ten users.

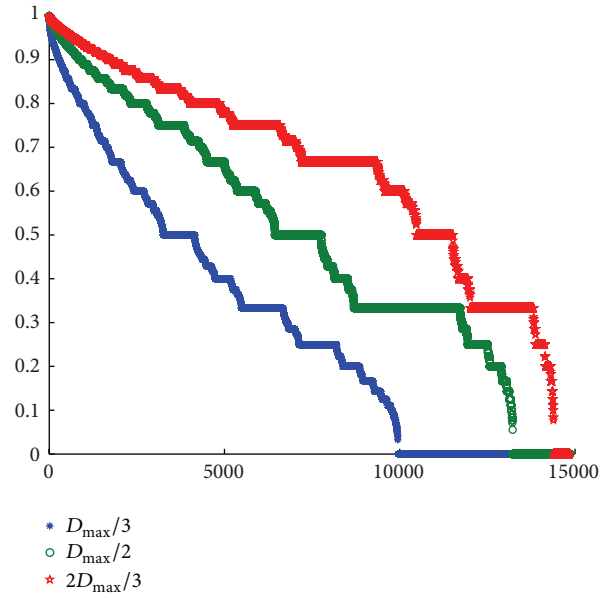


Figure 3: The percentage of businesses reviewed by 14583 users in three circle regions ($N > 3$).

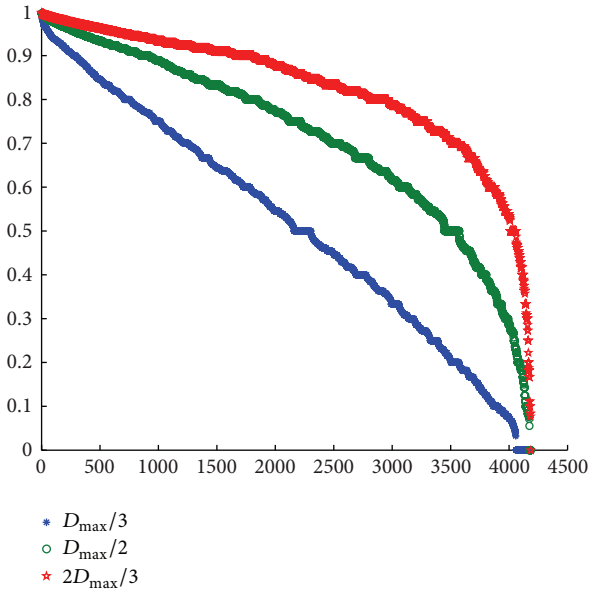


Figure 4: The percentage of businesses reviewed by 4185 users in three circle regions ($N > 10$).

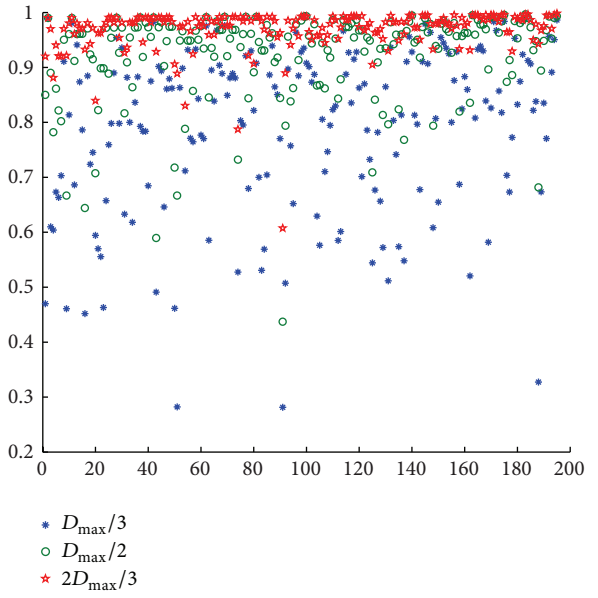


Figure 5: The percentage of businesses reviewed by 195 users in three circle regions ($N > 100$).

whole movement region. If the test business appears in the recommendation list, we call it a hit, and then the Hit ratio can be calculated as follows:

$$\text{Hit\_Ratio} = \frac{\sum_{u_k} h_{u_k}(i) \mid i \in R(u_k, R_{k*})}{N}. \quad (5)$$

We compared the Top-N recommendation performance of our method with several existing methods: popularity-based (Pop@1), item-based collaborative filtering (ItemKNN@1), user-based collaborative filtering (UserKNN@1), and their extended methods under condition of users' movement regions divided into two parts (Pop@2, ItemKNN@2, and UserKNN@2).

Popularity-Based (Pop) Method. Popularity-based method generates a ranking list based on the popularity of businesses in the training dataset. It is not a personalized method

and consequently generates the same list of recommended businesses for every user.

Item-Based Collaborative Filtering (ItemKNN) Method. Item-based collaborative filtering method finds the N businesses which are similar to some businesses reviewed by each user. The similarity between two businesses can be calculated as follows:

$$\text{sim}(i_m, i_n) = \frac{|s_m \cap s_n|}{|s_m \cup s_n|}. \quad (6)$$

User-Based Collaborative Filtering (UserKNN) Method. User-based collaborative filtering method finds the N similar users to the active user and aggregates the ratings of the similar users. The similarity between two users and aggregation function are as follows:

$$\text{sim}(u, u') = \cos(u, u') = \frac{\sum_{i \in R(u) \cap R(u')} r_{u,i} \cdot r_{u',i}}{\sqrt{\sum_{i \in R(u) \cap R(u')} r_{u,i}^2 \cdot \sum_{i \in R(u) \cap R(u')} r_{u',i}^2}}, \quad (7)$$

$$r_{u,i} = \overline{r_u} + \frac{\sum_{u' \in NB} (r_{u',i} - \overline{r_{u'}}) \cdot \text{sim}(u', u)}{\sum_{u' \in NB} |\text{sim}(u', u)|}. \quad (8)$$

6.3. Experimental Results. In this section, we illustrate the results of all methods and show performance of our method for all cold users who do not review any business information. We will firstly investigate the impact of parameter η and then compare the Hit ratios of the four methods recommending cold users.

6.3.1. The Impact of Parameter η. We focus on analyzing parameter η which governs the influence of users' local mobility preference formed by their daily activity and long-distance mobility preference formed by their social network ties in location-based businesses recommendation. When tuning η, the results of how Hit ratios change against all algorithms are shown in Figure 6. Pop@1, ItemKNN@1, and UserKNN@1 do not have parameter η; thus, their Hit ratios are drawn as a straight line. The results show that Pop@2, ItemKNN@2, and UserKNN@2, respectively, outperform Pop@1, ItemKNN@1, and UserKNN@1, and our method always outperforms others whatever parameter η is. Moreover, when we set parameter η to 3/2, we can get the best results of Hit ratio for most of the methods, so we simply fix parameter η to 3/2 in the following experiments.

6.3.2. Make Recommendation for Cold Users. One challenge for most of existing methods is that the recommendation accuracy suffers when the user-business matrix is very sparse. From Table 2, we can see that over half users reviewed just one business in the dataset. Traditional user-based collaborative filtering cannot recommend any business to these users because of the fact that it is difficult to find out the nearest neighbors for these users in it. We regard the location of business reviewed by a user as the center of the user's local region and the average $D_{\max}/2$ of other users as the radius of his local region. We use (6) to calculate $\text{sim}(u_p, u_q)_{R_{p*} \cap R_{qk}}$, and thus our method possesses the advantages of user-based collaborative filtering and item-based collaborative filtering. Our method can alleviate the sparsity problem by exploiting movement region to find out users' local neighbors and remote friends. To verify this hypothesis, we use four methods that recommend some businesses in their local region to all cold users, and the results are shown in Table 3. Apparently, our method has better performance than other methods.

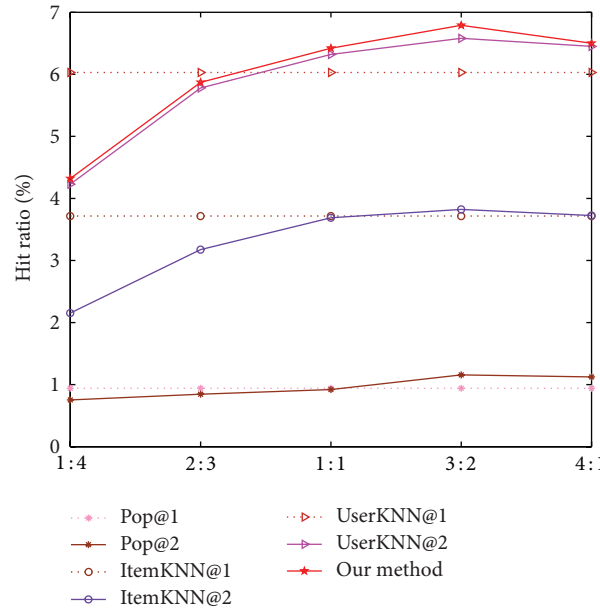


Figure 6: Hit ratios of all algorithms with different η.

Table 2: The number of users who reviewed N businesses ($1 \le N \le 5$).

$N = 5$	$N = 4$	$N = 3$	$N = 2$	$N = 1$
1625	2303	3617	7088	21932

7. Conclusion

User mobility often exhibits long- and short-distance factors which, respectively, formed daily activity and social network ties. Tracking and leveraging these factors for location-based business information recommendation pose great challenges. In this paper, we construct a region-based location graph (RLG), which can combine with user short-ranged mobility formed by daily activity and long-distance mobility formed by social network ties and sequentially can recommend local business information and long-distance business information to users. Moreover, it can combine user-based collaborative filtering with item-based collaborative filtering and can be successful in generating recommendation for cold start users, and, consequently, it can alleviate cold start problem which traditional recommender systems often suffer from.

Table 3: Hit ratios of algorithms for cold users ($\eta = 3 : 2$).

Method	Hit ratio (%)	Improvement
UserKNN@2	0	0
Pop@2	0.94	—
ItemKNN@2	1.428	51.9%
Our method	2.58	80%

The experiments on real dataset confirm that the effectiveness of the proposed method is better than that of other methods.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (no. 60872051) and the Mutual Project of Beijing Municipal Education Commission, China.

References

[1] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee, "Exploiting geographical influence for collaborative point-of-interest recommendation," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*, pp. 325–334, July 2011.

[2] J. Sang, T. Mei, and J. Sun, "Probabilistic sequential POIs recommendation via check-in data," in *Proceedings of 20th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 402–405, November 2012.

[3] T. H. Dao, S. R. Jeong, and H. Ahn, "A novel recommendation model of location-based advertising: context-aware collaborative filtering using GA approach," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3731–3739, 2012.

[4] C. Cheng, H. Yang, I. King, and M. R. Lyu, "Fused matrix factorization with geographical and social influence in location-based social networks," in *Proceedings of 26th AAAI Conference on Artificial Intelligence*, pp. 17–23, July 2012.

[5] C. Cheng, H. Yang, M. R. Lyu, and I. King, "Where you like to go next: successive point-of-interest recommendation," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI '13)*, pp. 2605–2611, August 2013.

[6] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*, pp. 1082–1090, August 2011.

[7] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proceedings of the 18th International World Wide Web Conference (WWW '09)*, pp. 791–800, April 2009.

[8] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Collaborative location and activity recommendations with GPS history data," in *Proceedings of the 19th International World Wide Web Conference (WWW '10)*, pp. 1029–1038, April 2010.

[9] V. W. Zheng, B. Cao, Y. Zheng, X. Xie, and Q. Yang, "Collaborative filtering meets mobile recommendation: a user-centered approach," in *Proceedings of the 24th AAAI Conference on Artificial Intelligence and the 22nd Innovative Applications of Artificial Intelligence Conference (AAAI '10)*, pp. 236–241, July 2010.

[10] K. W.-T. Leung, D. L. Lee, and W.-C. Lee, "CLR: a collaborative location recommendation framework based on co-clustering," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*, pp. 305–314, July 2011.

[11] X. Yu, A. Pan, L.-A. Tang, Z. Li, and J. Han, "Geo-friends recommendation in GPS-based cyber-physical social network," in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM '11)*, pp. 361–368, July 2011.

[12] Y. Zheng and X. Xie, "Learning travel recommendations from user-generated GPS traces," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 1, article 2, 2011.

[13] X. Cao, G. Cong, and C. S. Jensen, "Mining significant semantic locations from GPS data," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 1009–1020, 2010.

[14] M. Ye, P. Yin, and W. C. Lee, "Location recommendation for location-based social networks," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 458–461, 2010.

[15] J. J. Ying, E. H. Lu, and W. Kuo, "Urban point-of-interest recommendation by mining user check-in behaviors," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing (KDD '12)*, pp. 63–70, August 2012.

[16] J. Bao, Y. Zheng, and M. F. Mokbel, "Location-based and preference-aware recommendation using sparse geo-social networking data," in *Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '12)*, pp. 199–208, November 2012.

[17] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme, "Tag recommendations in social bookmarking systems," *AI Communications*, vol. 21, no. 4, pp. 231–247, 2008.

[18] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang, "Towards mobile intelligence: learning from GPS history data for collaborative recommendation," *Artificial Intelligence*, vol. 184-185, pp. 17–37, 2012.

[19] L. Xiang, Q. Yuan, S. Zhao et al., "Temporal recommendation on graphs via long- and short-term preference fusion," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*, pp. 723–731, July 2010.

[20] S. Lee, S.-I. Song, M. Kahng, D. Lee, and S.-G. Lee, "Random walk based entity ranking on graph for multidimensional recommendation," in *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11)*, pp. 93–100, October 2011.

[21] Z. Zhang, D. D. Zeng, A. Abbasi, J. Peng, and X. Zheng, "A random walk model for item recommendation in social tagging systems," *ACM Transactions on Management Information Systems*, vol. 4, no. 2, article 8, 2013.

[22] J. Weston, A. Elisseeff, D. Zhou, C. S. Leslie, and W. S. Noble, "Protein ranking: from local to global structure in the protein similarity network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 17, pp. 6559–6563, 2004.

[23] B. Hu and E. Martin, "Spatial topic modeling in online social media for location recommendation," in *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*, pp. 25–32, October 2013.

[24] G. Karypis, "Evaluation of item-based top-N recommendation algorithms," in *Proceedings of the 10th International Conference on Information and Knowledge Management*, pp. 247–254, November 2001.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 212639, 10 pages
http://dx.doi.org/10.1155/2015/212639

Research Article

Development of a Hand Gestures SDK for NUI-Based Applications

Seongjo Lee, Sohyun Sim, Kyhyun Um, Young-Sik Jeong, Seung-won Jung, and Kyungeun Cho

Department of Multimedia Engineering, Dongguk University-Seoul, Seoul 100-715, Republic of Korea

Correspondence should be addressed to Kyungeun Cho; cke@dongguk.edu

Received 18 August 2014; Accepted 18 October 2014

Academic Editor: Jong-Hyuk Park

Concomitant with the advent of the ubiquitous era, research into better human computer interaction (HCI) for human-focused interfaces has intensified. Natural user interface (NUI), in particular, is being actively investigated with the objective of more intuitive and simpler interaction between humans and computers. However, developing NUI-based applications without special NUI-related knowledge is difficult. This paper proposes a NUI-specific SDK, called "Gesture SDK," for development of NUI-based applications. Gesture SDK provides a gesture generator with which developers can directly define gestures. Further, a "Gesture Recognition Component" is provided that enables defined gestures to be recognized by applications. We generated gestures using the proposed SDK and developed a "Smart Interior," NUI-based application using the Gesture Recognition Component. The results of experiments conducted indicate that the recognition rate of the generated gestures was 96% on average.

1. Introduction

The major advantage of ubiquitous computing is that it enables users to use computers and networks in natural and intuitive ways. Furthermore, ubiquitous computing is characterized by the use of human-focused interfaces, as opposed to computer-focused interfaces using existing input devices. Accordingly, research related to ubiquitous concepts has been playing a significant role in realizing the future of computing, in which computers unobtrusively support humans in everyday life.

In this context, human computer interaction (HCI) is being actively investigated to facilitate the implementation of human-focused computer environments. In particular, research into approaches that enable computers to recognize and understand users' inputs and provide corresponding services has been gaining attention. Heretofore, general interaction between humans and computers has primarily been achieved using intermediary devices such as keyboard and mouse. However, one of the major disadvantages of interaction using such input devices is the fact that users have to learn how to use them. In addition, spatial limitations exist because the devices have to be directly connected to a computer. Consequently, natural user interface (NUI) research is being actively conducted to supplement existing interaction approaches [1, 2].

NUI is human-focused interface that uses the human body as an input, using sensors or cameras [3]. In the NUI approach, users interact with computers via natural human traits such as gestures, viewpoints, and face tracking and recognition [4, 5].

The interaction approach using gestures facilitates a more natural interface and enables users to intuitively use computer systems [6]. Furthermore, it also facilitates intuitive use of the system such that users do not have to learn how to use the interface and input devices [7]. Subsequent to the development of the Kinect sensor, which recognizes depth information and human joint information, by Microsoft's "Project Natal" at the end of 2010, research into the recognition of gestures using depth information and human joint information has been actively pursued [8, 9]. Further, research is also being conducted into how human centric interfaces can be used to recognize situations and intentions for a flexible interface between humans and machines, such as conversations between human beings [10].

The hand gesture SDK proposed in this paper identifies hand gestures using vector chain code. The screen area is divided into a grid to identify the vector chain code and gestures are recognized using the identified vector chain code and HMM. The proposed approach shows the identification area on the screen and thereby enables users to obtain feedback on the execution of a gesture by watching the screen. Furthermore, this system is suitable for application development because gestures are quickly recognized.

NUI-based application development currently requires special NUI-related knowledge. Consequently, developers who have no basic knowledge of NUI have difficulty developing NUI-based applications. This paper proposes a hand gesture SDK that facilitates development of NUI-based applications without any specific NUI knowledge. We present the results obtained using the proposed SDK to develop a "Smart Interior" system comprising a virtual environment and contents.

The remainder of this paper is organized as follows: Section 2 introduces existing research related to the recognition of hand and arm gestures. Section 3 outlines the overall structure of the gesture SDK proposed in this paper. Section 4 presents and discusses our implementation results for the "Smart Interior" application using the proposed gesture SDK. Finally, Section 5 concludes this paper.

2. Related Work

This paper proposes a hand gesture SDK that uses joint information from the Kinect sensor. Approaches for recognizing hand gestures have been extensively investigated over the years. Further, the use of hand gestures as input is being studied for a variety of applications and games in various fields, including the medical field.

For example, Kim et al. [1] proposed an algorithm that controls multimedia content solely by recognizing human gestures. The method detects the hands on the basis of the depth image of Kinect and YCbCr color image. The hand trace is then translated into eight-direction chain code and the hand gestures are consequently recognized by hidden Markov model (HMM) algorithm.

Park et al. [3] proposed a hand gesture recognition approach that uses visual images and depth information. In the proposed approach, the hand area is detected in the depth image obtained from a Kinect sensor. Then, the hand area in the depth information and the visual image are mapped. Consequently, hand gestures are recognized by defining the number of fingers unfolded. They assessed the performance of their proposed system by investigating its ability to control a medical image on a monitor.

Heo et al. [6] presented an approach based on HMM algorithm that recognizes one-arm gestures in real time environments. In their approach, they configure a vector using the coordinates of arm joints detected by Kinect. Next, they execute a feature transformation process to convert the vector into the angle value between arm joints. After dimension reduction and discretization through k-means clustering, the arm gestures are then recognized using HMM. Lee and Choi [11] applied mathematical morphology to gesture recognition. Their approach traces the trajectory of the hand centers in hand gesture sequences containing important data related to the form of the hand gesture. The hand gesture sequences are then recognized by acquiring eight-direction chain code edge vectors from the traced trajectory of centers.

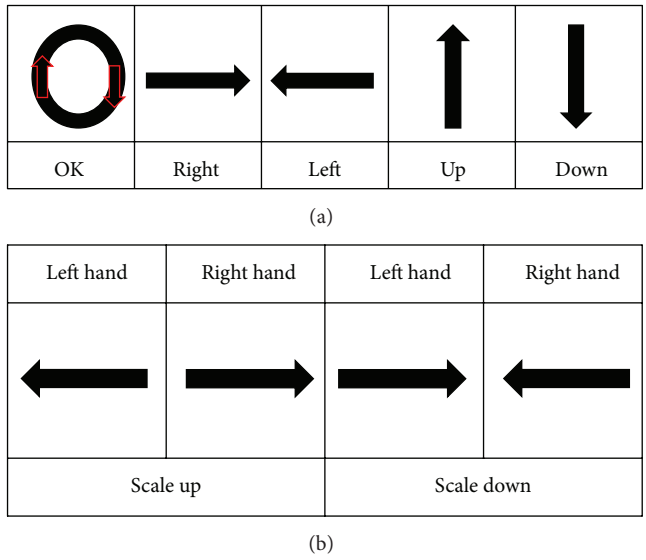


Figure 1: Gestures predefined by the proposed system.

The NUI approaches presented above use human gestures as input data in applications, games, and the medical field. However, they did not develop an SDK, which could aid in the development of programs, for recognition of hand or arm gestures. This paper proposes a gesture SDK that enables developers to develop NUI-based applications more easily.

3. The Gesture SDK

The gesture SDK proposed in this paper comprises the Gesture Recognition Component, which recognizes gestures in applications, and the Custom-Defined Gesture Creator, which generates the gestures used in applications.

3.1. The Custom-Defined Gesture Creator. The gesture SDK proposed in this paper classifies gestures as either one-handed or two-handed gestures. It provides five kinds of basic one-handed gestures and two kinds of two-handed gestures as predefined gestures. Figures 1(a) and 1(b) show the various predefined one-handed and two-handed gestures, respectively.

The Kinect Interaction Module in this system classifies the beginning and end of a gesture into grip and nongrip and gesture and nongesture. Consequently, this system can classify the gestures entered by a user and differentiate them from other gestures.

The Custom-Defined Gesture Creator enables developers to directly generate custom gestures for implementing a variety of NUI-based applications. It extracts feature vectors from human gestures and generates an HMM to recognize the gesture using the extracted feature vector.

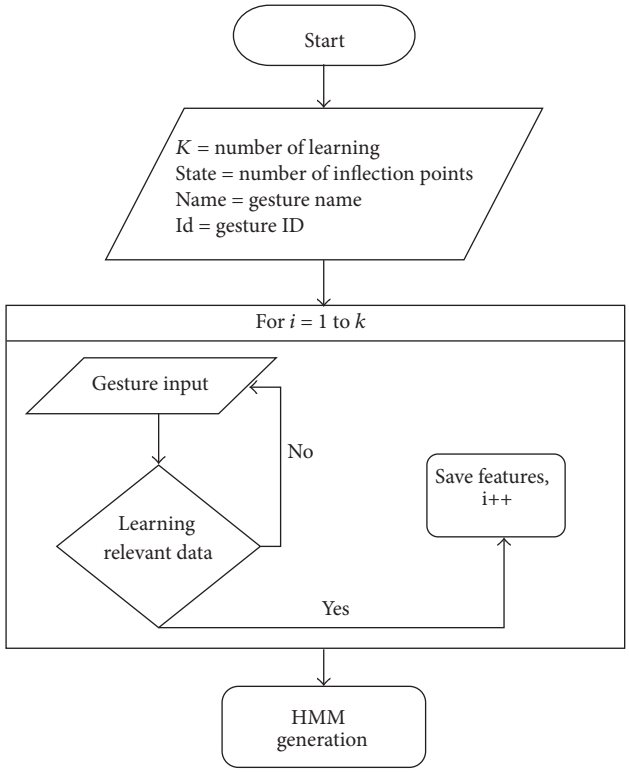


Figure 2: Custom-Defined Gesture Creator flowchart.

It comprises a Kinect Interaction Module and a Gesture Maker Module. The Kinect Interaction Module loads the human hand data captured by Kinect and extracts the features of the gesture from the loaded data. The Gesture Maker Module then generates an HMM to recognize the gesture using the features extracted by the Kinect Interaction Module.

The program generates gestures on receiving the gesture ID number required for defining the gesture, the gesture name, the number of inflection points classified into eight directions, and the gesture learning number. Figure 2 presents the one-handed gesture HMM generation flowchart using the Custom-Defined Gesture Creator program.

3.1.1. The Kinect Interaction Module. To recognize gestures, the proposed system classifies human motions into gestures and nongestures. A gesture is a hand motion for controlling a program, whereas a nongesture is meaningless motion between gestures. This module classifies motions using the hand status data from the Kinect SDK and the Kinect sensor. In the system, a clenched fist is classified as a gesture, whereas an open hand is deemed a nongesture.

The Kinect Interaction Module converts the trajectory of gestures into eight-directional vector chain code and trains the HMM using the chain code as the features for recognizing the gestures. It acquires joint location data from Kinect in order to obtain the trajectory of the hand, which is used as the features of the gestures. After getting the data to extract the features of gestures from Kinect, the Kinect Interaction Module extracts a feature vector to train the HMM to recognize gestures. The Kinect Interaction Module is also utilized to carry out this function in the Gesture Recognition Component.

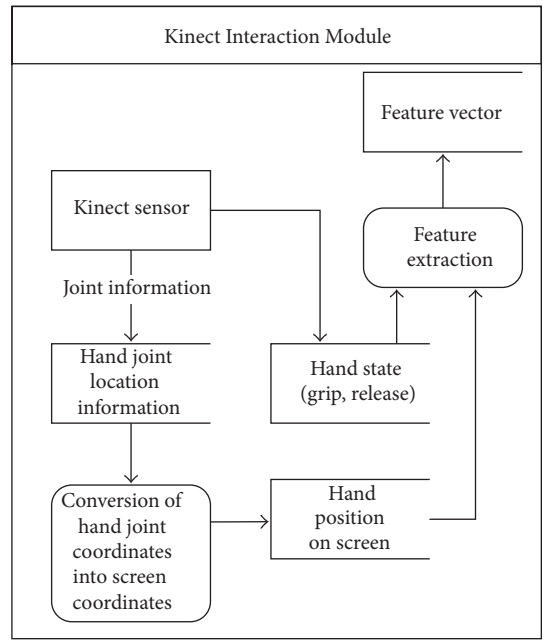


Figure 3: Data flow in the Kinect Interaction Module.

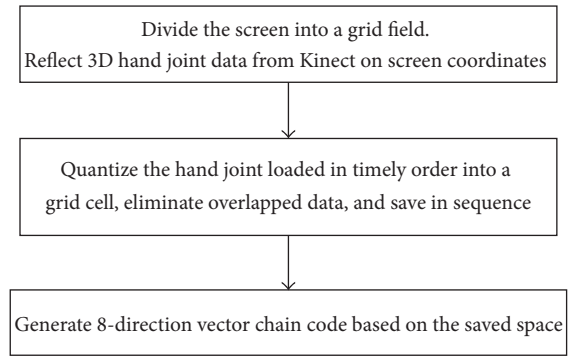


Figure 4: Eight-direction vector chain code extraction flow.

To effectively use the features of applications using screens, the proposed method adopts the screen for recognizing the gestures. The Kinect Interaction Module extracts the feature vector by converting hand coordinates into screen coordinates and dividing the screen space into several grids. Next, the hand coordinates on the screen are quantized on grids. The features of the gestures are then extracted using the quantized grids. The hand gesture recognition field can be moved to the screen using the proposed approach. Further, users can give feedback on the gestures while watching the hands on the screen. Figure 3 illustrates the data flow in the Kinect Interaction Module.

Eight-direction vector codes range from one to eight in a clockwise direction, with the top vector as the basis. Extraction of the eight-direction vector chain code comprises preparation for extracting the features, storage of the input data, and extraction of features. Figure 4 presents the three steps used to extract the features.

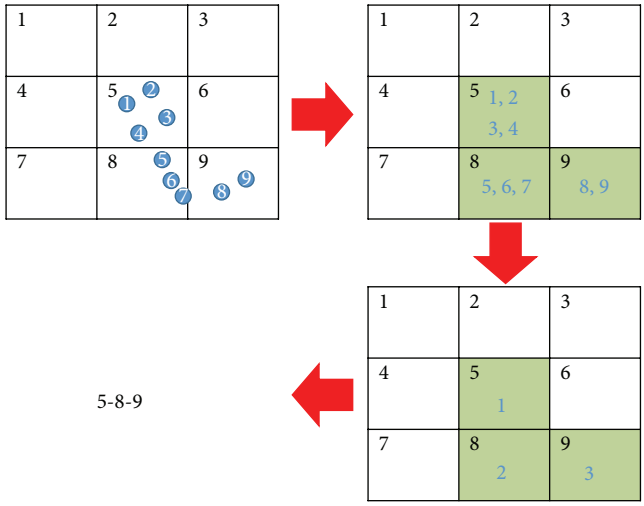


Figure 5: Quantization of hand joints on grid cells.

First, the screen is segmented into grid cells; in this case, the screen is segmented into 10 × 6 grid cells. The location data of the 3D hand joint is then projected onto 2D screen coordinates.

Next, the system quantizes the hand coordinates entered in sequence into the segmented grid cells. If the grid cell for the hand coordinates in the previous frame is the same as that in the present frame, the present grid cell is excluded to eliminate overlapping data in the quantized grid cell. The quantized grid cells are then saved in sequence following elimination of the overlapped data. Figure 5 shows an example on a 3 × 3 grid field that explains the second process outlined above. The dots in the first figure are 2D hand joint coordinates in accordance with the passage of time. The shaded grid cells in the second figure show the grid cells through which the hand joints have passed. Grid cells 5, 8, and 9 are saved in sequence.

Next, the relative vectors of the grid cells saved in sequence are calculated in relation to the previous location and converted to vector chain code. During the calculation of the relative vector using the rectangular grid field, the upward and downward movement and movement to the left and right can be accurately extracted because the surface area resulting from the grid cell meeting other grid cells above it, below it, on the left, and on the right is large. However, the grid cells located diagonally meet only at each grid's vertex. Consequently, accurate extraction of diagonal movement is difficult. Figure 6 illustrates the movement of a hand in the grid field in the diagonal direction.

As shown in Figure 6, when a hand moves to the vector in the number 2 direction, the grid cells through which the hand joint coordinates pass are saved in the order of 5, 2, and 3 or 5, 6 and 3 not in the order of 5 and 3.

In this case, while the hand actually moved to the vector in the number 2 direction, the vector for the relative location in the grid field is extracted as the combination of two vectors, 1-3 or 3-1. To rectify this problem, when a hand moves in the diagonal direction, the vector corresponding to the relevant diagonal direction conducts the extraction by considering

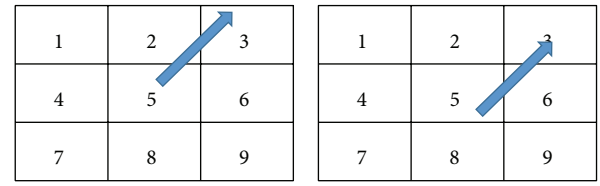


Figure 6: Movement of the vector in the number 2 direction in the grid field.

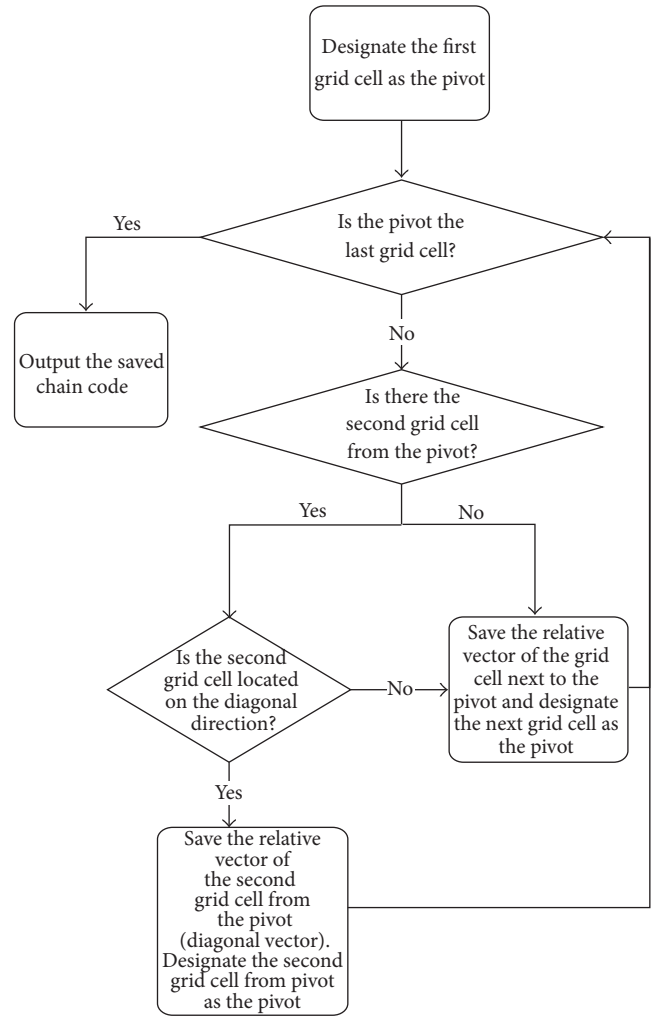


Figure 7: Flowchart for identifying vector chain code using saved grid cells.

the grid cell next to the cell in question. Figure 7 presents a flowchart for extraction of the eight-direction vector using the relative location of the grid cell saved.

3.1.2. The Gesture Maker Module. HMM is ideal for dealing with sequential data and is consequently frequently used to recognize gestures and voice in NUI fields. HMM has a variety of models depending on the direction of state transition. These include the Ergodic model and the Left-Right model. This paper adopts the Left-Right HMM to

recognize the hand gestures using the feature vector extracted by the Kinect Interaction Module.

The definition of the number of Left-Right HMM states varies according to the gesture patterns. The number of predefined gesture states ranges between two and eight and is equal to the number of inflection points. The number of HMM states is represented by N, and S represents a set of states:

$$S = \{s_1, s_2, \ldots, s_N\}. \tag{1}$$

HMM has three parameters:

$$\theta = (A, B, \pi), \tag{2}$$

in which A is a state transition probability matrix:

$$\begin{aligned} A &= \left|a_{ij}\right|, \quad (1 \le i, j \le N), \\ a_{ij} &= P\left(q_{t+1} = s_j \mid q_t = s_i\right), \\ \sum_{j=i}^{N} a_{ij} &= 1, \end{aligned} \tag{3}$$

where q_t is the state of the model at t.

B is an observation symbol probability matrix:

$$\begin{aligned} B &= \left|b_j(v_k)\right|, \quad v_k \in V \\ O &= \{o_1, o_2, \ldots, o_t, \ldots, o_L\}, \quad (1 \le t \le L) \\ b_j(v_k) &= P\left(o_t = v_k \mid q_t = s_j\right) \\ \sum_{k=1}^{8} b_j(v_k) &= 1, \end{aligned} \tag{4}$$

where V is a set of 8-direction vectors, O is a vector chain code, L is the length of O, and o_t is an observation vector at t.

π is a vector of initial state probabilities:

$$\begin{aligned} \pi &= \{\pi_i\}, \quad (1 \le i \le N), \\ \pi_i &= P(q_1 = s_i). \end{aligned} \tag{5}$$

To generate the HMM, the three HMM parameters are optimized using extracted 8-directional vector chain codes. The learning algorithm used in this paper is the Baum-Welch algorithm. The Baum-Welch algorithm, an expectation and maximization (EM) algorithm, is typically used to train HMMs. The algorithm uses two latent variables; it comprises estimating latent variables (E step) and updating parameters, A, B, and π (M step) [12].

However, when this algorithm is utilized, the possibility exists that the three parameters may become local optimums. To prevent this from happening, we generate the initial values of the parameters using a random number generator and repeat the process 150 times. When the random number generator assigns $A = |a_{ij}|$, the condition of A is given by

$$a_{ij} = 0, \quad (i > j). \tag{6}$$

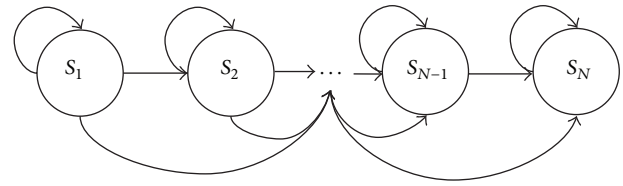


Figure 8: State transition diagram for Left-Right HMM.

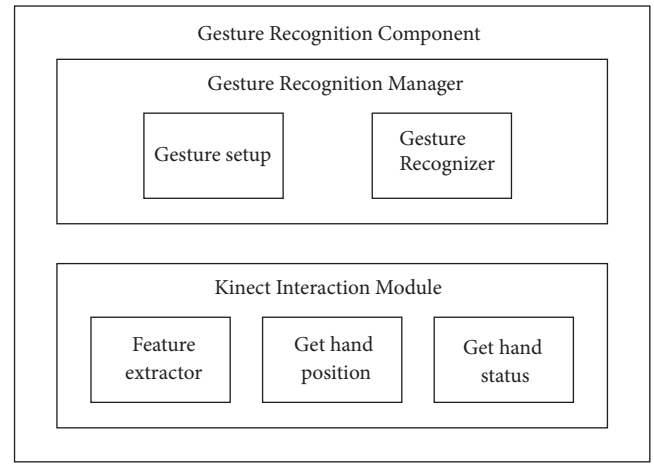


Figure 9: Architecture of the Gesture Recognition Component.

As stated above, we adopt Left-Right HMM. Figure 8 presents the state transition diagram for Left-Right HMM.

Fifty learning datasets were used per gesture to train the predefined gesture HMM; recognition accuracy increases with the amount of learning data used.

The models learned by the Gesture Maker Module are saved in "Gesture DB"—a repository containing the gestures that the proposed SDK can recognize. The models saved in "Gesture DB" are used to recognize the user gestures in the Gesture Recognition Component.

The proposed system supports both two-handed and one-handed gestures. The process used to generate two-handed gestures does not require the generation of a new model. Because a two-handed gesture is recognized as a combination of two one-handed gestures, a new ID is defined for a two-handed gesture by inputting the one-handed gesture ID for the left hand and that for the right hand.

3.2. The Hand Gesture Recognition Component. The Gesture Recognition Component acts as an intermediary between the application layer and users. It comprises the Kinect Interaction Module and the Gesture Recognition Manager Module.

The Kinect Interaction Module is the same as that described for the Custom-Defined Gesture Creator above. The Gesture Recognition Manager Module recognizes users' gestures and set the gestures the application can recognize. Figure 9 depicts the architecture of the Gesture Recognition Component.

The Gesture Setup Module in the Gesture Recognition Manager sets gestures that can be recognized in the application. Developers can use "Gesture Setup" to change the kinds of gestures depending on the application state. The "Gesture Recognizer DB" registers the models of gestures for recognition in the application. When a developer enters the list of gestures to recognize some gestures using "Gesture Setup," the HMM of the gestures registered in "Gesture DB" in the Custom-Defined Gesture Creator is registered in "Gesture Recognizer DB."

With the "Gesture Recognizer DB," only the HMM of the gesture to be recognized is compared. This approach reduces the frequency of false recognition by the Gesture Recognizer as compared to the comparison containing the HMM, even for gestures not used in the application. In consideration of the features of the "Gesture Recognizer DB," one-handed and two-handed gestures are saved in different places in the repository. Furthermore, even the repository for the two-handed gesture is divided into two spaces—one for the left hand and the other for the right hand.

Gesture Recognizer classifies the gestures when a user's gesture is entered and generates the corresponding events. In the HMMs of the gesture registered in the "Gesture Recognizer DB," the likelihoods are computed using the vector chain extracted by the Kinect Interaction Module.

Each model evaluates the probability with which a vector chain code identified from a user's gesture is entered. The gesture model with the highest such probability is classified as the gesture of the identified vector chain code. This is given by the expression

$$q = \arg \max_i P\left(O \mid \theta_{g^i}\right), \tag{7}$$

where O is the vector chain of a hand motion, θ_{g^i} is the HMM of the gesture learned, q is an estimation gesture index, and $P(O \mid \theta)$ is calculated using the Forward algorithm [12].

When the left and right hand IDs entered when defining two-handed gestures in the Custom-Defined Gesture Creator match, the relevant two-handed gesture is recognized.

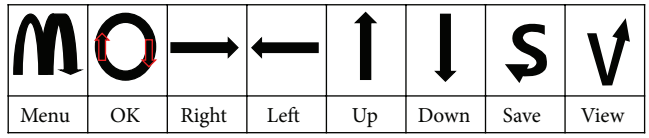
When left and right hand IDs do not match, input gesture is rejected.

Subsequently, the event function containing the result of gesture recognition is called. Developers develop applications using the results of gesture recognition by registering the callback function to the event.

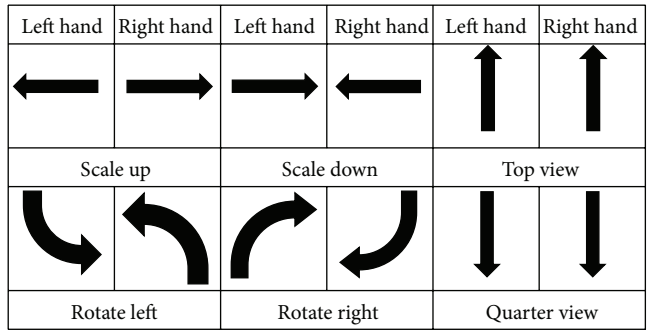
4. Experiment and Analysis

This section introduces "Smart Interior," a NUI-based application developed using the gesture SDK proposed in this paper. Further, the performance of the gesture SDK in the "Smart Interior" application is evaluated. This SDK is developed using C# language.

4.1. Smart Interior. Smart Interior is used to decorate a virtual house by arranging furniture or accessories or changing floor materials, ceiling, or wallpaper. Further, users can enter and look around the house in a first person point of view. Additional gestures were generated using the Custom-Defined Gesture Creator, the SDK tool, for the gestures required in "Smart Interior." Figure 10 shows the gestures used in "Smart Interior," while Figure 11 illustrates the steps used to add the gestures using the Custom-Defined Gesture Creator.



(a) Types of one-handed gestures



(b) Types of two-handed gestures

Figure 10: Types of gestures used in "Smart Interior."

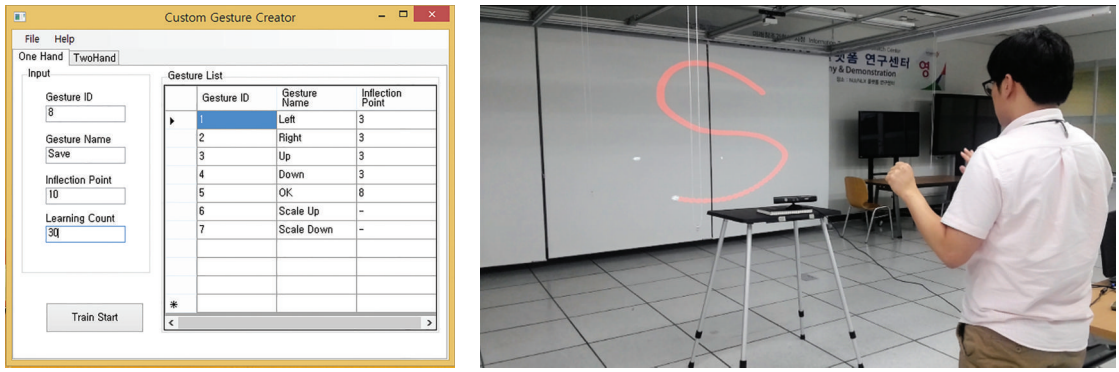
Table 1: Gestures used in "Start Scene" and actions for gestures.

Grip hand	Gesture	Actions
Left	Left	Go to house structure scene
Left	Right	Load the previous house

The application classifies the purpose of gestures using the status of the left and right hands and whether a hand is open or closed, when recognizing one-handed gestures. The gesture made with the left hand closed is the System Gesture, which controls system components such as Confirm, Cancel, Undo, Redo, Save, and View Mode. The gesture made with the right hand closed is the Menu Control Gesture to control the list of items on the interior menu. Two-handed gestures are used to control the camera view and edit the interior items. It executes scale and rotate for items and sets the camera view to TopView or QuarterView.

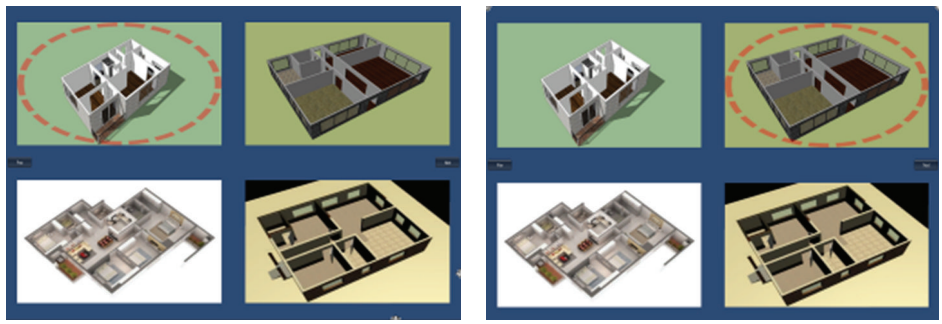
The Smart Interior application has four kinds of scenes for using gestures: "Start Scene," "Select House Model," "Load," and "Edit Scene." "Start Scene" is used to build a new house or load a previously saved house. "Select House Model" is used to select the house structure for new houses. "Load" is used to select and load previously decorated houses. "Edit Scene" is used to decorate houses. Because each scene uses different gestures, the gestures to be used in the relevant scene were defined using Gesture Setup in the Gesture Recognition Manager. Tables 1, 2, and 3 list the commands used to control Smart Interior.

The results above show that NUI-based applications can be developed using the gesture SDK proposed in this paper.



(a) Input (b) Training

Figure 11: Custom-Defined Gesture Creator.



(a) Before (b) After

Figure 12: Interaction with the application using the left gestures in "Select Model House."

Table 2: Gestures used in "Select House Model" and "Load" and actions for gestures.

Grip hand	Gesture	Actions
Left	OK	Load the selected house
Right	Left	Go to the house on the left
Right	Right	Go to the house on the right

Figures 12 and 13 show the interaction with the application using the gestures defined in this application.

4.2. Performance Assessment. We used 50 images per gesture in our assessment of the hand gestures recognition performance. The Kinect sensor used to obtain the images was installed at a height of 1 m. The distance between the Kinect sensor and a human being was 1.5 m–2 m when the pictures were being taken.

Table 4 presents the performance calculated using the approach proposed in this paper. N is the number of input gestures, N_{rej} is the number of rejected gestures, N_E is the number of gestures with recognition errors, N_{Hit} is the number of gestures recognized properly, and R is the recognition ratio:

$$R = \frac{N_{\text{Hit}}}{N} \times 100\%. \tag{8}$$

R_{ext} is the accuracy of the gestures actually detected, excluding the rejected gestures:

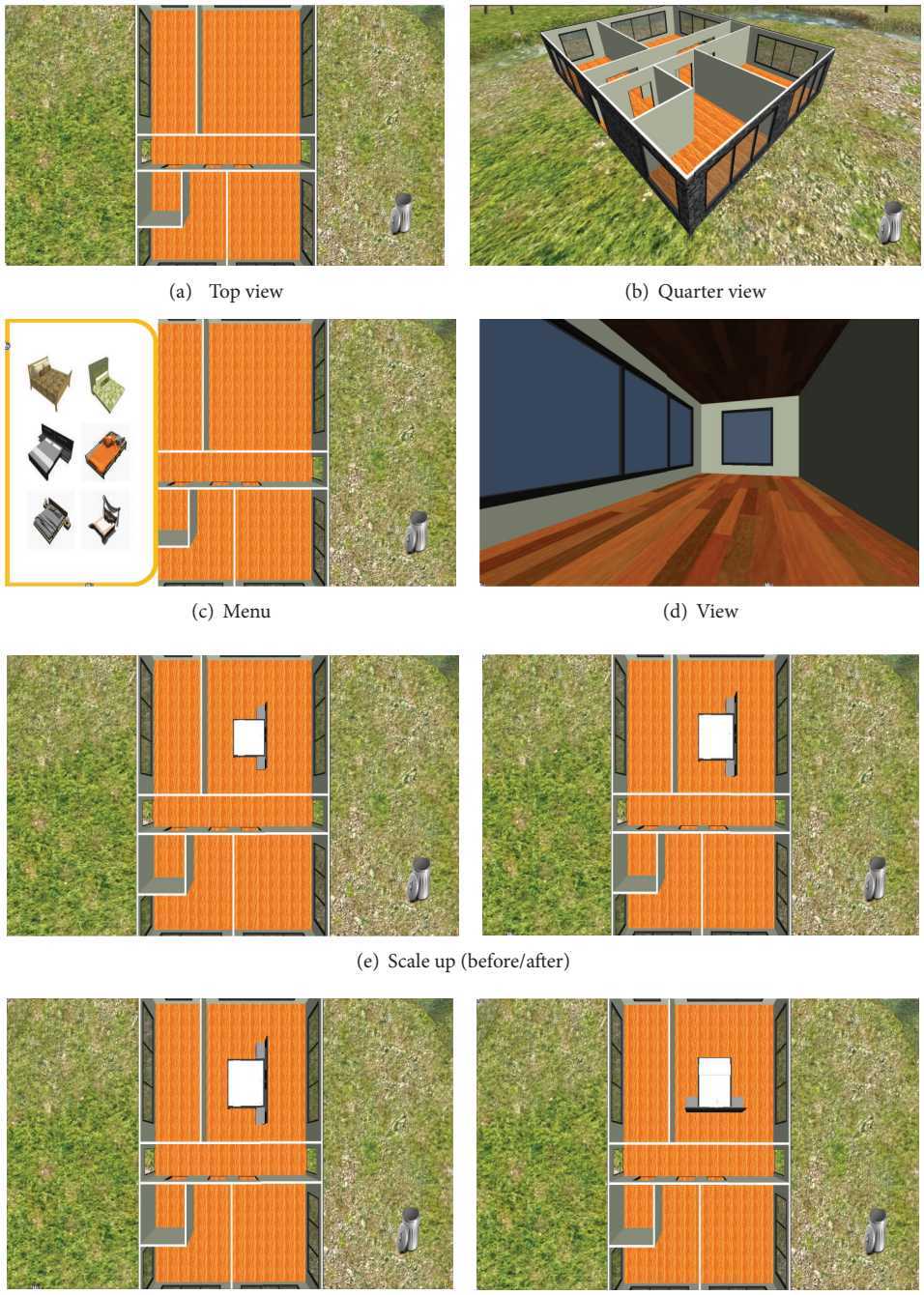
$$R_{\text{ext}} = \frac{N_{\text{Hit}}}{N - N_{\text{Rej}}} \times 100\%. \tag{9}$$

To verify the real time interaction with a computer using the recognition of gestures proposed in this paper, the recognition time of each gesture used in "Smart Interior" was measured 20 times. Table 5 presents the average performance time.

The default value for monitor refresh rate is predominantly 60 Hz. Thus, when the program executes the vertical synchronization in accordance with the default value above, the output time for one frame is 16.6667 ms. This result confirms that the program is applicable to real time HCI.

5. Conclusion

This paper proposed "Gesture SDK" to facilitate the development of gesture recognition-based applications without special knowledge. Although various gesture recognition-related proposals exist, they are primarily limited to identification of and recognition of gestures, in contrast to our proposed SDK for developing NUI-based applications. Using the proposed SDK, developers can directly define gestures



(a) Top view (b) Quarter view

(c) Menu (d) View

(e) Scale up (before/after)

(f) Rotate (before/after)

Figure 13: Interaction with the application using the left gestures in "Edit Scene."

and simply apply the defined gestures to their applications. Our gesture recognition experiment for the "Smart Interior" developed using the SDK indicates an average recognition ratio of 94%. The results suggest that NUI-based applications can be developed using the proposed "Gesture SDK."

The gesture generator, Custom-Defined Gesture Creator, proposed in this paper currently cannot identify duplicated gestures. Further study is therefore needed in order to develop a method for recognizing duplicated gestures. Furthermore, the algorithm that automatically sets the state value when defining the gestures also requires more research. A method of differentiating gestures from user's motion without using grip and release is also needed.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Table 3: Gestures used in "Edit Scene" and actions for gestures.

Grip hand	Gesture	Actions
Right	Menu	Open the list of furniture/items
Right	Right	Open the list of furniture/items on the right
Right	Left	Open the list of furniture/items on the left
Right	Up	Go to the next page of the list of furniture/items
Right	Down	Go to the previous page of the list of furniture/items
Left	Left	Undo
Left	Right	Redo
Left	View	Look around the house in the process of decoration
Left	Save	Save the house in the process of decoration
Left/right	Scale up	Increase the size of the selected furniture/item
Left/right	Scale down	Reduce the size of the selected furniture/item
Left/right	Rotate left	Rotate the selected furniture counterclockwise
Left/right	Rotate right	Rotate the selected furniture/item clockwise
Left/right	Top view	Change the camera view to top
Left/right	Quarter view	Change the camera view to quarter

Table 4: Gesture performance assessment.

	N	N_{rej}	N_E	N_{Hit}	R	R_{ext}
Right	50	0	0	50	100%	100%
Left	50	0	5	45	90%	90%
Up	50	0	0	50	100%	100%
Down	50	0	0	50	100%	100%
OK	50	0	0	50	100%	100%
Menu	50	0	1	49	98%	98%
Save	50	0	3	47	94%	94%
View	50	0	0	50	100%	100%
Scale up	50	1	0	49	98%	100%
Scale down	50	1	0	49	98%	100%
Top view	50	2	0	48	96%	100%
Quarter view	50	2	1	47	94%	95.92%
Rotate right	50	3	0	47	94%	100%
Rotate left	50	4	0	46	92%	100%

Acknowledgments

This research was supported by the Ministry of Science, ICT, and Future Planning (MSIP), Korea, under the Information Technology Research Center (ITRC) support program (NIPA-2014-H0301-14-1021) supervised by the National IT Industry Promotion Agency (NIPA).

Table 5: Average gesture recognition performance time.

Gesture	T_{ext} (ms)	T_{recog} (ms)	T_{total} (ms)
Right	6.0026	1.6967	7.6993
Left	7.4087	1.0049	8.4136
Up	7.8124	2.0002	9.8126
Down	7.0095	2.3000	9.3095
OK	5.3028	1.4005	6.7033
Menu	7.9049	1.7017	9.6066
Scale up	10.9095	1.1981	12.1076
Scale down	10.9058	1.4999	12.4057
Rotate left	11.9065	1.7012	13.6077
Rotate right	11.8101	1.6002	13.4103
Top view	13.2083	2.3996	15.6079
Quarter view	13.6069	1.7022	15.3091
Total	9.5657	1.68377	11.2494

References

[1] Y. Kim, S. Park, S. Ok, S. Lee, and E. Lee, "Human gesture recognition technology based on user experience for multimedia contents control," *Journal of Korea Multimedia Society*, vol. 15, no. 10, pp. 1196–1204, 2012.

[2] S. Cho, H. Byun, H. Lee, and J. Cha, "Arm gesture recognition for shooting games based on Kinect sensor," *Journal of KIISE: Software and Applications*, vol. 39, no. 10, pp. 796–805, 2012.

[3] K. Park, D. Lee, and Y. Park, "Hand gesture recognition using depth information and visual image," *Journal of KIIT*, vol. 11, no. 7, pp. 57–65, 2013.

[4] D. Ghimire and J. Lee, "A robust face detection method based on skin color and edges," *Journal of Information Processing Systems*, vol. 9, no. 1, pp. 141–156, 2013.

[5] X. Yang, G. Peng, Z. Cai, and K. Zeng, "Occluded and low resolution face detection with hierarchical deformable model," *Journal of Convergence*, vol. 4, no. 2, pp. 11–14, 2013.

[6] S. Heo, Y. Shin, H. Kim, and I. Kim, "Design of an arm gesture recognition system using feature transformation and hidden Markov models," *KIPS Transactions on Software and Data Engineering*, vol. 2, no. 10, pp. 723–730, 2013.

[7] M. K. Sohn, S. H. Lee, D. J. Kim, B. Kim, and H. Kim, "3D hand gesture recognition from one example," in *Proceedings of the IEEE International Conference on Consumer Electronics (ICCE '13)*, pp. 171–172, January 2013.

[8] K. Biswas and S. Basu, "Gesture recognition using Microsoft Kinect," in *Proceedings of the 5th International Conference on Automation, Robotics and Applications (ICARA '11)*, pp. 100–103, December 2011.

[9] Y. Wang, C. Yang, X. Wu, S. Xu, and H. Li, "Kinect based dynamic hand gesture recognition algorithm research," in *Proceedings of the 4th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC '12)*, pp. 274–279, August 2012.

[10] N. Howard and E. Cambria, "Intention awareness: improving upon situation awareness in human-centric environments," *Human-Centric Computing and Information Sciences*, vol. 3, no. 1, pp. 1–17, 2013.

[11] K. Lee and J. Choi, "Hand gesture sequence recognition using morphological chain code edge vector," *Journal of The Korea Society of Computer and Information*, vol. 9, no. 4, pp. 85–91, 2004.

[12] X. Li, M. Parizeau, and R. Plamondon, "Training hidden markov models with multiple observations-a combinatorial method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, pp. 371–377, 2000.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 807213, 16 pages
http://dx.doi.org/10.1155/2015/807213

Research Article

Provable Secure and Efficient Digital Rights Management Authentication Scheme Using Smart Card Based on Elliptic Curve Cryptography

Yuanyuan Zhang,[1] Muhammad Khurram Khan,[2] Jianhua Chen,[1] and Debiao He[1]

[1] *School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China*
[2] *Center of Excellence in Information Assurance, King Saud University, Riyadh 12372, Saudi Arabia*

Correspondence should be addressed to Muhammad Khurram Khan; mkhurram@ksu.edu.sa

Received 30 June 2014; Accepted 28 July 2014

Academic Editor: Jong-Hyuk Park

Since the concept of ubiquitous computing is firstly proposed by Mark Weiser, its connotation has been extending and expanding by many scholars. In pervasive computing application environment, many kinds of small devices containing smart cart are used to communicate with others. In 2013, Yang et al. proposed an enhanced authentication scheme using smart card for digital rights management. They demonstrated that their scheme is secure enough. However, Mishra et al. pointed out that Yang et al.'s scheme suffers from the password guessing attack and the denial of service attack. Moreover, they also demonstrated that Yang et al.'s scheme is not efficient enough when the user inputs an incorrect password. In this paper, we analyze Yang et al.'s scheme again, and find that their scheme is vulnerable to the session key attack. And, there are some mistakes in their scheme. To surmount the weakness of Yang et al.'s scheme, we propose a more efficient and provable secure digital rights management authentication scheme using smart card based on elliptic curve cryptography.

1. Introduction

In 1991, ubiquitous computing was firstly proposed by Mark Weiser, who thought that ubiquitous computing technology could provide users service with a variety of equipment in environment which would be disappeared from the user's consciousness [1]. Later, IBM Corporation scientists also raised the idea in 1999, and they forecasted that pervasive computing can be a way to compute everywhere, anytime, and anywhere [2, 3]. Since the computer and internet technology development, multimedia contents (image, document, music, movie, video, etc.) have been greatly enriched all the time, and all of them can be easily redistributed, copied, and downloaded on the internet without authorization. This drawback results in rampant piracy and causes huge revenue to lose to the electronic commerce [4]. As a result, in pervasive computing application environment, the protection of digital publication copyright becomes more and more important. Digital rights management (DRM) technology is developed to overcome the problem [5]. Normally, DRM is only software which usually restricts the usage of the content to protect copy and distributed contents [6–9]. The DRM system manages the procedure of the digital contents including protection, distribution, and authorization. Using DRM technology, intellectual property is respected and protected by data encryption, so it can only be accessed by authorised users without limitless distribution [10, 11].

In 2009, the first three-role based DRM implementation scenario authentication scheme using smart card was proposed by Zhang et al. [12]. Then, Yang et al. showed that Zhang et al.'s scheme was vulnerable to the insider attack and the stolen smart card attack [10]. Due to surmounting the weaknesses of Zhang et al.'s scheme, Yang et al. proposed an enhanced digital rights management authentication scheme based on smart card. They demonstrated that their scheme could preclude all the weaknesses existing in Zhang et al.'s

scheme. Recently, Mishra and Mukhopadhyay cryptanalyzed Yang et al.'s scheme and found that their scheme cannot resist the password guessing attack and the denial of service attack. Moreover, they also pointed out that Yang et al.'s scheme is not efficient enough when the user inputs an incorrect password, and this drawback may cause a denial of service attack [13]. Except for the attacks mentioned by Mishra et al., we find out that Yang et al.'s scheme does not resist the session key attack. In addition to this, we also discover that there are some mistakes in their scheme.

We proposed a new efficient and provable secure digital rights management authentication scheme using smart card based on elliptic curve cryptography [14–16]. To demonstrate the scheme is provable secure, we introduce a security model AFP05 [17, 18] and analyze our scheme in this model. In the following, we will give the proof that our proposed scheme is secure in the AFP05 model. As known to all, one-way hash function is more efficient than the operation of scalar multiplication and pairings [19–21]. Moreover, the pairing operation costs much more than the scalar multiplication operation. The effort of evaluating one pairing operation is approximately three times the effort of evaluating one scalar multiplication operation. So, we cut down some pairings operation of point on elliptic curve and use hash function instead to increase the scheme's efficiency.

The structure of this paper is arranged as follows. In Section 2, we introduce the notations and definitions used in this paper. Section 3 reviews Yang et al.'s scheme, and Section 4 discusses its weakness analysis. We show the scheme details we propose in Section 5. Section 6 shows a formal security proof of the scheme, while Section 7 demonstrates the security analysis of our proposed scheme. In Section 8, we compare our proposed scheme with Yang et al.'s and Zhang et al.'s scheme. Section 9 concludes the paper.

2. Notations and Definitions

Let G_1 be an additive group with an elliptic curve by the generator P and G_2 a multiplicative cyclic group by the generator g. And both of them have the prime order q. Let e denote a computable bilinear map $e : G_1 \times G_1 \mapsto G_2$ satisfying the following three properties [10, 12]:

(i) *Computability.* Given $R, Q \in G_1$, there is an efficient algorithm to compute $e(R, Q)$.

(ii) *Bilinear.* $e(aR, bQ) = e(R, Q)^{ab}$, where $R, Q \in G_1$ and $a, b \in Z_q^*$.

(iii) *Nondegenerate.* Let P be the generator of G_1, $e(P, P) \neq 1_{G_2}$.

Several commonly used notations and their descriptions are described after the Conclusions Section to facilitate the following references.

3. Review of Yang et al.'s Scheme

There are three phases in their scheme; they are, respectively, registration phase, mutual authentication and key agreement phase, and password update phase.

3.1. Registration Phase

3.1.1. User's Registration Section. In this part, a user U requests to be a legal user and the server S conducts the next operations.

U1 ($U \rightarrow S$: $\{\mathrm{ID}_U, \mathrm{PW}D_U\}$). The user U generates his/her own identity ID_U and password PW_U freely. Then, U chooses a nonce N_U randomly and computes

$$\mathrm{PW}D_U = H_2\left(\mathrm{PW}_U \oplus N_U\right). \tag{1}$$

After that, U sends $\{\mathrm{ID}_U, \mathrm{PW}D_U\}$ to the server securely.

U2 ($S \rightarrow U$: $\{S_U, H_2(\cdot)\}$). After obtaining the message sent by U, the server S begins to compute

$$\begin{aligned} K_U &= H_2\left(sH_1\left(\mathrm{ID}_U\right)\right), \\ S_U &= K_U \oplus \mathrm{PW}D_U, \\ T_U &= \mathrm{PW}D_U \oplus K_U. \end{aligned} \tag{2}$$

Then, S stores $\{T_U\}$ in the verification table. Afterward, the server issues a smart card containing $\{S_U, H_2(\cdot)\}$ and transmits it to U through a secure channel.

U3. The N_U is input into the smart card by the U, which contains $\{S_U, N_U, H_2(\cdot)\}$, finally.

3.1.2. Device's Registration Section. In this section, the device V requests to be authorized by the S, and the following steps should be performed together with the server.

D1 ($V \rightarrow S$: $\{\mathrm{ID}_V\}$). The device V transmits its identity $\{\mathrm{ID}_V\}$ to the server S through a secure channel.

D2 ($S \rightarrow V$: $\{S_V\}$). After obtaining the message sent by V, the S begins to compute

$$S_V = sH_1\left(\mathrm{ID}_V\right). \tag{3}$$

And it is sent to V via a secure channel. Afterward, the device's public key and secret key are $P_V = H_1(\mathrm{ID}_V)$ and S_V.

3.2. Mutual Authentication and Key Agreement Phase

M1 ($U \rightarrow V$: $\{M_U\}$). After the U inserts his/her smart card into a smart card reader and inputs his/her identity ID_U and password PW_U, the smart card randomly chooses a secret number r_U and a nonce N'_U. And it computes

$$\begin{aligned} \mathrm{PW}D'_U &= H_2\left(\mathrm{PW}_U \oplus N'_U\right), \\ M_{L1} &= \mathrm{PW}D'_U \oplus S_U, \\ M_{L2} &= H_2\left(S_U \oplus \mathrm{PW}D'_U\right), \end{aligned}$$

$$\begin{aligned} \mathrm{PW}D_U &= H_2\left(\mathrm{PW}_U \oplus N_U\right), \\ K_U &= S_U \oplus \mathrm{PW}D_U. \end{aligned} \tag{4}$$

Next U transmits the message $M_U = \{E_{K_U}(r_U), M_{L1}, M_{L2}\}$ to V.

M2 ($V \rightarrow S$: $\{\mathrm{ID}_V, \mathrm{ID}_U, X, Y, M_U\}$). Upon obtaining the message sent by U, the V chooses a number r_V randomly and computes

$$\begin{aligned} X &= r_V P, \\ H_V &= H_3\left(\mathrm{ID}_V \| \mathrm{ID}_U \| X \| M_U\right), \\ Y &= r_V P_V + H_V S_V. \end{aligned} \tag{5}$$

Then, V transmits the message $\{\mathrm{ID}_V, \mathrm{ID}_U, X, Y, M_U\}$ to S.

M3 ($S \rightarrow V$: $\{\mathrm{ID}_S, \mathrm{ID}_V, \mathrm{ID}_U, M_{S1}, M_{S2}, M_{S3}, M_{S4}, M_{S5}, M_{S6}\}$). When the server receives the message, it computes

$$\begin{aligned} K_U &= H_2\left(sH_1\left(\mathrm{ID}_U\right)\right), \\ \mathrm{PW}D_U &= T_U \oplus K_U, \\ \mathrm{PW}D'_U &= M_{L1} \oplus K_U \oplus \mathrm{PW}D_U \end{aligned} \tag{6}$$

and checks whether M_{L2} is equal to $H_2(K_U \oplus \mathrm{PW}D_U \oplus \mathrm{PW}D'_U)$. If this holds, the S will authenticate U and update T_U with $T'_U = \mathrm{PW}D'_U \oplus K_U$. Otherwise, this authentication request is rejected. Then, S computes

$$\begin{aligned} r'_U &= D_{K_U}\left(E_{K_U}\left(r_U\right)\right), \\ H'_V &= H_3\left(\mathrm{ID}_V \| \mathrm{ID}_U \| X \| M_U\right) \end{aligned} \tag{7}$$

and checks whether

$$e(P, Y) = e\left(P_V, X + H'_V P_S\right). \tag{8}$$

If this holds, the device V is authenticated by S. Otherwise, this authentication request is rejected. After that, the server generates two random strings r_{S1}, r_{S3}, and a random number r_{S2}. S computes

$$\begin{aligned} M_{S1} &= E_{K_U}\left(r_{S1}\right), \\ M_{S2} &= E_{r_{S1}}\left(H_2\left(\mathrm{ID}_S \| \mathrm{ID}_V \| r_{S1}\right) \oplus r'_U\right), \\ M_{S3} &= r_{S2} P, \\ k_{SV} &= H_2\left(r_V r_{S2} P\right), \\ k_{SU} &= H_2\left(r'_U \| r_{S1}\right), \\ M_{S5} &= E_{k_{SV}}\left(r_{S3}\right), \\ M_{S6} &= E_{k_{SU}}\left(r_{S3}\right), \\ h_S &= H_3\left(\mathrm{ID}_S \| \mathrm{ID}_V \| \mathrm{ID}_U \| M_{S1} \| M_{S2} \| M_{S3} \| M_{S5} \| M_{S6}\right), \\ M_{S4} &= r_{S2}\left(P_S + P_U\right) + sh_S P. \end{aligned} \tag{9}$$

At last, the S replies with the message $\{\mathrm{ID}_S, \mathrm{ID}_V, \mathrm{ID}_U, M_{S1}, M_{S2}, M_{S3}, M_{S4}, M_{S5}, M_{S6}\}$ to V.

The correctness of (8) is shown as follows:

$$\begin{aligned} e(P, Y) &= e\left(P, r_V \cdot P_V + H_V \cdot S_V\right) \\ &= e\left(P, r_V \cdot P_V\right) e\left(P, H_V \cdot S_V\right) \\ &= e\left(r_V \cdot P, P_V\right) e\left(sP, H_V \cdot P_V\right) \\ &= e\left(P_V, X\right) e\left(P_V, H_V \cdot P_S\right) \\ &= e\left(P_V, X + H_V \cdot P_S\right) \\ &= e\left(P_V, X + H'_V \cdot P_S\right). \end{aligned} \tag{10}$$

M4 ($V \rightarrow U$: $\{\mathrm{ID}_S, \mathrm{ID}_V, M_{S1}, M_{S2}, M_{S6}, E_{r'_{S3}}(r_{V_1})\}$). After receiving the message from U, V computes

$$h'_S = H_3\left(\mathrm{ID}_S \| \mathrm{ID}_V \| \mathrm{ID}_U \| M_{S1} \| M_{S2} \| M_{S3} \| M_{S5} \| M_{S6}\right) \tag{11}$$

and checks whether

$$e\left(P, M_{S4}\right) = e\left(P_S + P_U, M_{S3}\right) e\left(h'_S P, P_S\right). \tag{12}$$

If this holds, the server S is authenticated by the device V. Otherwise, this authentication procedure fails. Then, V generates a random string r_{V_1} and computes

$$\begin{aligned} k_{SV} &= H_2\left(r_V M_{S3}\right), \\ r'_{S3} &= D_{k_{SV}}\left(M_{S5}\right), \\ & E_{r'_{S3}}\left(r_{V_1}\right). \end{aligned} \tag{13}$$

Finally, V sends the message $\{\mathrm{ID}_S, \mathrm{ID}_V, M_{S1}, M_{S2}, M_{S6}, E_{r'_{S3}}(r_{V_1})\}$ to the user U.

The correctness of (12) is shown as follows:

$$\begin{aligned} e\left(P, M_{S4}\right) &= e\left(P, r_{S2}\left(P_S + P_U\right) + sh_S \cdot P\right) \\ &= e\left(P, r_{S2}\left(P_S + P_U\right)\right) e\left(P, sh_S \cdot P\right) \\ &= e\left(r_{S2} \cdot P, P_S + P_U\right) e\left(sP, h_S \cdot P\right) \\ &= e\left(P_S + P_U, M_{S3}\right) e\left(h_S \cdot P, P_S\right) \\ &= e\left(P_S + P_U, M_{S3}\right) e\left(h'_S \cdot P, P_S\right). \end{aligned} \tag{14}$$

M5 ($U \rightarrow V$: $\{E_{r_{S3}}(r_{U1})\}$). The user U computes

$$\begin{aligned} r'_{S1} &= D_{K_U}\left(M_{S1}\right) \\ S'_U &= S_U \oplus \mathrm{PW}D_U \oplus \mathrm{PW}D'_U \end{aligned} \tag{15}$$

and checks whether $D_{r'_{S1}}(M_{S2}) \oplus r_U$ is equal to $H_2(\mathrm{ID}_S \| \mathrm{ID}_V \| r'_{S1})$. If it holds, the S is authenticated by the U. Otherwise, this authentication procedure fails. Then, U updates N_U and S_U with N'_U and S'_U, which are saved in

the smart card. After that, U generates a random string r_{U1} and computes

$$\begin{aligned} k_{SU} &= H_2\left(r_U \| r_{S1}\right), \\ r_{S3} &= D_{k_{SU}}\left(M_{S6}\right), \\ r_{V_1} &= D_{r_{S3}}\left(E_{r'_{S3}}\left(r_{V_1}\right)\right), \\ & E_{r_{S3}}\left(r_{U1}\right). \end{aligned} \tag{16}$$

Finally, U generates the session key shared with V by computing $k_{UV} = H_2(r_{U1} \| r_{V_1})$ and sends the message $\{E_{r_{S3}}(r_{U1})\}$ back to the device V.

M6. After obtaining the message sent by U, the device V computes

$$\begin{aligned} r_{U1} &= D_{r_{S3}}\left(E_{r_{S3}}\left(r_{U1}\right)\right), \\ k_{UV} &= H_2\left(r_{U1} \| r_{V_1}\right). \end{aligned} \tag{17}$$

The k_{UV} is the session key between U and device V.

3.3. Password Update Phase. When the user requests to change the password PW_U to a new one ($\mathrm{PW}_{\mathrm{new}}$), he/she should perform the next procedures.

P1 ($U \rightarrow S$: $\{M_C\}$). After the U inserts the smart card into a smart card reader and inputs his/her identity ID_U, old password PW_U and new password $\mathrm{PW}_{\mathrm{new}}$. Then the smart card chooses a secret nonce N'_U randomly to compute the following results:

$$\begin{aligned} \mathrm{PW}D'_U &= H_2\left(\mathrm{PW}_U \oplus N'_U\right), \\ M_{C1} &= \mathrm{PW}D'_U \oplus S_U, \\ \mathrm{PW}D_{\mathrm{new}} &= H_2\left(\mathrm{PW}_{\mathrm{new}} \oplus N'_U\right), \\ M_{C2} &= H_2\left(S_U \oplus \mathrm{PW}D'_U \oplus \mathrm{PW}D_{\mathrm{new}}\right), \\ M_{C3} &= \mathrm{PW}D_{\mathrm{new}} \oplus S_U. \end{aligned} \tag{18}$$

Then, U transmits $M_C = \{M_{C1}, M_{C2}, M_{C3}\}$ to S.

P2 ($S \rightarrow U$: $\{M_{C4}\}$). Once obtaining the message sent by U, the S computes

$$\begin{aligned} K_U &= H_2\left(sH_1\left(\mathrm{ID}_U\right)\right), \\ \mathrm{PW}D_U &= T_U \oplus K_U, \\ \mathrm{PW}D'_U &= M_{C1} \oplus K_U \oplus \mathrm{PW}D_U, \\ \mathrm{PW}D_{\mathrm{new}} &= M_{C3} \oplus K_U \oplus \mathrm{PW}D_U, \\ M_{C4} &= H_2\left(\mathrm{PW}D_{\mathrm{new}} \oplus K_U \oplus \mathrm{PW}D_U\right) \end{aligned} \tag{19}$$

and checks whether M_{C2} is equal to $H_2(K_U \oplus \mathrm{PW}D_U \oplus \mathrm{PW}D'_U \oplus \mathrm{PW}D_{\mathrm{new}})$. If it holds, the S will accept the user's request and update the verifier $T'_U = \mathrm{PW}D_{\mathrm{new}} \oplus K_U$. Otherwise, this request procedure is rejected. Then, the server S transmits the message $\{M_{C4}\}$ back to the user U.

P3. Once the user obtains the message sent by the server, he/she checks whether M_{C4} is equal to $H_2(\mathrm{PW}D_{\mathrm{new}} \oplus S_U)$. If this holds, the user replaces N_U and S_U with N'_U and S'_U, where $S'_U = S_U \oplus \mathrm{PW}D_U \oplus \mathrm{PW}D_{\mathrm{new}}$.

4. Cryptanalysis of Yang et al.'s Scheme

Recently, Mishra and Mukhopadhyay demonstrated Yang et al.'s scheme is vulnerable to the password guessing attack and the denial of service attack. Moreover, they also pointed that Yang et al.' scheme does not present efficient login and password change phases such that smart card executes the session in case of incorrect input [13]. Based on the attacks and problems mentioned in Mishra et al.'s article, we analyze Yang et al.'s scheme and find out that the scheme does not resist the session key attack. In addition to this, we also discover that there are some mistakes in their scheme. We will introduce our new discoveries in the following.

4.1. Session Key Attack. If an attacker intercepted the message $\{\mathrm{ID}_S, \mathrm{ID}_V, \mathrm{ID}_U, M_{S1}, M_{S2}, M_{S3}, M_{S4}, M_{S5}, M_{S6}\}$ which was sent from the server to the device and modified some data in it, the user and the device may establish different session key. So the attacker A can realize the session key attack as the following steps.

(1) The attacker A can intercept and capture the message $\{\mathrm{ID}_S, \mathrm{ID}_V, \mathrm{ID}_U, M_{S1}, M_{S2}, M_{S3}, M_{S4}, M_{S5}, M_{S6}\}$ sent to device by the server. Next, A generates a number r_0 randomly and computes

$$\begin{aligned} M_{S3_0} &= r_0 P, \\ h_{S_0} &= H_3\left(\mathrm{ID}_S \| \mathrm{ID}_V \| \mathrm{ID}_U \| M_{S1} \| M_{S2} \| M_{S3_0} \| M_{S5} \| M_{S6}\right), \\ M_{S4_0} &= r_0\left(P_S + P_U\right) + h_{S_0} P_S. \end{aligned} \tag{20}$$

Then, the message $\{\mathrm{ID}_S, \mathrm{ID}_V, \mathrm{ID}_U, M_{S1}, M_{S2}, M_{S3_0}, M_{S4_0}, M_{S5}, M_{S6}\}$ is sent to the device V by the attacker A.

(2) After receiving the message, V computes

$$h'_{S_0} = H_3\left(\mathrm{ID}_S \| \mathrm{ID}_V \| \mathrm{ID}_U \| M_{S1} \| M_{S2} \| M_{S3_0} \| M_{S5} \| M_{S6}\right) \tag{21}$$

and checks the equation

$$e\left(P, M_{S4_0}\right) = e\left(P_S + P_U, M_{S3_0}\right) e\left(h'_{S_0} P, P_S\right). \tag{22}$$

Obviously, they are equal. Then, V generates a random string r_{V_1} and computes

$$\begin{aligned} k_{SV_0} &= H_2\left(r_V M_{S3_0}\right), \\ r_{S3_0} &= D_{k_{SV_0}}\left(M_{S5}\right), \\ & E_{r_{S3_0}}\left(r_{V_1}\right). \end{aligned} \tag{23}$$

Finally, V sends the message $\{\mathrm{ID}_S, \mathrm{ID}_V, M_{S1}, M_{S2}, M_{S6}, E_{r_{S3_0}}(r_{V_1})\}$ to the user U. In this step, the string r_{S3_0} is not equal to the random string r_{S3} generated by the server.

The correctness of (22) is shown as follows:

$$\begin{aligned} e\left(P, M_{S4_0}\right) &= e\left(P, r_0\left(P_S + P_U\right) + h_{S_0} \cdot P_S\right) \\ &= e\left(P, r_0\left(P_S + P_U\right)\right) e\left(P, h_{S_0} \cdot P_S\right) \\ &= e\left(r_0 \cdot P, P_S + P_U\right) e\left(h_{S_0} \cdot P, P_S\right) \\ &= e\left(P_S + P_U, M_{S3_0}\right) e\left(h'_{S_0} \cdot P, P_S\right). \end{aligned} \tag{24}$$

(3) Because A did not modify the message $\{\mathrm{ID}_S, \mathrm{ID}_V, M_{S1}, M_{S2},\}$, $D_{r'_{S1}}(M_{S2}) \oplus r_U$ is equal to $H_2(\mathrm{ID}_S \| \mathrm{ID}_V \| r'_{S1})$. Then, U updates N_U and S_U. After that, U generates a random string r_{U1} and computes

$$\begin{aligned} k_{SU} &= H_2\left(r_U \| r_{S1}\right), \\ r'_{S3} &= D_{k_{SU}}\left(M_{S6}\right), \\ r_{V_{10}} &= D_{r'_{S3}}\left(E_{r_{S3_0}}\left(r_{V_1}\right)\right), \\ & E_{r'_{S3}}\left(r_{U1}\right). \end{aligned} \tag{25}$$

Finally, U generates the session key $k_{UV} = H_2(r_{U1} \| r_{V_{10}})$ shared with the device and sends the message $\{E_{r'_{S3}}(r_{U1})\}$ back to the device V. In this step, the string r'_{S3} is equal to the random string r_{S3} generated by the server and not equal to the string r_{S3_0} computed by the device.

(4) Once obtaining the message sent by the user, the device computes

$$\begin{aligned} r_{U1_0} &= D_{r_{S3_0}}\left(E_{r'_{S3}}\left(r_{U1}\right)\right), \\ k_{UV_0} &= H_2\left(r_{U1_0} \| r_{V_1}\right), \end{aligned} \tag{26}$$

k_{UV_0} is the session key shared between the user U and the device V. Obviously, the session key k_{UV} computed by the user is different from the session key k_{UV_0} computed by the device. So Yang et al.'s scheme suffers from the session key attack.

4.2. Some Mistakes. In mutual authentication and key agreement phase, the identity ID_U has not been sent to the device V. But, when the device computes $H_V = H_3(\mathrm{ID}_V \| \mathrm{ID}_U \| X \| M_U)$, it already knows the user's identity. According to the common sense, if the user has not sent identity to the device, the device cannot obtain the user's identity. So there is a mistake in this phase. What is more, this mistake also exists in the password update phase of Yang et al.'s scheme.

5. Our Proposed Scheme

Based on Yang et al.'s scheme, our protocol also contains four phases: the registration phase, the login phase, the key agreement phase, and the password update phase. Algorithm 1 describes our scheme's registration phase. The login phase and the key agreement phase will be shown in Algorithm 2. At last, we show the password update phase in Algorithm 3. The detail is shown as the following.

5.1. Registration Phase. In our proposed scheme, the registration phase also can be divided into two parts: the user's registration phase and the device's registration phase. Our device's registration phase is the same as the device's registration phase in Yang et al.'s scheme. We will describe our user's registration phase as follows.

R1 ($U \rightarrow S$: $\{\mathrm{ID}_U, \mathrm{PW}D_U, \mathrm{ID}D_U\}$). An identity ID_U and password PW_U are chosen by user U freely. Then, U generates a nonce r_U randomly and computes

$$\begin{aligned} \mathrm{PW}D_U &= H_2\left(\mathrm{PW}_U \oplus r_U\right), \\ \mathrm{ID}D_U &= H_2\left(\mathrm{ID}_U \oplus r_U\right). \end{aligned} \tag{27}$$

After that, U sends $\{\mathrm{ID}_U, \mathrm{PW}D_U, \mathrm{ID}D_U\}$ to the server S via a secure channel.

R2 ($S \rightarrow U$: $\{S_U, P_U, M_U, P, H_2(\cdot), E_K(\cdot)/D_K(\cdot)\}$). After obtaining the message sent by U, S computes

$$\begin{aligned} K_U &= H_2\left(sH_1\left(\mathrm{ID}D_U\right)\right), \\ Q_U &= H_2\left(sH_1\left(\mathrm{PW}D_U\right)\right), \\ S_U &= K_U \oplus \mathrm{PW}D_U \oplus H_2\left(\mathrm{ID}_U\right), \\ M_U &= Q_U \oplus \mathrm{ID}_U, \\ P_U &= H_2\left(K_U \oplus Q_U\right). \end{aligned} \tag{28}$$

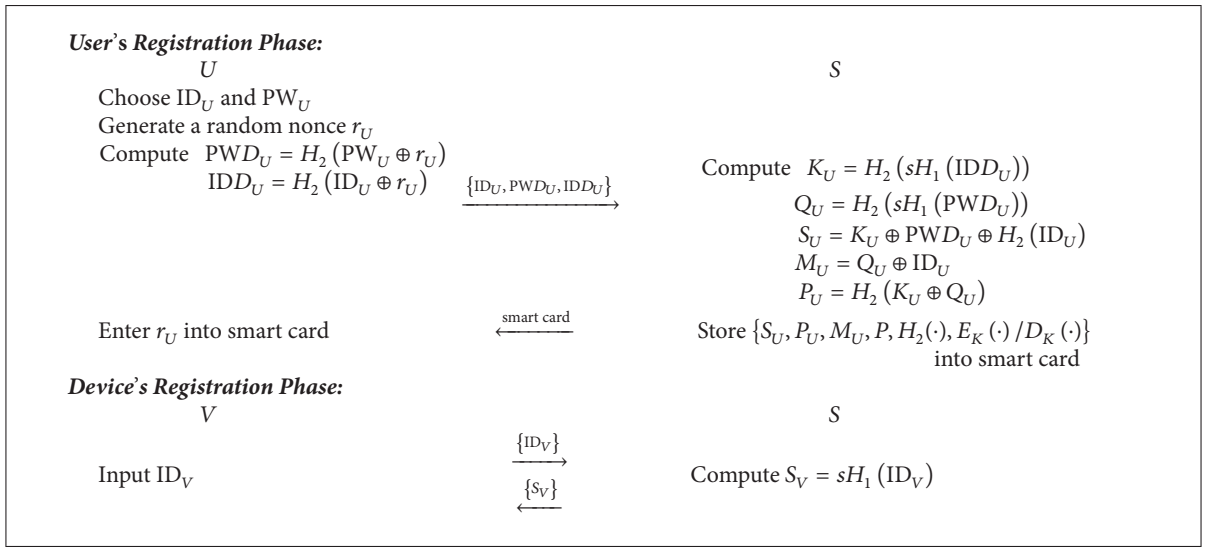
Then, the server U issues a smart card containing $\{S_U, P_U, M_U, P, H_2(\cdot), E_K(\cdot)/D_K(\cdot)\}$ and passes it to U securely.

R3. U enters r_U into the smart card, so it contains $\{r_U, S_U, P_U, M_U, P, H_2(\cdot), E_K(\cdot)/D_K(\cdot)\}$.

5.2. Login Phase. U inserts his/her smart card into a smart card reader and inputs his/her identity ID_U and password PW_U. Then the smart card begins to compute

$$\begin{aligned} \mathrm{PW}D_U &= H_2\left(\mathrm{PW}_U \oplus r_U\right), \\ K'_U &= S_U \oplus \mathrm{PW}D_U \oplus H_2\left(\mathrm{ID}_U\right), \\ Q'_U &= M_U \oplus \mathrm{ID}_U \end{aligned} \tag{29}$$

and checks whether P_U is equal to $H_2(K'_U \oplus Q'_U)$. If this holds, it will authenticate the identity and password of the user. Otherwise, this user's request procedure is rejected.



Algorithm 1: The registration phase.

5.3. Key Agreement Phase. There are six steps and five messages during each run of the proposed protocol. The details are as follows.

A1 ($U \rightarrow V$: $\{M_{U2V_1}\}$). Then, the smart card generates a secret string r_{U1} randomly and computes

$$\begin{aligned} M_{U1} &= H_2\left(K'_U \oplus Q'_U \oplus r_{U1}\right), \\ \mathrm{ID}D_U &= H_2\left(\mathrm{ID}_U \oplus r_U\right). \end{aligned} \tag{30}$$

Next U transmits the message $M_{U2V_1} = \{\mathrm{ID}D_U, E_{K'_U}(r_{U1}), E_{r_{U1}}(\mathrm{PW}D_U), M_{U1}\}$ to V.

A2 ($V \rightarrow S$: $\{M_{V_2S}\}$). Upon obtaining the message sent by U, the V generates a number r_{V_1} randomly and computes

$$\begin{aligned} X &= r_{V_1} \cdot P, \\ H_{V_1} &= H_3\left(\mathrm{ID}_V \| \mathrm{ID}D_U \| X \| M_{U2V_1}\right), \\ Y &= r_{V_1} \cdot P_V + H_{V_1} \cdot S_V. \end{aligned} \tag{31}$$

Then, V transmits the message $M_{V_2S} = \{\mathrm{ID}_V, X, Y, M_{U2V_1}\}$ to the server S.

A3 ($S \rightarrow V$: $\{M_{S2V}\}$). When the server received the message, it computes

$$H'_{V_1} = H_3\left(\mathrm{ID}_V \| \mathrm{ID}D_U \| X \| M_{U2V_1}\right) \tag{32}$$

and checks whether $e(P, Y)$ is equal to $e(P_V, X + H'_{V_1} \cdot P_S)$. If this holds, the device V is authenticated by S. Otherwise, this authentication request is rejected. Then, S computes

$$\begin{aligned} K_U &= H_2\left(sH_1\left(\mathrm{ID}D_U\right)\right), \\ r'_{U1} &= D_{K_U}\left(E_{K'_U}\left(r_{U1}\right)\right), \\ \mathrm{PW}D'_U &= D_{r'_{U1}}\left(E_{r_{U1}}\left(\mathrm{PW}D_U\right)\right), \\ Q_U &= H_2\left(sH_1\left(\mathrm{PW}D'_U\right)\right) \end{aligned} \tag{33}$$

and checks whether M_{U1} is equal to $H_2(K_U \oplus Q_U \oplus r'_{U1})$. If this holds, the user U is authenticated by S. Otherwise, this authentication request is rejected. After that, the server S generates a random string r_{S1}. Then S computes

$$\begin{aligned} k_{SV} &= s \cdot X = \left(k_{SVx}, k_{SVy}\right), \\ S_V &= s \cdot H_1\left(\mathrm{ID}_V\right) = \left(S_{Vx}, S_{Vy}\right), \\ M_{S1} &= k_{SVx} \oplus k_{SVy} \oplus r_{S1}, \\ M_{S2} &= H_2\left(r'_{U1} \| K_U\right) \oplus r_{S1}, \\ M_{S3} &= H_2\left(r'_{U1} \oplus K_U \oplus M_{S2}\right), \\ M_{S4} &= H_2\left(k_{SVx} \| k_{SVy} \| S_{Vx} \| S_{Vy} \| M_{S1} \| M_{S3}\right). \end{aligned} \tag{34}$$

Finally, the S replies with the message $M_{S2V} = \{M_{S1}, M_{S2}, M_{S3}, M_{S4}\}$ to the device V.

A4 ($V \rightarrow U$: $\{M_{V_2U}\}$). After receiving the message, V computes

$$k'_{SV} = r_{V_1} \cdot P_S = \left(k'_{SVx}, k'_{SVy}\right) \tag{35}$$

U	V	S
Login Phase:		
Input ID_U and PW_U		
Compute $\mathrm{PW}D_U = H_2(\mathrm{PW}_U \oplus r_U)$		
$K'_U = S_U \oplus \mathrm{PW}D_U \oplus H_2(\mathrm{ID}_U)$		
$Q'_U = M_U \oplus \mathrm{ID}_U$		
Check $P_U \stackrel{?}{=} H_2(K'_U \oplus Q'_U)$		
Key agreement Phase:		
Generate a random string r_{U1}		
Compute $M_{U1} = H_2(K'_U \oplus Q'_U \oplus r_{U1})$		
$\mathrm{ID}D_U = H_2(\mathrm{ID}_U \oplus r_U)$		
$M_{U2V_1} = \{\mathrm{ID}D_U, E_{K'_U}(r_{U1}), E_{r_{U1}}(\mathrm{PW}D_U), M_{U1}\}$		
	$\xrightarrow{\{M_{U2V_1}\}}$ Generate a random number r_{V_1}	
	Compute $X = r_{V1} \cdot P$	
	$H_{V_1} = H_3(\mathrm{ID}_V \Vert \mathrm{ID}D_U \Vert X \Vert M_{U2V_1})$	
	$Y = r_{V1} \cdot P_V + H_{V_1} \cdot s_V$	
	$M_{V_2S} = \{\mathrm{ID}_V, X, Y, M_{U2V_1}\}$	
		$\xrightarrow{\{M_{V_2S}\}}$ Compute $H'_{V_1} = H_3(\mathrm{ID}_V \Vert \mathrm{ID}D_U \Vert X \Vert M_{U2V_1})$
		Check $e(P,Y) \stackrel{?}{=} e(P_V, X + H'_{V_1} \cdot P_S)$
		Compute $K_U = H_2(sH_1(\mathrm{ID}D_U))$
		$r'_{U_1} = D_{K_U}(E_{K'_U}(r_{U_1}))$
		$\mathrm{PW}D'_U = D_{r'_{U1}}(E_{r_{U1}}(\mathrm{PW}D_U))$
		$Q_U = H_2(sH_1(\mathrm{PW}D'_U))$
		Check $M_{U1} \stackrel{?}{=} H_2(K_U \oplus Q_U \oplus r'_{U1})$
		Generate a random string r_{S1}
		Compute $k_{S_V} = s \cdot X = (k_{S_Vx}, k_{S_Vy})$
		$S_V = s \cdot H_1(\mathrm{ID}_V) = (S_{Vx}, S_{Vy})$
		$M_{S1} = k_{SVx} \oplus k_{SVy} \oplus r_{S1}$
		$M_{S2} = H_2(r'_{U1} \Vert K_U) \oplus r_{S1}$
		$M_{S3} = H_2(r'_{U1} \oplus K_U \oplus M_{S2})$
	$\xleftarrow{\{M_{S2V}\}}$	$M_{S4} = H_2(k_{SVx} \Vert k_{SVy} \Vert S_{Vx} \Vert S_{Vy} \Vert M_{S1} \Vert M_{S3})$
		$M_{S2V} = \{M_{S1}, M_{S2}, M_{S3}, M_{S4}\}$
	Compute $k'_{SV} = r_{V_1} \cdot P_S = (k'_{SVx}, k'_{SVy})$	
	Check $M_{S4} \stackrel{?}{=} H_2(k'_{SVx} \Vert k'_{SVy} \Vert S_{SVx} \Vert S_{SVy} \Vert M_{S1} \Vert M_{S3})$	
	Generate a random number r_{V_2}	
	Compute $r'_{S1} = M_{S1} \oplus k'_{SVx} \oplus k'_{SVy}$	
Check $M_{S3} \stackrel{?}{=} H_2(r_{U1} \oplus K'_U \oplus M_{S2})$ $\xleftarrow{\{M_{S2V}\}}$	$R_V = r_{V_2} \cdot P = (R_{Vx}, R_{Vy})$	
	$H_{V_2} = H_2(R_{Vx} \Vert R_{Vy} \Vert r'_{S1})$	
Compute $r''_{S1} = M_{S2} \oplus H_2(r_{U1} \Vert K'_U)$	$M_{V_2U} = \{R_V, H_{V_2}, M_{S2}, M_{S3}\}$	
Check $H_{V_2} \stackrel{?}{=} H_2(R_{Vx} \Vert R_{Vy} \Vert r''_{S1})$		
Generate a random number r_{U2}		
Compute $R_U = r_{U2} \cdot P = (R_{Ux}, R_{Uy})$		
$H_U = H_2(R_{Ux} \Vert R_{Uy} \Vert r''_{S1})$		
Compute $R_{UV} = r_{U2} \cdot R_V$ $\xrightarrow{\{M_{U2V_2}\}}$	Check $H_U \stackrel{?}{=} H_2(R_{Ux} \Vert R_{Uy} \Vert r'_{S1})$	
$k_{UV} = H_2(R_U \Vert R_V \Vert R_{UV})$	Compute $R'_{UV} = r_{V_2} \cdot R_U$	
$M_{U2V_2} = \{R_U, H_U\}$	$k'_{UV} = H_2(R_U \Vert R_V \Vert R'_{UV})$	

Algorithm 2: The login phase and the key agreement phase.

U		S
Input ID_U, PW_U and $\mathrm{PW}_{\mathrm{new}}$		
Compute $\mathrm{PW}D_U = H_2(\mathrm{PW}_U \oplus r_U)$		
$K'_U = S_U \oplus \mathrm{PW}D_U \oplus H_2(\mathrm{ID}_U)$		
$Q'_U = M_U \oplus \mathrm{ID}_U$		
Check $P_U \stackrel{?}{=} H_2(K'_U \oplus Q'_U)$		
Generate a random string r_{C1}		
Compute $\mathrm{ID}D_U = H_2(\mathrm{ID}_U \oplus r_U)$		
$\mathrm{PW}D_{\mathrm{new}} = H_2(\mathrm{PW}_{\mathrm{new}} \oplus r_U)$		
$M_{UC1} = \mathrm{PW}D_{\mathrm{new}} \oplus S_U \oplus H_2(\mathrm{ID}_U)$		
$M_{UC2} = H_2(K'_U \oplus Q'_U \oplus r_{C1} \oplus \mathrm{PW}D_{\mathrm{new}})$		
$M_{U2S} = \{\mathrm{ID}D_U, E_{K'_U}(r_{C1}), E_{r_{C1}}(\mathrm{PW}D_U), M_{UC1}, M_{UC2}\}$	$\xrightarrow{\{M_{U2V_1}\}}$	
		Compute $K_U = H_2(sH_1(\mathrm{ID}D_U))$
		$r'_{C1} = D_{K_U}(E_{K'_U}(r_{C1}))$
		$\mathrm{PW}D'_U = D_{r'_{C1}}(E_{r_{C1}}(\mathrm{PW}D_U))$
		$Q_U = H_2(sH_1(\mathrm{PW}D'_U))$
		$\mathrm{PW}D'_{\mathrm{new}} = M_{UC1} \oplus K_U \oplus \mathrm{PW}D'_U$
		Check $M_{UC2} \stackrel{?}{=} H_2(K_U \oplus Q_U \oplus r'_{U1} \oplus \mathrm{PW}D'_{\mathrm{new}})$
		Compute $Q_{U_{\mathrm{new}}} = H_2(sH_1(\mathrm{PW}D'_{\mathrm{new}}))$
		$M_{SC1} = Q_{U_{\mathrm{new}}} \oplus H_2(K_U \oplus r'_{C1})$
	$\xleftarrow{\{M_{S2V}\}}$	$M_{SC2} = H_2(Q_{U_{\mathrm{new}}}) \oplus H_2(\mathrm{PW}D'_{\mathrm{new}} \| r'_{C1})$
Compute $Q'_{U_{\mathrm{new}}} = M_{SC1} \oplus H_2(K'_U \oplus r_{C1})$		$M_{S2U} = \{M_{SC1}, M_{SC2}\}$
Check $M_{SC2} \stackrel{?}{=} H_2(Q'_{U_{\mathrm{new}}}) \oplus H_2(\mathrm{PW}D_{\mathrm{new}} \| r_{C1})$		
Compute $S_{U_{\mathrm{new}}} = K'_U \oplus \mathrm{PW}D_{\mathrm{new}} \oplus H_2(\mathrm{ID}_U)$		
$P_{U_{\mathrm{new}}} = H_2(K'_U \oplus Q'_{U_{\mathrm{new}}})$		
$M_{U_{\mathrm{new}}} = Q'_{U_{\mathrm{new}}} \oplus \mathrm{ID}_U$		
Replace S_U, P_U, M_U with $S_{U_{\mathrm{new}}}, P_{U_{\mathrm{new}}}, M_{U_{\mathrm{new}}}$		

Algorithm 3: The password update phase of our scheme.

and checks whether M_{S4} is equal to $H_2(k'_{SVx} \| k'_{SVy} \| S_{Vx} \| S_{Vy} \| M_{S1} \| M_{S3})$. If this holds, the S is authenticated by the device V. Otherwise, this authentication procedure fails. Then, V generates a random number r_{V_2} and computes

$$\begin{aligned} r'_{S1} &= M_{S1} \oplus k'_{SVx} \oplus k'_{SVy}, \\ R_V &= r_{V_2} \cdot P = \left(R_{Vx}, R_{Vy}\right), \\ H_{V_2} &= H_2\left(R_{Vx} \| R_{Vy} \| r'_{S1}\right). \end{aligned} \tag{36}$$

Finally, the device V transmits the message $M_{V_2U} = \{R_V, H_{V_2}, M_{S2}, M_{S3}\}$ to the user U.

A5 ($U \rightarrow V$: $\{M_{U2V_2}\}$). The user U checks whether M_{S3} is equal to $H_2(r_{U1} \oplus K'_U \oplus M_{S2})$. If this holds, the user U authenticates the S. Otherwise, this authentication procedure fails. After that, U computes

$$r''_{S1} = M_{S2} \oplus H_2\left(r_{U1} \| K'_U\right). \tag{37}$$

Then, the user U checks whether H_{V_2} is equal to $H_2(R_{Vx} \| R_{Vy} \| r''_{S1})$. If this holds, the device V is authenticated by the user U. Otherwise, this authentication procedure fails. U generates a random number r_{U2} and computes

$$\begin{aligned} R_U &= r_{U2} \cdot P = \left(R_{Ux}, R_{Uy}\right), \\ H_U &= H_2\left(R_{Ux} \| R_{Uy} \| r''_{S1}\right), \\ R_{UV} &= r_{U2} \cdot R_V, \\ k_{UV} &= H_2\left(R_U \| R_V \| R_{UV}\right), \end{aligned} \tag{38}$$

where k_{UV} is the session key shared between the user U and the device V. Finally, U sends the message $M_{U2V_2} = \{R_U, H_U\}$ back to the device V.

A6. After obtaining the message sent by the user, the device V checks whether H_U is equal to $H_2(R_{Ux} \| R_{Uy} \| r'_{S1})$. If this holds, the user U is authenticated by the device V. Otherwise, this authentication procedure fails. Then, V computes

$$\begin{aligned} R'_{UV} &= r_{V_2} \cdot R_U, \\ k'_{UV} &= H_2\left(R_U \| R_V \| R'_{UV}\right). \end{aligned} \tag{39}$$

The k'_{UV} is the session key between the user U and the device V.

5.4. Password Update Phase. When the user requests to update his/her password PW_U to a new $\mathrm{PW}_{\mathrm{new}}$, he/she should perform the next procedures.

C1 ($U \rightarrow S : \{\mathrm{ID}D_U, E_{K'_U}(r_{C1}), E_{r_{C1}}(\mathrm{PW}D_U), M_{UC1}, M_{UC2}\}$). U inserts his/her smart card into a smart card reader and inputs his/her identity ID_U, old password PW_U, and new password $\mathrm{PW}_{\mathrm{new}}$. Then the smart card checks whether P_U is equal to $H_2(S_U \oplus H_2(\mathrm{PW}_U \oplus r_U) \oplus M_U \oplus \mathrm{ID}_U)$. If this holds, it will authenticate the identity and password of the user. Otherwise, this user's password update request procedure is rejected. After this, the smart card generates a secret string r_{C1} randomly and computes

$$\begin{aligned} K'_U &= S_U \oplus H_2\left(\mathrm{PW}_U \oplus r_U\right) \oplus H_2\left(\mathrm{ID}_U\right), \\ Q'_U &= M_U \oplus \mathrm{ID}_U, \\ \mathrm{ID}D_U &= H_2\left(\mathrm{ID}_U \oplus r_U\right), \\ \mathrm{PW}D_{\mathrm{new}} &= H_2\left(\mathrm{PW}_{\mathrm{new}} \oplus r_U\right), \\ M_{UC1} &= \mathrm{PW}D_{\mathrm{new}} \oplus S_U \oplus H_2\left(\mathrm{ID}_U\right), \\ M_{UC2} &= H_2\left(K'_U \oplus Q'_U \oplus r_{C1} \oplus \mathrm{PW}D_{\mathrm{new}}\right). \end{aligned} \tag{40}$$

Next U transmits the message $\{\mathrm{ID}D_U, E_{K'_U}(r_{C1}), E_{r_{C1}}(\mathrm{PW}D_U), M_{UC1}, M_{UC2}\}$ to the S.

C2 ($S \rightarrow U$: $\{M_{SC1}, M_{SC2}\}$). Once obtaining the message sent by the user, the S computes

$$\begin{aligned} K_U &= H_2\left(sH_1\left(\mathrm{ID}D_U\right)\right), \\ r'_{C1} &= D_{K_U}\left(E_{K'_U}\left(r_{C1}\right)\right), \\ \mathrm{PW}D'_U &= D_{r'_{C1}}\left(E_{r_{C1}}\left(\mathrm{PW}D_U\right)\right), \\ Q_U &= H_2\left(sH_1\left(\mathrm{PW}D'_U\right)\right), \\ \mathrm{PW}D'_{\mathrm{new}} &= M_{UC1} \oplus K_U \oplus \mathrm{PW}D'_U \end{aligned} \tag{41}$$

and checks whether M_{UC2} is equal to $H_2(K_U \oplus Q_U) \oplus r'_{C1} \oplus \mathrm{PW}D'_{\mathrm{new}})$. If this holds, S will accept the user's request. Otherwise, this request procedure is rejected. Then, the server S computes

$$\begin{aligned} Q_{U_{\mathrm{new}}} &= H_2\left(sH_1\left(\mathrm{PW}D'_{\mathrm{new}}\right)\right), \\ M_{SC1} &= Q_{U_{\mathrm{new}}} \oplus H_2\left(K_U \oplus r'_{C1}\right), \\ M_{SC2} &= H_2\left(Q_{U_{\mathrm{new}}}\right) \oplus H_2\left(\mathrm{PW}D'_{\mathrm{new}} \| r'_{C1}\right) \end{aligned} \tag{42}$$

and sends the message $\{M_{SC1}, M_{SC2}\}$ back to the user U.

C3. When the user received the message from the server, he/she computes

$$Q'_{U_{\mathrm{new}}} = M_{SC1} \oplus H_2\left(K'_U \oplus r_{C1}\right) \tag{43}$$

and checks whether M_{SC2} is equal to $H_2(Q'_{U_{\mathrm{new}}}) \oplus H_2(\mathrm{PW}D_{\mathrm{new}} \oplus r_{C1})$. If not, the request is rejected. Otherwise, the user computes

$$\begin{aligned} S_{U_{\mathrm{new}}} &= K'_U \oplus \mathrm{PW}D_{\mathrm{new}} \oplus H_2\left(\mathrm{ID}_U\right), \\ P_{U_{\mathrm{new}}} &= H_2\left(K'_U \oplus Q'_{U_{\mathrm{new}}}\right), \\ M_{U_{\mathrm{new}}} &= Q'_{U_{\mathrm{new}}} \oplus \mathrm{ID}_U \end{aligned} \tag{44}$$

and replaces S_U, P_U, and M_U with $S_{U_{\mathrm{new}}}$, $P_{U_{\mathrm{new}}}$, and $M_{U_{\mathrm{new}}}$, which are all saved in the smart card.

6. Security Model and Proof

In this part, the provable secure method will be employed to prove that our proposed protocol is provable secure in the models in [18].

6.1. Security Model. In 2005, Abdalla et al. proposed a security model AFP05, which is suitable for the three-party authenticated key agreement scenario. It contains two types of participants, such as the client and the trusted server [18]. But there are three types of participants in our proposed protocol, a user, a trusted device, and a trusted server. So we add a query *SendDevice* in our security model. During the execution of the protocol, U and V have many instances, respectively. U_i and V_j denote the ith instance of U and the jth instance of V. There exists one state of accept, reject, and $\perp$ in an oracle. If the oracle gets correct message, it turns the accept state; otherwise, it turns reject. $\perp$ means that no decision has been reached or no result has been returned. The adversary A, which is abstracted as a probabilistic polynomial time Turing Machine, interacts with other participants through a bounded number of queries which model the capabilities of the adversary in an actual attack. The queries are listed as follows.

SendClient(U_i, m). After receiving message m sent by the adversary, the U_i generates a message and outputs as the result of this query. A query *SendClient*(U_i, Start) begins a new key agreement process.

SendDevice(V_j, m). After receiving message m sent by the adversary, the V_j generates a message and output.

SendServer(S, m). After receiving message m sent by the adversary, the S generates a message and output.

Reveal(U_i/V_j). If no session key is defined for instance U_i/V_j or if either U_i/V_j or its partner is asked a *Test* query, the result of this query is the invalid symbol $\perp$. Otherwise the session key generated by the instance U_i/V_j is returned.

Test(U_i/V_j). If no session key is defined for instance U_i/V_j or if either U_i/V_j or its partner is asked a *Reveal* query, the result of this query is the invalid symbol $\perp$. Otherwise, the oracle flips a coin b. If $b = 1$, the session key is output. Otherwise, a value randomly chosen from the distribution space of session key is output.

The queries defined in our improved AFP05 security model can be simulated using the *SendClient*, *SendDevice,* and *SendServer* queries repeatedly if we assume that there is at least one benign adversary which faithfully relays message flows. In our improved AFP05 security model, the notion of freshness is already embedded in the definition of the oracles. A Find-Then-Guess (FTG) model exists in our improved AFP05 security model, in which the semantic security is defined by a game with two phases. In the first phase, the adversary is able to adaptively execute *SendClient*, *SendDevice*, *SendServer*, *Reveal,* and *Test* query. In the second, A executes a single *Test* query and guesses a bit b' for b, where b is selected in the *Test* query. If $b' = b$, the adversary wins the game. Let Succ denote the event that the adversary correctly guesses the bit b, and the advantage of A that attacks the protocol S is defined as

$$\mathrm{Adv}_S^{\mathrm{FTG-3PAKA}}(A) = |2\Pr[\mathrm{Succ}] - 1|. \tag{45}$$

A 3PAKA protocol S is considered semantically secure in FTG model if and only if $\mathrm{Adv}_S^{\mathrm{FTG-3PAKA}}(t,Q) = \max_A\{\mathrm{Adv}_S^{\mathrm{FTG-3PAKA}}(A)\}$ is negligible, where the maximum time executed by all the adversaries with time-complexity at most t and the number of queries at most Q.

6.2. Mathematical Computational Problems [10]. Let p, r be two large prime numbers and $r \mid (p-1)$. Let G be a multiplicative subgroup of Z_p^*, with prime number r order and element g_0 generator.

Computational Diffie-Hellman Problem (CDH). Given $\{g_0, g_0^a, g_0^b \in G\}$ and $a, b \in_R Z_r^*$, it is hard to compute $g^{ab} \bmod p$.

The probabilistic polynomial time Turing Machine denoted as Δ, the probability of which could successfully solve CDH problem in G, is defined as

$$\mathrm{Succ}_G^{\mathrm{CDH}}(\Delta) = \Pr\left[\Delta\left(g^a, g^b\right) = g^{ab} \bmod \ p : a, b \in_R Z_r^*\right]. \tag{46}$$

CDH Assumption. For any probabilistic Turing Machine Δ, the probability of $\mathrm{Succ}_G^{\mathrm{CDH}}(\Delta)$ is negligible.

Elliptic Curve Computational Diffie-Hellman (ECCDH) Assumption. Let G_1 be an additive point group with an elliptic curve by the generator P and aP and bP are elements of G_1. A is able to compute the value of abP only with aP and bP in time t at most

$$\mathrm{Succ}_{G_1}^{\mathrm{ECCDH}}(\Delta, t) = \Pr\left[\Delta(aP, bP) = abP : a, b \in_R Z_q^*\right]. \tag{47}$$

The fact $\mathrm{Succ}_{G_1}^{\mathrm{ECCDH}}(\Delta, t)$ is negligible means that the ECCDH assumption holds.

6.3. Security Proof

Theorem 1. *The hash functions $H_1(\cdot)$, $H_2(\cdot)$, and $H_3(\cdot)$ are modeled as random oracles. Let A be an adversary against our protocol S within time t. We denote that q_1, q_2, and q_3, respectively, represent the number of $H_1(\cdot)$, $H_2(\cdot)$, and $H_3(\cdot)$ oracle queries executed by A, and q_{sc}, q_{sv}, and q_{ss} denote the number of SendClient queries, SendDevice queries, and SendServer queries. Then,*

$$\begin{aligned} Adv_S^{FTG-DRM}(t,Q) \le{}& \frac{q_1^2 + q_3^2 + q_{ss} + 10q_2 + 2q_3}{q-1} \\ &+ \frac{q_2^2 + 2q_{ss} + 4q_{sc} + 4q_{sv}}{2^l} \\ &+ 2q_2 Succ_{P,G_1}^{CDH}\left(t'\right), \end{aligned} \tag{48}$$

where $Q = \{q_1, q_2, q_3, q_{sc}, q_{sv}, q_{ss}\}$ and $t' \le t + (q_{sc} + q_{sv} + q_{ss})T_s$, with T_s denoting the computational time of one scalar multiplication in G_1.

Proof. We define several attack games from Game GM_0 to Game GM_6. For each game GM_i, Succ_i denotes the event that A has successfully guessed the bit b in the test session. The games are listed as follows.

Hash, Reveal, Test Queries.

(i) For a hash query $H(i, *, h)$, $i \in \{1, 2, 3\}$, we proceed as follows:

(a) Rule $H^{(1)}$: If a record $(i, *, h)$ exists on list L_h, h is returned. Otherwise, $h \leftarrow G_1(i = 1)/\{0,1\}^l(i = 2)/Z_q^*(i = 3)$ and h is returned. Record $(i, *, h)$ is added to list L_h. If the adversary directly issues this query, the record $(i, *, h)$ is added to list L_A.

(ii) For a query *Reveal* (U_i/V_j), we proceed as follows:

(a) If no session key is defined for instance U_i/V_j or if either U_i/V_j or its partner is asked a *Test* query, the output is $\perp$. Otherwise the output of this query is k_{UV} which is defined for the instance U_i/V_j.

(iii) For a query Test (U_i/V_j), we proceed as follows:

(a) If no session key is defined for instance U_i/V_j or if either U_i/V_j or its partner is asked a *Reveal* query, the output is $\perp$. Otherwise, the oracle flips a bit b. If $b = 1$, the session key is output. Otherwise, a value randomly chosen from the distribution space of session key is output.

Game GM_0. This is the actual attack game. According to the definition, we have

$$\mathrm{Adv}_P^{\mathrm{FTG-DRM}}(A) = \left|2\Pr\left[\mathrm{Succ}_0\right] - 1\right|. \tag{49}$$

Otherwise, we randomly generate a bit b' if the game aborts or stops without answer from A or A has not finished the game.

Game GM_1. We simulate all the oracles for each query and keep three lists to store the oracles answers. L_h stores answers

of random oracles $H_1(\cdot)$, $H_2(\cdot)$, and $H_3(\cdot)$. L_A is denoted for random oracle query asked by A. L_S is for the transcripts in the channel. We simulate the *SendClient*, *SendDevice*, *SendServer*, *Reveal,* and *Test* queries as in the actual attack. We list hash, *Reveal*, *Test* queries in *Hash, Reveal, Test Queries* and *SendClient*, *SendDevice*, *SendServer* queries in *SendClient*, *SendDevice*, *SendServer Queries*. Obviously, game GM_0 and game GM_1 are indistinguishable. So we have

$$\Pr\left[\mathrm{Succ}_1\right] = \Pr\left[\mathrm{Succ}_0\right]. \tag{50}$$

Game GM_2. In this game, all the oracles simulated are almost the same as in the game GM_1, but here, we avoid some collisions in the transcripts. The hash oracles $H_1(\cdot)$, $H_2(\cdot)$, and $H_3(\cdot)$ may collide with different input values. We use the following rule.

Rule $H^{(2)}$. If this query is directly issued by the adversary, and $(i, *, h)_{i\in\{1,2,3\}} \in L_A$, the game abort; Otherwise, h is returned.

Given that hash value h is selected from a random uniform distribution, the probability of collisions is at most $(q_1^2/2(q-1)) + (q_2^2/2^{l+1}) + (q_3^2/2(q-1))$ [22].

Game GM_2 and game GM_1 are perfectly indistinguishable unless the abovementioned rule causes the game abort. Hence,

$$\left|\Pr\left[\mathrm{Succ}_2\right] - \Pr\left[\mathrm{Succ}_1\right]\right| \le \frac{q_1^2}{2(q-1)} + \frac{q_2^2}{2^{l+1}} + \frac{q_3^2}{2(q-1)}, \tag{51}$$

Game GM_3. The game GM_3 is defined by aborting the executions in which the adversary has obtained a valid authenticator without asking the corresponding hash query by guessing Y, M_{U1}, M_{S4}, M_{S3}, H_{V_2}, or H_U. The following rules are used.

Rule $S1^{(3)}$: $H'_{V_1} \leftarrow H_3(\mathrm{ID}_V \| \mathrm{ID}D_U \| X \| M_{U2V_1})$. We check the equation $e(P, Y) = e(P_V, X + H'_{V_1} \cdot P_S)$. If it does not hold, S terminates without accepting; else we verify whether $\{(\mathrm{ID}D_U, *, *, *), (\mathrm{ID}_V, *, Y, *), (*, *, *, *), (*, *, *, *), (*, *)\} \in L_S$ or $(3, (\mathrm{ID}_V \| \mathrm{ID}D_U \| X \| M_{U2V_1}), H_{V_1}) \in L_A$. If the two tests fail, S rejects the authenticator and terminates without accepting; otherwise, S accepts and continues.

Rule $S2^{(3)}$: $K_U \leftarrow H_2(sH_1(\mathrm{ID}D_U))$, $r'_{U1} \leftarrow D_{K_U}(E_{K'_U}(r_{U1}))$, $\mathrm{PW}D'_U \leftarrow D_{r'_{U1}}(E_{r_{U1}}(\mathrm{PW}D_U))$, $Q_U \leftarrow H_2(sH_1(\mathrm{PW}D'_U))$. We check the equation $H_2(K_U \oplus Q_U \oplus r'_{U1}) = M_{U1}$. If it does not hold, S terminates without accepting; else we verify whether $(2, (K_U \oplus Q_U \oplus r'_{U1}), M_{U1}) \in L_A$ or $\{(\mathrm{ID}D_U, *, *, M_{U1}), (\mathrm{ID}_V, *, *, *), (*, *, *, *), (*, *, *, *), (*, *)\} \in L_S$. If the two tests fail, S rejects the authenticator and terminates without accepting; Otherwise, S accepts and continues.

SendClient, SendDevice, SendServer Queries.

(i) For a query *SendClient*(U_i, Start), we proceed as follows:

(a) Rule $U1^{(1)}$: $r_{U1} \leftarrow \{0,1\}^l$, $M_{U1} \leftarrow H_2(K'_U \oplus Q'_U \oplus r_{U1})$, $\mathrm{ID}D_U \leftarrow H_2(\mathrm{ID}_U \oplus r_U)$ and $M_{U2V_1} = \{\mathrm{ID}D_U, E_{K'_U}(r_{U1}), E_{r_{U1}}(\mathrm{PW}D_U), M_{U1}\}$ is returned. Then, instance U_i proceeds to an expecting state.

(ii) For a query *SendClient*(U_i, M_{V_2U}), we proceed as follows if instance U_i is in an expecting state:

(a) Rule $U2^{(1)}$: We check whether $H_2(r_{U1} \oplus K'_U \oplus M_{S2}) = M_{S3}$. If the equation does not hold, instance U_i terminates without accepting. Otherwise, instance U_i accepts and applies the following rule.

(b) Rule $U3^{(1)}$: $r''_{S1} \leftarrow M_{S2} \oplus H_2(r_{U1} \| K'_U)$, and we check whether $H_2(R_{Vx} \| R_{Vy} \| r''_{S1}) = H_{V_2}$. If the equation does not hold, instance U_i terminates without accepting. Otherwise, instance U_i accepts and applies the following rule.

(c) Rule $U4^{(1)}$: $r_{U2} \leftarrow Z_q^*$, $R_U \leftarrow r_{U2} \cdot P$, $H_U \leftarrow H_2(R_{Ux} \| R_{Uy} \| r''_{S1})$, and instance U_i accepts and applies the following rule.

(d) Rule $U5^{(1)}$: $R_{UV} \leftarrow r_{U2} \cdot R_V$, $k_{UV} \leftarrow H_2(R_U \| R_V \| R_{UV})$, and $M_{U2V_2} = \{R_U, H_U\}$ is returned. Then, instance U_i terminates and record $\{(\mathrm{ID}D_U, E_{K'_U}(r_{U1}), E_{r_{U1}}(\mathrm{PW}D_U), M_{U1}), (\mathrm{ID}_V, X, Y, M_{U2V_1}), (M_{S1}, M_{S2}, M_{S3}, M_{S4}), (R_V, H_{V_2}, M_{S2}, M_{S3}), (R_U, H_U)\}$ is added to list L_S.

(iii) For a query *SendDevice*(V_j, M_{U2V_1}), we proceed as follows:

(a) Rule $V1^{(1)}$: $r_{V_1} \leftarrow Z_q^*$, $X \leftarrow r_{V_1} \cdot P$, $H_{V_1} \leftarrow H_3(\mathrm{ID}_V \| \mathrm{ID}D_U \| X \| M_{U2V_1})$, $Y \leftarrow r_{V_1} \cdot P_V + H_{V_1} \cdot S_V$, and $M_{V_2S} = \{\mathrm{ID}_V, X, Y, M_{U2V_1}\}$ is returned. Then, instance V_j proceeds to an expecting state.

(iv) For a query *SendDevice*(V_j, M_{S2V}), we proceed as follows if instance V_j is in an expecting state:

(a) Rule $V2^{(1)}$: $k'_{SV} \leftarrow r_{V1} \cdot P_S = (k'_{SVx}, k'_{SVy})$, and we check whether $M_{S4} = H_2(k'_{SVx} \| k'_{SVy} \| S_{Vx} \| S_{Vy} \| M_{S1} \| M_{S3})$. If the equation does not hold, instance V_j terminates without accepting. Otherwise, instance V_j accepts and applies the following rule.

(b) Rule $V3^{(1)}$: $r_{V_2} \leftarrow Z_q^*$, $r'_{S1} \leftarrow M_{S1} \oplus k'_{SVx} \oplus k'_{SVy}$, $R_V \leftarrow r_{V_2} \cdot P$, $H_{V_2} \leftarrow H_2(R_{Vx} \| R_{Vy} \| r'_{S1})$, and $M_{V_2U} = \{R_V, H_{V_2}, M_{S2}, M_{S3}\}$ is returned. Then, instance V_j proceeds to an expecting state.

(v) For a query *SendDevice*(V_j, M_{U2V_2}), we proceed as follows if instance V_j is in an expecting state:

(a) Rule $V4^{(1)}$: We check whether $H_2(R_{Ux} \| R_{Uy} \| r'_{S1}) = H_U$. If the equation does not hold, instance V_j terminates without accepting. Otherwise, instance V_j accepts and applies the following rule.

(b) Rule $V5^{(1)}$: $R'_{UV} \leftarrow r_{V_2} \cdot R_U$, $k'_{UV} \leftarrow H_2(R_U \| R_V \| R'_{UV})$, and instance V_j terminates.

(vi) For a query $SendServer(S, M_{V_2S})$, we proceed as follows:

(a) Rule $S1^{(1)}$: $H'_{V_1} \leftarrow H_3(\mathrm{ID}_V \| \mathrm{ID}D_U \| X \| M_{U2V_1})$, and we check whether $e(P, Y) = e(P_V, X + H'_{V_1} \cdot P_S)$. If the equation does not hold, instance S terminates without accepting. Otherwise, instance S accepts and applies the following rule.

(b) Rule $S2^{(1)}$: $K_U \leftarrow H_2(sH_1(\mathrm{ID}D_U))$, $r'_{U1} \leftarrow D_{K_U}(E_{K'_U}(r_{U1}))$, $\mathrm{PW}D'_U \leftarrow D_{r'_{U1}}(E_{r_{U1}}(\mathrm{PW}D_U))$, $Q_U \leftarrow H_2(sH_1(\mathrm{PW}D'_U))$ and we check whether $H_2(K_U \oplus Q_U \oplus r'_{U1}) = M_{U1}$. If the equation does not hold, instance S terminates without accepting. Otherwise, instance S accepts and applies the following rule.

(c) Rule $S3^{(1)}$: $r_{S1} \leftarrow \{0,1\}^l$, $k_{SV} \leftarrow s \cdot X$, $S_V \leftarrow s \cdot H_1(\mathrm{ID}_V)$, $M_{S1} \leftarrow k_{SVx} \oplus k_{SVy} \oplus r_{S1}$, $M_{S2} \leftarrow H_2(r'_{U1} \| K_U) \oplus r_{S1}$, $M_{S3} \leftarrow H_2(r'_{U1} \oplus K_U \oplus M_{S2})$, $M_{S4} \leftarrow H_2(k_{SVx} \| k_{SVy} \| S_{Vx} \| S_{Vy} \| M_{S1} \| M_{S3})$, and $M_{S2V} = \{M_{S1}, M_{S2}, M_{S3}, M_{S4}\}$ is returned. Then, instance S terminates.

Rule $V2^{(3)}$: $k'_{SV} \leftarrow r_{V_1} \cdot P_S = (k'_{SVx}, k'_{SVy})$. We check the equation $M_{S4} = H_2(k'_{SVx} \| k'_{SVy} \| S_{Vx} \| S_{Vy} \| M_{S1} \| M_{S3})$. If it does not hold, V_j terminates without accepting; else we verify whether $(2, (k'_{SVx} \| k'_{SVy} \| S_{Vx} \| S_{Vy} \| M_{S1} \| M_{S3}), M_{S4}) \in L_A$ or $\{(\mathrm{ID}D_U, *, *, *), (\mathrm{ID}_V, *, *, *), (*, *, *, M_{S4}), (*, *, *, *), (*, *)\} \in L_S$. If the two tests fail, V_j rejects the authenticator and terminates without accepting; otherwise, V_j accepts and continues.

Rule $U2^{(3)}$. We check the equation $H_2(r_{U1} \oplus K'_U \oplus M_{S2}) = M_{S3}$. If it does not hold, U_i terminates without accepting; else we verify whether $\{(\mathrm{ID}D_U, *, *, *), (\mathrm{ID}_V, *, *, *), (*, *, M_{S3}, *), (*, *, M_{S2}, *), (*, *)\} \in L_S$ or $(2, (r_{U1} \oplus K'_U \oplus M_{S2}), M_{S3}) \in L_A$. If the two tests fail, U_i rejects the authenticator and terminates without accepting; otherwise, U_i accepts and continues.

Rule $U3^{(3)}$: $r''_{S1} \leftarrow M_{S2} \oplus H_2(r_{U1} \| K'_U)$. We check the equation $H_2(R_{Vx} \| R_{Vy} \| r''_{S1}) = H_{V_2}$. If it does not hold, U_i terminates without accepting; else we verify whether $\{(\mathrm{ID}D_U, *, *, *), (\mathrm{ID}_V, *, *, *), (*, *, *, *), (*, H_{V_2}, *, *), (*, *)\} \in L_S$ or $(2, (R_{Vx} \| R_{Vy} \| r''_{S1}), H_{V2}) \in L_A$. If the two tests fail, U_i rejects the authenticator and terminates without accepting; otherwise, U_i accepts and continues.

Rule $V4^{(3)}$. We check the equation $H_2(R_{Ux} \| R_{Uy} \| r'_{S1}) = H_U$. If it does not hold, V_j terminates without accepting; else we verify whether $\{(\mathrm{ID}D_U, *, *, *), (\mathrm{ID}_V, *, *, *), (*, *, *, *), (*, *, *, *), (*, H_U)\} \in L_S$ or $(2, (R_{Ux} \| R_{Uy} \| r'_{S1}), H_U) \in L_A$. If the two tests fail, V_j rejects the authenticator and terminates without accepting; otherwise, instance V_j accepts and continues.

Game GM_3 and Game GM_2 are almost indistinguishable only if the rules mentioned above cause the instance to reject a valid authenticator. Because the adversary deduces the authenticator without issuing the corresponding hash queries, the probability of guessing is at most $(q_{ss}/2(q-1)) + (q_{ss}/2^l) + (q_{sv}/2^l) + (q_{sc}/2^l) + (q_{sc}/2^l) + (q_{sv}/2^l)$. Hence,

$$\left|\Pr\left[\mathrm{Succ}_3\right] - \Pr\left[\mathrm{Succ}_2\right]\right| \le \frac{q_{ss} + 2q_{sc} + 2q_{sv}}{2^l} + \frac{q_{ss}}{2(q-1)}. \tag{52}$$

Game GM_4. We define game GM_4 by aborting the executions in which the adversary may have obtained valid authenticator $Y, M_{U1}, M_{S4}, M_{S3}, H_{V_2}$, or H_U by guessing the corresponding secret messages and querying the corresponding hash function. We use the following rules.

Rule $S1^{(4)}$: $H'_{V_1} \leftarrow H_3(\mathrm{ID}_V \| \mathrm{ID}D_U \| X \| M_{U2V_1})$. We check the equation $e(P, Y) = e(P_V, X + H'_{V_1} \cdot P_S)$. If it does not hold, S terminates without accepting; else we confirm whether $\{(\mathrm{ID}D_U, *, *, *), (\mathrm{ID}_V, *, Y, *), (*, *, *, *), (*, *, *, *), (*, *)\} \in L_S$. If the test fails, S rejects the authenticator and terminates without accepting; otherwise, we check whether $(3, (\mathrm{ID}_V \| \mathrm{ID}D_U \| X \| M_{U2V_1}), H_{V_1}) \in L_A$. If this is the case, we abort the game.

Rule $S2^{(4)}$: $K_U \leftarrow H_2(sH_1(\mathrm{ID}D_U))$, $r'_{U1} \leftarrow D_{K_U}(E_{K'_U}(r_{U1}))$, $\mathrm{PW}D'_U \leftarrow D_{r'_{U1}}(E_{r_{U1}}(\mathrm{PW}D_U))$, $Q_U \leftarrow H_2(sH_1(\mathrm{PW}D'_U))$. We check the equation $H_2(K_U \oplus Q_U \oplus r'_{U1}) = M_{U1}$. If it does not hold, S terminates without accepting; else we confirm whether $\{(\mathrm{ID}D_U, *, *, M_{U1}), (\mathrm{ID}_V, *, *, *), (*, *, *, *), (*, *, *, *), (*, *)\} \in L_S$. If the test fails, S rejects the authenticator and terminates without accepting; otherwise, we check whether $(2, (K_U \oplus Q_U \oplus r'_{U1}), M_{U1}) \in L_A$. If this is the case, we abort the game.

Rule $V2^{(4)}$: $k'_{SV} \leftarrow r_{V_1} \cdot P_S = (k'_{SVx}, k'_{SVy})$. We check the equation $M_{S4} = H_2(k'_{SVx} \| k'_{SVy} \| S_{Vx} \| S_{Vy} \| M_{S1} \| M_{S3})$. If it does not hold, V_j terminates without accepting; else we confirm whether $\{(\mathrm{ID}D_U, *, *, *), (\mathrm{ID}_V, *, *, *), (*, *, *, M_{S4}), (*, *, *, *), (*, *)\} \in L_S$. If the test fails, V_j rejects the authenticator and terminates without accepting; otherwise, we check whether $(2, (k'_{SVx} \| k'_{SVy} \| S_{Vx} \| S_{Vy} \| M_{S1} \| M_{S3}), M_{S4}) \in L_A$. If this is the case, we abort the game.

Rule $U2^{(4)}$. We check the equation $H_2(r_{U1} \oplus K'_U \oplus M_{S2}) = M_{S3}$. If it does not hold, U_i terminates without accepting; else we confirm whether $\{(\mathrm{ID}D_U, *, *, *), (\mathrm{ID}_V, *, *, *), (*, *, M_{S3}, *), (*, *, M_{S2}, *), (*, *)\} \in L_S$. If the test fails, U_i rejects the authenticator and terminates without accepting; otherwise, we check whether

$(2, (r_{U1} \oplus K'_U \oplus M_{S2}), M_{S3}) \in L_A$. If this is the case, we abort the game.

Rule $U3^{(4)}$: $r''_{S1} \leftarrow M_{S2} \oplus H_2(r_{U1} \| K'_U)$. We check the equation $H_2(R_{Vx} \| R_{Vy} \| r''_{S1}) = H_{V_2}$. If it does not hold, U_i terminates without accepting; else we confirm whether $\{(\mathrm{ID}D_U, *, *, *), (\mathrm{ID}_V, *, *, *), (*, *, *, *), (*, H_{V_2}, *, *), (*, *)\} \in L_S$. If the test fails, U_i rejects the authenticator and terminates without accepting; otherwise, we check whether $(2, (R_{Vx} \| R_{Vy} \| r''_{S1}), H_{V_2}) \in L_A$. If this is the case, we abort the game.

Rule $V4^{(3)}$. We check the equation $H_2(R_{Ux} \| R_{Uy} \| r'_{S1}) \stackrel{?}{=} H_U$. If it does not hold, V_j terminates without accepting; else we confirm whether $\{(\mathrm{ID}D_U, *, *, *), (\mathrm{ID}_V, *, *, *), (*, *, *, *), (*, *, *, *), (*, H_U)\} \in L_S$. If the test fails, V_j rejects the authenticator and terminates without accepting; otherwise, we check whether $(2, (R_{Ux} \| R_{Uy} \| r'_{S1}), H_U) \in L_A$. If this is the case, we abort the game.

Games GM_4 and GM_3 are almost indistinguishable only if the rules mentioned above cause the game to abort. Because the secret messages corresponding to valid authenticators $M_{U1}, Y, M_{S4}, M_{S2}, H_{V_2}$, and H_U are deduced by the adversary, the probability of guessing is at most $(q_3 + q_2 + q_2 + q_2 + q_2 + q_2)/(q - 1)$. Hence,

$$\left|\Pr\left[\mathrm{Succ}_4\right] - \Pr\left[\mathrm{Succ}_3\right]\right| \le \frac{5q_2 + q_3}{q - 1}. \tag{53}$$

Game GM_5. Here we simulate a private hash function $H_4(\cdot) : \{0,1\}^* \mapsto \{0,1\}^l$ and take place of random oracle $H_2(\cdot)$ with $H_4(\cdot)$. Then, we do not use the R_{UV} or R'_{UV} to generate session key k_{UV} or k'_{UV}. As the result, the session key is completely independent of $H_2(\cdot)$ and either R_{UV} or R'_{UV}.

Rule $U5^{(5)}$: $R_{UV} \leftarrow r_{U2} \cdot R_V$, $k_{UV} \leftarrow H_2(R_U \| R_V)$, and $M_{U2V_2} = \{R_U, H_U\}$ is returned. Then, instance U_i terminates and record $\{(\mathrm{ID}D_U, E_{K'_U}(r_{U1}), E_{r_{U1}}(\mathrm{PW}D_U), M_{U1}), (\mathrm{ID}_V, X, Y, M_{U2V_1}), (M_{S1}, M_{S2}, M_{S3}, M_{S4}), (R_V, H_{V_2}, M_{S1}, M_{S2}), (R_U, H_U)\}$ is added to list L_S.

Rule $V5^{(5)}$: $R'_{UV} \leftarrow r_{V_2} \cdot R_U$, $k'_{UV} \leftarrow H_2(R_U \| R_V)$, and instance V_j terminates.

Games GM_5 and GM_4 are almost indistinguishable only if the following event Ask$H5$ occurs: A queries the hash function $H_2(\cdot)$ on $(R_U \| R_V \| R_{UV})$ or on $(R_U \| R_V \| R'_{UV})$. Because the $H_4(\cdot)$ is a private oracle, the probability that A correctly guesses the value of b is $1/2$. Hence,

$$\begin{aligned} \left|\Pr\left[\mathrm{Succ}_5\right] - \Pr\left[\mathrm{Succ}_4\right]\right| &\le \Pr\left[\mathrm{Ask}H5\right], \\ \Pr\left[\mathrm{Succ}_5\right] &= \frac{1}{2}. \end{aligned} \tag{54}$$

Game GM_6. In this game, the random self-reducibility of the CDH problem is used to simulate the executions. Given a CDH instance (A, B), where $A, B \in G_1$, the next rules are listed.

Rule $V3^{(6)}$: $a \leftarrow Z_q^*$, $r'_{S1} \leftarrow M_{S1} \oplus k'_{SVx} \oplus k'_{SVy}$, $R_V \leftarrow a \cdot A$, $H_{V_2} \leftarrow H_2(R_{Vx} \| R_{Vy} \| r'_{S1})$. The message $M_{V_2U} = \{R_V, H_{V_2}, M_{S2}, M_{S3}\}$ is returned. Then, instance V_j proceeds to an expecting state.

Rule $U4^{(6)}$: $b \leftarrow Z_q^*$, $R_U \leftarrow b \cdot B$, $H_U \leftarrow H_2(R_{Ux} \| R_{Uy} \| r''_{S1})$. Instance U_i accepts and proceeds.

By the definition of event Ask$H5$, event Ask$H6$ means that the adversary has issued a query to random oracle on $(R_U \| R_V \| \mathrm{CDH}(R_U, R_V))$. The number of records, such as $(2, *, *)$ in the list L_A, is q_2, and the probability of obtaining the $\mathrm{CDH}(R_U, R_V)$ value from list L_A is $1/q_2$. Hence, the accumulated probability is $(1/q_2) \times \Pr[\mathrm{Ask}H6]$. Besides, the CDH problem is random characteristics of the self-reducibility, and the equation $\mathrm{CDH}(R_U, R_V) = \mathrm{CDH}(A, B)^{ab}$ holds. Thus,

$$\begin{aligned} \Pr\left[\mathrm{Ask}H5\right] &= \Pr\left[\mathrm{Ask}H6\right], \\ \Pr\left[\mathrm{Ask}H6\right] &= q_2 \mathrm{Succ}_{P,G_1}^{\mathrm{CDH}}\left(t'\right). \end{aligned} \tag{55}$$

Finally, we summarize all the relationships and complete the proof. □

From (50) to (53), we could derive

$$\begin{aligned} \left|\Pr\left[\mathrm{Succ}_0\right] - \Pr\left[\mathrm{Succ}_4\right]\right| \le{}& \frac{q_1^2 + q_3^2 + q_{ss} + 10q_2 + 2q_3}{2(q-1)} \\ &+ \frac{q_2^2 + 2q_{ss} + 4q_{sc} + 4q_{sv}}{2^{l+1}}. \end{aligned} \tag{56}$$

From (54) to (55), we obtain

$$\left|\Pr\left[\mathrm{Succ}_4\right] - \frac{1}{2}\right| \le q_2 \mathrm{Succ}_{P,G_1}^{\mathrm{CDH}}\left(t'\right). \tag{57}$$

From (56) and (57), we have

$$\begin{aligned} \left|\Pr\left[\mathrm{Succ}_0\right] - \frac{1}{2}\right| \le{}& \frac{q_1^2 + q_3^2 + q_{ss} + 10q_2 + 2q_3}{2(q-1)} \\ &+ \frac{q_2^2 + 2q_{ss} + 4q_{sc} + 4q_{sv}}{2^{l+1}} \\ &+ q_2 \mathrm{Succ}_{P,G_1}^{\mathrm{CDH}}\left(t'\right). \end{aligned} \tag{58}$$

From (49) and (58), we obtain (48) and the Theorem 1.

7. Security Analysis of Our Scheme

To get over the problems existing in Yang et al.'s scheme, we proposed a provable secure and efficient authentication scheme using smart card based on elliptic curve cryptography. In this part, we will show that the scheme we proposed is secure against various attacks [23, 24].

7.1. Smart Card Loss Attack. The smart card of user U contains $\{r_U, S_U, P_U, M_U, P, H_2(\cdot), E_K(\cdot)/D_K(\cdot)\}$. If the smart card of the user U is stolen by the attacker, he/she could only get the secret data r_U, S_U, P_U, and M_U from it; other data in the smart card is public to all clients. However, he/she does not know the identity ID_U or the password PW_U of U. As a result, he/she cannot use the secret data r_U, S_U, P_U, and M_U to impersonate the user U to pass the authentication of the server and the device. As the result, our scheme can resist the smart card loss attack.

7.2. Denial of Service Attack. In some schemes [10, 25–27], both of the server and user need to update some shared data in their smart card or verifier table after the key agreement phase or the authentication phase. The attacker can eavesdrop, intercept, and modify any transmitted messages on the public channel. And the behavior of the attacker may cause the difference of the shared data between the user and the server. So, these schemes cannot resist the denial of service attack. In our proposed scheme, the user and the server have not needed to update some data in their smart card or verifier table. Thus, the attacker no longer can perform the denial of service attack.

7.3. Efficient Login Phase. To improve the efficiency of our proposed scheme, before the key agreement phase, the smart card checks the correctness of user's identity and password. In the login phase of our scheme, when the user U inputs his/her identity ID_U and password PW_U, the smart card checks the correctness of ID_U and PW_U firstly through the equation $P_U = H_2(S_U \oplus H_2(\mathrm{PW}_U \oplus r_U) \oplus M_U \oplus \mathrm{ID}_U)$. If it does not hold, the smart card reject U's request. Otherwise, it authenticates the legality of the user U and turns to the key agreement phase.

7.4. Session Key Attack. Firstly, the security of the session key in our scheme is based on the computational Diffie-Hellman problem. Secondly, the session key is generated by the random numbers, which are randomly selected by the user and the device, respectively. At last, before computing a session key, both of them must authenticate each other [28]. Based on the reasons mentioned above, the attacker cannot perform the session key attack.

7.5. Insider Attack. Because of without verifier table in this system, the insider cannot acquire any secret data from the server's system. In addition, the insider cannot obtain K_U without the server's private key. And the adversary has no idea to derive the secret data K_U from all messages he/she can achieve. Thus, he/she cannot impersonate a legal user U' to pass the authentication of the server and the device, or the server S' to deceive a legal user. Therefore, the proposed scheme is able to withstand the insider attack.

7.6. Replay Attack. If the U's message M_{U2V_1} is intercepted and resent to the device by the attacker, the message M_{V_2S} is computed and sent to the server by the device. Obviously the user and the device can through the server's certification, and the server responses the message M_{S2V} to the device. After that, the device can authenticate the server and sends the message M_{V_2U} to the attacker. The attacker can acquire nothing from M_{V_2U}. So, he/she cannot send a legality message M_{U2V_2} to the device. The device must not establish a session key with the attacker [29]. In the same way, if the attacker replays the device's message, he/she also cannot pass the user's authentication.

8. Performance Comparisons

In this part, our proposed scheme's performance will be evaluated compared with some other schemes [10, 12]. The comparison is summarized in Table 1. We define seven parameters of time complexity which are adopted in the schemes mentioned above as follows.

(i) T_{H1}: The time complexity of executing a hash function $H_1(\cdot)$.

(ii) T_{H2}: The time complexity of executing a hash function $H_2(\cdot)$.

(iii) T_{H3}: The time complexity of executing a hash function $H_3(\cdot)$.

(iv) T_P: The time complexity of executing a pairings operation of point on elliptic curve.

(v) T_M: The time complexity of executing a scalar multiplication operation of point on elliptic curve.

(vi) T_A: The time complexity of executing an addition operation of point on elliptic curve.

(vii) T_S: The time complexity of executing a symmetric key computation.

A comparison of our proposed scheme and that of Zhang et al. and Yang et al. is summarized in Table 1. It is known to all that one-way hash function is more efficient than the operation of scalar multiplication. Moreover, the pairing operation costs much more than the scalar multiplication operation. The effort of evaluating one pairing operation is approximately three times the effort of evaluating one scalar multiplication operation. Therefore our proposed scheme performs better than Zhang et al.'s scheme and Yang et al.'s scheme. Consequently, our proposed scheme is much more suitable for practical applications.

9. Conclusions

We have analyzed the scheme of Yang et al. and pointed out, except the attacks mentioned in Mishra et al. paper, their scheme suffers from the session key attack and has some mistakes. We propose a new provable secure and efficient digital rights management authentication scheme using smart card based on elliptic curve cryptography to surmount the problems in Yang et al.'s. And we demonstrate that the new scheme is provable secure under the model AFP05 introduced in this paper. Because hash function is used to replace the operations of point on elliptic curve and the symmetric key computation in our scheme, our scheme

Table 1: Performance comparisons.

	Our proposed scheme	Yang et al.'s scheme	Zhang et al.'s scheme
Registration phase			
U	$2T_{H2}$	$1T_{H2}$	—
S	$3T_{H1} + 3T_{H2} + 3T_M$	$2T_{H1} + 1T_{H2} + 2T_M$	$2T_{H1} + 4T_{H2} + 2T_M$
Login phase and key agreement phase			
U	$9T_{H2} + 2T_M + 2T_S$	$6T_{H2} + 6T_S$	$7T_{H2} + 7T_S$
V	$4T_{H2} + 1T_{H3} + 6T_M + 1T_A$	$2T_{H2} + 2T_{H3} + 3T_P + 6T_M + 2T_A + 3T_S$	$3T_{H2} + 2T_{H3} + 3T_P + 6T_M + 2T_A + 3T_S$
S	$3T_{H1} + 6T_{H2} + 1T_{H3} + 1T_P + 5T_M + 1T_A + 2T_S$	$1T_{H1} + 5T_{H2} + 2T_{H3} + 2T_P + 7T_M + 3T_A + 5T_S$	$5T_{H2} + 2T_{H3} + 2T_P + 7T_M + 3T_A + 6T_S$
Password update phase			
U	$9T_{H2} + 2T_S$	$4T_{H2}$	—
S	$3T_{H1} + 7T_{H2} + 3T_M + 2T_S$	$1T_{H1} + 3T_{H2}$	$4T_{H2}$
Total	$9T_{H1} + 40T_{H2} + 2T_{H3} + 1T_P + 19T_M + 2T_A + 4T_S$	$4T_{H1} + 22T_{H2} + 4T_{H3} + 5T_P + 15T_M + 5T_A + 8T_S$	$2T_{H1} + 23T_{H2} + 4T_{H3} + 5T_P + 15T_M + 5T_A + 9T_S$

is more efficient than Yang et al.'s scheme. As a result, our proposed scheme is more suitable for practical applications in ubiquitous computing.

Notations

U: The user
S: The server
V: The device
A: The attacker
ID_U: The user U's identity
s: The secret key of the server S
P_S: The public key of the server S and $P_S = sP$
PW_U: The user U's password
P_U: The public key of the user U and $P_U = sH_1(\mathrm{ID}_U)$
$H_1(\cdot)$: $\{0,1\}^* \mapsto G_1$, a one-way hash function maps an arbitrary length bit string into a member of group G_1.
$H_2(\cdot)$: $\{0,1\}^* \mapsto \{0,1\}^l$, a one-way hash function maps an arbitrary length bit string into a l-bits string.
$H_3(\cdot)$: $\{0,1\}^* \mapsto Z_q^*$, a one-way hash function maps an arbitrary length bit string into a random member in group Z_q^*.
$E_K(\cdot)/D_K(\cdot)$: The symmetric encryption/decryption algorithm using key K
$\oplus$: The bitwise XOR operation
$\|$: String concatenation operation.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors would like to extend their sincere appreciation to the Deanship of Scientific Research at King Saud University for its funding of this research through the Research Group Project no. RGP-VPP-288.

References

[1] M. Weiser, "The computer for the 21st century," *Scientific American*, pp. 94–104, 1991, reprinted in IEEE Pervasive Computing, 2002: 19–25.

[2] B. Alomair and R. Poovendran, "Efficient authentication for mobile and pervasive computing," in *Information and Communications Security*, vol. 6476 of *Lecture Notes in Computer Science*, pp. 186–202, Springer, Berlin, Germany, 2010.

[3] D. H. Seo and I. Y. Lee, "A study on RFID system with secure service availability for ubiquitous computing," *Journal of Information Processing Systems*, vol. 1, no. 1, pp. 96–101, 2005.

[4] T. M. Thanh and M. Iwakiri, "A proposal of digital rights management based on incomplete cryptography using invariant Huffman code length feature," *Multimedia Systems*, vol. 20, no. 2, pp. 127–142, 2014.

[5] S. W. Park and I. Y. Lee, "Anonymous authentication scheme based on NTRU for the protection of payment information in NFC mobile environment," *Journal of Information Processing Systems*, vol. 9, no. 3, pp. 461–476, 2013.

[6] D. Kirovski, M. Peinado, and F. A. P. Petitcolas, "Digital rights management for digital cinema," in *Proceedings of the International Society for Optical Engineering*, pp. 105–120, August 2001.

[7] S. Emmanuel and M. S. Kankanhalli, "A digital rights management scheme for broadcast video," *Multimedia Systems*, vol. 8, no. 6, pp. 444–458, 2003.

[8] H. Chang and M. J. Atallah, "Protecting software code by guards," in *DRM: ACM CCS-8 Workshop on Security and Privacy in Digital Rights Management*, pp. 160–175, Springer, Berlin, Germany, 2002.

[9] A. Seki and W. Kameyama, "A proposal on open DRM system coping with both benefits of rights-holders and users," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '03)*, vol. 7, pp. 4111–4115, December 2003.

[10] H. W. Yang, C. C. Yang, and W. Lin, "Enhanced digital rights management authentication scheme based on smart card," *IET Information Security*, vol. 7, no. 3, pp. 189–194, 2013.

[11] U. J. Jang, H. Lim, and Y. Shin, "A license audit model for secure DRM systems in IP-based environments," *Journal of Information Processing Systems*, vol. 6, no. 2, pp. 253–260, 2010.

[12] Y. C. Zhang, L. Yang, P. Xu, and Y. S. Zhan, "A DRM authentication scheme based on smart-card," in *Proceedings of the International Conference on Computational Intelligence and Security (CIS '09)*, pp. 202–207, Beijing, China, December 2009.

[13] D. Mishra and S. Mukhopadhyay, "Cryptanalysis of Yang et al.'s digital rights management authentication scheme based on smart card," in *Recent Trends in Computer Networks and Distributed Systems Security*, vol. 420 of *Communications in Computer and Information Science*, pp. 288–297, 2014.

[14] C. C. Lee, M. S. Hwang, and W. P. Yang, "A flexible remote user authentication scheme using smart cards," *ACM Operating Systems Review*, vol. 36, no. 3, pp. 45–52, 2002.

[15] J. J. Yuan, "An enhanced two-factor user authentication in wireless sensor networks," *Telecommunication Systems*, vol. 55, no. 1, pp. 105–113, 2014.

[16] J. Petit and Z. Mammeri, "Authentication and consensus overhead in vehicular ad hoc networks," *Telecommunication Systems*, vol. 52, no. 4, pp. 2699–2712, 2013.

[17] H. M. Yang, Y. X. Zhang, and Y. Z. Zhou, "Provably secure three-party authenticated key agreement protocol using smart cards," *Computer Networks*, vol. 58, pp. 29–38, 2014.

[18] M. Abdalla, P. A. Fouque, and D. Pointcheval, "Password based authenticated key exchange in the three-party setting," in *Proceedings of the International Workshop on Practice and Theory in Public Key Cryptography (PKC '05)*, pp. 65–84, 2005.

[19] C. Tsai, C. Lee, and M. Hwang, "Password authentication schemes: current status and key issues," *International Journal of Network Security*, vol. 3, no. 2, pp. 101–115, 2006.

[20] J. W. K. Gnanaraj, K. Ezra, and E. B. Rajsingh, "Smart card based time efficient authentication scheme for global grid computing," *Human-Centric Computing and Information Sciences*, vol. 3, no. 1, pp. 1–14, 2013.

[21] X. Wang, W. Guo, W. Zhang, M. K. Khan, and K. Alghathbar, "Cryptanalysis and improvement on a parallel keyed hash function based on chaotic neural network," *Telecommunication Systems*, vol. 52, no. 2, pp. 515–524, 2013.

[22] J. Katz and Y. Lindell, *Introduction to Modern Cryptography*, Chapman & Hall/CRC Press, 2007.

[23] C. Lee, T. Lin, and R. Chang, "A secure dynamic ID based remote user authentication scheme for multi-server environment using smart cards," *Expert Systems with Applications*, vol. 38, no. 11, pp. 13863–13870, 2011.

[24] C. Lee, C. Chen, P. Wu, and T. Chen, "Three-factor control protocol based on elliptic curve cryptosystem for universal serial bus mass storage devices," *IET Computers & Digital Techniques*, vol. 7, no. 1, pp. 48–55, 2013.

[25] L. Cao and W. Ge, "Analysis and improvement of a multi-factor biometric authentication scheme," *Security and Communication Networks*, 2014.

[26] Y. An, "Security analysis and enhancements of an effective biometric-based remote user authentication scheme using smart cards," *Journal of Biomedicine and Biotechnology*, vol. 2012, Article ID 519723, 6 pages, 2012.

[27] M. K. Khan and J. Zhang, "An efficient and practical fingerprint-based remote user authentication scheme with smart cards," in *Information Security Practice and Experience*, pp. 260–268, Springer, Berlin, Germany, 2006.

[28] T.-T. Truong, M.-T. Tran, and A.-D. Duong, "Improvement of the more efficient & secure ID-based remote mutual authentication with key agreement scheme for mobile devices on ECC," *Journal of Convergence*, vol. 3, no. 2, pp. 19–30, 2012.

[29] Y. Chung, S. Choi, and D. Won, "Lightweight anonymous authentication scheme with unlinkability in global mobility networks," *Journal of Convergence*, vol. 4, no. 4, pp. 23–29, 2013.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 217568, 11 pages
http://dx.doi.org/10.1155/2015/217568

Research Article

Partially Occluded Facial Image Retrieval Based on a Similarity Measurement

Sohee Park,[1] Hansung Lee,[1] Jang-Hee Yoo,[1] Geonwoo Kim,[1] and Soonja Kim[2]

[1]*SW·Content Research Laboratory, ETRI, Daejeon 305-700, Republic of Korea*
[2]*Department of Electronics Engineering, Kyungpook National University, Daegu 702-701, Republic of Korea*

Correspondence should be addressed to Sohee Park; parksh@etri.re.kr

Received 22 September 2014; Revised 31 March 2015; Accepted 30 April 2015

Academic Editor: Aime' Lay-Ekuakille

We present a partially occluded facial image retrieval method based on a similarity measurement for forensic applications. The main novelty of this method compared with other occluded face recognition algorithms is measuring the similarity based on Scale Invariant Feature Transform (SIFT) matching between normal gallery images and occluded probe images. The proposed method consists of four steps: (i) a Self-Quotient Image (SQI) is applied to input images, (ii) Gabor-Local Binary Pattern (Gabor-LBP) histogram features are extracted from the SQI images, (iii) the similarity between two compared images is measured by using the SIFT matching algorithm, and (iv) histogram intersection is performed on the SIFT-based similarity measurement. In experiments, we have successfully evaluated the performance of the proposed method with the commonly used benchmark database, including occluded facial images. The results show that the correct retrieval ratio was 94.07% in sunglasses occlusion and 93.33% in scarf occlusion. As such, the proposed method achieved better performance than other Gabor-LBP histogram-based face recognition algorithms in eyes-hidden occlusion of facial images.

1. Introduction

CCTV cameras have been widely deployed for various purposes such as surveillance, crime investigation, and disaster monitoring in the past decade. Facial images captured by CCTV cameras or mobile devices are utilized as clues in the investigation of criminal suspects in forensic applications. The state-of-the-art commercial face recognition systems have achieved perfect accuracy in matching facial images acquired in controlled scenarios, but face identification and retrieval are still processed by human examiners. They are very inefficient in terms of performance and time, because the investigator has to match the captured facial images one by one with the criminal face database in order to find the criminal suspect. Therefore an automated facial image retrieval system has become a recent research issue for forensic application [1–3]. Facial images in the real world have many challenging factors such as facial expression, illumination, pose, aging, and occlusion. These factors can decrease the true acceptance rate from 99% to below 60%, so the topic of face recognition algorithms is studied in order to overcome the aforementioned drawbacks [4]. Furthermore, one of the most significant issues in the surveillance environment where people are not cooperating with the system is occlusion, which involves both intentional occlusion such as a hat, sunglasses, or scarf and natural occlusion such as a beard, hand, or makeup. Figure 1 shows examples of partially occluded facial images from the benchmark database. The occlusion causes falling-off of the performance of face recognition, because it distorts the appearance of the face and reduces the information to represent the facial image, as shown in the figure.

Face recognition algorithms can be categorized as holistic and local feature-based approaches. The well-known conventional holistic approaches including Principal Component Analysis (PCA) [5], Linear Discriminant Analysis (LDA) [6], and Independent Component Analysis (ICA) [7] have a major weakness, which is sensitivity to the loss of the distorted area, because these kinds of algorithm utilize the whole face information [8, 9]. An alternative method is local feature-based approaches that are more robust to occlusion than the holistic approaches. Most face recognition

(a) (b)

Figure 1: Examples of partially occluded facial images from the FDDB [38]: (a) intentional occlusion and (b) natural occlusion.

algorithms dealing with occlusion are based on the local features, such as LBP [10], Modified Census Transform (MCT) [11], and Local Gradient Pattern (LGP) [12]. Local feature-based face recognition for occluded facial images has been studied in [13–16], in which the matching scores of local images are fused or voted for improving recognition. Kim et al. [17] have proposed the effective part-based locally salient features for face recognition, which is robust to local distortion and partial occlusion. Occluded face recognition for single training image per person has been presented by Tan et al. [18]. Occlusion-invariant face recognition using selective local nonnegative matrix factorization (S-LNMF) based images has been proposed by Oh et al. [19]. PCA is used in order to detect occlusion, and S-LNMF is used for face recognition. PCA-based face recognition algorithms for handling occlusion are compared in [20]. The face recognition approach based on a Support Vector Machine (SVM) [21] has been studied in [22], in which SVM is used to extract local features invariant to partial occlusion. Sharma et al. [23] have proposed an efficient partially occluded face recognition system, which makes use of Eigen faces and Gabor wavelet filters. Min et al. [24] have presented an occlusion detection algorithm using Gabor wavelet, PCA, and SVM, while recognition of the nonoccluded facial part is performed using LBP. Most occluded face recognition algorithms have employed statistical approaches to deal with the occluded region of the facial image. This is a complicated mechanism to detect the occluded region and acquire information on the distorted region of the face. In addition, the learning mechanism is difficult to apply in applications, because the face database for forensics is large scale. Also, learning and updating is a time-consuming process.

In this paper, a facial image retrieval method for partially occluded facial images, which focuses on the similarity measurement based on SIFT matching [25], is proposed. The goal of the proposed method is to provide a simple and nonstatistical approach that is able to detect the occluded region of probe images, to find the same subregions between original gallery images and partially occluded probe images, and to compensate the face misalignment. The similarity measurement is performed for detection of the occluded part in a probe facial image and selection of the same subregions in two compared images. The preprocessing method is adopted to compensate illumination changes and extract more reliable SIFT key-points. The proposed method also employs a Gabor-LBP histogram for extraction of representative face features [26]. The Gabor-LBP histogram based on a nonstatistical approach is independent of the characteristics of the training database. Therefore it is very suitable for facial image retrieval for real applications. Histogram intersection is performed only in the selected subregions, which contain SIFT-matched key-points for facial image retrieval. To validate the performance of the proposed method, the experiments are conducted using a benchmark database. The experimental results show that the proposed method improves the face retrieval rate, particularly with occlusion by sunglasses, and it is feasible for forensic application in terms of accuracy and simplicity. The rest of this paper is organized as follows. Section 2 presents the proposed facial image retrieval method. Experimental results are described to evaluate the proposed method on the benchmark database in Section 3. Finally, Section 4 concludes this paper, with remarks on future work.

2. Methodology

The proposed method takes into account the fact that SIFT matching [25] can be used to detect the occluded part caused by any kind of obstacle (sunglasses, scarf, hair, hand, etc.). It measures the nonmetric partial similarity between two images (a gallery image and a probe image) using the SIFT matching algorithm. This partial similarity can help extract the intrapersonal features for face retrieval. The overall architecture is described in Figure 2.

2.1. SQI Preprocessing. The illumination influences gray intensity, and a change of lighting makes it difficult to recognize the face precisely. Therefore the preprocessing step can be improved to rectify the illumination for face

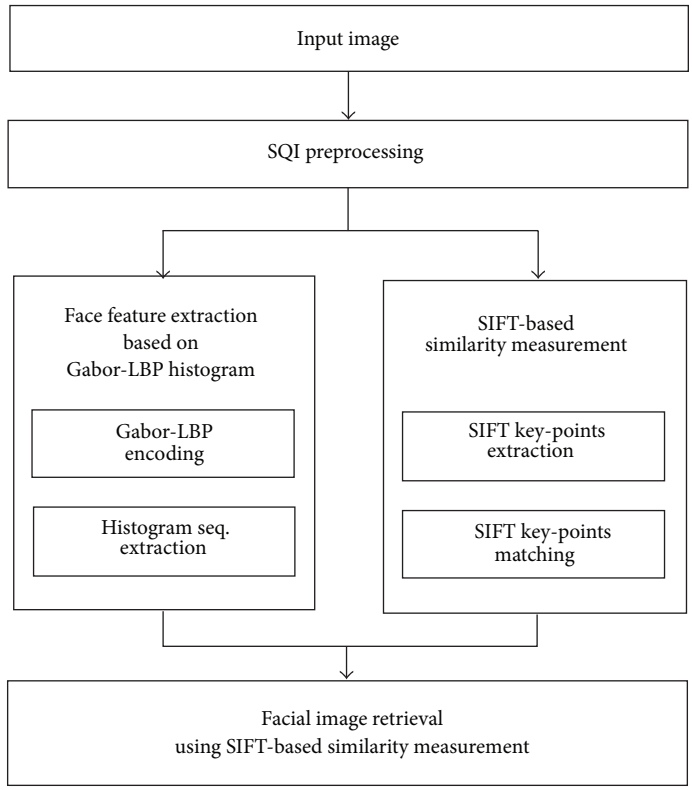


Figure 2: The overall architecture of the proposed facial image retrieval method.

recognition. Here, SQI as a type of high pass filter is adopted in order to be robust against illumination variance [27]. In SQI, Q is defined as

$$Q = \frac{I}{\widehat{I}} = \frac{I}{F * I}, \tag{1}$$

where $\widehat{I}$ is the low frequency image of an input image I, F is the Gaussian kernel, and $*$ denotes the convolution operation. The Gaussian kernel F is calculated by multiplication of weight W and Gaussian filter G as

$$\begin{aligned} F(i,j) &= W(i,j) \cdot G(i,j) \\ W(i,j) &= \begin{cases} 0, & I(i,j) < \tau \\ 1, & \text{otherwise} \end{cases} \\ \tau &= \text{mean}(I_\Omega), \end{aligned} \tag{2}$$

where $I(i, j)$ is the intensity of a pixel (i, j), τ is the mean value of intensity of the filtering region, and Ω is the kernel size. $G(i, j)$ denotes a Gaussian function with standard deviation. Figure 3 shows illumination-rectified images applied by the SQI operation to original images. As shown in Figure 3(b), SQI rectifies the illumination variation and retains the face features.

2.2. Face Feature Extraction. Gabor filters are adopted to extract face features since they have shown the decomposition power to extract representative information of facial images for face recognition in [28–31]. Gabor filters are defined as [32]

$$G_{\mu,\nu}(z) = \frac{\left\| k_{\mu,\nu} \right\|^2}{\sigma^2} e^{\left(-\|k_{\mu,\nu}\|^2 \|z\|^2 / 2\sigma^2\right)} \left[e^{i k_{\mu,\nu} z} - e^{-\sigma^2/2} \right], \tag{3}$$

where $z = (x, y)$ is the pixel and $\| \cdot \|$ denotes the norm operator. μ and ν are the orientation and scale of the Gabor filter and $k_{\mu,\nu}$ is

$$k_{\mu,\nu} = k_\nu e^{i\phi_\mu} \quad \text{with } k_\nu = \frac{k_{\max}}{\lambda^\nu},\ \phi_\mu = \mu \cdot \frac{\pi}{8}, \tag{4}$$

where k_ν and ϕ_μ define the orientation and scale of the Gabor wavelets, $k_{\max}$ is the maximum frequency, and λ is the spacing factor between kernels in the frequency domain. Gabor filters are used with five scales ($\nu = 0, 1, \ldots, 4$) and eight orientations ($\mu = 1, 2, \ldots, 8$) with $\sigma = 5$ in this study. σ is the Gaussian distribution parameter and determines the kernel size of the filter. This means that the kernel size of the Gabor filter used is 5×5 pixels to extract the distinctive contours of facial images [33].

The Gabor-LBP histogram sequence is constructed to be robust against local variance. Each Gabor-LBP image is divided into multiple nonoverlapping subregions, for example, 6×6 and 10×10, and Gabor-LBP histograms in each region are extracted. Then the Gabor-LBP histograms are concatenated into a single feature histogram sequence in

(a) (b)

Figure 3: Examples of (a) original images and (b) SQI images.

order to represent the facial image [34]. The histogram h of a subregion of gray image $f(x, y)$ is formulated as

$$h_i = \sum_{x,y} I\{f(x, y) = i\}, \quad i = 0, 1, \ldots, 255, \tag{5}$$

where i is the gray value (0~255), h_i is the number of pixels with gray value i, and I is defined as

$$I\{f(x, y) = i\} = \begin{cases} 1, & \text{if } f(x, y) = i \text{ is true} \\ 0, & \text{if } f(x, y) = i \text{ is false.} \end{cases} \tag{6}$$

The Gabor-LBP histogram sequence of a single facial image can be calculated by concatenating each subregion histogram as

$$H_s = (h_0, h_1, h_2, \ldots, h_{m-1}), \tag{7}$$

where H_s is the Gabor-LBP histogram sequence of a single facial image and h_i is the histogram of the ith subregion. m is the number of subregions. Finally, the Gabor-LBP histogram sequence $\mathscr{H}_{\text{GLBP}}$ as a face representation is acquired by concatenating each Gabor-LBP histogram sequence which obtained the Gabor filtered facial images as

$$\mathscr{H}_{\text{GLBP}} = (H_0, H_1, \ldots, H_i, \ldots, H_{39}), \tag{8}$$

where H_i is the ith Gabor-LBP histogram sequence and $i = 0, 1, \ldots, 39$. Histogram intersection is used to measure the similarity between the gallery and probe images:

$$D(H_G, H_P) = \sum_{i=0}^{m-1} \min(H_G(i), H_P(i)), \tag{9}$$

where H_G and H_P denote Gabor-LBP histograms of a gallery image and a probe image, respectively, i represents the ith subregion, and m is the number of subregions.

2.3. SIFT-Based Similarity Measurement. To handle the occlusion for face retrieval, it is necessary to detect the occluded facial part and measure the similarity in the nonoccluded region between two compared images. SIFT key-points represent maxima or minima of the difference-of-Gaussian function in the scale-space. Let $I(x, y)$ be an image. $G(x, y, \sigma)$ denotes a variable-scale Gaussian function with standard deviation σ. The scale-space of an image is defined as

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \tag{10}$$

where $*$ denotes the convolution and

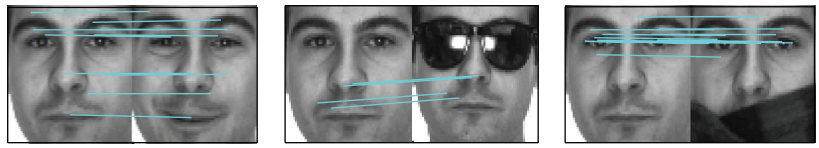
$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-((x^2+y^2)/2\sigma^2)}. \tag{11}$$

The difference-of-Gaussian function $D(x, y, \sigma)$ is formulated as

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma), \tag{12}$$

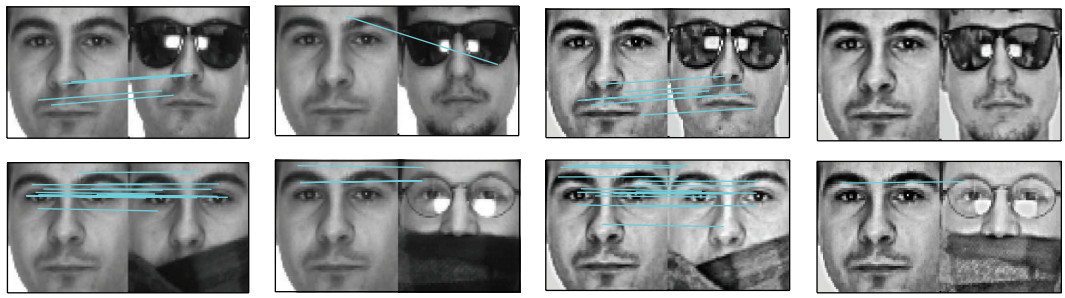
where k is a separation factor of two scales. Local maxima and minima of $D(x, y, \sigma)$ are obtained by comparison of the sample point, its eight neighboring points in the current scale image, and the nine neighboring points in the above scale and below scale images. If the pixel (x, y) is a local maximum or minimum, then it is selected as a key-point.

For SIFT key-point matching, each key-point descriptor of a probe image is matched independently against all key-point descriptors of a gallery image. If the distance ratio of two key-point descriptors is below the specific threshold, then two key-points are matched. Otherwise, the match is rejected, and the key-point is removed. The SIFT matching ability for face recognition is shown in Figure 4, which represents the matching results between a normal face and various types of facial images—facial expression and occlusion. As shown in Figures 4(b) and 4(c), the SIFT matching shows good performance over the occlusion (sunglasses and scarf). It has the good characteristics of detecting occluded regions of facial images. SIFT matching can be employed to measure the similarity between a normal facial image and an occluded facial image in the proposed method. In addition, SQI



(a) (b) (c)

Figure 4: Examples of SIFT matching for face recognition: SIFT matching between (a) a normal face and an expressive face, (b) a normal face and a face occluded by sunglasses, and (c) a normal face and a face occluded by a scarf.

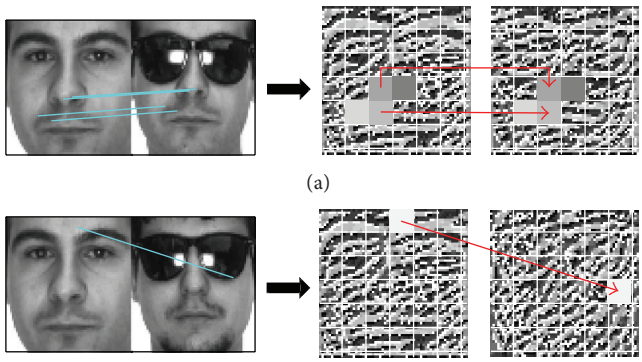


(a) (b)

Figure 5: Performance comparison of SIFT matching: (a) SIFT matching in the original images and (b) SIFT matching in SQI images.

preprocessing can be used to obtain more trustworthy SIFT key-points than original images. This is the most important property to acquire reliable key-points for matching in the SIFT algorithm. In particular, the occluded facial image does not have sufficient region, from which it is possible to extract key-points. SQI preprocessing is performed for SIFT key-point extraction in addition to illumination compensation. Figure 5 shows the performance comparison of SIFT key-point extraction between original images and SQI images. The SIFT matching of SQI images is more precise than that of the original images in imposter matching.

2.4. Facial Image Retrieval Using SIFT-Based Similarity. SIFT-based similarity measurement is necessary not only to detect the occluded region but also to compensate the error that occurs in the process of face alignment. If the facial image is misaligned, the face information in the same subregion is not necessarily the same face information even if in the same position. As shown in Figures 4 and 5, the measured similarity using SIFT represents nonoccluded regions in the face and the corresponding positions between a gallery image and a probe image. After SIFT-based similarity measurement, facial image retrieval performs a selected histogram intersection in a gallery image and a probe image. The selected histogram is a subregion histogram, which includes matched SIFT key-points between two compared images. The facial image retrieval method compares histograms not in the same numbered subregion but in a subregion including a matched key-point. Figure 6 shows a selected histogram intersection using SIFT-based similarity measurement in detail. In Figure 6(a), SIFT matching (genuine face matching) of two facial images has extracted 4 matched key-points. The 4 matched key-points are included in the specific subregion,



(a)

(b)

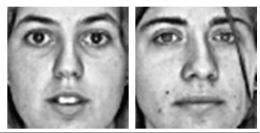
Figure 6: Face matching using selected histogram intersection: facial image retrieval for (a) genuine matching and (b) imposter matching.

and the specific subregions become selected subregions. The histogram of each selected subregion is intersected with the histogram of the corresponding subregion in order to measure the distance. Then other histogram intersections of the remaining selected subregions perform iteratively. Figure 6(b) shows imposter face matching. In the case of imposter matching, the number of matched key-points is smaller than in genuine matching. Moreover, the selected subregions are not the same numbered subregions in the facial images. Therefore the final similarity of the imposter is less than that with genuine matching.

The score matching function of the proposed facial image retrieval is

$$D\left(H_G, H_p\right) = \sum_{k=1}^{n} \min\left(H_G\left(R_{G,k}\right), H_P\left(R_{P,k}\right)\right), \quad (13)$$

Table 1: Examples of gallery and probe images in the test dataset.

Subject	Gallery images	Probe images	
	Neutral	Sunglasses	Scarf
Male			
Female			

where $H_G(R_{G,k})$ and $H_P(R_{P,k})$ are the histograms of $R_{G,k}$ and $R_{P,k}$ subregions, respectively. $R_{G,k}$ and $R_{P,k}$ are the selected subregions including SIFT-matched key-points of a gallery image and a probe image, respectively. k is the ith matched key-point, n is the number of matched key-points, and

$$R_{G,k} = g, \quad \text{if } \left(x_{g,k}, y_{g,k}\right) \in R_g\left(x, y\right), \\ R_{P,k} = p, \quad \text{if } \left(x_{p,k}, y_{p,k}\right) \in R_p\left(x, y\right), \tag{14}$$

where $R_{G,k}$ and $R_{P,k}$ are the gth and pth subregions including the kth matched key-point in a gallery image and a probe image, respectively, $(x_{g,k}, y_{g,k})$ are the kth matched key-point coordinates in a gallery image, and $(x_{p,k}, y_{p,k})$ are the kth matched key-point coordinates in a probe image. The score matching approach in this study uses histogram intersection of selected subregions. The face features extraction and matching procedures which involve Gabor-LBP histogram sequence extraction and SIFT-based similarity measurement have no statistical and learning stages. Thus it can be simpler and faster than the conventional Gabor-LBP histogram algorithm.

3. Experimental Results

3.1. Database. The performance of the proposed method has been evaluated on the AR face database [37]. The AR face database is a commonly used benchmark database for face recognition and especially occluded face recognition. It consists of a total of 3,510 images of 135 different subjects (76 males and 59 females), and it can be divided into facial expressions (neutral, smile, angry, and scream), various types of illumination (right, left, and full), and an occlusion (sunglasses and scarf) section. There are no restrictions on glasses, hairstyle, moustache, and beard. The test dataset used in this experiment can be summarized as shown in Table 1. The normalized facial image which is obtained in our experiments is a gray-scale image of 64 × 64 pixels, and the distance between the eye centers is 32 pixels. The SQI image is made from a gray-scale normalized facial image by preprocessing. The gallery images are 135 neutral images with no illumination, and the probe images are 270 occluded images with no illumination, consisting of 135 sunglasses and 135 scarf images.

3.2. Experiment 1: Preprocessing and Number of Subregions. The first experiment was conducted to validate the performance, according to the preprocessing and the number of subregions, in the conventional Gabor-LBP histogram approach [26]. The x-axis of the graph represents the rank, which means that the subject is searched at the nth rank. Rank-1 implies that the subject is found at the first rank, and the matching score is the highest in the test dataset. The y-axis of the graph is the retrieval rate, which sums up the matching score of each subject at the nth rank. The images in the experiment are used without preprocessing images and SQI images. Face features were extracted using Gabor-LBP histograms, and the facial image retrieval was performed in the sunglasses and scarf datasets, respectively. The other parameter in this experiment was the number of subregions. Gabor-LBP histograms were extracted in each small region so as to be robust against local variance. $n \times m$ subregions represent that a facial image is divided into n subregions of row by m subregions of column. 10 × 10 and 6×6 subregions were tested in the experiment. Figure 7 shows that the performance of SQI images is better than the original images and that 6 × 6 subregions achieve better performance than 10×10 subregions. This shows that local variance affects the performance of face retrieval and that preprocessing and the number of subregions are important factors in Gabor-LBP-based face recognition algorithms.

3.3. Experiment 2: Comparison with Other Face Recognition Algorithms. In the case of sunglasses occlusion, the performance of the proposed method was compared with other algorithms as described in Table 2. The Gabor-LBP histogram achieved 75.56% accuracy at Rank-1 in the 6 × 6 and 10 × 10 subregions, and the SIFT total weight and SIFT probe weight achieved lower accuracy than the original Gabor-LBP histogram. Analysis shows that the Gabor-LBP histogram and SIFT total weight use the occlusion region of the face, and the occlusion region causes noise in the recognition. The SIFT probe weight is reflected on the occlusion using SIFT key-points, but the similarity between the gallery and probe images is not considered.

The proposed method achieved a retrieval rate at Rank-1 of 77.78% and 92.60% in the 6 × 6 and 10 × 10 subregions, respectively. The performance was about 20% higher than other algorithms [26, 35, 36]. The performance of the

Table 2: Performance comparison of the proposed method for partially occluded facial image retrieval in the case of occlusion by sunglasses.

Algorithm	Image format used in SIFT matching	Number of subregions	Retrieval rate
Gabor-LBP [26]	Without SQI	6 × 6	75.56%
		10 × 10	75.56%
	With SQI	6 × 6	75.56%
		10 × 10	75.56%
SIFT total weight [35]	Without SQI	6 × 6	68.89%
		10 × 10	45.93%
	With SQI	6 × 6	77.04%
		10 × 10	71.11%
SIFT probe weight [36]	Without SQI	6 × 6	59.26%
		10 × 10	45.93%
	With SQI	6 × 6	64.44%
		10 × 10	53.33%
Proposed method	Without SQI	6 × 6	77.78%
		10 × 10	92.60%
	With SQI	6 × 6	**94.07%**
		10 × 10	92.59%

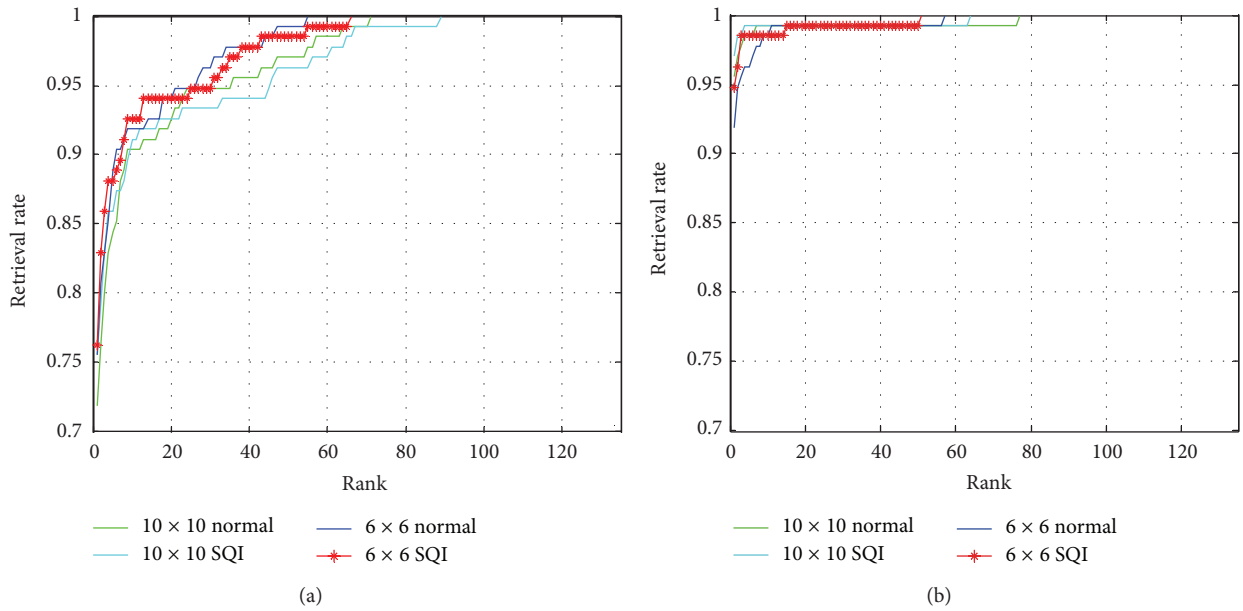


Figure 7: Performance comparison of image format (original images and SQI images) and the number of subregions (6 × 6 and 10 × 10) in cases of (a) sunglasses and (b) scarf.

proposed method in the case of SQI images used in SIFT matching achieved greater accuracy than original images. The proposed method with SQI images achieved 94.07% and 92.59% at Rank-1 in the 6 × 6 and 10 × 10 subregions. Other algorithms showed similar results—that SQI images for SIFT matching improved performance. This validates that SQI images have the ability to extract more reliable SIFT key-points than original images. In addition, 10 × 10 subregions give more precise face-representative information than 6 × 6 subregions in the case of original images for SIFT matching. However, the experiment for SQI images used in SIFT matching showed the opposite result. It is assumed that if it is able to extract more precise and reliable key-points, the histogram of more neighboring points of each key-point is more robust against local variance. The performance comparison of sunglasses occlusion with 6×6 subregions and 10 × 10 subregions in the case of SQI images used in SIFT matching is shown in Figure 8.

Table 3 describes the performance comparison in case of scarf occlusion. The performance of the Gabor-LBP histogram achieved 97.94% at Rank-1 when we used SQI images in SIFT matching and 10 × 10 subregions. This is the best performance in this experiment. The extensions of the Gabor-LBP histogram achieved from 89% to 95% approximately. The

Table 3: Performance comparison of the proposed method for partially occluded facial image retrieval in the case of occlusion by a scarf.

Algorithm	Image format used in SIFT matching	Number of subregions	Retrieval rate
Gabor-LBP [26]	Without SQI	6 × 6	91.85%
		10 × 10	97.04%
	With SQI	6 × 6	91.85%
		10 × 10	**97.94%**
SIFT total weight [35]	Without SQI	6 × 6	94.07%
		10 × 10	94.07%
	With SQI	6 × 6	92.60%
		10 × 10	95.56%
SIFT probe weight [36]	Without SQI	6 × 6	94.07%
		10 × 10	94.07%
	With SQI	6 × 6	93.33%
		10 × 10	88.89%
Proposed method	Without SQI	6 × 6	87.41%
		10 × 10	91.11%
	With SQI	6 × 6	**93.33%**
		10 × 10	91.11%

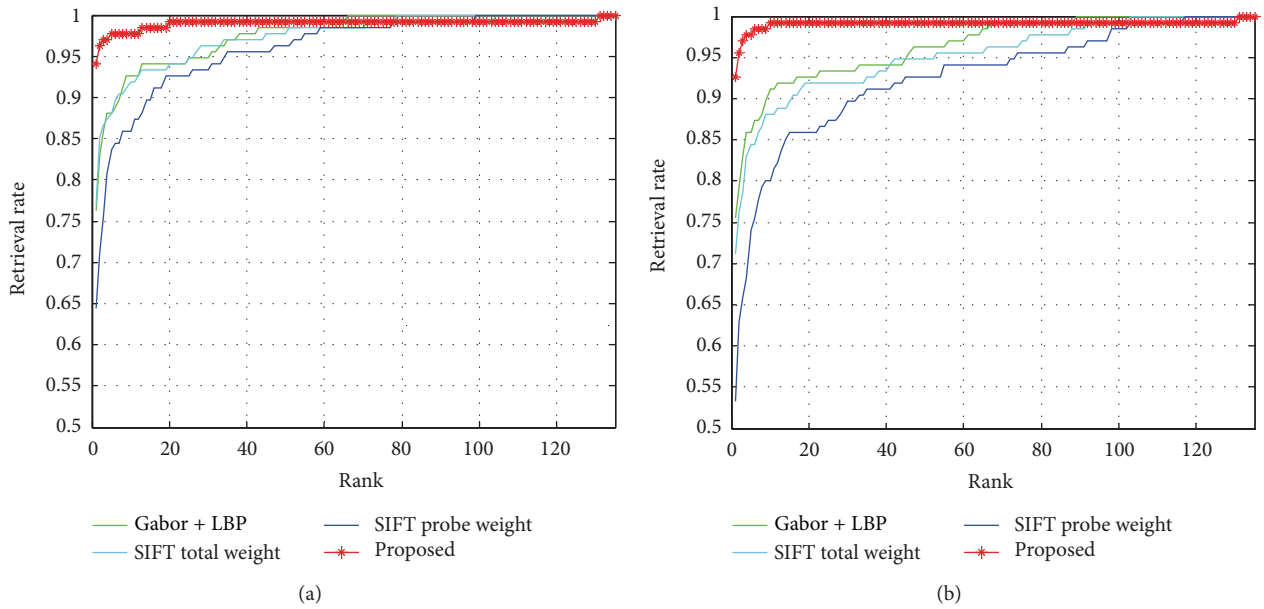


Figure 8: Performance comparison in sunglasses occlusion of the proposed facial image retrieval with other Gabor-LBP histogram-based face recognition algorithms in the case of SQI images used in SIFT matching: (a) 6 × 6 subregions and 10 × 10 subregions.

proposed method achieved 93.33% and 91.11% accuracy in the 6 × 6 and 10 × 10 subregions. The performance of the proposed method was similar to or less than the performance of other algorithms. This indicates that occlusion in the lower area of the face has less influence on the retrieval accuracy than other occlusion problems. It can be deduced that the discriminative face features are distributed in the region around the eyes, and the high performance can be achieved to utilize more information about the eye area. Gabor-LBP histogram-based face recognition algorithms used all the subregions of the upper area of the facial image, whereas the proposed method used only the selected subregions including SIFT matched key-points. The combination of SQI images for SIFT matching and 10 × 10 subregions shows the best performance in the experiment for scarf occlusion. The performance comparison for scarf occlusion is shown in Figure 9, which shows the performance in the case of SQI images used in SIFT matching.

Consequently, the proposed method has achieved better performance, where the parameters selected are 6 × 6 subregions and SIFT matching using SQI facial images; the performance of sunglasses and scarf occlusion was 94.07% and 93.33%, respectively. The results of the proposed method are shown to perform better compared to [26, 35, 36] in the

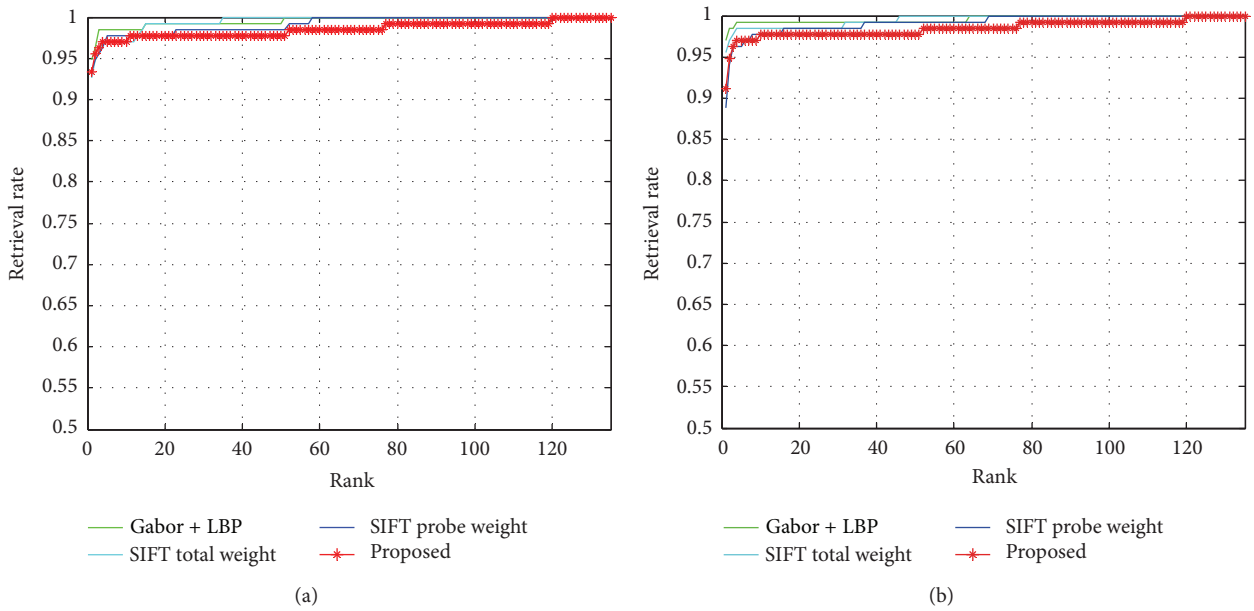


Figure 9: Performance comparison in scarf occlusion of the proposed facial image retrieval with other Gabor-LBP histogram-based face recognition algorithms in the case of SQI images used in SIFT matching: (a) 6×6 subregions and 10×10 subregions.

sunglasses occlusion. However, the Gabor-LBP histogram-based face recognition algorithms are better than the proposed method in the case of occlusion by scarf. The analysis shows that eyes and their environs are a significant area for face recognition. The lower part of the face is less important than the upper, and it is difficult to deal with occlusions which conceal the upper region of the face (including eyes) for face recognition. In that respect, the proposed method can contribute to eye-hidden occlusion such as sunglasses and hats for facial image retrieval.

4. Conclusions

We have described a facial image retrieval method using SIFT-based similarity measurement for partially occluded facial images. The proposed method achieved a positive contribution to facial image retrieval for partially occluded facial images using similarity measurement based on SIFT matching. The similarity measurement through SIFT matching was used in order to (i) detect the occluded region of the face, (ii) find the same positions of the face, and (iii) compensate for face misalignment. The SIFT-based similarity measurement can detect the occluded region of probe images and find similar positions between gallery and probe images without complex image processing. For facial image retrieval, the proposed method adopted Gabor-LBP histogram-based face features, and SIFT matching was employed to measure the similarity between a gallery image and a probe image. The proposed method has achieved retrieval rates of 94.07% and 93.33% with sunglasses and scarf occlusion, respectively. The retrieval accuracy of the conventional Gabor-LBP histogram was nearly 75% and 98% in the case of sunglasses and scarf occlusion. This shows that occlusions that hide the eyes and their environs have a greater influence on retrieval accuracy than other occlusion problems.

The proposed method enhanced the performance by about 20% compared with other Gabor-LBP histogram-based face recognition algorithms in the case of sunglasses occlusion. The average performance of facial image retrieval for scarf occlusion was almost 93%. Analysis shows that the most discriminative face features are in the environs of the eyes, and high performance can be achieved by utilizing more information on this region. The performance of the proposed method was similar to or less than other algorithms in the case of scarf occlusion, because the proposed method used insufficient face features for face recognition. The proposed method was robust against eyes-hidden occlusion such as sunglasses and hat and could be applied to a facial image retrieval system for alignment-free facial images. Furthermore, the proposed facial image retrieval method was based on a nonstatistical approach, which did not need a supervised learning process. This is another advantage for the facial image retrieval system because it is able to solve the problems of the large scale and updating of the face database. However, the proposed method has the weakness that it is not possible to match facial images if the matched key-points are not extracted or are insufficient. Therefore our future work will be focused on research on a more reliable and sufficient fiducial key-points detector and matching approach, instead of SIFT.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] U. Park and A. K. Jain, "Face matching and retrieval using soft biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 406–415, 2010.

[2] A. K. Jain, B. Klare, and U. Park, "Face matching and retrieval in forensics applications," *IEEE Multimedia in Forensics, Security, and Intelligence*, vol. 19, no. 1, pp. 20–28, 2012.

[3] J. Klontz and A. Jain, "A case study on unconstrained facial recognition using the boston marathon bombings suspects," Tech. Rep. MSU-CSE-13-4, 2013.

[4] J. C. Klontz, B. F. Klare, S. Klum, A. K. Jain, and M. J. Burge, "Open source biometric recognition," in *Proceedings of the 6th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS '13)*, pp. 1–8, October 2013.

[5] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–591, June 1991.

[6] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *Journal of the Optical Society of America A: Optics and Image Science, and Vision*, vol. 14, no. 8, pp. 1724–1733, 1997.

[7] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450–1464, 2002.

[8] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: a literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[9] A. Azeem, M. Sharif, M. Raza, and M. Murtaza, "A survey: face recognition techniques under partial occlusion," *International Arab Journal of Information Technology*, vol. 11, no. 1, pp. 1–10, 2014.

[10] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[11] B. Froba and A. Ernst, "Face detection with the modified census transform," in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition (FGR '04)*, pp. 91–96, Seoul, Republic of Korea, May 2004.

[12] B. Jun and D. Kim, "Robust face detection using local gradient patterns and evidence accumulation," *Pattern Recognition*, vol. 45, no. 9, pp. 3304–3316, 2012.

[13] K. Ohba and K. Ikeuchi, "Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 9, pp. 1043–1048, 1997.

[14] C.-Y. Huang, O. I. Camps, and T. Kanungo, "Object recognition using appearance-based parts and relations," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pp. 877–883, June 1997.

[15] A. M. Martínez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 748–763, 2002.

[16] Y. Zhang and A. M. Martínez, "A weighted probabilistic approach to face recognition from multiple images and video sequences," *Image and Vision Computing*, vol. 24, no. 6, pp. 626–638, 2006.

[17] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective representation using ICA for face recognition robust to local distortion and partial occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1977–1981, 2005.

[18] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, "Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft k-NN ensemble," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 875–886, 2005.

[19] H. J. Oh, K. M. Lee, and S. U. Lee, "Occlusion invariant face recognition using selective local non-negative matrix factorization basis images," *Image and Vision Computing*, vol. 26, no. 11, pp. 1515–1523, 2008.

[20] A. Rama, F. Tarres, L. Goldmann, and T. Sikora, "More robust face recognition by considering occlusion information," in *Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition (FG '08)*, pp. 1–6, IEEE, Amsterdam, Netherlands, September 2008.

[21] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–201, 2001.

[22] K. Hotta, "Robust face recognition under partial occlusion based on support vector machine with local Gaussian summation kernel," *Image and Vision Computing*, vol. 26, no. 11, pp. 1490–1498, 2008.

[23] M. Sharma, S. Prakash, and P. Gupta, "An efficient partial occluded face recognition system," *Neurocomputing*, vol. 116, pp. 231–241, 2013.

[24] R. Min, A. Hadid, and J.-L. Dugelay, "Improving the recognition of faces occluded by facial accessories," in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG '11)*, pp. 442–447, IEEE, Santa Barbara, Calif, USA, March 2011.

[25] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[26] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor Binary Pattern Histogram Sequence (LGBPHS): a novel non-statistical model for face representation and recognition," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 786–791, October 2005.

[27] H. Wang, S. Z. Li, and Y. Wang, "Face recognition under varying lighting conditions using self quotient image," in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition (FGR '04)*, pp. 819–824, May 2004.

[28] L. Shen and L. Bai, "A review on Gabor wavelets for face recognition," *Pattern Analysis and Applications*, vol. 9, no. 2-3, pp. 273–292, 2006.

[29] S. Xie, S. Shan, X. Chen, and J. Chen, "Fusing local patterns of gabor magnitude and phase for face recognition," *IEEE Transactions on Image Processing*, vol. 19, no. 5, pp. 1349–1361, 2010.

[30] T. Gao and M. He, "A novel face description by local multichannel Gabor histogram sequence binary pattern," in *Proceedings of the International Conference on Audio, Language and Image Processing (ICALIP '08)*, pp. 1240–1244, Shanghai, China, July 2008.

[31] S. Shan, W. Gao, Y. Chang, B. Cao, and P. Yang, "Review the strength of gabor features for face recognition from the angle of its robustness to mis-alignment," in *Proceedings of the IEEE 17th International Conference on Pattern Recognition (ICPR '04)*, pp. 338–341, August 2004.

[32] M. Yang, L. Zhang, S. C. K. Shiu, and D. Zhang, "Gabor feature based robust representation and classification for face recognition with Gabor occlusion dictionary," *Pattern Recognition*, vol. 46, no. 7, pp. 1865–1878, 2013.

[33] P. Kruizinga and N. Petkov, "Non-linear operator for oriented texture," *IEEE Transactions on Image Processing*, vol. 8, no. 10, pp. 1395–1407, 1999.

[34] T. Ahonen, A. Hadid, and M. Pietikainen, "Face recognition with local binary patterns," in *Computer Vision—ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004. Proceedings, Part I*, vol. 3021 of *Lecture Notes in Computer Science*, pp. 469–481, Springer, Berlin, Germany, 2004.

[35] B. Da and N. Sang, "Local binary pattern based face recognition by estimation of facial distinctive information distribution," *Optical Engineering*, vol. 48, no. 11, Article ID 117203, 2009.

[36] S. Park and J. Yoo, "Real-time face recognition with SIFT-based local feature points for mobile device," in *Proceedings of the 1st International Conference on Artificial Intelligence, Modelling and Simulation (AIMS '13)*, pp. 304–308, Kota Kinabalu, Malaysia, December 2013.

[37] A. Martinez and R. Benavente, "The AR face database," CVC Technical Report 24, 1998.

[38] V. Jain and E. Miller, "FDDB: a benchmark for face detection in unconstrained settings," Tech. Rep. UM-CS-2010-009, Department of Computer Science, University of Massachusetts, Amherst, Mass, USA, 2010.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 838929, 7 pages
http://dx.doi.org/10.1155/2015/838929

Research Article

Performance Improvement of Collision Warning System on Curved Road Based on Intervehicle Communication

Hong Cho[1] and Byeong-woo Kim[2]

[1]*Graduate School of Electrical Engineering, University of Ulsan, 93 Daehak-ro, Ulsan 680-749, Republic of Korea*
[2]*School of Electrical Engineering, University of Ulsan, 93 Daehak-ro, Ulsan 680-749, Republic of Korea*

Correspondence should be addressed to Byeong-woo Kim; bywokim@ulsan.ac.kr

Received 20 September 2014; Accepted 15 February 2015

Academic Editor: Jong-Hyuk Park

The vehicle on-board sensor based Advanced Driver Assistant System possesses limitations on a small road with a small radius of curvature because of the sensor's inability to operate in nondetectable domains. This study suggests an Improved Cooperative Collision Warning System (ICCWS) that considers the curvature of the road and is based on intervehicle communication. To predict the radius of curvature of the road, the Arc Relative Distance (ARD), the real relative distance to a preceding vehicle on a curved road has been used. The risk of collision with the preceding vehicle is decided by calculating an index of the risk of collision on a curved road using the computed ARD. The effects of ICCWS, proposed through this simulation, have been reviewed, and the improvement in performance in following a preceding vehicle has been analyzed quantitatively via comparative analysis with the conventional forward collision warning system. Accordingly, if the estimating algorithm for curvature developed in this study is applied to a real system, the performance of following a preceding vehicle can be improved without any specific changes to the system.

1. Introduction

The number of road-traffic accidents is rising proportionally in line with an increase in the number of vehicles on the road. Based on data reported by the National Highway Traffic Safety Administration (NHTSA) [1], about 80% of traffic accidents are due to the carelessness of drivers. Recently, studies related to an Advanced Driver Assistant System (ADAS) have actively been carried out, not only to reinforce legal regulations of vehicle safety, but also to increase awareness of safety devices available to consumers and decrease the possibility of the traffic accidents related to driver carelessness. The ADAS is a representative vehicle safety system that detects stopping or moving objects using sensors such as camera and radar and classifies them to reduce the possibility of an accident [2]. The major longitudinal ADAS is composed of a forward collision warning system (FCWS) and an Adaptive Cruise Control System (ACCS). The FCWS is a system that analyzes the risk of collision with a preceding vehicle and operates by using a vehicle on-board sensor to warn the driver of a collision. However, the vehicle on-board sensor based on ADAS only works within a measurable range of the sensor and possesses blind spots in relation to areas that it is unable to detect, such as at crossroads and on curved roads. In order to overcome these limitations, studies introducing a communication-connected safety system using Vehicle to Vehicle (V2V) communication and Vehicle to Infra (V2I) communication have actively been planned in relation to further developments made in communication technology [3–5].

The FCWS has already been established as an international standard by the International Organization for Standardization (ISO) [6]. However, the existing FCWS uses vehicle-mounted sensors that cause a malfunctioning of the system when the object in front enters a blind spot, thus escaping the measurable range of the sensor [7]. In order to overcome the limitations related to such blind spots, studies on the Cooperative Collision Warning System (CCWS) have been carried out by combining V2V and V2I technology [8]. However, the CCWS proposed thus far is not suitable for

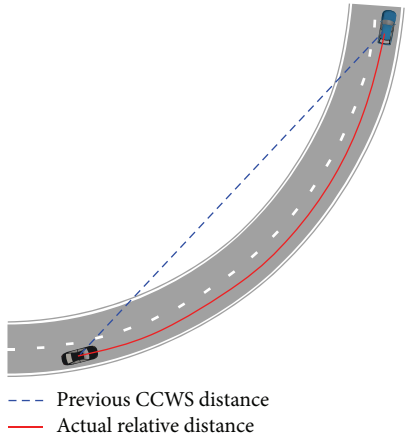


Figure 1: Limitations of CCWS.

narrow curved roads with a small radius of curvature, and it can only consider straight roads and slightly curved narrow roads. It is therefore considered that, to overcome such limitations, the development of a collision warning system that considers the curvature of curved roads is required. Thus, this study proposes an Improved Cooperative Collision Warning System (ICCWS) to overcome the problems of the conventional FCWS and CCWS without adding devices to a vehicle.

2. Improved Cooperative Collision Warning System Design

CCWSs proposed in previous studies use V2V communication to overcome the limitations of vehicle-mounted sensors and to warn the driver in advance by detecting the risk of collision with preceding vehicles, regardless of the existence of obstacles [9]. However, as shown in Figure 1, CCWSs proposed in previous studies are not suitable for curved roads with a small radius of curvature. Therefore, this paper proposes an ICCWS, which takes into account the curvature of the road. Figure 2 shows a block diagram of the ICCWS proposed in this study, which firstly calculates the relative distance (RD) and relative angle (RA) between nearby vehicles using the vehicle surroundings monitoring system according to whether the road is curved or straight and then applies the ICCWS to an ego-vehicle.

2.1. Vehicle Surroundings Monitoring System. As shown in Figure 3, the system generates coordinates based on the ego-vehicle and calculates the RA and RD to any surrounding vehicles. In this study, a new Cartesian coordinate system, the CS_{ego}, was defined using the ego-vehicle's current position (x_{ego}, y_{ego}), as the starting point. The CS_{ego} defined here is a coordinate system in which the longitudinal direction is the x-axis and the lateral direction is the y-axis, based on the traveling direction of the ego-vehicle. In the CS_{ego}, the positions of the surrounding vehicles are indicated as relative coordinates (x_n, y_n) in the four quadrants of the CS_{ego} by comparing the positions of the ego-vehicle and the surrounding vehicles that are received through V2V communication. Here, n is vehicle ID. RA_n is then calculated using a comparison with the azimuth, φ, which is related to the traveling direction of the ego-vehicle. As shown in Figure 3, based on RA_n, which is the relative angle that changes according to the four quadrants, the vehicle surroundings monitoring system can detect the position of surrounding vehicles. As shown in Figure 4, φ changes according to the heading angle of the vehicle, and East is designated as 0°. The CS_{ego}, as shown in Figure 4, rotates its axes depending on any changes made to the vehicle's φ. Considering this axis rotation, the longitudinal vehicle travelling direction was set to align with the x-axis at all times and is represented by

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{1}$$

2.2. Curved Road Collision Detection System. In this study, the time to collision (TTC) was used as the collision risk index to determine the risk of a collision with a preceding vehicle. TTC refers to the time remaining before impact, which is determined by the ratio of the distance to the preceding vehicle and the relative speed, in accordance with

$$\text{TTC (s)} = \frac{\text{Relative distance}}{\text{Relative speed}}. \tag{2}$$

Under curved road conditions, RD, which is the relative distance between the ego-vehicle and the forward vehicle calculated by the vehicle surroundings monitoring system, is not suitable for calculating TTC, the longitudinal collision risk index. Therefore, the actual relative distance needs to be defined by taking the conditions of the curved road into consideration. In this study, as shown in Figure 5, the RD and turning radius (TR_{ego}) of the ego-vehicle were used to calculate the actual relative distance (the Arc Relative Distance (ARD)) for a curved road, as defined using the following:

$$ARD_n = \frac{\theta_n}{180} \cdot \pi \cdot TR_{ego}, \tag{3}$$

where ARD_n is the relative distance from the forward vehicle, n, on a curved road; TR_{ego} is the ego-vehicle's turning radius; θ_n is the mid angle when the ego-vehicle's turning radius is TR_{ego} and the ego-vehicle's preceding vehicle, n, is distanced by ARD_n. The vehicle's TR can be determined through

$$TR_n = \frac{V_n}{\dot{\psi}}, \tag{4}$$

where V_n is the speed of the vehicle and $\dot{\psi}_n$ is the yaw rate of the vehicle.

The ICCWS can only operate on preceding vehicles in the same lane. For this, when the distance between TR_{ego} and TR_n is less than half the width of the lane, it was determined that the ego-vehicle and the preceding vehicle, n, are in the same lane. When TR_n could not be determined because the preceding vehicle, n, had stopped, it was estimated through the following (Figure 6) [10]:

$$TR_n = \sqrt{\left(TR_{ego} - RD_n \cdot \sin\left(RA_n\right)\right)^2 - \left(RD_n \cdot \cos\left(RA_n\right)\right)^2}. \tag{5}$$

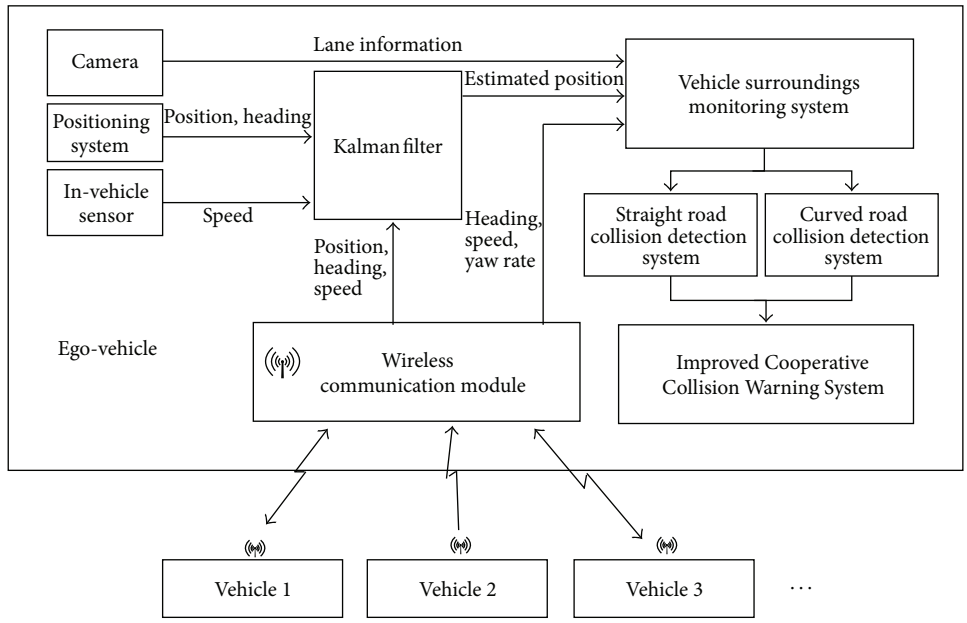


Figure 2: Block diagram for ICCWS.

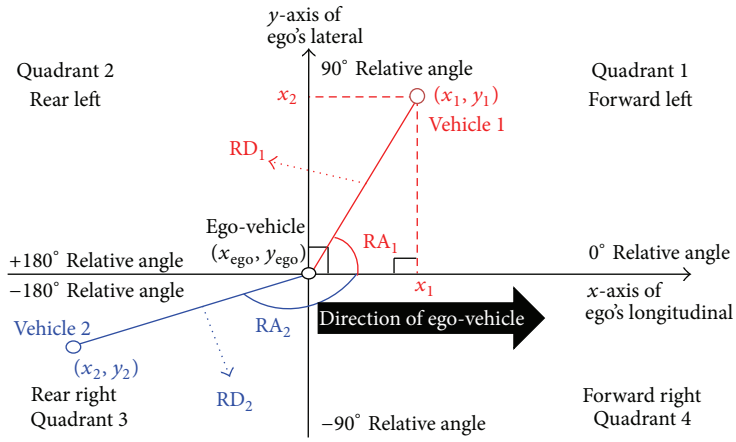


Figure 3: Principles of the vehicle surroundings monitoring system.

3. Simulation and Results

Commercial programs such as PreScan, CarSim, and MATLAB/Simulink were used to conduct the simulation. PreScan was used to perform the following modeling operations: (1) configuration of the V2V communication environment; (2) vehicle-mounted camera and radar sensor modeling; (3) modeling of vehicle surroundings; and (4) modeling in relation to driving conditions. Furthermore, a CarSim vehicle model was used for expressing the detailed kinetic characteristics of the ego-vehicle. Finally, the system was configured after interfacing the PreScan and CarSim through the MATLAB/Simulink.

3.1. Simulation Scenario. The simulation scenario is shown in Figure 7. The ICCWS warning times for an ego-vehicle travelling on a curved road were observed for a case when Vehicle 1 was travelling ahead of the ego-vehicle. In this context, the initial and driving speeds of the ego-vehicle were standard speeds corresponding to the curvature radius of the curved road with a superelevation of 6%, but in the established scenario the driving speed was exceeded by 5 km/h and 10 km/h, respectively. Table 1 presents the standard vehicle speeds on the curved road corresponding to various curvature radii, with a superelevation of 6%. The warning levels were defined based on the TTC; Table 2 lists the warning level standards corresponding to variations in the TTC [7]. The warning level was divided into a total of three levels: Level 1 was defined as the lowest level of risk, Level 2 as intermediate risk, and Level 3 as the highest risk.

3.2. Simulation Results. The proposed ICCWS was verified by comparing the ideal TTC_x with the TTC_x calculated from the ICCWS. The ideal TTC_x can be calculated using the

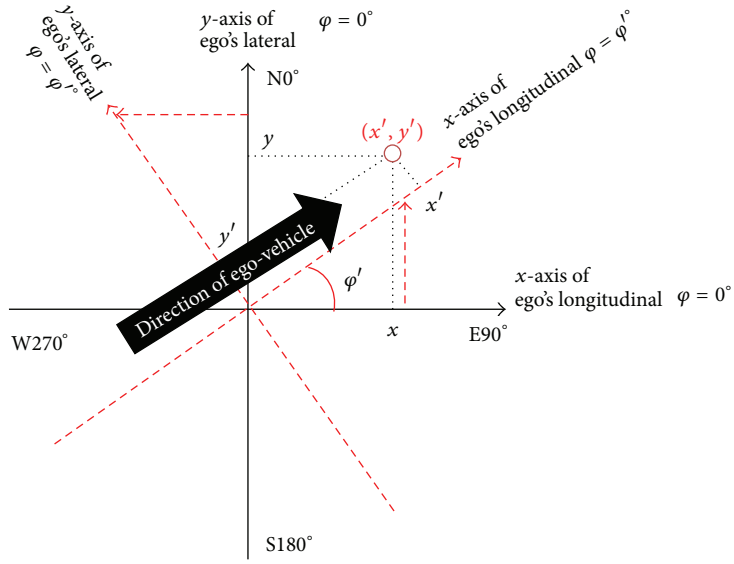


Figure 4: Rotations of CS_{ego} due to changes in the azimuth.

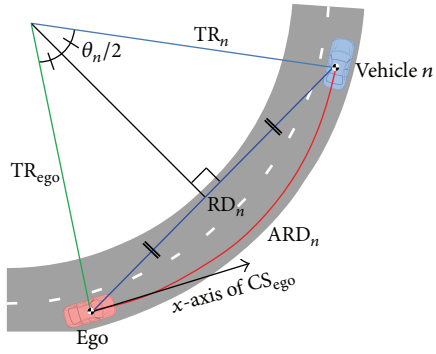


Figure 5: Arc Relative Distance (ARD) on a curved road.

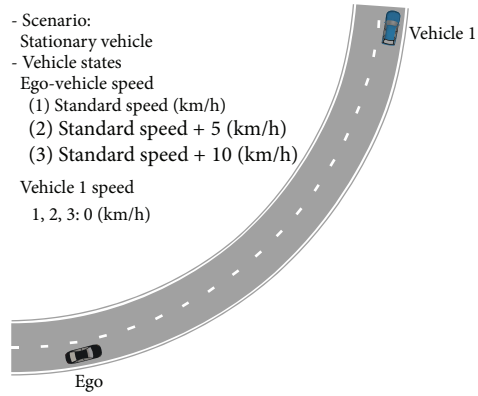


Figure 7: Simulation scenario.

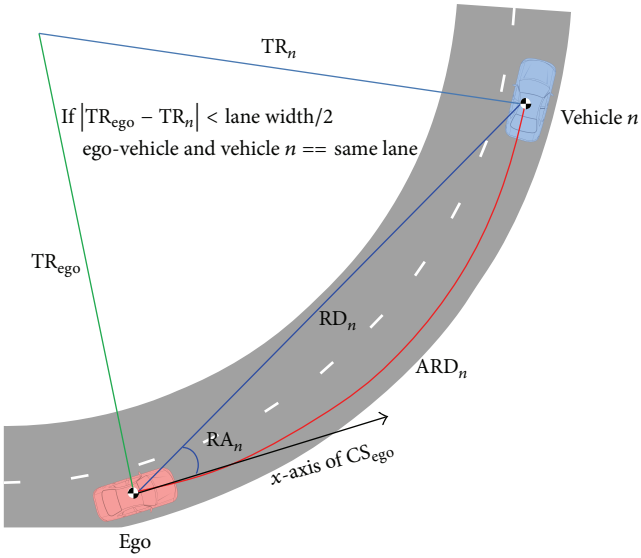


Figure 6: Estimation of radius of rotation and decision of same lane with vehicle n.

Table 1: Standard speed of vehicle according to the radius of curvature of a curved road.

Scenario number	1	2	3	4
Radius of curvature [m]	15	30	60	90
Standard speed [km/h]	20	30	40	50
Superelevation [%]	6	6	6	6
Initial relative distance [m]	42	85	145	175

Table 2: Definition of warning level according to changes in TTC_x.

Degree of risk		Low←——→high	
TTC_x [sec]	≤2.7	≤1.7	≤0.8
Warning level	1	2	3

difference between the time of collision and the simulated time, because the vehicle is being driven at a constant speed in the scenario. Figure 8 shows the margin of error used when comparing the $TTC_{x,ideal}$ (which is the ideal longitudinal

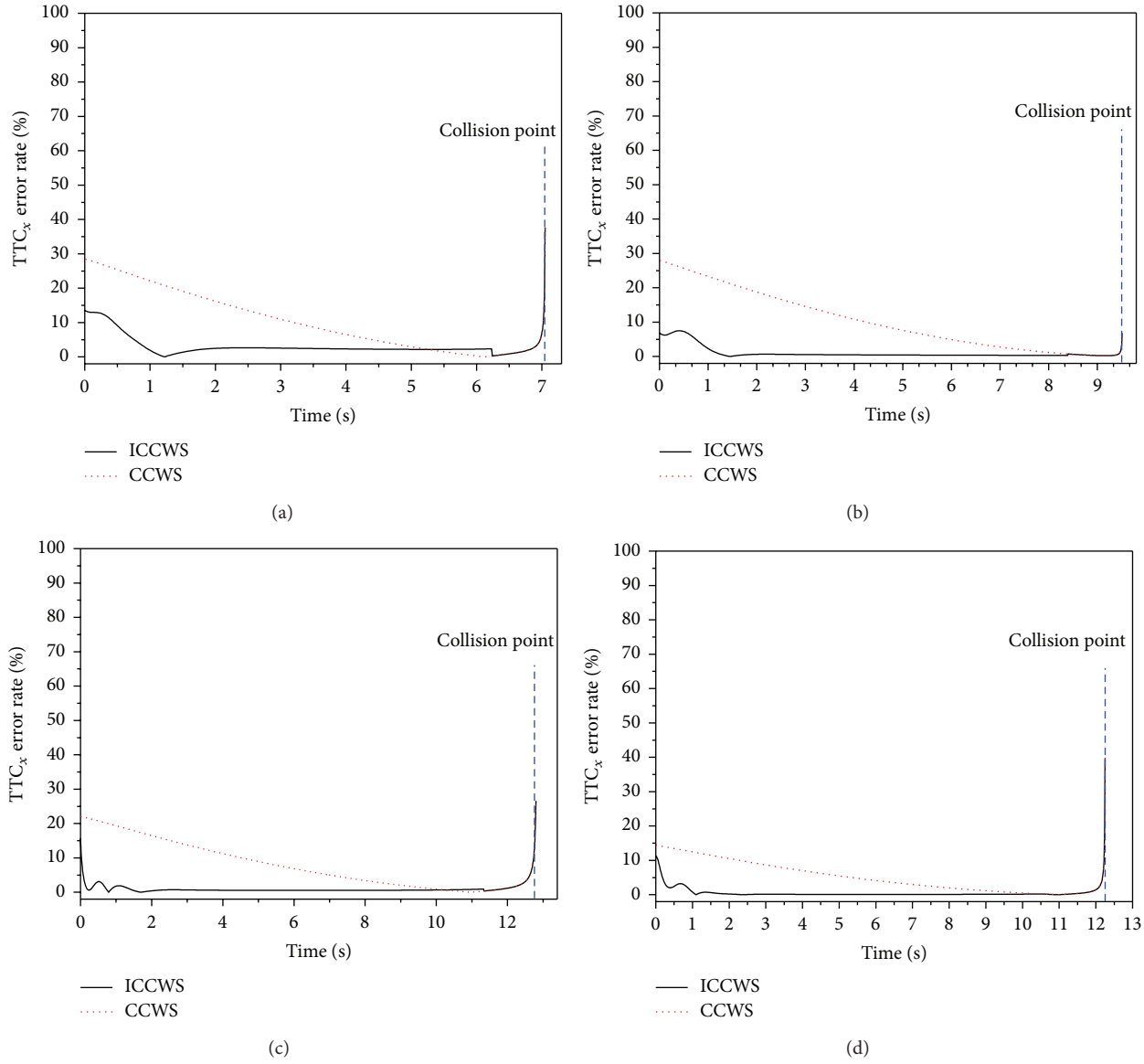


Figure 8: Margin of error of collision risk index of ICCWS and CCWS: (a) Scenario 1; (b) Scenario 2; (c) Scenario 3; and (d) Scenario 4.

collision risk index) with each TTC_x, calculated using the proposed ICCWS and the existing CCWS, after simulation of identical scenarios (Scenario 1, 2, or 3). The results of $TTC_{x,ICCWS}$, which is the longitudinal collision risk index of the proposed ICCWS corresponding to each scenario, indicate the following approximate average margins of error: for Scenario 1 (Figure 8(a)) = 3.18%, Scenario 2 (Figure 8(b)) = 1.06%, Scenario 3 (Figure 8(c)) = 1.15%, and Scenario 4 (Figure 8(d)) = 0.59%. These results confirm that the calculation of collision risk is relatively accurate. The reason for a $TTC_{x,ICCWS}$ error in the initial simulation is related to an error in the vehicle's TR owing to the rapidly changing value of the yaw rate when the ego-vehicle enters the curved road. However, the margin of error declines as the ego-vehicle begins a stable turn (after approximately 1.3 s), and thus the correct estimation of the collision risk can be confirmed.

The CCWS showed the following approximate average margins of error of the CCWS corresponding to each scenario: Scenario 1 (Figure 8(a)) = 10.67%, Scenario 2 (Figure 8(b)) = 10.32%, Scenario 3 (Figure 8(c)) = 8.17%, and Scenario 4 (Figure 8(d)) = 5.23%. This confirms a difference of approximately 7.10% compared to the proposed ICCWS. The reason for this difference is that the existing CCWS calculates the $TTC_{x,CCWS}$, which is the longitudinal collision risk index, without considering the curvature of the road. Table 3 presents the results of a comparative analysis simulation using the application of the ICCWS, CCWS, and FCWS (based on vehicle-mounted sensors) on the scenarios defined in Table 1. The simulation results listed in Table 3 show the margins of error, presented according to the scenarios. According to the results, the FCWS that uses a vehicle on-board sensor has a limited measurable range on curved roads. In other words,

Table 3: Margin of error, ICCWS, CCWS, and FCWS.

	Standard speed [km/h]	Collision warning system	Margin of error [%]
Scenario 1	20	ICCWS	3.18
		CCWS	10.67
		FCWS	83.40
	25	ICCWS	2.40
		CCWS	10.00
		FCWS	81.73
	30	ICCWS	2.96
		CCWS	10.48
		FCWS	81.76
Scenario 2	30	ICCWS	1.06
		CCWS	10.32
		FCWS	78.20
	35	ICCWS	0.78
		CCWS	10.17
		FCWS	77.92
	40	ICCWS	2.65
		CCWS	12.38
		FCWS	80.07
Scenario 3	40	ICCWS	1.15
		CCWS	8.17
		FCWS	79.30
	45	ICCWS	1.94
		CCWS	8.66
		FCWS	79.88
	50	ICCWS	2.20
		CCWS	8.68
		FCWS	79.80
Scenario 4	50	ICCWS	0.59
		CCWS	5.23
		FCWS	82.58
	55	ICCWS	1.33
		CCWS	5.70
		FCW	82.96
	60	ICCWS	1.11
		CCWS	5.53
		FCWS	83.19

it exhibits a margin of error that is greater than 79% in all of the scenarios. On the other hand, the ICCWS proposed in this study exhibits a margin of error that is, at most, less than 3% in all of the scenarios and is similar to the actual performance of following a preceding vehicle. Moreover, compared to CCWS, which does not consider the curvatures of roads, ICCWS reduces the margin of error to a maximum of approximately 7% for roads with large curvatures. By improving the performance of following a preceding vehicle, safe driving can be ensured because the driver can recognize accurate risk-warning signals on curved roads.

4. Conclusion

This research proposes an ICCWS that considers a small curvature radius on curved roads. ARD, which is the actual relative distance between the ego-vehicle and the preceding vehicle on the curved road, was calculated by utilizing the turning radius of the ego-vehicle and data obtained from the vehicle surroundings monitoring system. As per the results, we reduced a maximum of approximately 7% margin of error compared to CCWS and 82% compared to FCWS. With the improvement in following a preceding vehicle using ICCWS proposed in this study, more accurate risk-warning signals were provided to drivers on curved roads, and, thus, driver resistance to the system was minimized.

In future research, the application of the proposed ICCWS to the primary collision evasion system, autonomous emergency braking (AEB), is expected, in addition to studies on the implementation of the ICCWS on various roads and within multivehicle environments.

In conclusion, this study proposed an ICCWS to overcome the problems of the conventional FCWS and CCWS, without adding devices to the vehicle.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the 2013 Research Fund of University of Ulsan.

References

[1] NHTSA, "The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic driving study data," NHTSA Report, National Highway Traffic Safety Administration, 2006.

[2] K. Goswami, G. Hong, and B. Kim, "A novel mesh-based moving object detection technique in video sequence," *Journal of Convergence*, vol. 4, no. 3, pp. 20–24, 2013.

[3] D. Caveney and W. B. Dunbar, "Cooperative driving: beyond V2V as an ADAS sensor," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV '12)*, pp. 529–534, June 2012.

[4] Z. C. Taysi and A. G. Yavuz, "ETSI compliant GeoNetworking protocol layer implementation for IVC simulations," *Human-centric Computing and Information Sciences*, vol. 3, article 4, 2013.

[5] R. Hussain and H. Oh, "Cooperation-aware VANET clouds: providing secure cloud services to vehicular ad hoc networks," *Journal of Information Processing Systems*, vol. 10, no. 1, pp. 103–118, 2014.

[6] ISO, "Transport information and control systems forward vehicle collision warning systems performance requirements and test procedures," ISO 15623, International Organization for Standardization, London, UK, 2013.

[7] K. D. Kusano and H. C. Gabler, "Safety benefits of forward collision warning, brake assist, and autonomous braking systems in rear-end collisions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1546–1555, 2012.

[8] A. Sebastian, M. Tang, Y. Feng, and M. Looi, "Multi-vehicles interaction graph model for cooperative collision warning system," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 929–934, IEEE, Xi'an, China, June 2009.

[9] H. Cho, G.-E. Kim, and B.-W. Kim, "Usability analysis of collision avoidance system in vehicle-to-vehicle communication environment," *Journal of Applied Mathematics*, vol. 2014, Article ID 951214, 10 pages, 2014.

[10] E. Ryu, S. Jo, J. Kwon, T. Hong, and K. Park, "A study on the development of risk situation awareness algorithm and performance verification with PreScan," in *Proceedings of the Conference for Korea Society of Automotive Engineers*, pp. 605–613, May 2013.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 902043, 10 pages
http://dx.doi.org/10.1155/2015/902043

Research Article

Qualitative Spatial Reasoning with Directional and Topological Relations

Sangha Nam and Incheol Kim

Department of Computer Science, Kyonggi University, San94-6, Yiui-Dong, Yeongtong-Gu, Suwon-Si 443-760, Republic of Korea

Correspondence should be addressed to Incheol Kim; kic@kgu.ac.kr

Received 17 June 2014; Accepted 22 December 2014

Academic Editor: Jong-Hyuk Park

A wide range of application domains from cognitive robotics to intelligent systems encompassing diverse paradigms such as ambient intelligence and ubiquitous computing environments require the ability to represent and reason about the spatial aspects of the environment within which an agent or a system is functional. Many existing spatial reasoners share a common limitation that they do not provide any checking functions for cross-consistency between the directional and the topological relation set. They provide only the checking function for path-consistency within a directional or topological relation set. This paper presents an efficient spatial reasoning algorithm working on a mixture of directional and topological relations between spatial entities and then explains the implementation of a spatial reasoner based on the proposed algorithm. Our algorithm not only has the checking function for path-consistency within each directional or topological relation set, but also provides the checking function for cross-consistency between them. This paper also presents an application system developed to demonstrate the applicability of the spatial reasoner and then introduces the results of the experiment carried out to evaluate the performance of our spatial reasoner.

1. Introduction

A wide range of application domains from cognitive robotics to intelligent systems encompassing diverse paradigms such as ambient intelligence and ubiquitous computing environments [1, 2] require the ability to represent and reason about the spatial aspects of the environment within which an agent or a system is functional [3]. Qualitative reasoning is concerned with capturing everyday commonsense knowledge of the physical world with a limited set of symbols and relationships and manipulating it in a nonnumerical manner. The subfield of qualitative reasoning that is concerned with representation and reasoning of space is called qualitative spatial reasoning (QSR) [4].

The main aim of research in qualitative spatial reasoning (QSR) is to develop powerful representation formalisms that account for the multimodality of space in a cognitively acceptable way. CSD- (cone-shaped directional-) 9 [5] and RCC- (region connection calculi-) 8 [6] are well-known qualitative spatial formalisms for representing and reasoning about the directional and the topological relationships between spatial entities. Some implementations of spatial reasoner based on these formalisms include SOWL [7], PelletSpatial [8], and CHOROS [9]. SOWL is a framework for handling spatiotemporal knowledge in OWL, the standard semantic web language. Both RCC-8 topological and CSD-9 directional relations are integrated in SOWL. The SOWL reasoner is capable of inferring new relations and checking their consistency. The reasoner is realized by a set of SWRL [10] rules operating on spatiotemporal relations. PelletSpatial is a spatial reasoner that adopts a highly efficient algorithm for checking path-consistency in a RCC-8 topological relation set. CHOROS is an extension of PelletSpatial to add CSD-9 directional reasoning facility. However, all these existing spatial reasoners share a common limitation that they do not provide any checking functions for cross-consistency between the CSD-9 directional relation set and the RCC-8 topological relation set. They provide only the checking function for path-consistency within a directional or topological relation set.

In this paper, we present an efficient spatial reasoning algorithm working on a mixture of CSD-9 directional

and RCC-8 topological relations, and then we explain the implementation of a spatial reasoner based on the proposed algorithm. Our algorithm not only has the checking function for path-consistency within each directional or topological relation set, but also provides the checking function for cross-consistency between them. We also present an application system developed to demonstrate the applicability of the spatial reasoner and then introduce the results of the experiment carried out to evaluate the performance of our spatial reasoner.

2. Design of Spatial Reasoner

2.1. Spatial Knowledge Representation. The first requirement in the design of a spatial reasoner is to decide how to represent the spatial knowledge required for the inference. This is the spatial knowledge representation. In this paper, we assume that a spatial knowledge base used for reasoning is represented by triple statements (subject, predicate, and object) according to RDF/OWL, a semantic web-standard ontology language. Each location covered by the knowledge base is defined as an element that belongs to the GeoInstance class.

Figure 1 shows a statement or fact that builds the spatial knowledge base. Relations of direction, boundary, and containment between two spaces or locations (GeoInstance) are represented using the spatial property defined in CSD-9 and RCC-8. For example, the directional relation between two spaces, "the state of Oregon in the USA is located to the north of the state of California," can be represented by (Oregon north of California).

The spatial reasoning of CSD-9 and RCC-8 assumes a knowledge type that represents relations between spaces. The CSD- (cone-shaped directional-) 9 theory assumes that a direction relation between two arbitrary points on a two-dimensional space can be represented, if one point is set, by one of nine directions: East (E), West (W), South (S), North (N), North East (NE), North West (NW), South East (SE), South West (SW), and Identical (O) for the direction of the other point as shown in Figure 2.

The RCC- (region connection calculi-) 8 theory assumes that boundary and containment relations between two arbitrary regions on a two dimensional space can be represented by one of eight relations: disconnected (DC), externally connected (EC), partially overlapping (PO), equal (EQ), tangential proper part (TPP), tangential proper part inverse (TPPi), nontangential proper part (NTPP), and nontangential proper part inverse (NTPPi) as shown in Figure 3. That is, the CSD-9 spatial knowledge expresses a directional relation between two spaces in terms of points whereas the RCC-8 spatial knowledge expresses a relation of boundary and containment between two spaces in terms of regions. Most real-world spaces or locations can be interpreted as either a single point or single region that represents multifaceted characteristics. Thus, CSD-9 and RCC-8 spatial knowledge can be utilized complementarily to express and infer diverse relations between real-world spaces.

2.2. Spatial Inference Rule. The CSD-9 spatial inference rules, which are applied to the spatial knowledge base represented by nine directional relations, can be summarized by the composition table shown in Table 1.

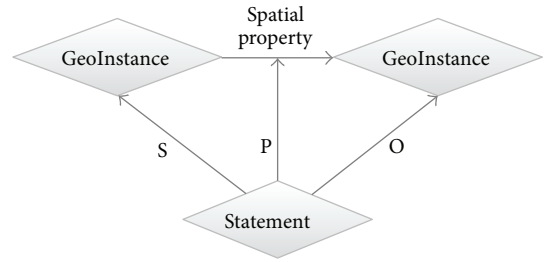


Figure 1: Triple statement representation.

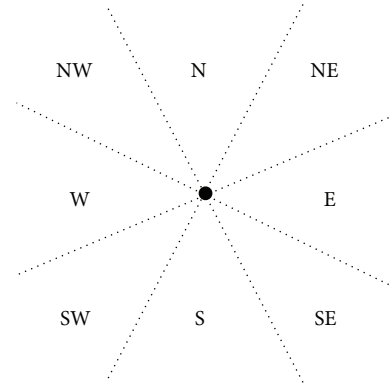


Figure 2: The set of CSD-9 directional relations.

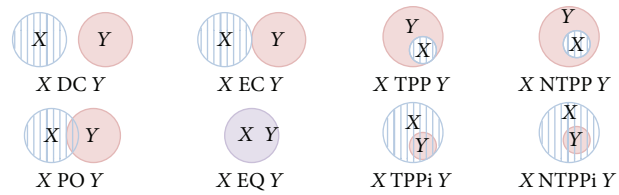


Figure 3: The set of RCC-8 topological relations.

Table 1 suggests that if items in the horizontal row and vertical column are simultaneously true, a new composition can be produced from the facts that are listed in the cell where the corresponding row and column intersect. For example, if location "A" is located at north of location "B" (A N B) and "B" is located in the northeast of location "C" (B NE C), then "A" may be located north or northeast of "C" (A [N, NE] C). This is inferred as a new fact. When a relation between two spaces is defined by a definite relation as shown above, for example, (A N B) and (B NE C), these facts are called "defined relations." On the other hand, when a relation between two spaces cannot be definitely defined, for example (A [N, NE] C) as shown above, this relation is called a "disjunctive relation." The RCC-8 spatial inference rules, which use a similar method, can be summarized by the composition table shown in Table 2. For example, when "A" is bordered with "B" (A EC B) and "B" contains "C" completely without intersection (B NTPPi C), a new fact can be inferred that "A" and "C" are separated from each other (A DC C).

Table 1: Composition table for CSD-9 relations.

	N	**NE**	E	SE	S	SW	W	NW	O
N	N	**N, NE**	N, NE, E	N, NE, E, SE	N, NE, E, SE, S, SW, W, NW, O	N, SW, W, NW	N, W, NW	N, NW	N, O
NE	N, NE	NE	NE, E	NE, E, SE	NE, E, SE, S	N, NE, E, SE, S, SW, W, NW, O	N, NE, W, NW	N, NE, NW	NE, O
E	N, NE, E	NE, E	E	E, SE	E, SE, S	E, SE, S, SW	N, NE, E, SE, S, SW, W, NW, O	N, NE, E, NW	E, O
SE	N, NE, E, SE	NE, E, SE	E, SE	SE	SE, S	SE, S, SW	SE, S, SW, W	N, NE, E, SE, S, SW, W, NW, O	SE, O
S	N, NE, E, SE, S, SW, W, NW, O	NE, E, SE, S	E, SE, S	SE, S	S	S, SW	S, SW, W	S, SW, W, NW	S, O
SW	N, SW, W, NW	N, NE, E, SE, S, SW, W, NW, O	E, SE, S, SW	SE, S, SW	S, SW	SW	SW, W	SW, W, NW	SW, O
W	N, W, NW	N, NE, W, NW	N, NE, E, SE, S, SW, W, NW, O	SE, S, SW, W	S, SW, W	SW, W	W	W, NW	W, O
NW	N, NW	N, NE, NW	N, NE, E, NW	N, NE, E, SE, S, SW, W, NW, O	S, SW, W, NW	SW, W, NW	W, NW	NW	NW, O
O	N, O	NE, O	E, O	SE, O	S, O	SW, O	W, O	NW, O	N, NE, E, SE, S, SW, W, NW, O

Table 2: Composition table for RCC-8 relations.

	DC	EC	PO	TPP	NTPP	TPPi	**NTPPi**	EQ
DC	DC, EC, PO, TPP, NTPP, TPPi, NTPPi, EQ	DC, EC, PO, TPP, NTPP	DC, EC, PO, TPP, NTPP	DC, EC, PO, TPP, NTPP	DC, EC, PO, TPP, NTPP	DC	DC	DC
EC	DC, EC, PO, TPPi, NTPPi	DC, EC, PO, TPP, TPPi, EQ	DC, EC, PO, TPP, NTPP	EC, PO, TPP, NTPP	PO, TPP, NTPP	DC, EC	**DC**	EC
PO	DC, EC, PO, TPPi, NTPPi	DC, EC, PO, TPPi, NTPPi	DC, EC, PO, TPP, NTPP, TPPi, NTPPi, EQ	PO, TPP, NTPP	PO, TPP, NTPP	DC, EC, PO, TPPi, NTPPi	DC, EC, PO, TPPi, NTPPi	PO
TPP	DC	DC, EC	DC, EC, PO, TPP, NTPP	TPP, NTPP	NTPP	DC, EC, PO, TPP, TPPi, EQ	DC, EC, PO, TPPi, NTPPi	TPP
NTPP	DC	DC	DC, EC, PO, TPP, NTPP	NTPP	NTPP	DC, EC, PO, TPP, NTPP	DC, EC, PO, TPP, NTPP, TPPi, NTPPi, EQ	NTPP
TPPi	DC, EC, PO, TPPi, NTPPi	EC, PO, TPPi, NTPPi	PO, TPPi, NTPPi	PO, TPP, TPPi, EQ	PO, TPP, NTPP	TPPi, NTPPi	NTPPi	TPPi
NTPPi	DC, EC, PO, TPPi, NTPPi	PO, TPPi, NTPPi	PO, TPPi, NTPPi	PO, TPPi, NTPPi	PO, TPP, NTPP, TPPi, NTPPi, EQ	NTPPi	NTPPi	NTPPi
EQ	DC	EC	PO	TPP	NTPP	TPPi	NTPPi	EQ

Table 3: Conversion table between CSD-9 directional and RCC-8 topological relations.

Directional relations	Topological relations
O	EQ, PO, TPP, NTPP, TPPi, NTPPi
N, NE, E, SE, S, SW, W, NW	DC, EC, PO

Inherently, CSD-9 and RCC-8 are independent theories that provide distinct spatial knowledge representations and inference methods from a different viewpoint. However, as explained earlier, in many real-world spaces and locations, it is necessary that the directional relations of CSD-9 and containment relations of RCC-8 be represented and inferred simultaneously. In such cases, a spatial reasoning algorithm that requires integrated inference must determine whether a new fact can be inferred in terms of RCC-8 containment relations from the facts that represent the CSD-9 directional relations or whether the inference must be produced conversely. In this paper, the conversion rules between CSD-9 and RCC-8 relations were determined via analysis of a special knowledge base representing various cases as shown in Table 3. In Table 3, a single fact that represents a CSD-9 O relation suggests a fact that satisfies one relation from {EQ, PO, TPP, NTPP, TPPi, NTPPi} of the RCC-8 relations. Similarly, an item that represents CSD-9 relations of {N, NE, E, SE, S, SW, W, NW} suggests that one of the RCC-8 relations {DC, EC, PO} can be satisfied. Conversely, relations of {EQ, PO, TPP, NTPP, TPPi, NTPPi} from RCC-8 suggest a CSD-9 O relation. Relations of {DC, EC, PO} from RCC-8 suggest that one of the CSD-9 relations of {N, NE, E, SE, S, SW, W, NW} can be satisfied. Thus, once defined or disjunctive relations that correspond to each case are determined from a knowledge base, they can be converted to newly suggested defined relations or disjunctive relations. For example, when the state of California completely contains the city of LA (California NTPPi LA), the state of California cannot be assumed to be located at any of the eight directions therefore the state of California and LA are interpreted as an Identical relation (California O LA). If two regions satisfy a PO relation, that is, partially overlapped, depending on how the centers of the two regions are located, their relation can be interpreted as being either Identical (O) or one of the remaining eight directions.

2.3. Spatial Reasoning Algorithm. In this paper, we design a spatial reasoning algorithm based on the spatial inference rules summarized in Tables 1, 2, and 3.

Algorithm 1 summarizes the overall process of the spatial reasoning algorithm. This algorithm consists of detailed steps that check the consistency of Set N of CSD-9 and RCC-8 spatial relations contained in the knowledge base. If Set N is an empty set or all spatial relations satisfy consistency, the algorithm is terminated successfully (Lines 1~4). The inverseComplete(N) and equalsComplete(N) steps process N to infer all the inverse and equals relations. For example, with regard to a single spatial relation of (A N B), the inverse relation (B S A) and equal relations (A O A) and (B O B) are inferred (Lines 5~6). For all the spatial relations created, the checks IsPathConsistent(N, R_{ab}) and IsCrossConsistent(N, R_{ab}) are performed iteratively based on the spatial relation R_{ab}. If an inconsistency of the knowledge base is detected, inference is interrupted (Lines 7~11). IsPathConsistent(N, R_{ab}) performs a path-consistency check on the relation sets of CSD-9 and RCC-8 using Tables 1 and 2, respectively. And then IsCrossConsistent(N, R_{ab}) performs the cross-consistency checks between CSD-9 and RCC-8 relations using Table 3.

```
SpatialReasoning(N)
(1)  require N a set of defined relations
(2)  ensure N consistent then true; else false
(3)  if N = ∅ then
(4)      return true
(5)  N ← inverseComplete(N)
(6)  N ← equalsComplete(N)
(7)  for R_ab ∈ N do
(8)      if !isPathConsistent(N, R_ab) then
(9)          return false
(10)     if !isCrossConsistent(N, R_ab) then
(11)         return false
(12) return true
```

Algorithm 1: Algorithm for spatial reasoning.

```
IsPathConsistent(N, R_ab)
(1)  require N a set of relations, R_ab ∈ N
(2)  ensure R_ab path-consistent then true; else false
(3)  if isCSDRelation(R_ab) then
(4)      M ← getCSDRelations(N)
(5)  else
(6)      M ← getRCCRelations(N)
(7)  for S_bc ∈ M do
(8)      T_ac ← composeRelations(R_ab, S_bc)
(9)      (N, consistency) ← addRelations(N, T_ac)
(10)     if !consistency then
(11)         return false
(12) return true
```

Algorithm 2: Algorithm for checking path-consistency.

Algorithm 2 shows the IsPathConsistent(N, R_{ab}) algorithm, which performs a path-consistency check on the relation sets of CSD-9 and RCC-8 based on relation R_{ab}, respectively. If the isCSDRelation(R_{ab}) step returns the result that R_{ab} is CSD-9 knowledge, all the items of the CSD-9 knowledge from Set N are stored in Set M, while all items of the RCC-8 knowledge from Set N are stored in Set M (Lines 3~6). Then, the composeRelations(R_{ab}, S_{bc}) and AddRelations(N, T_{ac}) steps are performed iteratively with regard to the items of knowledge (S_{bc}) that enable composition inference with R_{ab} from M. During the composeRelations(R_{ab}, S_{bc}) step, a new spatial relation is derived through the composition-inference process of R_{ab} and S_{bc} as mentioned in Section 2.2. Next, during the AddRelations(N, T_{ac}) step, while a new spatial relation is stored in Set N, it is determined whether a new spatial relation satisfies the

consistency with the existing relations (Lines 7~12). Using the path-consistency check, a new spatial relation is derived that can be composed from the existing spatial relations.

Algorithm 3 shows the IsCrossConsistent(N, R_{ab}) algorithm that performs the cross-consistency check between the relations of CSD-9 and RCC-8 based on the relation R_{ab}. During the convertRelation(R_{ab}) step, new cross-spatial relations are derived from the implications of the single spatial relation R_{ab} using conversion Table 3 (Line 3). For example, a new topological relation (A [DC, EC, PO] B) that is implied by the directional relation (A N B) between the two spaces "A" and "B" is derived. Then, Set U_{ab} containing the newly derived spatial relations is integrated to Set N through AddRelations(N, U_{ab}) followed by another consistency check (Lines 4~6).

Algorithm 4 shows the AddRelations(N, V_{ab}) algorithm that adds relation V_{ab} to Set N of the spatial relations and simultaneously checks the consistency between the relations. Relation V_{ab} corresponds to a new spatial relation obtained through the composition or conversion process. If the certainty of this spatial relation is 0%, it is determined that it is not necessary to add this relation to Set N and thus the operation is terminated (Lines 2~3). Otherwise, it determines what spatial relation existed previously between "a" and "b" from Set N. If an existing spatial relation is not found, relation V_{ab} is added to Set N (Lines 4~6). If an existing spatial relation is found, the high-certainty spatial relations are derived through the intersection of the existing spatial relations and V_{ab}. The high-certainty spatial relations are added to Set N and the existing spatial relations are removed. However, if the above intersection is an empty set, which means that the consistency is not met between the existing spatial relations and new spatial relations, inconsistency of the spatial relations is returned (Lines 7~15). For example, where the directional relation of (A [N, E] B) between two spaces "A" and "B" exists in the existing Set N, and a new directional relation of (A [N, NE] B) is to be added to it, the high-certainty spatial relation (X_{ab}), which is (A N B), is derived through an intersection between the two. Finally, X_{ba}, which is an inverse relation of X_{ab}, is added to Set N using a recursive call of the function.

The spatial reasoning algorithm, in which a cross-consistency check step is extended in this paper, has three advantages compared to existing spatial reasoning algorithms. First, the amount of knowledge derived from inference increases. For example, in "LA is located northwest of San Diego (LA NW SD)" where spatial knowledge from a directional viewpoint exists although spatial knowledge of the topological viewpoint between LA and San Diego is not indicated, the following knowledge can be inferred using the conversion table (LA [DC, EC, PO] SD). Moreover, the certainty of the inferred knowledge increases. For example, the following two cases "LA is located southeast or identical of San Francisco (LA [SE, O] SF)" where the spatial knowledge of the direction viewpoint exists and "LA and San Francisco are separated (LA DC SF)" where the spatial knowledge of the topological viewpoint exists are assumed. Here, certainty can be added to the spatial knowledge of uncertainty from a directional viewpoint using the cross-consistency check

IsCrossConsistent(N, R_{ab})
(1) **require** N a set of relations, $R_{ab} \in N$
(2) **ensure** R_{ab} cross-consistent then **true**; else **false**
(3) $U_{ab} \leftarrow convertRelation(R_{ab})$
(4) (N, *consistency*) $\leftarrow addRelations(N, U_{ab})$
(5) **if** !*consistency* **then**
(6) **return** false
(7) **return** true

Algorithm 3: Algorithm for checking cross-consistency.

AddRelations(N, V_{ab})
(1) **require** N, V_{ab} relation
(2) **if** $V = \tilde{\forall}$ **then**
(3) **return**
(4) $W_{ab} \leftarrow \{R_{ij} \mid i = a, j = b, R_{ij} \in N\}$
(5) **if** $\nexists W_{ab}$ **then**
(6) $X_{ab} \leftarrow V_{ab}$
(7) **else**
(8) $X_{ab} \leftarrow V_{ab} \cap W_{ab}$
(9) **if** $X_{ab} = \emptyset$ **then**
(10) *consistency* = *false*
(11) **return**
(12) **if** $W_{ab} = X_{ab}$ **then**
(13) **return**
(14) $N \leftarrow N \setminus \{W_{ab}\}$
(15) $N \leftarrow N \cup \{X_{ab}\}$
(16) *addRelations* (N, X_{ba})

Algorithm 4: Algorithm for adding new relations.

(LA SE SF). Finally, additional validation on incorrect spatial knowledge can be performed. For example, let us assume that "San Francisco is located northwest of San Diego (SF NW SD)" where the spatial knowledge from a directional viewpoint exists and "San Francisco contains San Diego (SF TPPi SD)" where spatial knowledge from a topological viewpoint exists. Here, using the cross-consistency check, incorrect spatial knowledge can be detected.

3. Implementation and Application

3.1. Implementation. Based on the spatial reasoning algorithm presented before, a qualitative spatial reasoner was implemented using the Java programming language. As shown in Figure 4, the reasoner consists of a spatial reasoning engine that derives new pieces of spatial knowledge from existing knowledge bases and a pellet-reasoning engine that derives new pieces of RDF/OWL knowledge. The spatial reasoning engine is then composed of a path-consistency checker and a cross-consistency checker. The former is responsible for consistency checking of the spatial relations sets CSD-9 and RCC-8, respectively.

The latter is responsible for cross-consistency checking between them. Previously defined composition and conversion tables are used to perform the consistency check.

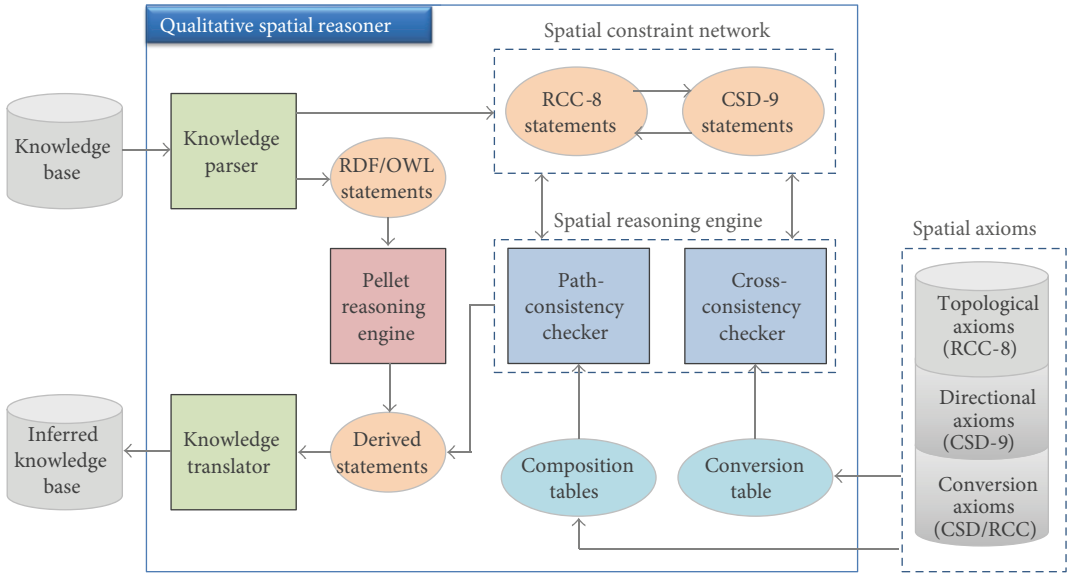


Figure 4: The qualitative spatial reasoner.

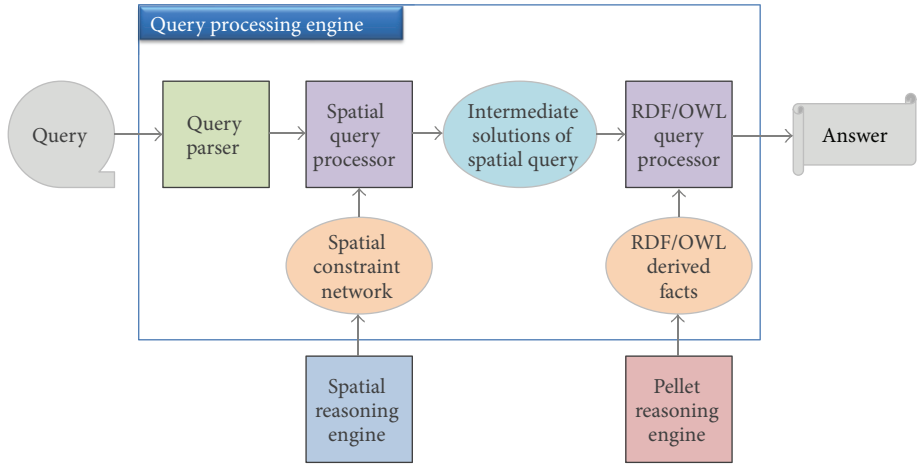


Figure 5: The query processing engine.

The knowledge base is divided into CSD-9/RCC-8 spatial relations and general RDF/OWL relations using a knowledge parser. The CSD-9/RCC-8 spatial relations are utilized to derive new spatial relations via the consistency check of the spatial reasoning engine. The general RDF/OWL relations are utilized to derive new RDF/OWL relations using the pellet-reasoning engine.

The architecture of the query-processing engine for response to spatial queries is shown in Figure 5. When a user submits a SPARQL-type spatial query including one or more spatial relations to the query-processing engine, it is transferred to the spatial query processor using a query parser. The spatial query processor attempts to determine an answer to the query using the spatial reasoning engine and obtains intermediate solutions for the spatial query. Then, the general RDF/OWL query processor seeks a solution to the query using the pellet-reasoning engine. The solution is combined with the spatial query processing results that are then returned to the user.

3.2. Application. In order to investigate the applicability of our spatial reasoner, we construct a sample spatial knowledge base using an ontology editor called Protégé. This spatial knowledge base contains some geographic information on the United States of America (USA), consisting of a total of nine classes, 127 individuals, and 1,900 spatial statements. Table 4 shows the number of individuals included in the sample knowledge base.

More than 80% of the individuals consist of states, counties, and cities. Several mountains, lakes, and rivers are also included to represent various pieces of spatial knowledge.

Figure 6 shows an example of the spatial knowledge. It is represented in the RDF statements. Each statement describes

Table 4: Number of individuals for each class.

Continent	Country	State	County	City
1	3	42	30	34
Ocean	Mountain	Lake	River	
2	2	7	6	

the ID, class type, name, description, and spatial relationships of the CSD-9/RCC-8 with other individuals.

Figure 7 shows a screenshot of the geographic information system explorer using the spatial reasoner and spatial knowledge base. A user can select a region on the left side of the screen. The right side of the screen is configured for a user to select a spatial property to query the selected region in the Google map. For example, the query "Where is a place that is bordered by Texas and is located northwest of Texas?" would respond with "New Mexico."

4. Experiment

An experiment was conducted to evaluate the performance of the spatial reasoning algorithm proposed in this paper using the spatial reasoner and spatial knowledge base described previously. In each experiment, the performances of the CHOROS mode spatial reasoning algorithm (CSD/RCC) that performs only path consistency on CSD-9 and RCC-8 and the proposed algorithm (CSD/RCC+CC) that introduces the cross-consistency check in addition to the CSD/RCC were compared.

In the first experiment, the two algorithms were compared in terms of the amount of new knowledge derived from the inference results from a quantitative aspect. The experiment result is shown in Figure 8. The graphs clearly demonstrate that the volume of new knowledge generated by the reasoning algorithm (CSD/RCC+CC) proposed in this paper has increased to a substantially greater extent, compared to the CHOROS mode spatial reasoning algorithm (CSD/RCC), as the scale of spatial knowledge base increases according to the reasoning inputted. This result indicates that the proposed algorithm is considerably better in terms of inference capability that derives new knowledge compared to the CHOROS mode algorithm.

In the second experiment, the two algorithms were compared in terms of reasoning response time. The experiment result is shown in Figure 9. As shown in the figure, the proposed spatial reasoning algorithm (CSD/RCC+CC) reasoning response time increased faster compared to the CHOROS mode spatial reasoning algorithm (CSD/RCC) as the size of the spatial knowledge base increased. This result can be explained by the following. First, the proposed algorithm requires more computation effort for the cross-consistency check compared to the CHOROS mode algorithm. Further, as the amount of spatial knowledge derived from inference increases, the composition and conversion between items of knowledge must be performed more frequently. As a result the total number of iterations performed by the algorithm increases.

```
<owl:NamedIndividual
rdf:about="&spatial;101001011">
<rdf:type rdf:resource="&spatial;State"/>
<spatial:hasName
rdf:datatype="&xsd;string">California
</spatial:hasName>
<rdfs:comment>California</rdfs:comment>
<spatial:westOf
rdf:resource="&spatial;101001012"/>
<spatial:disconnectedFrom
rdf:resource="&spatial;101001016"/>
```

Figure 6: Example of spatial knowledge.

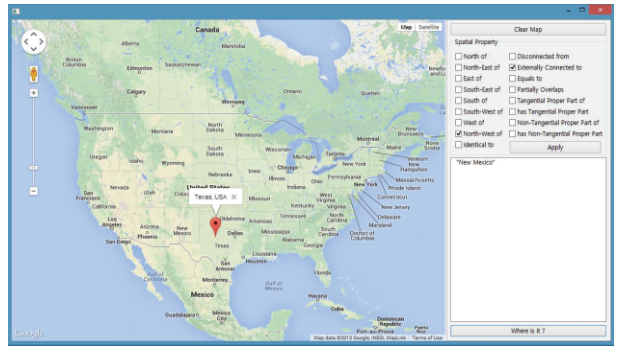
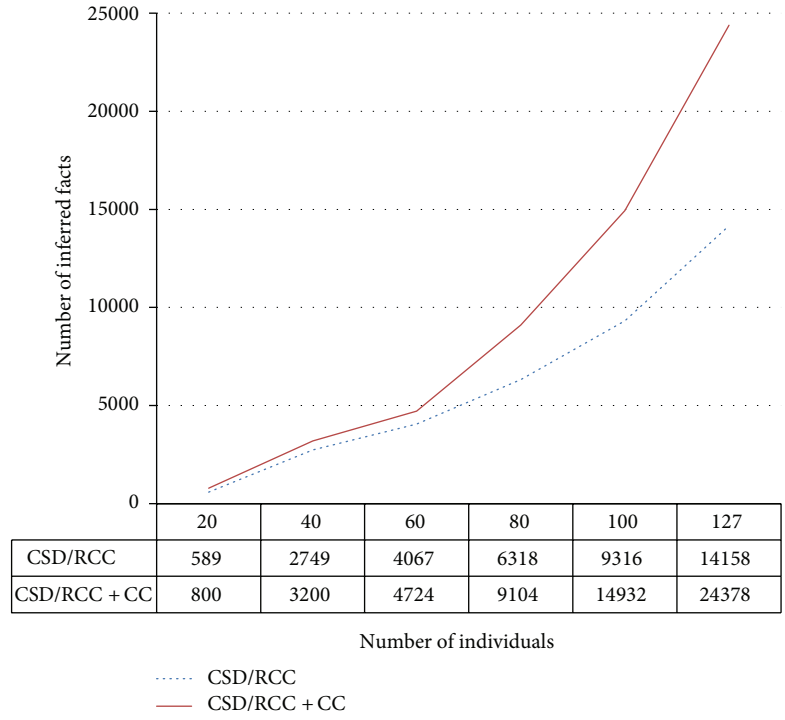


Figure 7: A screenshot of the GIS Explorer.

In the third experiment, the two algorithms were compared in terms of certainty of the inferred knowledge. The experiment result is shown in Figure 10. Certainty of the inferred knowledge was measured by a ratio of defined relations to the derived relations. As shown in the figure, the proposed spatial reasoning algorithm (CSD/RCC+CC) had a higher level of certainty than the CHOROS mode spatial reasoning algorithm (CSD/RCC). This result shows that the proposed algorithm has superior inference capability compared to the CHOROS mode algorithm even though more pieces of knowledge were derived.

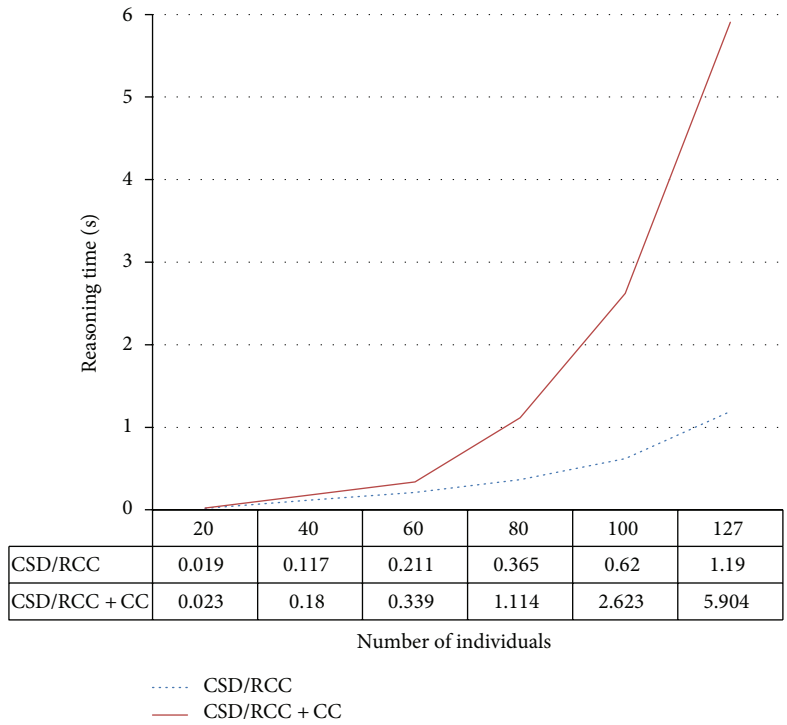
5. Conclusion

CSD- (cone-shaped directional-) 9 and RCC- (region connection calculi-) 8 are well-known qualitative spatial formalisms for representing and reasoning about the directional and the topological relationships between spatial entities. Based upon CSD-9 and RCC-8, we proposed an efficient algorithm for reasoning about directional and topological relationships between spatial entities. In contrast to previous ones, our spatial reasoning algorithm contains not only path-consistency checking over the CSD-9 directional or RCC-8 topological relation set, but also cross-consistency checking between two different sets, to produce the complete entailments. In order to investigate the applicability of our spatial reasoner, we developed a geographic information system (GIS) explorer as an application. In this application, our reasoner performs effective *knowledge materialization* to enrich the initial spatial knowledge base and then to enable more rich answers provided to the given queries. We also



	20	40	60	80	100	127
CSD/RCC	589	2749	4067	6318	9316	14158
CSD/RCC + CC	800	3200	4724	9104	14932	24378

Figure 8: Number of inferred facts.



	20	40	60	80	100	127
CSD/RCC	0.019	0.117	0.211	0.365	0.62	1.19
CSD/RCC + CC	0.023	0.18	0.339	1.114	2.623	5.904

Figure 9: Reasoning time.

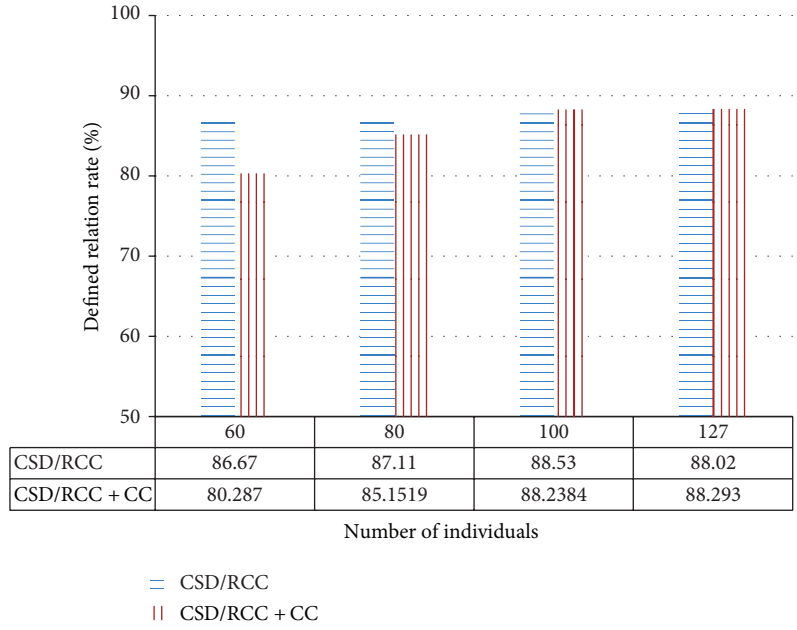


Figure 10: Ratio of defined relations.

ensured the high performance of our spatial reasoner through quantitative experiments performed on a large knowledge base. In the experiments, our reasoner showed better inference capability that derives new knowledge compared to the existing ones, such as CHOROS.

In future, our research highlights would be focused on the following points. On the one hand, we will optimize and scale up our spatial reasoner so as to further improve the performance. On the other hand, our system will be extended for many useful applications such as knowledge management [11], Semantic Web [12, 13], geographical information systems (GIS) [14], intelligent tutoring [15], robot navigation, and spatial planning.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the IT R&D program of MSIP/KEIT (10044494, WiseKB: Big data based self-evolving knowledge base and reasoning platform).

References

[1] O. Akcay and O. Altan, "Spatial relations and inferences for context aware visualization," in *Proceedings of the 14th International Symposium for Spatial Data Handling*, pp. 484–486, 2010.

[2] I. Seifert, S. Bertel, and T. Barkowsky, "Assistance for spatio-temporal planning in ubiquitous computing environments based on mental models," in *Proceedings of the International Workshop on Artificial Intelligence in Mobile Systems*, pp. 9–14, 2004.

[3] F. Dylla and M. Bhatt, "Qualitative spatial scene modeling for ambient intelligence environments," in *Proceedings of the 1st International Conference on Intelligent Robotics and Applications*, pp. 716–725, 2008.

[4] J. Chen, A. G. Cohn, D. Liu, S. Wang, J. Ouyang, and Q. Yu, "A survey of qualitative spatial representations," *The Knowledge Engineering Review*, vol. 30, no. 1, pp. 106–136, 2015.

[5] D. J. Peuquet and Z. Ci-Xiang, "An algorithm to determine the directional relationship between arbitrarily-shaped polygons in the plane," *Pattern Recognition*, vol. 20, no. 1, pp. 65–74, 1987.

[6] J. Renz, "Maximal tractable fragments of the region connection calculus: a complete analysis," in *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI '99)*, pp. 448–454, August 1999.

[7] S. Batsakis and E. G. Petrakis, "SOWL: a framework for handling spatio-temporal information in OWL 2.0," in *Proceedings of the 5th International Symposium on Rule-Based Reasoning, Programming, and Applications (RuleML '11)*, pp. 242–249, 2011.

[8] M. Stocker and E. Sirin, "PelletSpatial: a hybrid RCC-8 and RDF/OWL reasoning and query engine," in *Proceedings of the 6th International Workshop on OWL Experiences and Directions*, 2009.

[9] G. Christodoulou, E. G. Petrakis, and S. Batsakis, "Qualitative spatial reasoning using topological and directional information in OWL," in *Proceedings of IEEE the 24th International Conference on Tools with Artificial Intelligence*, pp. 596–602, 2012.

[10] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean, "SWRL: a semantic web rule language combining OWL and RuleML," W3C member submission , 2004.

[11] M. Brahami, B. Atmani, and N. Matta, "Dynamic knowledge mapping guided by data mining: application on healthcare," *Journal of Information Processing Systems*, vol. 9, no. 1, pp. 1–30, 2013.

[12] H. Hsueh, C. Chen, and K. Huang, "Generating metadata from web documents: a systematic approach," *Human-Centric Computing and Information Sciences*, vol. 3, no. 7, pp. 1–17, 2013.

[13] R. Y. Shtykh and Q. Jin, "A human-centric integrated approach to web information search and sharing," *Human-Centric Computing and Information Sciences*, vol. 1, no. 2, pp. 1–37, 2011.

[14] S.-Y. Noh and S. K. Gadia, "A model comparison for spatiotemporal data in ubiquitous environments: a case study," *Journal of Information Processing Systems*, vol. 7, no. 4, pp. 635–652, 2011.

[15] E. Elsayed, K. Eldahshan, and S. Tawfeek, "Automatic evaluation technique for certain types of open questions in semantic learning systems," *Human-Centric Computing and Information Sciences*, vol. 3, article 19, 2013.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 108286, 9 pages
http://dx.doi.org/10.1155/2014/108286

Research Article

Framework of Resource Management for Intercloud Computing

Mohammad Aazam and Eui-Nam Huh

Computer Engineering Department, Kyung Hee University, Suwon 446-701, Republic of Korea

Correspondence should be addressed to Eui-Nam Huh; johnhuh@khu.ac.kr

Received 19 June 2014; Revised 18 August 2014; Accepted 18 August 2014; Published 11 September 2014

Academic Editor: Jong-Hyuk Park

There has been a very rapid increase in digital media content, due to which media cloud is gaining importance. Cloud computing paradigm provides management of resources and helps create extended portfolio of services. Through cloud computing, not only are services managed more efficiently, but also service discovery is made possible. To handle rapid increase in the content, media cloud plays a very vital role. But it is not possible for standalone clouds to handle everything with the increasing user demands. For scalability and better service provisioning, at times, clouds have to communicate with other clouds and share their resources. This scenario is called Intercloud computing or cloud federation. The study on Intercloud computing is still in its start. Resource management is one of the key concerns to be addressed in Intercloud computing. Already done studies discuss this issue only in a trivial and simplistic way. In this study, we present a resource management model, keeping in view different types of services, different customer types, customer characteristic, pricing, and refunding. The presented framework was implemented using Java and NetBeans 8.0 and evaluated using CloudSim 3.0.3 toolkit. Presented results and their discussion validate our model and its efficiency.

1. Introduction

Digital media have convincingly surpassed traditional media, as a result of which this trend makes big and possibly long-term changes to the content being exchanged over the Internet. The global Internet video traffic had surpassed global peer-to-peer (P2P) traffic in 2010 [1]. Excluding the amount of video exchanged through P2P file sharing, at the time being, Internet video is 40 percent of consumer Internet traffic. Since 2012, it has surpassed the mark of 50 percent and will reach 62 percent by the end of 2015. If all forms of videos are counted, the number will be approximately 90 percent by 2015 [2]. To meet the great opportunities and challenges coming along with media revolution, sophisticated technology and better facilities with more powerful capabilities have become the most urgent demands.

Cloud computing recently has emerged and advanced rapidly as a promising and inevitable technology. Cloud computing platform provides highly scalable, manageable, and schedulable virtual servers, storage, computing power, virtual networks, and network bandwidth, according to user's requirement and affordability [3, 4]. Therefore, it can provide solution package for the media revolution, if wisely designed for media cloud and deployed and integrated with the advanced technologies on media processing, transmission, and storage, keeping in view the industrial and commercial trends and models as well [5]. An average user generates content very quickly, until he/she runs out of storage space. Most of the content may be used frequently by the user, which requires to be accessed easily. Media management is among the key aspects of cloud computing, since cloud makes it possible to store, manage, and share large amounts of digital media.

Cloud computing is a handy solution for processing content in distributed environments. Cloud computing provides ubiquitous access to the content [6], without the hassle of keeping large storage and computing devices. Sharing large amount of media content is another feature that cloud computing provides [7]. Other than social media, traditional cloud computing provides additional features of collaboration and editing of content. Also, if content is to be shared, downloading individual files one by one is not easy. Cloud computing caters to this issue, since all the content can be accessed at once by other parties, with whom the content is being shared.

The increasing demands in cloud computing arena has resulted in more heterogeneous infrastructure, making interoperability an area of concern. Due to this, it becomes a challenge for cloud customers to select appropriate cloud service provider (CSP) and hence it ties them to a particular CSP. This is where intercloud computing comes into play. Although intercloud computing is still in its infancy, its purpose is to allow smooth interoperability between clouds, regardless of their underlying infrastructure. This allows users to migrate their workloads across clouds easily. Cloud brokerage is a promising aspect of intercloud computing [8, 9].

Most of the data-intensive applications are now deployed on the clouds. These applications, storage, and data resource are so diversely located that they have to reach even cross-continental networks. Due to this, performance degradation in networks affects the performance of cloud systems and user requests. To ensure service quality, especially for bulk-data transfer, resource reservation and utilization become a critical issue [10].

Previous works mainly focus on integrated and collaborative uses of resources to meet application requirements. They do not focus on bulk-data transfer consistency and efficiency. They assume that all resources are connected by high-speed stable networks. Continuously growing cloud market faces new challenges now. Even though users have well collaborated end systems and resources are allocated according to their needs, still, bulk-data transfer for cross-continental users in remote places might create performance bottleneck. For instance, multimedia services like IP Television (IPTV) rely on availability of sufficient network resources and hence they have to be operated within the limitation of time constraints [10–12].

The findings in the study presented by Deelman et al. support the fact that right amount of resource allocation can reduce significant amount of cost, without affecting the performance [13]. In user's perspective, fairness is a key concern in resource allocation and pricing. Current pay-as-you-go billing mechanisms typically charge the users on hourly basis, which incurs unfairness [14]. As cloud computing is an economically driven paradigm, fairness becomes an important feature in a pricing scheme. In terms of economics, pricing fairness has two types: personal fairness and social fairness. Personal fairness is subjective and means that the pricing should be affordable for the users. Social fairness, on the other hand, means how much fairness was adopted among the users, using the same service. Having the same cost for the same services utilized is known as social fairness. An unfair pricing develops dissatisfaction in users and eventually the service provider loses its customers [14].

Cloud computing still faces some open challenges, but to provide better reliability, availability, cost-efficiency, and QoS, intercloud computing has already been envisioned. Research on intercloud computing is still not mature enough, but its effectiveness cannot be denied by any means [15]. CSPs have their customers dispersed all around the globe. To serve them optimally, CSPs have to set up many of their data centers at different geographical locations. Existing systems are not capable enough of coordinating dynamically the load distribution among data centers to determine optimal location for hosting services to achieve desired performance. Furthermore, users' geographical distribution cannot be predicted as well. Hence, load coordination as well as service distribution has to be done automatically. Intercloud computing is meant to counter this problem. It provides scalable provisioning of services with consistent performance, under variable workload and dynamically changing requirements. It supports dynamic expansion and contraction of resources to handle abrupt variations in service demands [9]. Broker's responsibility is to identify appropriate CSP, according to the needs of its customer, through cloud exchange. Broker negotiates with the gateway to allocate resources, according to user and service requirements [16].

When there is no broker and user interaction with the CSP directly, user has to decide what and how many instances to be reserved. When there is an intermediary (broker) between CSP and the user, then broker asks its users to decide on resource reservation and helps them with this decision making [17]. With cloud federation, where multiple independent cloud service providers cooperate with each other, a customer's request from one CSP is entertained by another CSP through the mediation of an intermediary, termed as cloud broker [18]. The broker has to manage resource allocation, which can be tricky at times. This resource allocation can be ad hoc or advanced. But broker has to intelligently decide what to do based on customer's behavior/characteristic and type and price of service(s) [19]. For some services, broker performs ad hoc reservation of resources, while, in some cases, advanced reservation of resources is to be done. In this case, broker has to decide what resources are to be reserved and to what extent [19]. CSPs prefer to have the information about resource reservation in advance. This allows them to adapt to the demands of service and user and allows them to do capacity planning in a better way [17]. For on-demand reservation of resources, CSPs have to manage everything at the time the demand has been made. This makes management of resources difficult for the CSPs and hence reservation is more costly for the users.

There is no complete brokerage model present so far that could handle all the important tasks, from resource prediction, reservation, billing, and refunding. For example, Amazon Elastic Compute Cloud (EC2) charges its users on hourly basis. Thus, if a resource is used for few minutes, it would still charge the user on hourly basis, as if that instance was running for an hour. This is certainly not appreciated by the users. This becomes more of an issue when billing cycle is even longer, like, daily basis (e.g., VPS.NET [20]) [17]. Since the demands are dynamic and may change at any time, it remains a challenge for the broker to decide what amount of resources has to be reserved on-demand [17]. In this paper, we address this issue by presenting a resource management model for intercloud broker.

In rest of the paper, Section 2 is on the already done works in this regard and their shortcomings. Section 3 explains intercloud broker. Section 4 presents broker's resource management model. Section 5 is on performance evaluation. Section 6 concludes the paper.

2. Related Work

Intercloud computing is still in its start due to which there is no standard architecture available for data communication, media storage, compression, and media delivery. Already done studies mainly focus on presenting architectural blueprints for this purpose. In Intel-HP Viewpoint paper [21], industrial overview of the media cloud is presented. The authors state that media cloud is the solution to suffice the dramatically increasing trends of media content and media consumption. For media content delivery, Quality of Service (QoS) is going to be the main concern. We discuss it in detail by presenting end to end QoS provisioning mechanism using flow label of IPv6 and multiprotocol label switching (MPLS) in [22] and impact of IPv6 tunneling in clouds in [23]. To reduce delay and jitter of media streaming, better QoS is required, for which [24] proposes media-edge cloud (MEC) architecture. The authors present in this paper that an MEC is a cloudlet which locates at the edge of the cloud. MEC is composed of storage space, central processing unit (CPU), and graphics processing unit (GPU) clusters. The MEC stores, processes, and transmits media content at the edge, thus incurring a shorter delay. In turn the media cloud is composed of MECs, which can be managed in a centralized or peer-to-peer (P2P) way.

Ferretti et al. [25] present an approach to use a pair of proxies, a client proxy at the user's side and a server proxy at the cloud side, to integrate the cloud seamlessly to the wireless applications. Diaz-Sanchez et al. [26] and Díaz-Sánchez et al. [27] also present proxy as a bridge, for sharing the contents of home cloud to other home clouds and to the outside, public media clouds. This proxy can do additional task of indexing the multimedia content, allowing public cloud to build search database and content classification. Media cloud can then provide discovery service to the users to search the content of their interest. Huang et al. [28] also present a proxy scheme for transcoding and delivery of media. On the other hand, Jin and Kwok [29] propose usage of P2P for delivering media stream outside the media cloud. In both cases, it builds a hybrid architecture, which includes P2P as well as media cloud.

Transcoding and compression of media content require a lot of resources. Pereira et al. present an architecture in [30, 31], in which MapReduce model is applied for this purpose, in private and public clouds.

Feng et al. [32] have proposed the concept of stream-oriented cloud and stream-oriented object. The authors introduce stream-oriented cloud with a high-level description. Mahajan et al. [33], Lim and Lee [34], and Ahmed et al. [35] discuss load balancing among virtual machines and applications.

Rogers and Cliff present [19] a resource allocation mechanism, but resource prediction and detailed billing, along with refunding issue, are not considered (see Algorithms 1 and 2). Ki-Woong et al. [36] present a billing system with some security features. To resolve different types of disputes in the future, a mutually verifiable billing system is presented. Their work only focuses on the reliability of transactions made in purchasing and consuming resources. They do not focus on the overall resource management, pricing, refunding, or similar important features of cloud broker.

Wang et al. [17] propose a brokerage service for reservation of instances. The authors propose a brokerage service for on-demand reservation of resources, for IaaS clouds. Their work is limited to only on-demand jobs and they do not present anything beyond that. Jrad et al. present a generic architecture of broker. They present how broker handles service level agreement (SLA) management and interoperability of resources [8, 9]. Yang et al. present resource allocation algorithm in a simplistic way [10]. Deelman et al. present performance tradeoffs of different resource provisioning plans. They also present tradeoffs in terms of storage fee of Amazon S3 [13]. Their work does not take into account pricing strategies and other resource management tasks. Ibrahim et al. present the concept of fairness in pricing in respect of microeconomics [14]. Grozev and Buyya present basic taxonomies for intercloud architecture [15], not discussing the way broker handles resources, services, and customers. Rajkumar et al. present architectural fundamental of intercloud computing [16]. Villegas et al. present [18] how multiple clouds are influenced by creating a cloud federation environment. Their work also lacks any further discussion on brokerage.

3. Intercloud Broker

As discussed different types of media content are rapidly increasing and so are users' demands. A single cloud cannot always fulfill the requests or provide required services. There comes a situation when two or more clouds have to communicate with each other, or another intermediary comes into play and federates the resources of two or more clouds. In intercloud terminology, that intermediary is known as "cloud broker" or simply "broker." Broker is the entity which introduces the cloud service customer (CSC) to the cloud service provider (CSP) and vice versa.

Cloud broker provides a single interface through which multiple clouds can be managed and share resources. Cloud broker operates outside of the clouds and controls and monitors those clouds. The main purpose of the broker is assisting the customer to find the best provider and service, according to customer's needs, with respect to specified SLA and providing the customer with a uniform interface to manage and observe the deployed services. Cloud broker earns its profit by fulfilling requirements of both parties. Cloud broker uses a variety of methods, such as a repository for data sharing and integration across data sharing services, to develop a commendable service environment and achieve the best possible deal and SLA between two parties (i.e., CSP and CSC) [30]. Broker typically makes profit either by taking remuneration from the completed deal or by varying the broker's spread or some combination of both [37]. The spread is the difference between the price at which a broker buys from seller (provider) and the price at which it sells to the buyer (customer).

4. Broker's Resource Management Framework

One of the cloud computing core attributes is pay-as-you-go billing model. It enables the customers to scale their capacity

(1) $totalUsers = n$
(2) **for** all users i **do**
$relProb[i] \leftarrow P_i(L \mid H)$
$userCh\,[i] \leftarrow \mu_i$
(3) **Input** $pCost[i]$ for service s
(4) $rRes[i] \leftarrow c * d$
$c \leftarrow \left(pCost[i] * relProb\,[i] * n\right)$
$d \leftarrow (1 - userCh[i])$ for D_s
(5) **end for**
(6) $thresh \leftarrow \theta$
(7) **for** all users i and $relProbk$ **do**
$probSum \mathrel{+}= relProb\,[i]$
$avgRelProb \leftarrow probSum/k$
(8) **end for**
(9) **if** $thresh > avgRelProb$
$rAlloc$
(10) **end if**

Algorithm 1: Resource estimation.

(1) **for all** $relProb == l$ **do**
$$userChL \leftarrow \frac{pCost - relProbL}{pCost * (1 - relProbL)}$$
$totalPriceL \leftarrow (pCost + userChL + userCh)$
(2) **end for**
(3) **for all** $relProb == h$ **do**
$$userChH \leftarrow \frac{pCost - relProbH}{pCost * (1 - relProbH)}$$
$totalPriceH \leftarrow (pCost + userChH + userCh)$
(4) **end for**

Algorithm 2: Cost estimation.

according to their changing requirements and pay for the consumed resources.

CSCs contact broker to acquire the required service(s) at best price. Broker performs the negotiation and SLA tasks with CSP. Once the contract is settled, the service is provided to the customer. In this regard, not only does broker provide services on ad hoc basis, but also it has to predict consumption of resources, so that they can be allocated in advance, allowing more efficiency and fairness at the time of consumption. This prediction as well as preallocation of resources also depends upon user's behavior and its probability of using those resources in future.

To handle commercial services, broker has a cost management system. Broker includes application programming interfaces (APIs) and a standard abstract API, which is used to manage cloud resources from different cloud providers. Broker holds another abstract API for the negotiation of cloud service facilities with the customer. Different modules perform a specific task in broker's architecture; for example, registration of new services is handled by service registration manager. Deployment of services and making them available is done by deployment manager. Similarly, each module has its own specific utility.

For this intermediary service, broker performs pricing and billing, which is presented in this section.

We formulate the estimation of required service as

$$R_{\text{res}} = \sum_{i=0}^{n} \begin{cases} \left(C_{m_i} * n * P_i\,(L \mid H)\right) * \left(1 - \mu_i\right) * D_s \\ 0, \end{cases} \tag{1}$$

where R_{res} represents required resources mapped value, C_m is the maximum cost a particular user can afford or is willing to pay, n is the number of cloud customers, and $P(L \mid H)$ is the probability of a particular customer of giving up resources. For simplicity, we have categorized it into two probabilities, low (L) or high (H) probability. Consider

$$\begin{aligned} 0 > L \leq 0.5, \\ 0.5 > H \leq 1, \end{aligned} \tag{2}$$

where "μ" is user characteristic, a constant decision variable value, which is assigned by the broker to each user, according to its characteristic or history and D_s is duration for which a particular service has been requested.

With this formulation, cloud broker can determine future resource requirements. It is important for cloud broker to

rightly decide when to reserve resources and not to waste precious cloud resources. It will also help power consumption management, which is becoming a concern for cloud datacenters.

After predicting the future resource requirement, next task is price allocation and billing. Pricing is not a straight forward matter. There are different pricing strategies available.

Time-Based Pricing. A dynamic pricing strategy, based on a large number of data gathered from customers about their behavior and trends.

Value-Based Pricing. Pricing a service on the basis of the value it holds for the customer, instead of the cost of production.

Target-Pricing. Selling the services at a planned profit rate.

Psychological Pricing. Selling at a rate having positive psychological impact; for example, a service of \$10 is priced at \$9.99 or \$9.95.

Price Discrimination. Setting up different prices in different segments of markets, based on class, age, behavior, and so forth.

Premium Pricing. Artificially setting up the prices higher. This practice exploits customers' tendency of having perception that higher priced products are of better quality.

We formulate broker's generic pricing method as follows:

$$\begin{aligned} \rho_{S_{P(L)}} &= \int_0^t C_m + \left(\frac{\mu_L}{D_s}\right) + \left(\frac{\mu}{D_s}\right), \\ \rho_{S_{P(H)}} &= \int_0^t C_m + \left(\frac{\mu_H}{D_s}\right) + \left(\frac{\mu}{D_s}\right), \end{aligned} \tag{3}$$

where $\rho_{S_{P(L)}}$ is the price for service S, which has been requested with relinquish probability P_L. Similarly, $\rho_{S_{P(H)}}$ is the price with relinquish probability P_H, $\mu_L$is the decision variable for user P_L, calculated through (4), and μ_H is the decision variable for user P_H, calculated through (5). Consider

$$\mu_L = \frac{C_m * P_L}{\delta}, \tag{4}$$

$$\mu_H = \frac{C_m * P_H}{\delta}, \tag{5}$$

where δ represents average profit earned so far from the currently requesting CSC.

When n number of customers have requested a service S, broker has to decide whether or not to register the service. Based on the relinquish probabilities of each customer, broker makes this decision. For every service, depending upon the type, duration, and cost of service, broker sets a threshold value "θ". It then accumulates the relinquish probabilities of each subscriber (P_T). If the accumulated probability is greater than or equal to the threshold value, it registers the service; otherwise, the service is not provided, because with some services, a particular minimum number of customers are required to register. Otherwise, that service becomes very costly, either not affordable for the customer(s) or leaving broker and service provider with a very low margin of profit. This task is formulated as follows:

$$P_T = \overline{x}\left(\sum_{i=0}^{n} P(L \mid H)_i\right). \tag{6}$$

Service is provided if $P_T > \theta$.

When the service is being utilized, customer can decide to discontinue at any stage. At that time, broker has to halt the service and refund the remaining amount to the customer. In this case, broker has to take into account the utilized resources, or consumed services, and the remaining service value of the decided total initial service. This can be formulated through the following equations:

$$\Upsilon_t = \Upsilon_{\text{un}} + \Upsilon_{\text{deg}}, \tag{7}$$

$$R_{\text{un}} = 1 - \frac{\alpha}{100}, \tag{8}$$

where Υ_t is the total amount to be refunded, Υ_{un} is the refund amount of unutilized resources, and Υ_{deg} is the refund amount to be paid on quality degradation. During service delivery, it is not always possible to deliver the service exactly according to the promise made during SLA. Υ_{un} is calculated through (9) to (12) and Υ_{deg} is further calculated through (13).

In (8), R_{un} represents unutilized resources, while α represents utilized resources.

For those customers who have used more service, broker and service provider have earned more money from them. Therefore, when they quit the service, they can be provided with some appreciation amount $\Upsilon_{\text{un}A}$ while refunding. We call it appreciation index ω. For example, customers who have used 60% or more of the service are eligible for this. In that case, the refund amount should linearly increase, encouraging the customer, which in turn works as an appealing factor and allows customer to return to that service provider again and again and consume more services. In this case, the formula would be

$$\Upsilon_{\text{un}A} = R_{\text{un}} * C_m + \omega, \tag{9}$$

$$\omega = \log \alpha, \tag{10}$$

$$\Upsilon_{\text{un}D} = R_{\text{un}} * C_m + \varepsilon, \tag{11}$$

$$\varepsilon = \ln\left(\frac{\alpha}{100}\right). \tag{12}$$

ε is the depreciation index, which deducts some amount, based on business policy, from those customers who used fewer services. Currently in our model, we apply depreciation index when resource utilization is less than 60%:

$$\Upsilon_{\text{deg}} = \left(\frac{Q_{\text{SLA}}}{Q_a} * \frac{\beta}{100}\right) * \left(C_m - \Upsilon_{\text{un}}\right), \tag{13}$$

where Q_a is the acquired QoS, Q_{SLA} is the promised QoS, during SLA, and β is the ratio of refund amount, set by the broker, based on business contract and condition (e.g., 10% of the total amount).

Table 1: Simulation setup.

System	Intel(R) Core(TM) i5-M430, 2.27 GHz
Memory	4 GB
Simulator	CloudSim 3.0.3
Operating system (OS)	Window 7 Home Premium

Table 2: Parameters' setting for evaluation.

Parameters	Range
Service level agreement (Q_{SLA})	0.9
Acquired service quality (Q_a)	0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1
Service price (ρ)	100, 150, 200, 250, . . . , 1000
Service utilization	40%, 50%, 60%, 70%, 80%, 90%
Refund (Υ)	30%~90%
SLA based refund (Υ_{deg})	0.8~0.1
Number of registered services	10
Service duration (D_s) in months	1, 2, 3, 4, 5, 6

5. Efficiency and Performance Evaluation

In this section, we present efficiency of our proposed refundable service model. We defined our service model through authentic algorithm to evaluate the effectiveness in cloud computing business. Our main objective is to observe the influence of performance factors on the systems and test the feasibility of our method.

5.1. Experimental Setup. The experiment environment for our model's evaluation is shown in Table 1. We have considered different parameters to estimate the required resources, pricing, and billing for two types of users and refund amount calculation. Table 2 shows the parameters' setting.

5.2. Resource Prediction according to User Type for Different Services. When different CSCs are requesting a particular service, the CSP or broker has to analyze what number of resources has to be allocated for that service, based on the type of customer. For CSCs having low relinquishing probability, priority in resource allocation is given. Figure 1 shows the unit of resources being predicted for both types of customers, for all types of registered services.

The unit is greater for L customers, while it is smaller for H customers, because of their behavior. Since there are more chances of an H customer to relinquish the service(s), as a result, more priorities and quality are provided to the more loyal customer, having L probability. Resources increase as the service price increases.

5.3. Service Price Based on Customer's Characteristic. According to the cloud resource consumer's characteristic, service price might be different. Those users who have a trend to give up their services, their final service price would be higher, compared to the other types of customers who always utilized

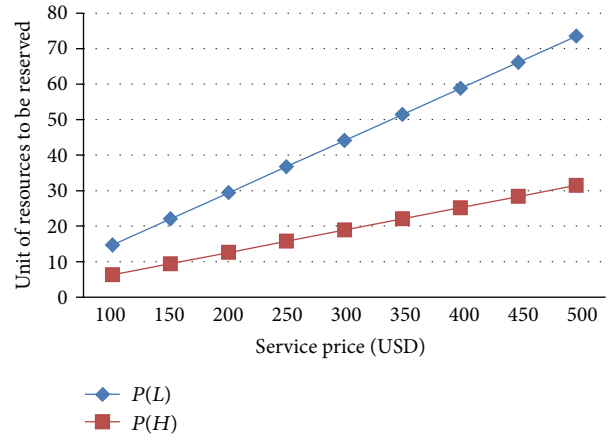


Figure 1: Resource prediction for different types of CRCs, for different requested services.

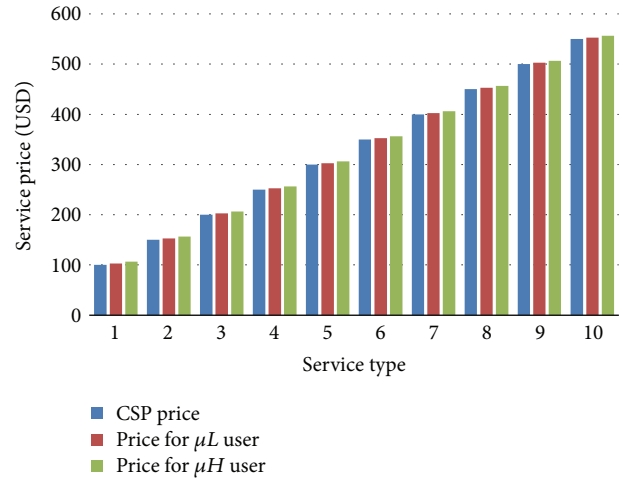


Figure 2: Service price based on user's characteristic.

their reserved resources. From Figure 2, we can see the actual service price from CSP is USD 100. However, users L (utilizing most of their reserved services) need to pay around USD 102, since they will get some sort of advantage (e.g., refundable service) from cloud broker. In addition, the users H (higher tendency to give up their service) need to pay around USD 106, because cloud broker has enough risk to pay back for unutilized resources. Therefore, those users are assigned a greater amount of the service price. Similarly, prices vary for both types of users, for each service ranging from USD 100 to USD 400. In every case, H users' prices have always been greater than the actual price as well as L price.

5.4. Refund Amount Based on Service Utilization. In this part of evaluation, reserved services' prices range from USD 100 to USD 500 and service utilization ranges from 50% to 90% of total reservation. Service quality is kept constant here at 0.9 (SLA fulfilled).

Figure 3 shows refund amount for the service reserved, on the basis of resources consumed. As the service price

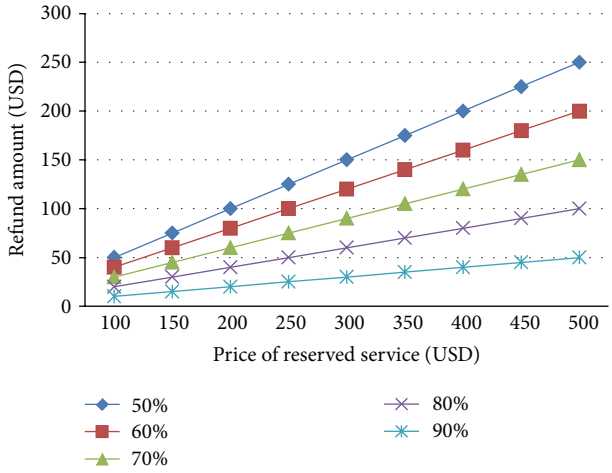


Figure 3: Reserved service price and refund amount.

increases, the refund amount also increases linearly. Similarly, the more the resource consumption is, the less the refund amount to be paid back to the customer would be. The graph shows that, for a USD 100 service, refund amount decreases on the basis of the amount of resources consumed. For 50% resource consumption, refund amount is higher, for 60%, it decreases comparatively, and so on. Similarly, for a more expensive service, the refund amount is increasing.

5.5. Appreciated Refund Amount for Different Resource Consumption with Fixed Service Quality and Resource Reservation. In this part, service utilization ranges from 30% to 90%. With fixed resource reservation and reasonable service quality (0.7-0.8 achieved SLA), impact on refund amount is measured. When cloud broker and CSP reserve resources or provide services to the CRC, they expect that the service would be completely utilized. Otherwise, they have to pay the customer back the remaining amount, which is a kind of loss for them. In this case, CSP and broker have to appreciate those customers who utilized more of their subscription and depreciate those who did not. In this part, we have set the level of decision making for appreciation or depreciation at 60% resource consumption. If a CRC consumes less than 60% of its reserved resources, it will get a depreciated refund amount. If it consumes 60% or more, it will receive appreciated refund amount.

As shown in Figure 4, CSC consuming 30% resources of its reserved service which values USD 100 will receive around USD 68, instead of USD 70. This is because the utilized service is less than 60%, since it was depreciated. Similarly, for 40% resource consumption, refund amount is USD 59. On the other hand, for 60% resource consumption, refund amount is around USD 42, instead of USD 40, since it has been appreciated. Similarly, for 90% resource consumption, refund is around USD 12, instead of USD 10. By this, broker motivates its customers to consume more resources or complete the subscription.

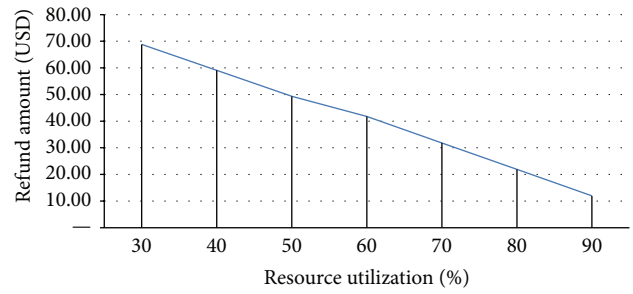


Figure 4: Appreciated refund amount according to resource utilization.

6. Conclusion and Future Work

Intercloud computing as well as brokerage is still in its infancy. With rapidly increasing multimedia content in the cloud, QoS, efficiency, and user's satisfaction are becoming a crucial task. The future is of intercloud computing, in which two or more clouds would be interacting with each other for resource management and service provisioning. Brokerage is one of the key aspects of intercloud computing. Broker has to deal with all types of resource management and service availability tasks for its customers. In this case, a complete model that takes care of efficient resource allocation, pricing, billing, and refunding is not yet available. We have tried our best to put all these things under one umbrella and presented a framework. Based on 6 parameters (resource prediction according to user characteristic, pricing according to user characteristic, pricing according to service type, refund according to utilization, appreciated refund, and refund according to service type), the model is evaluated and results are presented. The tested evaluation of this model shows the validity and efficient performance of our proposed model. We intend to extend this part now and work on more varied parameters, under more heterogeneous environment, where different types of devices are being used by customers and diverse services are requested.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2014 (H0301-14-1020)) supervised by the NIPA (National IT Industry Promotion Agency). The corresponding author is Professor Eui-Nam Huh. This research was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (no. NRF-2013R1A1A2013620). The corresponding author is Professor Eui-Nam Huh.

References

[1] M. Tan and X. Su, "Media cloud: when media revolution meets rise of cloud computing," in *Proceedings of the 6th IEEE International Symposium on Service-Oriented System Engineering (SOSE '11)*, pp. 251–261, Irvine, Calif, USA, December 2011.

[2] "Cisco-White-Paper," Cisco Visual Networking Index—Forecast and Methodology, 2010–2015, June 2011.

[3] S.-H. Lee and I.-Y. Lee, "A secure index management scheme for providing data sharing in cloud storage," *Journal of Information Processing Systems*, vol. 9, no. 2, pp. 287–300, 2013.

[4] K.-C. Huang, T.-C. Huanga, M.-J. T. H.-Y. Chang, and Y.-H. Tung, "Moldable job scheduling for HPC as a service with application speedup model and execution time information," *Journal of Convergence*, vol. 4, no. 4, 2013.

[5] J. C. Tsai and N. Y. Yen, "Cloud-empowered multimedia service: an automatic video storytelling tool," *Journal of Convergence*, vol. 4, no. 3, pp. 13–19, 2013.

[6] J. W. K. Gnanaraj and K. E. E. B. Rajsingh, "Smart card based time efficient authentication scheme for global grid computing," *Human-centric Computing and Information Sciences*, vol. 3, article 16, 2013.

[7] K. Chorianopoulos, "Collective intelligence within web video," *Human-Centric Computing and Information Sciences*, vol. 3, p. 10, 2013.

[8] F. Jrad, J. Tao, and A. Streit, "SLA based service brokering in intercloud environments," in *Proceedings of the 2nd International Conference on Cloud Computing and Services Science (CLOSER '12)*, pp. 76–81, Porto, Portugal, April 2012.

[9] F. Jrad, J. Tao, and A. Streit, "Simulation-based evaluation of an intercloud service broker," in *Proceedings of the 3rd International Conference on Cloud Computing, GRIDs, and Virtualization*, Nice, France, 2012.

[10] Y. Yang, Y. Zhou, L. Liang, D. He, and Z. Sun, "A sevice-oriented broker for bulk data transfer in cloud computing," in *Proceedings of the 9th International Conference on Grid and Cloud Computing (GCC '10)*, pp. 264–269, Nanjing, China, November 2010.

[11] G. Christou, "A comparison between experienced and inexperienced video game players' perceptions," *Human-Centric Computing and Information Sciences*, vol. 3, article 15, 2013.

[12] H. Yo-Sung, "Challenging technical issues of 3D video processing," *Journal of Convergence*, vol. 4, no. 1, 2013.

[13] E. Deelman, G. Singh, M. Livny, B. Berriman, and J. Good, "The cost of doing science on the cloud: the montage example," in *Proceedings of the ACM/IEEE Conference on Supercomputing*, IEEE Press, November 2008.

[14] S. Ibrahim, B. He, and H. Jin, "Towards pay-as-you-consume cloud computing," in *Proceedings of the IEEE International Conference on Services Computing (SCC '11)*, pp. 370–377, Washington, DC, USA, July 2011.

[15] N. Grozev and R. Buyya, "Inter-Cloud architectures and application brokering: taxonomy and survey," *Software: Practice and Experience*, vol. 44, no. 3, pp. 369–390, 2014.

[16] B. Rajkumar, R. Ranjan, and R. N. Calheiros, "Intercloud: utility-oriented federation of cloud computing environments for scaling of application services," in *Algorithms and Architectures for Parallel Processing*, Springer, Berlin, Germany, 2010.

[17] W. Wang, D. Niu, B. C. Li, and B. Liang, "Dynamic cloud resource reservation via cloud brokerage," in *Proceedings of the 33rd IEEE International Conference on Distributed Computing Systems (ICDCS '13)*, pp. 400–409, Philadelphia, Pa, USA, 2013.

[18] D. Villegas, N. Bobroff, I. Rodero et al., "Cloud federation in a layered service model," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1330–1344, 2012.

[19] O. Rogers and D. Cliff, "A financial brokerage model for cloud computing," *Journal of Cloud Computing*, vol. 1, no. 1, pp. 1–12, 2012.

[20] VPS.NET, http://vps.net.

[21] "Moving to the media cloud," Viewpoint Paper, Intel-HP, Palo Alto, Calif, USA, 2010.

[22] M. Aazam, A. M. Syed, and E.-N. Huh, "Redefining flow label in IPv6 and MPLS headers for end to end QoS in virtual networking for thin client," in *Proceedings of the 19th Asia-Pacific Conference on Communications (APCC '13)*, pp. 585–590, Bali, Indonesia, 2013.

[23] M. Aazam and E.-N. Huh, "Impact of ipv4-ipv6 coexistence in cloud virtualization environment," *Annals of Telecommunications*, 2013.

[24] W. Zhu, C. Luo, J. Wang, and S. Li, "Multimedia cloud computing," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 59–69, 2011.

[25] S. Ferretti, V. Ghini, F. Panzieri, and E. Turrini, "Seamless support of multimedia distributed applications through a cloud," in *Proceedings of the 3rd IEEE International Conference on Cloud Computing, (CLOUD '10)*, pp. 548–549, Miami, Fla, USA, July 2010.

[26] D. Diaz-Sanchez, F. Almenares, A. Marin, and D. Proserpio, "Media cloud: sharing contents in the large," in *Proceedings of the IEEE International Conference on Consumer Electronics (ICCE '11)*, pp. 227–228, Las Vegas, Nev, USA, January 2011.

[27] D. Díaz-Sánchez, F. Almenarez, A. Marín, D. Proserpio, and P. A. Cabarcos, "Media cloud: an open cloud computing middleware for content management," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 2, pp. 970–978, 2011.

[28] Z. Huang, C. Mei, L. E. Li, and T. Woo, "CloudStream: delivering high-quality streaming videos through a cloud-based SVC proxy," in *Proceeding of the 30th IEEE INFOCOM*, pp. 201–205, Shanghai, China, April 2011.

[29] X. Jin and Y.-K. Kwok, "Cloud assisted P2P media streaming for bandwidth constrained mobile subscribers," in *Proceedings of the 16th IEEE International Conference on Parallel and Distributed Systems (ICPADS '10)*, pp. 800–805, Shanghai, China, December 2010.

[30] R. Pereira, M. Azambuja, K. Breitman, and M. Endler, "An architecture for distributed high performance video processing in the cloud," in *Proceedings of the 3rd IEEE International Conference on Cloud Computing (CLOUD '10)*, pp. 482–489, Miami, Fla, USA, July 2010.

[31] R. Pereira and K. Breitman, "A cloud based architecture for improving video compression time efficiency: the split and merge approach," in *Proceedings of the Data Compression Conference (DCC '11)*, p. 471, Snowbird, Utah, USA, March 2011.

[32] J. Feng, P. Wen, J. Liu, and H. Li, "Elastic stream cloud (ESC): a stream-oriented cloud computing platform for Rich Internet Application," in *Proceedings of the International Conference on High Performance Computing and Simulation (HPCS '10)*, pp. 203–208, Caen, France, July 2010.

[33] K. Mahajan, A. Makroo, and D. Dahiya, "Round robin with server affinity: a VM load balancing algorithm for cloud based infrastructure," *Journal of Information Processing Systems*, vol. 9, no. 3, p. 379, 2013.

[34] M. Lim and Y. Lee, "A simulation model of object movement for evaluating the communication load in networked virtual environments," *Journal of Information Processing Systems*, vol. 9, no. 3, pp. 489–498, 2013.

[35] E. Ahmed, A. Akhunzada, A. Gani, S. H. Hamid, and R. Buyya, "Network-centric performance analysis of runtime application migration in mobile cloud computing," *Simulation Modelling Practice and Theory*, 2014.

[36] P. Ki-Woong, J. Han, J. W. Chung, and K. H. Park, "THEMIS: a mutually verifiable billing system for the cloud computing environment," *IEEE Transactions on Services Computing*, vol. 6, no. 3, pp. 300–313, 2013.

[37] M. Aazam and E.-N. Huh, "Inter-cloud architecture and media cloud storage design considerations," in *Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD '14)*, Anchorage, Alaska, USA, June -July 2014.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 395295, 8 pages
http://dx.doi.org/10.1155/2014/395295

Research Article

Development of Highly Interactive Service Platform for Social Learning via Ubiquitous Media

Gangman Yi[1] and Neil Y. Yen[2]

[1] *Department of Computer Science & Engineering, Gangneung-Wonju National University, Gangwon 220-711, Republic of Korea*
[2] *School of Computer Science and Engineering, University of Aizu, Tsuruga, Ikkimachi, Aizu-Wakamatsu, Fukushima 965-8580, Japan*

Correspondence should be addressed to Neil Y. Yen; neilyyen@u-aizu.ac.jp

Received 19 June 2014; Accepted 30 July 2014; Published 1 September 2014

Academic Editor: Jong-Hyuk Park

Several emerging issues concerning the development of interactive learning environment were left unsolved although e-learning has been applied for years. With several studies indicate that more interaction between students and systems increases students' level of interest and allows them to focus on learning support. Due to the way current interactive learning tools are designed, users have to wear or operate actual tools in order to carry out the required learning procedures. The use of tools for long durations of time results in user fatigue. Hence, this study incorporates the Microsoft Kinect as interactive tool for detecting gestures in the e-learning process. This study also uses the interaction method that we had developed on Facebook to interact with the proposed learning system. The experiments in this study are divided into five parts: system performance of the 3D web engine, gesture accuracy, system and gesture usability, system and gesture satisfaction, and learning satisfaction of the learner. Also, the gesture design was accepted by learners when they interacted with the learning system. Our research shows that our concept as well as the features of our system can fully support social learning and enhance interaction between users in learning environments.

1. Introduction

Social networks give structure to complex relations between individuals or organizations. There are many websites building social networks, such as Facebook, Twitter, and YouTube. Facebook is the most popular social network website in current internet world. Virtually everyone has a Facebook account and people use it every day to keep connected with their social circles [1]. Furthermore, social learning has always been the most natural way to learn new things. Through discussion, interaction, and collaboration between students, teachers, or student groups, students can construct related knowledge or experience via process of social interaction. However, traditional e-learning systems usually have less communication functionalities for supporting social learning [2]. They mainly play the role of a teaching assistant, which enables teachers to share learning content and manage classes as well as enables students to receive learning content from the system. Available methods for discussing problems or enhancing other types of interaction in traditional e-learning systems include only e-mail, short messaging, chat rooms, or forums in some advanced systems [3]. The interaction usually appears between teacher and students for questions and answers, seldom between students for problem discussions or idea brainstorming. It cannot provide an environment for fully supporting social learning. To enhance interaction and make the learning experience more diverse, we try to utilize the power of social networks. As an extension of traditional communication networks, social network sites enable users to share and discuss their interest and ideas in the network [4]. It makes users become closer to their friends and have more interaction in their daily lives. In order to utilize the power of social networks, we integrate our e-learning system with social networks to create a new learning environment, learning in social networks. From the perspective of learning, in our system, students and teachers can engage in learning activities as usual as well as share learning-related information on the social network sites through the social functionalities integrated into our system. From the perspective of social networks, user can see others' learning information as well as enhance interaction and discussion on learning topics, which achieves the goal of social learning. For

some advanced functionalities, we first collect users' social and learning information and then build a recommendation system based on that information. Then we build a tool for information visualization, called the Interactive 3D Social Graph, based on 3D and Kinematic technologies, which makes complex data more readable and clear.

The purpose of this research is to design a tool for visualizing social network data from the famous social networking website Facebook as well as propose a new interaction and navigation technique that uses Kinect to explore and interact with social networks [5]. The tool as well as the new method of interaction should help users interact and explore their social networks and also help learners experience better learning interaction. The purposes of this study are as follows:

(i) visualize data obtained from social network websites;

(ii) design a suitable and comfortable gesture for exploring and interacting using Kinect;

(iii) integrate gestures to enhance learners' level of interest and interaction in social networks.

In Section 1 of this paper, we briefly introduce the proposed interactive tool as well as its integration with e-learning. In Section 2, we introduce several types of gesture-based interactions. In Section 3, we discuss the methods used to visualize data, design interaction, and conduct exploration. In Section 4, we describe the evaluation as well as discussion of our tool and integration efforts. Section 5 is the conclusion and description of future works.

2. Related Works

2.1. Importance on Information Visualization for Interaction. Information visualization is the study of visual representations of abstract data to reinforce human cognition. The visualization of social network data as a graph has a long tradition in the social networks studies in social science. Freeman [2] gives the example of using shape, color, position, and the size of nodes to be the main features for presenting distinct view on the data. Many social networks visualizations researches applied for researcher analysis tool instead of provide it for end-users.

In recent years, with the growth of social network sites, projects about social networks visualization for the end-users showed up. Heer came out with their tool called Vizter tool [6]. Vizter tool does the visualization of one of the famous social network website, Friendster. They provided exploration tool for the users to play around their social network. The features such as connectivity highlighting, link-age views for viewing network content, X-ray mode, and profile search for exploring member profile data and visualization makes this tool is more explore-based play than analysis. The other project came from Matta and Pfeiffer [7]. They proposed social networks visualization tool for Facebook. In their research they studied the comparison between 3D visualization of social network and 2D visualization. The result showed that the visualizations of social networks can be more effective if 3D is utilized. We present tool for 3D visualization of integrated e-learning system and social network.

2.2. Application of Gestures Interaction. A gesture may be a motion of the body that contains several pieces of information [6]. It is a method of nonverbal communication in which visible bodily actions communicate particular messages, either in place of speech or together and in parallel with spoken words. Gestures include movement of the hands, face, or other parts of the body. Gestures differ from physical nonverbal communication that does not communicate specific messages, such as purely expressive displays, proxemics, or displays of joint attention [8]. Gestures allow individuals to communicate with a variety of feelings and thoughts, from contempt and hostility to approval and affection, often integrated with body language in addition to words when they speak [9]. There are two types of gestures, which are distinguished [10] below:

(i) offline gestures: those gestures that are processed after the user interaction with the object. They occur when a certain action has already occurred. An example is the gesture to activate a menu;

(ii) online gestures: direct manipulation gestures. They are used to scale or rotate a tangible object.

With the growth of technology nowadays, it is possible for us to perform gestural interaction even without physical contact with the device or object, which is called touch-less gestural interaction. Touch-less gestural interaction enables users to explore and control multimedia information space and/or digital devices through body movements and gestures, without the burden of physical contact. Interaction is said to be touch-less if it can take place without mechanical contact between the human and any part of the artificial system. Touch-less gestural interaction can be multimodal. Touch-less gestural interaction in-the-small has received marginal interest by the HCI community. When using touch-less interaction technologies, effort and time can be saved [11]. From all of those devices, Kinect is the most suitable and more natural for interaction with 3D space, due to the capability of Kinect to detect whole body movement through skeletal-based algorithms. Their own body movements and gestures combined with the degree of familiarity of humans with 3D space in real world makes it easier to control the 3D space in more natural ways. Right after the Kinect was introduced in 2011, Kinect gesture-based interaction research began to grow rapidly. It is used in many areas, such as medical [12], education [13], and rehabilitation [14]. There is already a great deal of research pertaining to interaction with 3D space itself [15], gesture interaction for architectural design in 3D, and using gestures to explore 3D space [16, 17].

2.3. Towards a Highly Interactive Social Learning Paradigm. Social learning theory states that social behavior (any type of behavior that we display socially) is learned primarily by observing and imitating the actions of others. In the process of learning social behavior, we may construct corresponding knowledge. Also, social constructivism, proposed by L. Vygotsky, says that knowledge is constructed while learner interacts or cooperates with other social members in the society, instead of one-way instruction. In light of this, while

we have social behavior, we are forming some knowledge at the same time.

Distributed constructionism [18] puts emphasis on the idea that, in the activities of design and construction, there needs to have multiparticipant to construct sharing and discussion, especially in the online learning environment, which takes knowledge and cognition as the result of the interaction between individual and elements in the online learning environment. Besides, Resnick thinks that, to learn effectively in social environment, it is necessary to have collaboration to design and construct meaningful knowledge and products. According to the theories, the core concept of social learning is having collaboration and interaction with others, and constructing related knowledge at the process of those social behaviors. The concept makes individuals absorb knowledge and learning experience in social behaviors, and, different from traditional e-learning, learner can receive complete learning participating experience from course constructing to after school assessment, especially in individual learning scenery. Other research [19] showed that interactive learning and collaborative learning can give user higher learning motivation and better learning efficiency than individual learning.

3. The Scenario

The scenario of proposed work can be simply summarized as follows:

(1) one open application powered by script language is applied to facilitate the sign-in process of user's own Facebook account;

(2) the same application is then sent an acknowledgement to trigger the Facebook Graph API, which is utilized to obtain user data such as personal information, friendship, and applications, after authentication;

(3) by using user ID and access token, system can make a call to get the data that we need for our system and visualized it through Unity3D engine on Web player;

(4) after performing the visualization of the Facebook data into 3D graph based user interface, user can interact and explore it by using Kinect hand gesture through Kinect skeleton tracking that provide by Kinect

3.1. Data Acquisition. In order to visualize the data on Facebook, we first need to get the data from Facebook. As we mentioned before, Facebook has launched a software environment that provides third-party developers with the ability to create their own applications and data services on Facebook. With this service, we can easily integrate our application with Facebook or allow our application to directly access data from Facebook. This platform offers a set of programming interfaces and tools, such as social plugins, login capabilities using Facebook, Open Graph, and an SDK for accessing Facebook APIs. In order to access and use the data on Facebook, we need to create an application for obtaining the App ID/API key as well as the App Secret. These two methods are required for the SDK to obtain permission to fetch data from Facebook. There are several SDKs provided by Facebook for accessing the Facebook API, such as the Facebook SDK for PHP, the Facebook SDK for JavaScript, the Facebook SDK for iOS, the Facebook SDK for Android, and SDKs for other platforms or languages. In this study, we use the Facebook SDK for JavaScript to access data from Facebook since our application is a web application and also because it will integrate with Unity3D for 3D visualization.

In order to visualize user social networks and implement various functions from Facebook in our system, we define the types of data we need for our system below:

(i) user data:

  (a) list of friends;
  (b) personal data (about users);
  (c) list of family members;
  (d) list of photos and albums;
  (e) data regarding users' posts, statuses, and feeds;

(ii) friend data:

  (a) personal data;
  (b) list of photos and albums;
  (c) data regarding friends' posts, statuses, and feeds.

In order to obtain the data that we have listed above, we will need the user to login through our application using their Facebook account and give permission (authentication) for us to access the data. It will show up in form of a web dialog interface right after we successfully log-in from our application. The permission list is as follows:

(1) email permissions: email is a protected property and access to it must be specifically asked for and granted;

(2) extended permissions give access to more sensitive info and provide the ability to publish and delete data;

(3) extended profile properties: nonoptional permissions for access to a user's data and that of their friends;

(4) open graph permission allows your app to publish actions to the Open Graph as well as retrieve actions published by other apps;

(5) page permissions: permission related to management of Facebook Pages;

(6) public profile and friend list are basic information types available to an app.

3.2. Design of Learning Support Mechanisms. One of the goals of information visualization is to visualize information, such as social information in our system, transforming textual descriptions to visual representations, thereby facilitating the perception and handling of hidden structures from underlying datasets. We choose a graph-based layout to visualize the information because it is widely used to depict data in which information is comprised of objects and relationships between those objects, such as relations between

users. In our system, we choose to use 3D-based graph layouts because it provides more space for future growth of the graph and because humans are more familiar with 3D in real world compared to 2D, which will make exploration easier. The drawing of graphs is another challenge in information visualization. We use the Spring-Embedding (force-directed) Algorithm to draw graphs. This algorithm calculates the layout of a graph using only information contained within the structure of the graph itself, rather than relying on domain-specific knowledge. Graphs drawn by this algorithm tend to be aesthetically pleasing, exhibit symmetries, and are free of crossings. Although this algorithm is for 2D graphs, with some modifications, it still can be used for 3D graph layouts. In this system, we design several interaction and navigation features, mouse-keyboard interactions, and Kinect interactions to explore graph. For both types of interactions, we use common actions to explore 3D space, such as zooming and rotating. With mouse-keyboard interaction, we use a common method for control, such as a scroll button for zooming, right-click-hold for rotating, left-click-hold for panning actions, and W, A, S, and D for keyboard actions. We use Kinect because by using our own body movements as a controller, it is easy for users to perform control. It also provides a more immersive and realistic experience while using our system.

The other reason why we use Kinect interaction is because, in the future, we want to combine this system and virtual smart classrooms with Holodeck technology [20]. We have designed several gestures for exploring and interacting with the 3D graph. These actions are as follows.

(i) *Zooming action:* the Kinect gesture for this action is conducted in two-handed mode, where both hands, left and right, are placed in front of the body and moved farther apart from each other on the x-axis to perform a zoom-out and moved closer to each other to perform the zoom-in action.

(ii) *Rotating action:* the Kinect gesture for this action is conducted in two-handed mode by holding the right hand higher than the waist and with the left hand moving on the x-axis to control the rotation. Moving to the left causes the graph to rotate clockwise and moving it to the right causes graph to rotate counterclockwise.

(iii) *Panning action:* to perform this interactive action on Kinect, the user must first be in one-handed mode, placing the left hand in front of user's body and holding it in place for a few seconds to enter panning mode. After entering panning mode, you can move your left hand up, down, right, and left to perform the panning action.

(iv) *Select object action:* to perform this interactive action on Kinect, the user can use the object picking action by moving one hand to the object and changing the moving axis from x-y axis to z-axis. When an object is selected, the content of the object will be shown on screen.

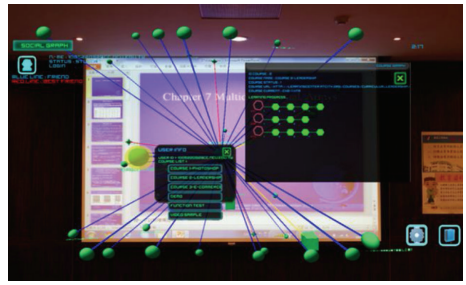


Figure 1: Social graph.

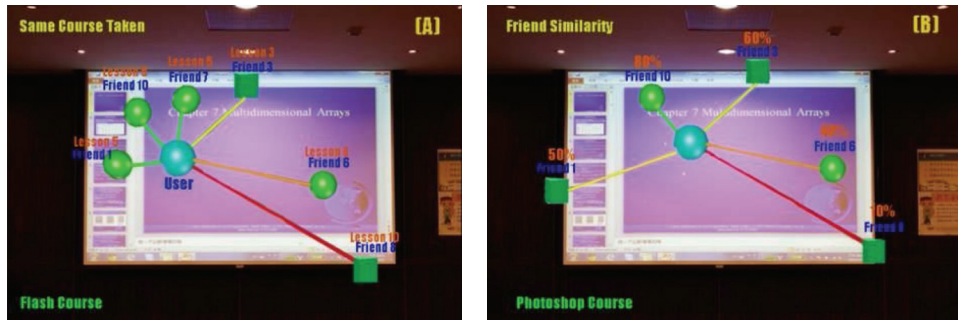
(v) *Operations and features:* our system provides several features. One of the features is called Social Graph, shown in Figure 1. This feature contains the graph of the current user as well as his social relationship. The current user node is displayed as a blue sphere, normal friends are presented as green sphere shapes with blue lines, acquaintances are displayed as box shapes with yellow lines, and best friend are represented as star shapes with red lines. Node information, including user ID, name, and taken-course list, are displayed and highlighted when the node is selected. Another feature is called Course Graph, which provides a course tree that contains learning history and scores for each lesson. This feature can help students to track their course records and histories.

We also provide the focus operation and the recommendation feature. The focus operation, shown in Figure 2, is basically the graph filter feature which can change the graph layout according to user's keywords. It makes the graph clearer to see and helps users find what they want. We designed three filter attributes for this feature, as described below.

(1) *The friend similarity operation* is a feature for finding similarities between a user and the user's friends on a learning domain according to keywords. The system will change the graph node's position and the greater the similarity the closer it will be to the user [21]. Nodes with no similarity will be removed. To make it clearer for the user, the line edge color of the user and their friends also changes based on similarity.

(2) *Same course operation*: using this feature, users can find friends with similar learning progress in a specific course. Nodes will be presented with different distances and colors based on learning progress. Friends that are currently on the same lesson as the user will be placed closer to the user. Other friends will be placed at the same distance and be given the same color line based on the lesson they took.

(3) *Friend recommendation* is a feature for helping users find new friends based on similarity of learning (courses taken, majors, and current learning progress). If people have equal or close similarity with the current user and are not friends with the current



(a) (b)

Figure 2: Example of (a) same course operation (b) friend similarity.

Figure 3: Example of recommendation operation.

user, they will show up in the list of friend recommendations.

(4) The goal of *course recommendation* is to give users suggestions regarding the next course that they can take based on their learning progress. This feature allows the user to get recommended courses from the system or from their friends. The list of course recommendations will be shown in the list (see Figure 3).

Figures, see Figures 4–8, are the proposed Kinect hand gesture implementation. As we discussed in the previous section, we divide the hand gesture interaction into two parts, exploration part and UI part. In the exploration part, user can perform zoom, panning, rotate, full screen, hide menu, mouse, click, and filter gesture. In the GUI mode user could only perform mouse, click, swipe up down, scroll up, down, and close gesture. To perform the gestures, first user needs to activate the Kinect mode by click the Kinect menu button (pink one) in the main view of our system. User needs to stand up in the front of Kinect device about 1,5 meters and system will detect and label all of the parts of human body by using skeleton tracking. Then user can start to perform the gesture to control the 3D graph and other UI in our system application. Here in Figures 4, 5, 6, 7, and 8 you can see our gestures implementation to explore and interact with the 3D graph based interface for social networks.

4. Implementation of a Facebook-Empowered Learning System

Current e-learning systems could only play the role of teaching assistance and seldom provide the social learning functionalities which makes learning not different from traditional classroom learning environment, where student feels

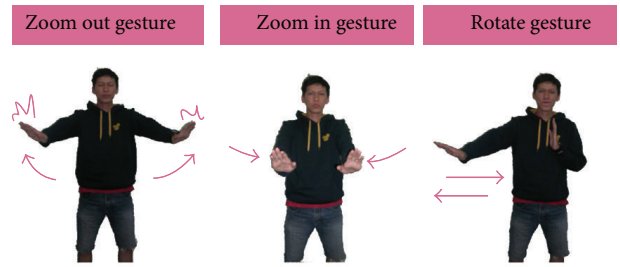


Figure 4: Gesture design of Kinect-1.

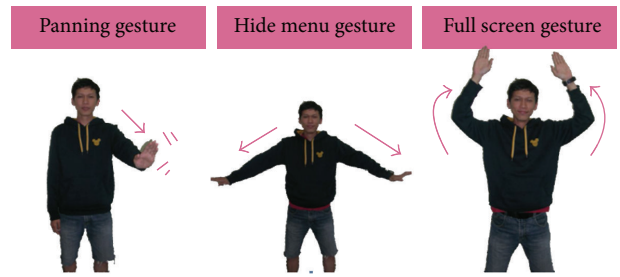


Figure 5: Gesture design of Kinect-2.

bored during the learning time. In this research, the main goal of our system is using the power of social network site to provide an open, online social learning environment that gives student the ability to have social interaction and learning behaviors on Internet at the same. The following are the specific goals of this research:

(i) integrate e-learning environment and Social Networks environment;

(ii) open course ware is a pure online learning environment, and the choice of courses or social relationships becomes one problem for user to face. We acquire and analyze user's learning portfolio from e-learning environment with social information and relationship from social networks environment, to build a recommendation system both in social networks and in learning activity parts and to help user to solve the problem of course and friendship choice.

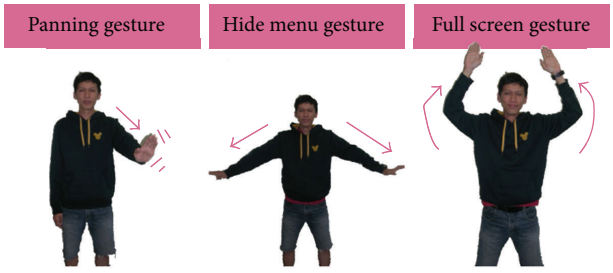


Figure 6: Gesture design of Kinect-3.

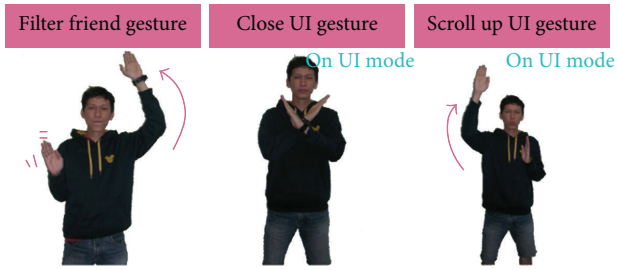


Figure 7: Gesture design of Kinect-4.

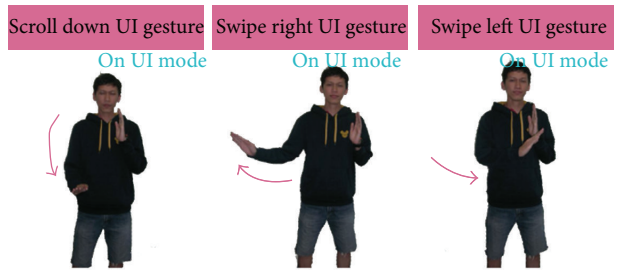


Figure 8: Gesture design of Kinect-5.

Table 1: System performance based on number of friends.

Number	Number of friends	Frame rate
1	<100	58–60 fps
2	100–300	55-56 fps
3	300–500	33–35 fps
4	>500	23–25 fps

In general, learning management system can be separated into two aspects.

4.1. Instructor Functionality. Instructor functionality is for teacher to manage learning related activities and corresponding content files; we can basically classify them into the following categories.

These functionalities can help teacher to manage the course or class state and give him the ability to monitor students or whole class learning situation, which can help teacher to adjust their instruction way or style, according to the condition of monitoring subjects.

4.2. Learner Functionality. Learner functionality is for student to participate learning activities and corresponding supportive and recording functionalities. We can basically classify them into the following categories.

These functionalities can help student access learning content, including course pages, assignment or assessment, and their related activities/behaviors support. One of the main functionalities of learning management system is the learning activities/behaviors monitoring and recording, and it will construct the learning portfolio of the student. With the help of learning portfolio, student can have self-examination on their learning result and adjust their learning attitude and skill to get better learning result or achieve the goal of learning topic, so do teacher in the same manner in instruction part.

The design of this learning management system focuses on the learner functionality aspect, and the functionalities are mainly to raise student's learning activities. We design the system with the e-learning standard SCORM (Sharable Content Object Reference Model), which prompts it to use the sharable SCORM learning content packages and display the learning content to student on the web browser directly. We separate the learning content to several categories that represents different learning domains, and each category contains several courses. Student could choose the category that he is interested in and read the containing courses. The courses in specific category are basically independent for the reason that, in the Open Course Ware online learning environment, the course content comes from different teacher in different universities and it is hard to integrate them into a complete learning procedure. The courses in the specific category are related to some specific learning domain but, basically, are independent from each other and seldom have before-and-after relationship, to give the students the most freedom to choose the course they are interested in.

5. Experiment and Empirical Studies

To evaluate our system, we divide several experiments as discussed below. At first, we evaluate the performance of 3D graph based social network visualization tool on the browser. We checked the performance by using different numbers of friends on Facebook and chose some users with 50 friends until 800 friends (the average Facebook users have at least 700) and implemented on Chrome browser. See Table 1.

Results indicated that the number of 3D nodes with more than 800 and edge with total of 3D object rendered showed 1600 3D objects; this results in good performance around 23–25 frame rate per second on the browser since normal rate for flash player around 30 fps for addition; it also runs with Kinect devices and interactions. The next evaluation was to evaluate the accuracy of hand gesture recognition of proposed system. We do testing by performing each of the gesture interactions for 60 times per gesture. The result of gesture recognition is evaluated as in Table 2.

From the results of the total recognition for our hand gesture is 91% of recognition. It indicated that select gesture is the lowest recognition because of position of the nodes

Table 2: Hand gesture recognition accuracy.

Number	Gesture	Success	Percentage
1	Zoom in	50/60	83%
2	Zoom out	54/60	90%
3	Rotate right	48/60	80%
4	Rotate left	50/60	83%
5	Panning	55/60	91%
6	Select	58/60	96%
7	Full screen	55/60	91%
8	Filter	58/60	96%
9	Close	55/60	91%
10	Swipe left	58/60	96%
11	Swipe right	60/60	100%
12	Scroll up	59/60	98%
13	Scroll down	59/60	98%
14	Hide menu	54/60	90%
15	Show menu	58/60	96%
	Average		91%

sometime it make it hard to perform the gesture for example, nodes position is to lower or in the corner of the screen.

We conducted the next experiment to check the user satisfaction of the system and the gesture design by system evaluation and questionnaire. The study was conducted at National Central University and collected from students. There are 83 participants that are Facebook users with at least using Facebook almost every day for 3 years.

After we test our system there are several results that we get from the experiment. From the experiment that we conduct, for the gesture part found, almost all of the gestures were easy to perform and easy to learn. But for some gestures, especially for the rotate gesture, it was found to be a little bit hard to perform. All of gestures were not awkward to use, but some of gestures made the user feel fatigue, for the zooming gesture. They said it was slow and needs to perform in the same position for a long time. The user said the gestures help the system well, especially for the filter gesture and close gesture. Those make it easy for the user, just like a shortcut. Overall they are satisfied with the gesture. Some of them also give some suggestions about the gesture design. For the system, overall, it is good, easy to use, comfortable, easy to learn, and they also really like the user interface theme. But because of the system only implement some few of the Facebook functions, they cannot stay in the system any longer. When we ask about the visualization can make them easy to explore their social network, they agreed. It makes it easy for them to see their social network in the physical ways. The 3D visualization combined with the gestures makes it more immersive, real, natural, and easy to explore their social network. Overall they satisfy the system, but some of them are confused about the use of the Kinect. When we explain it that in the future it will be improved and used in the smart room star trek holodeck like, they support and agreed that in the future they willing to use it.

6. Conclusion and Future Works

In this research, we proposed an interactive 3D Facebook visualization system using Kinect as a new interaction and using this interaction to integrate with the developed learning system, an open course ware, which is based on the integration of e-learning system and social network site Facebook, to provide a social learning environment. Our main objective of this paper was to provide a 3D visualization tool to social networking site (Facebook) and design the way of using new motion capture technology, Kinect, for interaction and exploration of our 3D visualization system. The results of the experiment indicated that our system had good using satisfaction and was easy to use, enhance learner's interest, and help to visualize the social interaction. Otherwise, the proposed 3D visualization makes the user more immersive and naturally interact with the system. Also, the gesture design also helped user to interact and explore the system. But some of the gestures made the user feel tired of fatigue problem (zooming gesture), and our research will improve gesture design to overcome the problems. Overall the system and the gesture design performed well; also we can say that we nearly achieved the goal of this research. For the social learning system, the social behavior is an important issue in social learning, and we can have further analysis here. Moreover, we can integrate more social functionalities in our learning management system, like Facebook club for learning group or online forum for discussion; then we can have more records about social learning behaviors. By doing this, we can take the social interaction result or outcome of the user, like participation rate and interaction times, to build a social learning evaluation model to analyze user's social learning behavior in a quantitative way, which can also be taken into consideration in the recommendation system. Adding this feature we can generate the recommended friends or users who are hospitable to help people and willing to join the social learning interaction.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] I. Herman, G. Melançon, and M. S. Marshall, "Graph visualization and navigation in information visualization: a survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 1, pp. 24–43, 2000.

[2] L. Freeman, "Visualizing social networks," *Journal of Social Structure*, vol. 1, 2000.

[3] B. Gretarsson, J. O'Donovan, S. Bostandjiev, C. Hall, and T. Höllerer, "SmallWorlds: visualizing social recommendations," *Computer Graphics Forum*, vol. 29, no. 3, pp. 833–842, 2010.

[4] M. Honeyman and G. Miller, "Agriculture distance education: a valid alternative for higher education," in *Proceedings of the 20th Annual National Agricultural Education Research Meeting*, pp. 67–73, 1993.

[5] R. Francese, I. Passero, and G. Tortora, "Wiimote and kinect: gestural user interfaces add a natural third dimension to HCI,"

in *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12)*, pp. 116–123, Naples, Italy, May 2012.

[6] J. D. B. Heer, "Vizter: Vizualizing online social networks," in *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS '05)*, pp. 32–39, 2005.

[7] N. Matta and T. Pfeiffer, "Relationships in social networks revealed: a Facebook App for social graphs in 3D based on X3DOM and WebGL," in *Proceedings of the IADIS International Conference Web Virtual Reality and Three-Dimensional Worlds*, pp. 269–276, IADIS Press, 2010.

[8] G. Kendon, *Visible Action as Utterance*, University Press, Cambridge, UK, 2004.

[9] J. Xu, P. J. Gannon, K. Emmorey, J. F. Smith, and A. R. Braun, "Symbolic gestures and spoken language are processed by a common neural system," *Proceedings of the National Academy of Sciences of the United States of America*, 2009.

[10] D. Kammer, "Taxonomy and overview of multi-touch frameworks: architecture, scope and features," in *Proceedings of the Workshop on Engineering Patterns for Multitouch Interface*, 2010.

[11] R. de la Barre, "Touchless interaction-novel chances and challenges," in *Human-Computer Interaction. Novel Interaction Methods and Techniques*, vol. 5611 of *Lecture Notes in Computer Science*, pp. 161–169, 2009.

[12] L. Gallo, A. P. Placitelli, and M. Ciampi, "Controller-free exploration of medical image data: experiencing the Kinect," in *Proceedings of the 24th International Symposium on Computer-Based Medical Systems (CBMS '11)*, June 2011.

[13] H.-M. J. Hsu, "The Potential of Kinect in education," *International Journal of Information and Education Technology*, vol. 1, no. 5, pp. 365–370, 2011.

[14] Y. J. Chang, S. F. Chen, and J. D. Huang, "A Kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities," *Research in Developmental Disabilities*, vol. 32, no. 6, pp. 2566–2570, 2011.

[15] M. F. Shiratuddin and K. W. Wong, "Non-contact multi-hand gestures interaction techniques for architectural design in a virtual environment," in *Proceedings of the International Conference on Information Technology and Multimedia: "Ubiquitous ICT for Sustainable and Green Living" (ICIM '11)*, November 2011.

[16] J.-W. Kang, D.-J. Seo, and D.-S. Jung, "A study on the control method of 3-dimensional space application using Kinect system," *International Journal of Computer Science and Network Security*, vol. 11, no. 9, pp. 55–59, 2011.

[17] R. Soro and R. Scateni, "Natural exploration of 3D models," in *Proceedings of the 9th ACM SIGCHI Italian Chapter International Conference on Computer-Human Interaction: Facing Complexity*, 2011.

[18] M. Resnick, "Distributed constructionism," in *Proceedings of the International Conference on Learning Sciences*, Evanston, Ill, USA, 1996.

[19] T. F. Stafford, "Understanding motivations for internet use in distance education," *IEEE Transactions on Education*, vol. 48, no. 2, pp. 301–306, 2005.

[20] D. R. Garrison and T. Anderson, *E-Learning in the 21st Century: A Framework for Research and Practice*, Routledge/Falmer, London, UK, 2003.

[21] T. Falkowski and M. Spiliopoulou, "Users in volatile communities: studying active participation and community evolution," in *User Modeling*, vol. 45 of *Lecture Notes in Computer Science*, pp. 47–56, 2007.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 846964, 8 pages
http://dx.doi.org/10.1155/2014/846964

Research Article

Subsurface Scattering-Based Object Rendering Techniques for Real-Time Smartphone Games

Won-Sun Lee, Seung-Do Kim, and Seongah Chin

Division of Multimedia Engineering, Xicom Lab, Sungkyul University, Anyang 430 742, Republic of Korea

Correspondence should be addressed to Seongah Chin; solideochin@gmail.com

Received 4 July 2014; Accepted 22 July 2014; Published 17 August 2014

Academic Editor: Jong-Hyuk Park

Subsurface scattering that simulates the path of a light through the material in a scene is one of the advanced rendering techniques in the field of computer graphics society. Since it takes a number of long operations, it cannot be easily implemented in real-time smartphone games. In this paper, we propose a subsurface scattering-based object rendering technique that is optimized for smartphone games. We employ our subsurface scattering method that is utilized for a real-time smartphone game. And an example game is designed to validate how the proposed method can be operated seamlessly in real time. Finally, we show the comparison results between bidirectional reflectance distribution function, bidirectional scattering distribution function, and our proposed subsurface scattering method on a smartphone game.

1. Introduction

Computer graphics on PC and console games that employ BRDF (bidirectional reflectance distribution function) rendering and advanced rendering techniques have made a considerable progress [1–3]. However, BSSRDF (bidirectional surface scattering reflectance distribution function) [4, 5] taking into account subsurface scattering does not seem to operate in smart phone games because it requires a considerable amount of computing. A technique for reducing power dissipation also was proposed when the smartphone is in a static state [6]. Recently, a new subsurface scattering model has been proposed called beam diffusion. The new model is a hybrid method-photon beam diffusion that combines Monte Carlo integration with the diffusion approximation. The rendering results are almost identical to traditional diffusion methods while it seems somewhat faster and stable. Even though this model seems accurate, the methods require minutes on a multicore PC as well [7]. Some image space subsurface scattering techniques in PC environments were studied, in which they required preprocessing to reduce running time. Also, natural scenes that seem to be highly complex could be rendered using a hardware accelerated fractal based method [8]. In the past, because of the limitations of the smartphone hardware, a few efforts were made to overcome the limitation of the graphics quality. Nowadays, the computational power of smartphones has continually advanced, and now, we can expect to solve the previous shortcomings of smartphones. However, subsurface rendering techniques that are optimized for smartphones have not been offered yet. Thus far, we have not found an approach to develop subsurface rendering techniques that can be realized in a real-time smartphone game as shown in Table 1.

In this paper, we present a real-time subsurface rendering technique without any preprocessing that is optimized for smartphone games that can be operated seamlessly. This contributes to enhancement of rendering in smartphones in which high quality of the graphics tends to draw users' interests. We focus on synthesizing a game character that has a skin texture associated with optical parameters for the proposed subsurface scattering for a smartphone. For this, Unity (http://unity3d.com/) shader is used for executing a subsurface scattering algorithm, and vertex and fragment shaders are used for optimizing the proposed method. The preliminary version of research was reported in UCES 2013 workshop. We have extended our research by developing our own subsurface color model that is crucial for real-time subsurface effect in games. Performance evaluation and user study have been carried out as well.

Table 1: Related studies.

Mobile rendering	Real-time Rendering	Subsurface rendering
Clockcycle [9]	Jensen et al. [4], Yang and Wang [10]	
	Yan et al. [11]	Doner and Wann Jensen [5]
Zhang and Shigeo [12]		
None		

The rest of this paper is organized as follows. Section 2 describes BRDF and BSSRDF rendering technologies related to this research. The critical flow of the proposed algorithm is described in Section 3. In Section 4, we explain the game design and development. Section 5 presents the conclusion.

2. Background and Literature Review

2.1. BRDF. The modeling is achieved through the light reflected from the surface by ambience processing, and the rendering equation is used for expressing the surface properties of the materials. The BRDF model, which is slightly more complex than the existing lighting equation, shows an improved result and is becoming the base for the current real-time rendering techniques [3]. There have been previous studies that enhance skin rendering using BRDF [13, 14]. In general, it is a function that shows how the light reflects on a surface, and the reflectance of light changes according to the position of the light and the point that is proportion to the direction of the surface and the tangent value. To be accurate, it shows the ratio between the outgoing radiance and the differential incoming irradiance.

2.2. BSSRDF. BSSRDF added to the BRDF, according to which the incidence and the reflection point express the same surface reflection, is an eight-dimensional function that describes the transport phenomena of the light between two points due to the scattering of the light at the subsurface [15]. Since the aforementioned BRDF is created on the basis of the assumption that the same light becomes the input and the output at the same point, it becomes the approximate value of BSSRDF. When BSSRDF is applied, not only does it express translucent objects, but it also enables a more realistic expression of the human skin [15, 16]. The human skin is a good object to apply the BSSRDF, which can even calculate the scattering phenomena, because it is a multiple-layer translucent object that undergoes a combination of transmission and reflection.

The BSSRDF utilizes the absorption coefficient, the reduced scattering coefficient, and the Fresnel index to express the effect caused by scattering and absorption within the material; the BSSRDF equation can be organized using the absorption coefficient, the reduced scattering coefficient, and the Fresnel index, which are parameters measured according to the materials. On the basis of this theory, many translucent objects have various reflectance values, and hence, their properties have to be expressed cautiously through the ratio of the diffusion approximation and the single scattering.

3. Rendering for Smartphones

3.1. BRDF Shader. Unity is widely known as a multiplatform engine, and a high-completion game can be developed by dealing with both the CPU and the GPU. GPU has to use ShaderLab, a script language, since it does not have a structure, such as the virtual machine of the CPU. In addition, forward pipeline, which calculates the lighting per pixel, is used for producing a game.

Among the Unity shader programs, the vertex fragment shader program was used for implementing the BRDF model. The important point here is that multiple shader languages are supported by ShaderLab, and since it supports multiple platforms, computer graphics have to be used for consistent processing, even on a mobile device.

We produced a special custom ramp that supports a customized light illumination model rather than default light models for the resolution of light treatment, and the shader was realized by using the Ward BRDF model [3] equation for the lighting ramp. Equation (1) is expressed as the tangent space; α_x and α_y denote the roughness of the surface; degrees $\theta_i, \theta_o, \phi_i$, and ϕ_o define the input and output of the light of the BRDF model:

$$f_{r(i,o)} = \frac{\rho_s}{4\pi\alpha_x\alpha_y\sqrt{\cos\theta_i\cos\theta_o}} \times e^{-\tan^2\theta_h((\cos^2\phi_i/\alpha_x^2)+(\sin^2\phi_o/\alpha_y^2))}, \quad (1)$$

where i is an incident direction of the light and o indicates outgoing direction. ρ_s controls the magnitude of the lobe, and θ_h is an angle between a normal and a half vector.

The formula is transformed by giving the scale value through ShaderLab and Ward BRDF model that can control the diffuse, ambient, and the specular light, and the scale value at the slider interface using the property phrase at the ShaderLab was produced. The ultimate data type of $f_r(i, o)$ was given as a fixed type for the optimization at the mobile device.

3.2. Scaled BSSRDF Shader. Because BSSRDF [4] takes minutes on a multicore PC, we modify it by scaling that can be running in real-time. The rendering equation has to be accurately implemented at the ShaderLab for the optical parameters to accurately express the peculiar properties of the objects. If the surface shader program is used, it is possible to produce a shader effectively and easily; however, BSSRDF was implemented using the vertex and fragment shader program and it is a rendering model that used Jensen's bipolar light source method. The diffusion term is achieved through formula (2):

$$S_d(x_i, \vec{w}_i; x_o, \vec{w}_o) = \frac{1}{\pi}F_t(\eta, \vec{w}_i)\, R_d(\|x_i - x_o\|)\, F_t(\eta, \vec{w}_o). \quad (2)$$

Formula (3) is the formula of $R_d(r)$ term of Jensen's BSSRDF model [4]:

$$R_d(r) = \frac{\alpha'}{4\pi}\left[(\sigma_{tr}d_r + 1)\frac{e^{-\sigma_{tr}d_r}}{\sigma_t' d_r^3} + z_v(\sigma_{tr}d_v + 1)\frac{e^{-\sigma_{tr}d_v}}{\sigma_t' d_v^3}\right], \tag{3}$$

where $\alpha' = \sigma_s'/\sigma_t'$, $\sigma_t' = \sigma_s' + \sigma_a$, $\sigma_{tr} = \sqrt{3\sigma_a\sigma_t'}$,

$$d_r(r) = \sqrt{z_r^2 + r^2}, \quad z_r = \frac{1}{\sigma_t'},$$

$$d_v(r) = \sqrt{z_v^2 + r^2}, \quad z_v = z_r + 4AD, \quad A = \frac{1 + F_{dr}}{1 - F_{dr}}, \tag{4}$$

$$F_{dr} = \frac{-1.44}{\eta^2} + \frac{0.71}{\eta} + 0.668 + 0.0636\eta, \quad D = \frac{1}{3\sigma_t'}.$$

The parameters used in formula (3) are all defined by σ_a and σ_s', and the $R_d(r)$ term's value is calculated using the light and camera's property provided by the update function of the Unity script. Original BSSRDF model takes running times about seven minutes on a multicore PC. Hence, instead of using the distance between the points of the object, the distance calculated on the basis of the location of each point on the object and the light source of the approximated model was used as parameter r. Furthermore, the distance was controlled by multiplying the ratio of the scale value in order to consider the distribution of the light's transmission and the size of the object in Unity.

3.3. Our Subsurface Scattering. The scaled BSSRDF mentioned in Section 3.2 seemed to be overly smooth that could be only suitable for rendering like a candle. It does not seem ideal for rendering human skin. Some advanced BSSRDF techniques are also proposed such as the hierarchical BSSRDF [17] that provides a rapid hierarchical rendering technique for translucent materials. However, we do not recognize that this BSSRDF can be suitable for mobile applications.

Authors propose a novel method that takes into consideration the drawbacks such as being overly smooth and taking longer running times. The proposed method needs to recover those limitations by employing our subsurface scattering color model [18] that takes optical parameters and return a set of final subsurface color values. We have a subsurface scattering color that can be computed in (5) that takes optical parameters

$$f_d = (1 - F_{dr,t})\left(\sigma_s' P(g)\left\{d - \sigma_a d^2 + \frac{1}{3}\sigma_a^2 d^3\right\} + F_{dr,b}\left(1 - \sigma_a d - \sigma_s' P(g)\, d\right)^2\right), \tag{5}$$

where σ_a and σ_s' are the absorption and reduced scattering coefficients, respectively. $P(g)$ is the integration of phase function for backward. d is the depth. $F_{dr,b}$ is the diffuse Fresnel reflectance at the bottom boundary. $F_{dr,t}$ represents the diffuse Fresnel reflectance at the top boundary.

Our final rendering color is defined as (6)

$$f_{r(i,o)} = f_d\left(\rho_d \cos\theta + \frac{20\rho_{ts}}{d^{2.5}}\right) + f_s L \cos\theta \times \left[\frac{\rho_s}{4\pi\alpha_x\alpha_y\sqrt{\cos\theta_i \cos\theta_o}} e^{-\tan^2\theta_h(\cos^2\phi_i/\alpha_x^2 + \sin^2\phi_o/\alpha_y^2)}\right], \tag{6}$$

where f_d is the subsurface surface color. f_s is ρ_d (diffuse coefficient), and ρ_s (specular coefficient) and ρ_{ts} (translucent coefficient) can be optimized using an experiment. L is the light intensity. d is the distance from the light source. (θ_i, ϕ_i) is the incident light vector, and (θ_o, ϕ_o) is the reflected light vector. f_d is obtained using optical parameters that is defined as formula (5). ρ_d (diffuse coefficient), ρ_s (specular coefficient), and ρ_{ts} (translucent coefficient) can be derived from the experiments meaning that the optimal parameters are determined as viewing rendering results. For the skin, we determine $\rho_d = 1$, $\rho_s = 0.2$, and $\rho_{ts} = 0.01$. The first term means diffusive reflectance and second term is for subsurface diffuse light that is used for subsurface effect for translucent materials. The last term determines the specular light.

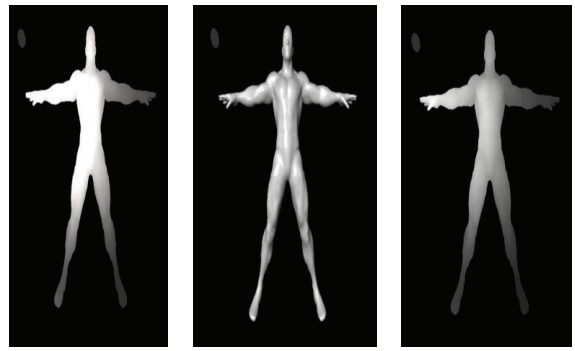
4. Experiments

Unique optical features of the material that we wish to render have to be considered when rendering a material. We implemented three rendering methods described in Section 3, using vertex and fragment program under Unity 4.3 ShaderLab. To reduce running times in GPU for mobile environments, we let some computations that are not necessary to run in GPU operate in CPU.

4.1. Rendering Results. Prior to comparing the rendering results, we carried out rendering of translucency that took into account optical parameters of a white coated material because the results of translucency can be easily observed. In Figure 1, we can observe the translucent effects of BSSRDF (a) and our method (c) rendering results. The rendering result for BRDF (b) does not show translucency but glitters because BRDF cannot consider subsurface scattering. In the first case, to express the absorption of the light and the scattering, the parameter of a white coating was used to express the feeling of a translucent material; the color of the skin was compared by using texture. A considerably softer and more translucent feeling at the fingertips and both feet can be seen in the case of BSSRDF and our model. However, the rendering result of BSSRDF seems rather translucent than our subsurface scattering (OSS).

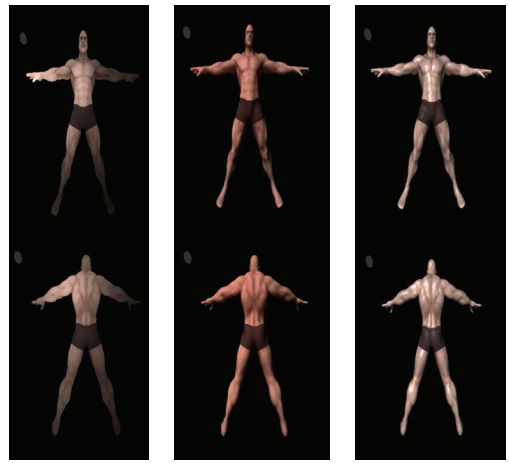
Levenberg-Marquardt method [19] was used to acquire optical parameters of the white coated material. In this simulation for Figure 1, we used $\sigma_s' = 1.94$ for red, $\sigma_s' = 1.94$ for green, $\sigma_s' = 1.874$ for blue, $\sigma_a = 0.000905$ for red, $\sigma_a = 0.001083$ for green, $\sigma_a = 0.002038$ for blue, and $\eta = 1.0936$ for Fresnel.

Then, after using skin 1's $\sigma_a = (0.032, 0.17, 0.48)$ and $\sigma_s' = (0.74, 0.88, 1.01)$ values to find $\alpha = \sigma_s'/(\sigma_s' + \sigma_a)$, the diffuse value was obtained by substituting the α value, acquired



(a) (b) (c)

Figure 1: The rendering results of a white translucent material. BSSRDF (a), BRDF (b), and OSS (c) are shown.



(a) (b) (c)

Figure 2: The rendering results of BSSRDF (a), our subsurface scattering (OSS) method (b), and BRDF (c).

using the parameters of skin instead of α, into the BRDF approximation. These optical parameters were measured by using our customized optical imaging equipment [20].

The comparison subject BRDF's texture was obtained by multiplying the subsurface scattering color value by the texture RGB value. Since the value of $R_d(r)$ determines the object's color in the BSSRDF model, the gray level's texture was used for comparing the three models under the same conditions.

Figure 2 shows three rendering results. The light source is located at upper (a). The screen shots in (c) represent the rendering result of BRDF that cannot convey scattering effects at all. However, the human skin should have been scattered. This seems like a plastic that mostly glitters. The BSSRDF rendering results shown in (a) seem more translucent than they should be.

The areas of around chest and head are much blurrier by subsurface scattering effect than they are. Comparing our method (shown in Figure 2(b)) with other two methods, our subsurface scattering (OSS) method seems reasonable because our rendering results show more relevant translucency than BSSRDF (a). In short, we can observe that the proposed rendering results seem more natural than others. For instance, more uniform shading distributions in the upper body and legs are observed in our rendering results as well.

Moreover, Figure 3 shows the comparison result of the BRDF, the BSSRDF, and OSS, which controlled the coefficients to be used in a game.

The rendering results from BSSRDF and OSS are smoother than one from BRDF. However, the BSSRDF is quite blurred and produces subsurface scattering that is not ideal to render human skin. OSS seems to produce somewhat medium effect between BSSRDF and BRDF since the algorithm is defined by conveying subsurface scattering parameters of the skins and BRDF effect as well. The proposed rendering results show acceptable rendering results superior to other methods in the screen shots of the smartphone game that we designed.

Also, we carried out a user study in order to evaluate reputations from users about three rendering results. The 50 subjects whose age ranges from 20 to 30 participated in our questionnaire. They are currently working for IT (information technology) related occupations or having experiences

(a) (b) (c)

Figure 3: BRDF (a), BSSRDF (b), and OSS (c) in each picture applied to the smartphone game.

with playing mobile games. The scores range 0–5 Likert scale points representing the higher scores indicate more reliable judgment than the lower scores. They were asked to answer two survey questions while viewing rendering results that were implemented in the smartphone game. For the first question, we asked how similar do you feel that it is like real skin.

The result scores in Figure 4 show BRDF (3.06), modified BSSRDF (2.91), and OSS (3.38). Also we asked them if they recognized that skin is translucent. Then, we again examined the question only if their responses were "yes."

More interesting analysis on the first question was found when we asked only a participant who recognized that skin is translucent. The score BRDF decreases by 2.69 from 3.06 whereas modified BSSRDF score becomes 3.21 from 2.91 and OSS score increases by 3.61 from 3.38 as shown in Figure 5.

This fact seems to support a prior study concerning comparison between experienced and inexperienced game player [21]. In short, our proposed method shows acceptable rendering results.

4.2. Game Development. In order to realize our proposed method into a real-time game, we have developed a game named "Xicom runner" using Unity. The game is composed of a main screen and a play screen, as shown in Figure 6. In the main screen, there is an option button to select a rendering method. Further, there is a start button to start the game. The score increases in the running process of the character, and the score is obtained when the character continues to run. If the character collides with obstacles, then the game is over. When the replay button that appears afterwards is clicked, the game returns to the play screen. The game considered in this study is easy and simple, since it was produced to test and display real-time rendering. Further, the position and the view of the camera were not fixed but were set to move flexibly in order to follow the flow of the game. The map was composed with sets of various objects to test the performance and the limits of the rendering.

4.3. Optimization Technique. For seamless real-time rendering of the shader, not only the optimization of the shader code but also the optimization of the game code is required.

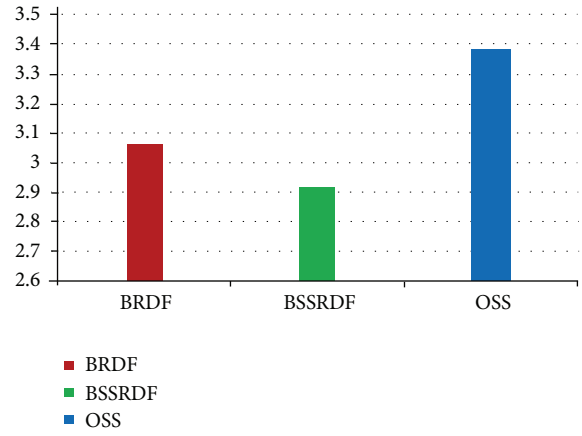


Figure 4: Rendering result scores on BRDF (left), BSSRDF (middle), and OSS (right).

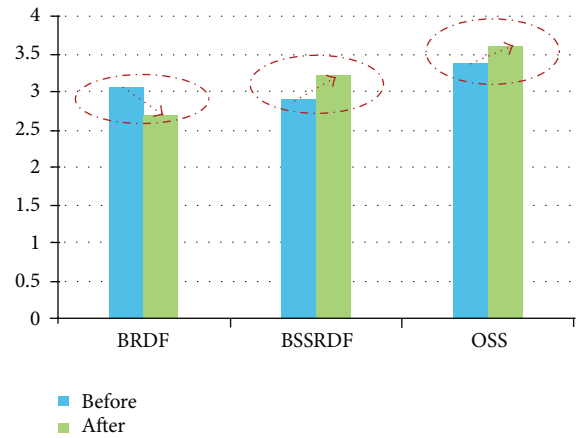


Figure 5: Score changes on BRDF (left), BSSRDF (middle), and OSS (right).

We developed the relevant code by considering the development engine Unity's properties and the following details for the optimization of the game code.

(a) (b)

Figure 6: Main screen (a) and play screen (b).

In order to avoid a bias that could happen to real-time game play among three rendering methods, we apply the following optimization techniques to all three rendering methods.

First, minimize the number of unnecessary variables and object creation. In other words, where variables and objects with the same attributes and functions are repeatedly created, they should be referred to instead of being constructed dynamically by declaring them as a member or global beforehand. By doing this, the process of initializing and deleting these attributes and functions every time is eliminated, and the burden of operations is reduced because it only has to refer to them.

Second, preprocess all the resources, such as texture, that are used in the game. Setting a resource in the memory requires a considerable number of operations and is time consuming. Therefore, if dynamically, each frame loads the resource, it will burden the system. To prevent this, the resource should be set in the memory beforehand and referred to when needed by preprocessing the resource loading.

Third, the GUI skin that is produced by Unity should not be used and should be self-produced and used. Because the GUI skin produced by Unity takes up more memory than expected, to minimize this, we self-produced and used the texture and minimized unnecessary memory use and operations.

Lastly, minimize the operations within the OnGUI() function. In Unity, to output text or textures on the screen, the functions related to the GUI in the OnGUI() function have to be used. However, caution must be taken when using the OnGUI() function, because each frame is called twice. Consequently, when control statements or expressions are inserted into this function, it becomes a problem as each calculation is run twice. Therefore, it is advantageous to exclude as many expressions and control statements as possible in the OnGUI() function.

4.4. Performance Test. In general, even in PC or console games, an advanced shader is applied only to the object that accounts for the largest proportion in the screen. This is because the application of an advanced shader to objects that have low importance and are not noticeable is a waste of memory and increases the number of unnecessary operations.

Table 2: GPU running times.

	GPU USAGE (ms)		
Sampled frame (30f)	BSSRDF	BRDF	Our method
1	0.179	0.146	0.111
2	0.153	0.148	0.109
3	0.174	0.133	0.113
4	0.181	0.137	0.114
5	0.175	0.144	0.112
6	0.165	0.155	0.109
7	0.174	0.14	0.115
8	0.182	0.159	0.109
9	0.184	0.115	0.133
10	0.217	0.158	0.122
Average	0.1784	0.1435	0.1147

We applied three shading codes to only the character model in the game we made that accounts for the biggest proportion in the screen. Other objects in the game such as tables, chairs, walls, and floors used the default illumination model in Unity to reduce GPU running times.

To verify the proposed method, we realized rendering algorithms under Unity 4.3 ShaderLab. We analyzed algorithm performance with respect to GPU running times and memory usage. To capture sample frames, we made the character that was implemented with shaders including BRDF, BSSRDF, and our method be animated. The 10 peak frames every 30 frames were captured.

For GPU performance test, we analyzed running times from 10 sampled frames using Unity profiler with GPU (*ms*). The average running times of GPU appear with 0.1784 for BSSRDF, 0.1435 for BRDF, and 0.1147 for our method, respectively.

In short, we found that modified BSSRDF and BRDF took longer operations than our proposed method as shown in Table 2. Even if three methods can quite work well in real time, the rendering results of our method are more outstanding than BSSRDF and BRDF. Moreover, our proposed method is superior to the other two methods with respect to GPU running time.

Table 3: Memory usage.

	Memory (kb)		
Memory	BSSRDF	BRDF	Our method
Shader memory	28.3	23.4	24.9
Skinned mesh renderer	0.6	0.6	0.6
Shader object	0.177	0.177	0.177
Scene memory	116.2	116	116.2

In addition, we performed memory usage tested and shown in Table 3 because the memory usage can affect somewhat game performance. If the more memory is occupied, then more access time is required. In particular, mobile games need to take into account small amount of memory if possible.

For memory test, we extracted some criteria such as shader memory, skinned mesh renderer, shader object, and scene memory. Shader memory represents the memory that is used for performing of BSSRDF, BRDF, and our method. Scene memory is composed of lots of components. Among them, we select three criteria. In shader memory, BRDF used the smallest portion of memory 23.4 (kb). Our proposed method used 24.9 (kb) and BSSRDF took 28.3 (kb). Other criteria show not a big difference. Our proposed method is acceptable in memory usage as well.

5. Conclusion

In this paper, we introduced the BRDF and modified BSSRDF and our proposed method for effective rendering taking into consideration the properties of object materials and optimized them for a smartphone environment. We applied these shaders to a smartphone game. As a result, we found that our proposed subsurface scattering method seemed slightly better than modified BSSRDF and BRDF with respect to rendering appearance and GPU running times. Also our method is superior to BSSRDF in the shader memory usage. Further, it was verified that the game could be played in real-time when the proposed shaders were used.

It was experimentally verified that our proposed method was more appropriate for rendering a character's translucent skin. The value of this research lies in the application of subsurface scattering, which could not be realized real-time on a smartphone thus far, by reproducing it in real time. In the proposed approach, we rather focus on rendering a character model. As future work, more materials need to be simulated simultaneously to achieve more enhanced quality of a scene for smart phone games that definitely comes with tradeoff between running time and visual appearance. We might estimate a fact that running time of a game is contingent on input factors such as the number of object, the number of polygon, the number of light, and the number of shader component. Precise analysis would be interesting. However, we have to determine optimal properties of the input factors for a real-time game.

Also, skin is used in many applications. A thresholding technique is critical when extracting human skin region that can be affected by various lights [22].

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the Korea Foundation for the Advancement of Science and Creativity and funded by the Korean Government. And this research was also partially funded by Korea Evaluation Institute of Industrial Technology, KEIT (CiMR no. 10043453).

References

[1] S. McAuley, S. Hill, N. Hoffman et al., "Practical physically-based shading in film and game production," in *Proceeding of the SIGGRAPH 2012, ACM SIGGRAPH Courses Article*, 2012.

[2] M. Kurt and D. Edwards, "A survey of BRDF models for computer graphics," *ACM SIGGRAPH Computer Graphics—Building Bridges—Science, the Arts & Technology*, vol. 43, no. 2, article 4, 2009.

[3] B. Walter, "Notes on the ward BRDF," Tech. Rep. PCG-05-06, 2005.

[4] H. W. Jensen, S. R. Marschner, M. Levoy, and P. Hanrahan, "A practical model for subsurface light transport," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)*, pp. 511–518, 2001.

[5] C. Doner and H. Wann Jensen, "A spectral BSSRDF for shading human skin," in *Proceedings of the 17th Eurographics Conference on Rendering Techniques*, pp. 409–417, 2006.

[6] H. Cho and M. Choi, "Personal mobile album/diary application development," *Journal of Convergence*, vol. 5, no. 1, pp. 32–37, 2014.

[7] R. Habel, P. H. Christensen, and W. Jarosz, "Photon beam diffusion: a hybrid Monte Carlo method for subsurface scattering," *Computer Graphics Forum*, vol. 32, no. 4, pp. 27–37, 2013.

[8] J. Divya Udayan, H. S. Kim, J. Lee, and J.-I. Kim, "Fractal based method on hardware acceleration for natural environments," *Journal of Convergence*, vol. 4, no. 3, pp. 6–12, 2013.

[9] Real-time BRDF on IOS, 2013, http://clockcycle.net/real-time-brdf-ios.

[10] M. Yang and K. Wang, "Real-time rendering for multi-layered translucent materials such as human skin under dynamic environment lighting," *Applied Mechanics and Materials*, vol. 274, pp. 423–426, 2013.

[11] C. Yan, T. Yue, J. Liu, and C. Zhao, "A novel method for dynamic surface modeling and real-time rendering," in *Proceedings of the 21st International Conference on Geoinformatics (GEOINFORMATICS '13)*, pp. 1–5, IEEE, Kaifeng, China, June 2013.

[12] Z. Zhang and M. Shigeo, "Real-time hair simulation on mobile device," in *Proceedings of the Motion on Games*, ACM, 2013.

[13] R. Stephen, H. Marschner Stephen, P. F. Eric, E. Kenneth, and P. Donald, "Image-based brdf measurement including human skin," in *Proceedings of 10th Eurographics Workshop on Rendering*, pp. 139–152, 1999.

[14] S. R. Marschner, S. H. Westin, E. P. F. Lafortune, K. E. Torrance, and D. P. Greenberg, "Reflectance measurement of human skin," Tech. Rep., Cornell University, 1999.

[15] H. Nakai, Y. Manabe, and S. Inokuchi, "Simulation and analysis of spectral distributions of human skin," in *Proceedings of 14th International Conference on Pattern Recognition*, vol. 2, pp. 1065–1067, 1998.

[16] N. Tsumura, N. Ojima, K. Sato et al., "Image-based skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin," in *Proceedings of the 30th Annual Conference on Computer Graphics and Interactive Technicques (SIGGRAPH '03)*, pp. 770–779, July 2003.

[17] H. W. Jensen and J. Buhler, "A rapid hierarchical rendering technique for translucent materials," in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '02)*, pp. 576–581, San Antonio, Tex, USA, July 2002.

[18] T. Choi and S. Chin, "Wound recovery synthesis on a 3d face using subsurface scattering with multi-layered skin features," in *Proceedings of the 26th International Conference on Computer Animation and Social Agents (CASA '13)*, Istanbul, Turkey, May 2013.

[19] D. M. Bates and D. G. Watts, *Nonlinear regression analysis and Its applications*, John Wiley & Sons, New York, 1988.

[20] T. Choi, S. Lee, and S. Chin, "A method of combining the spectrophotometer and optical imaging equipment to extract optical parameters for material rendering," *Journal of Sensors*. In press.

[21] G. Christou, "A comparison between experienced and inexperienced video game players," *Human-Centric Computing and Information Sciences*, vol. 3, article 15, 2013.

[22] Z. H. Al-Tairi, R. W. Rahmat, M. I. Saripan, and P. S. Sulaiman, "Skin segmentation using YUV and RGB color spaces," *Journal of Information Processing Systems*, vol. 10, no. 2, pp. 283–299, 2014.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 383270, 10 pages
http://dx.doi.org/10.1155/2014/383270

Research Article

Automatic 3D City Modeling Using a Digital Map and Panoramic Images from a Mobile Mapping System

Hyungki Kim,[1] Yuna Kang,[2] and Soonhung Han[1]

[1] *Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea*
[2] *1st R&D Institute, Agency for Defense Development, 488 Bugyuseong-daero, Yuseong-gu, Daejeon 305-152, Republic of Korea*

Correspondence should be addressed to Soonhung Han; shhan@kaist.ac.kr

Received 20 May 2014; Revised 16 July 2014; Accepted 16 July 2014; Published 13 August 2014

Academic Editor: Jong-Hyuk Park

Three-dimensional city models are becoming a valuable resource because of their close geospatial, geometrical, and visual relationship with the physical world. However, ground-oriented applications in virtual reality, 3D navigation, and civil engineering require a novel modeling approach, because the existing large-scale 3D city modeling methods do not provide rich visual information at ground level. This paper proposes a new framework for generating 3D city models that satisfy both the visual and the physical requirements for ground-oriented virtual reality applications. To ensure its usability, the framework must be cost-effective and allow for automated creation. To achieve these goals, we leverage a mobile mapping system that automatically gathers high-resolution images and supplements sensor information such as the position and direction of the captured images. To resolve problems stemming from sensor noise and occlusions, we develop a fusion technique to incorporate digital map data. This paper describes the major processes of the overall framework and the proposed techniques for each step and presents experimental results from a comparison with an existing 3D city model.

1. Introduction

Three-dimensional city models are widely used in applications in various fields. Such models represent either real or virtual cities. Virtual 3D city models are frequently used in movies or video games, where a geospatial context is not necessary. Real 3D city models can be used in virtual reality, navigation systems, or civil engineering, as they are closely related to our physical world. Google Earth [1] is a well-known 3D representation of real cities. It illustrates the entire earth using satellite/aerial images and maps superimposed on an ellipsoid, providing high-resolution 3D city models.

In the process of 3D city modeling, both the cost and the quality requirements must be considered. The cost can be estimated as the time and resource consumption of modeling the target area. The quality factor considers both visual quality and physical reliability. The visual quality is proportional to the degree of visual satisfaction, which affects the level of presence and reality. The physical reliability is the geospatial and geometrical similarity between the objects—in our case, mainly buildings—in the modeled and physical worlds. Generally, accomplishing a satisfactory level for both requirements is difficult.

Numerous techniques can be used for 3D city modeling. For instance, color and geometry data from LiDAR are mainly used if the application requires detailed building models for a small area. If the city model covers a large area and does not need detailed features, reconstruction from satellite/aerial images is more efficient [2]. This means that the effective approach can differ according to the target application using the 3D city model. The goal of our research is to propose a 3D city modeling method that can be applied in ground-oriented and interactive virtual reality applications, including driving simulators and 3D navigation systems, which require effective 3D city modeling methods for diverse areas.

2. Related Work

Existing 3D city modeling methods can be divided into manual modeling methods, BIM (Building Information Model) data-based methods, LiDAR data-based methods, and image-based methods.

The manual modeling method is highly dependent on the modeling experts. Older versions of Google Earth and Terra Vista employed this method in their modeling systems. Although current applications employ manual modeling because of its high quality, the method is also a high-cost, labor-intensive process. Hence, it is not efficient for urban environments that include numerous buildings. The BIM data-based method facilitates the use of building design data from the construction stage and is applied in city planning [3] and fire and rescue scenario [4]. However, this method is only efficient for applications in which the activity areas are strictly constrained, as gathering BIM data is problematic or even impossible given the large size of urban environments. Moreover, the BIM data should be postprocessed for the use of virtual reality applications. This is because the information in BIM does not contain the as-built 3D model; so the properties for the visual variables should be mapped.

To address these problems, remote sensing techniques are being aggressively adopted and studies of LiDAR data-based methods and image-based methods are increasingly common [5]. LiDAR is a device that samples precise data from the surrounding environment using laser scanning technology. In several studies (e.g., [6, 7]), high-quality 3D city models have been reconstructed using LiDAR data. However, as noted in other work [6], ground-level scanning has a limited data gathering range, meaning that redundant data collection is unavoidable in the modeling of diverse areas, whereas airborne LiDAR [8] is limited in terms of its cost and color data collection methods.

Image-based methods include those based on stereo matching and inverse procedural modeling approaches. In earlier research [7, 9], a method based on stereo matching was used to recover 3D data from the feature point matching between a series of images. This approach usually requires numerous images to satisfy the accuracy and robustness requirements of feature point matching. Several recent inverse procedural modeling approaches [10–12] have modeled buildings using relatively few (mainly one) images. This can overcome the difficulties of data-collection in stereo matching. This approach employs a plausible assumption; that is, that the shape of a building consists of a set of planes in three dimensions, to reconstruct individual 3D buildings without pixel-wise, 3D information. However, because image-based methods are not robust against instances of occlusion, user input or strong constraints are frequently necessary. This reduces their cost effectiveness and/or physical reliability.

In our research, the approach which preserves the cost-efficiency by using the existing image database while increasing physical reliability will be proposed. The image database is relatively easy to access than LiDAR database so cost on the data collection can be decreased. On the other hand the method based on the stereo matching requires numerous images on the large-scale modeling that decreases the universal applicability of method. Therefore inverse procedural modeling approach is preferred on our objective, while the physical reliability can be increased by combining accurate reference data [13].

3. Proposed Method

3.1. Mobile Mapping System and Digital Map. In this study, we propose a framework that uses a massive number of images gathered from a mobile mapping system (MMS). This addresses many problems in existing methods, which cannot simultaneously provide feasible levels of cost effectiveness, visual quality, or physical reliability. An MMS collects and processes data from sensor devices mounted on a vehicle. Services such as Google Street View [14], Naver Maps [15], and Baidu Maps [16] present information in the form of high-resolution panoramic images that include the geospatial position and direction of each image taken. The main focus of these services is to offer visual information about the surrounding environment at a given location. The advantages of data collected from MMS are as follows.

(1) Nationwide or even worldwide coverage following the development of remote sensing technologies and map services.

(2) Rich, visual, and omnidirectional information.

(3) Sensor information that allows geospatial coordination with the physical world.

Using these advantages, we can model a diverse city area for ground-oriented interactive systems in a cost-effective way with the existing image database. Moreover, high visual quality at ground level can be provided by high-resolution panoramic images [17]. However, there are currently several disadvantages in the data collected from MMS.

(1) Sensor data includes noise, which lowers its physical reliability.

(2) The number of images in a given area is limited and is insufficient for stereo matching-based reconstruction.

(3) Inclusion of an enormous amount of unnecessary visual information, including occlusions, cars, and pedestrians.

Noise is unavoidable in the sensing process. The amount of error this introduces differs according to the surrounding environment; a ± 5 m positional error and $\pm 6^{\circ}$ directional error have been reported in Google Street View data [18]. Such error levels can be problematic in the analysis required for 3D modeling. Moreover, the current service has an interval of ~10 m between images, which lowers the possibility of successful reconstruction using stereo matching. Additionally, the uncontrolled collection environment results in a severe disadvantage for inverse procedural modeling. MMS data also requires an additional process to classify individual buildings, unlike the inverse procedural modeling approaches.

To address these problems with MMS data, we propose a method that incorporates 2D digital map data. Digital maps have accurate geospatial information about various features in the physical world. For instance, the 1 : 5000 digital maps applied in our framework have a horizontal accuracy of 1 m, which is five times better than that of the MMS position data.

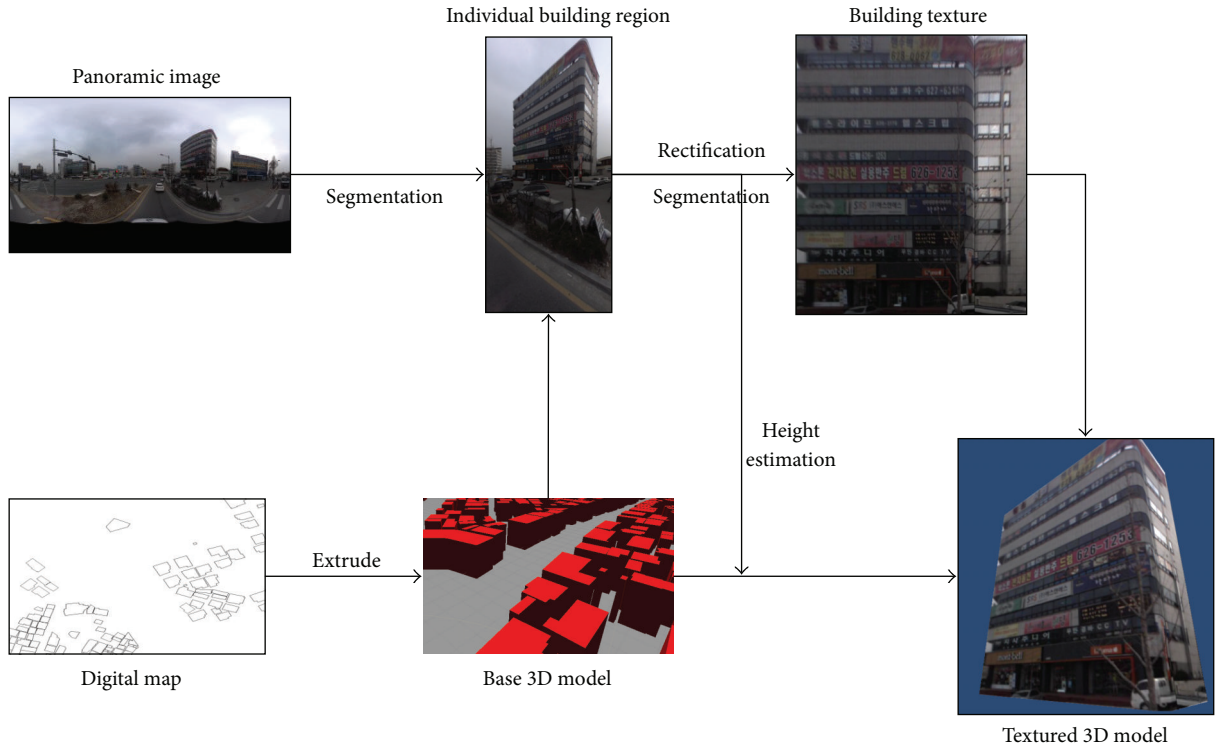


Figure 1: Overview of the proposed 3D city modeling framework.

Therefore, by combining data, the problems of sensor errors can be overcome and the selective use of visual information is possible. On the other hand, the geometrical characteristic of the building is restricted to a quasi-Manhattan world model. The quasi-Manhattan world model is the assumption that structures consist of vertical and horizontal planes, and is an extension of the Manhattan world model that assumes structures consist of vertical and horizontal planes orthogonal to each other.

3.2. Process Overview. The proposed framework is illustrated in Figure 1. The input data are the aforementioned digital maps, which contain building footprint information and panoramic images from the MMS system with sensor data. The base 3D model is generated from the footprint information of the buildings; the individual building regions are segmented and reprojected according to the combined GPS/INS (Inertial Navigation system) information. The reprojected region is further segmented and rectified to produce the texture image. Height estimation is possible by combining the building contour information from the texture image and the reprojected image. We can then obtain the textured 3D model by applying the height information to modify the height of the base 3D models.

The detailed process is illustrated in Figure 2. The entire modeling procedure can be divided into the following four stages.

(1) Image/building analysis.

(2) Error correction/compensation.

(3) Segmentation and validation.

(4) Texture mapping.

The error correction/compensation and segmentation and validation processes include a feedback loop to sequentially obtain the texture of individual buildings.

3.3. Image/Building Analysis. This stage analyzes the correlation between each image and the digital map. To do this, a base 3D model is generated from the footprint of the buildings by extending the model in the vertical direction. The footprint data consists of the geospatial coordinates of the building contour projected to the ground surface. This data should retain a certain level of accuracy and therefore contains precise information about the buildings.

The input image can then be positioned in the 3D environment according to the GPS/INS sensor information. Next, buildings are classified according to the three criteria listed below. The objective of this classification is to separate the buildings into texture-acquirable examples and others at a high resolution. The proposed criteria are as follows.

(1) The distance between the location from the GPS sensor corresponding to the image and the building.

(2) The occlusion between the buildings.

(3) The region occupied by the building in the image.

The distance criterion is quite straightforward: more distant buildings are less likely to appear in the image. The occlusion criterion is also reasonable, as an occluded building

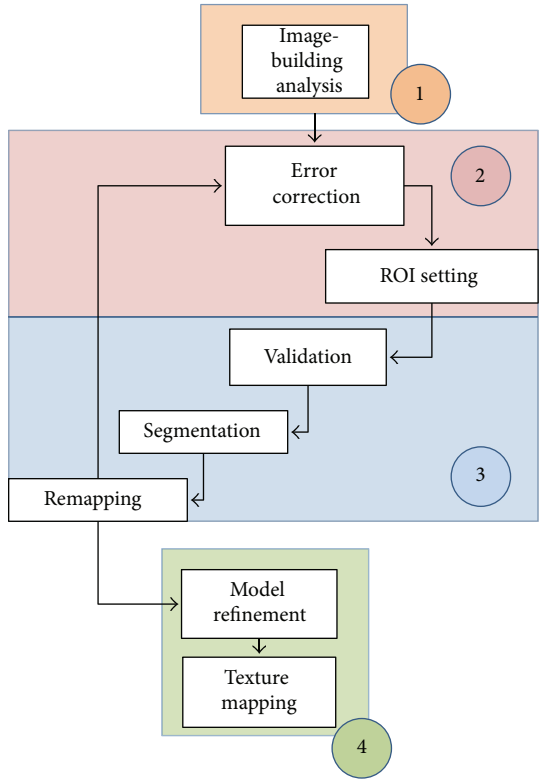


Figure 2: Detailed process of the proposed approach.

cannot appear in the image. The third criterion considers information about the brief texture resolution of each image by calculating the façade angle and width of the image, and then estimating the resolution of the texture. Assuming W is the width of the panoramic image that has a 360-degree field of view (FOV), and then the pixel per radian in the horizontal direction can be calculated as W/π. The location of the captured image can be expressed as $C \in \mathbb{R}^2$ in the 2D digital map and both ends of the façade footprints can be expressed as $P_l, P_r \in \mathbb{R}^2$. Then the pixel per meter, which is the texture resolution for each façade, can be calculated using the law of cosines

$$\text{Texture Resolution} = \frac{W \cdot \theta}{\pi},$$

where

$$\theta = \arccos\left(\frac{\left(\|C - P_l\|\right)^2 + \left(\|C - P_r\|\right)^2 - \left(\|P_r - P_l\|\right)^2}{2\left(\|C - P_l\|\right)\left(\|C - P_r\|\right)}\right). \tag{1}$$

These criteria are illustrated in Figure 3.

3.4. Error Correction/Compensation. The correction/compensation process turns the omnidirectional building detection problem into the simple problem of segmenting a single image. Noise in the GPS/INS sensor is the primary source of the mismatch between the image/building analysis results and the ground truth.

First, correction of the GPS/INS sensor error is performed using image-based localization methods. Image-based localization is a major research issue in robot vision and location-based services. The localization of a panoramic image can be conducted by applying semantic segmentation [18]. However, the authors assumed the existence of a detailed 3D cadastral model, but these are not often available in city areas. Other research [19] has proposed a localization method that uses images and digital maps, but this is strongly dependent on user input. Hence, their method is not appropriate for our framework, which deals with a large number of images. Instead, our framework utilizes a method that localizes the panoramic image based on the orientation descriptor [20]. The footprint orientation (FPO) descriptor encodes the relative angle between the lines emitted radially from a certain location on the map and the footprints of the buildings. In the same way, the FPO descriptor can be calculated from the panoramic image because the panoramic image has omnidirectional information so that by vanishing point estimation we can calculate the angle between the location of the image and the footprints of the visible buildings. By finding the minimum distance between the FPO descriptor calculated from the image and the sampled locations on the map, we can estimate where the panoramic image has been taken. Experiments showed that the error after estimation is less than 2 m, which is sufficient for our framework, and to proceed to the processing stage. Meanwhile, the 360-degree FOV panoramic image is preferred because of this error correction. As can be seen in earlier research, a single image from a normal lens contains a limited amount of visual information about the

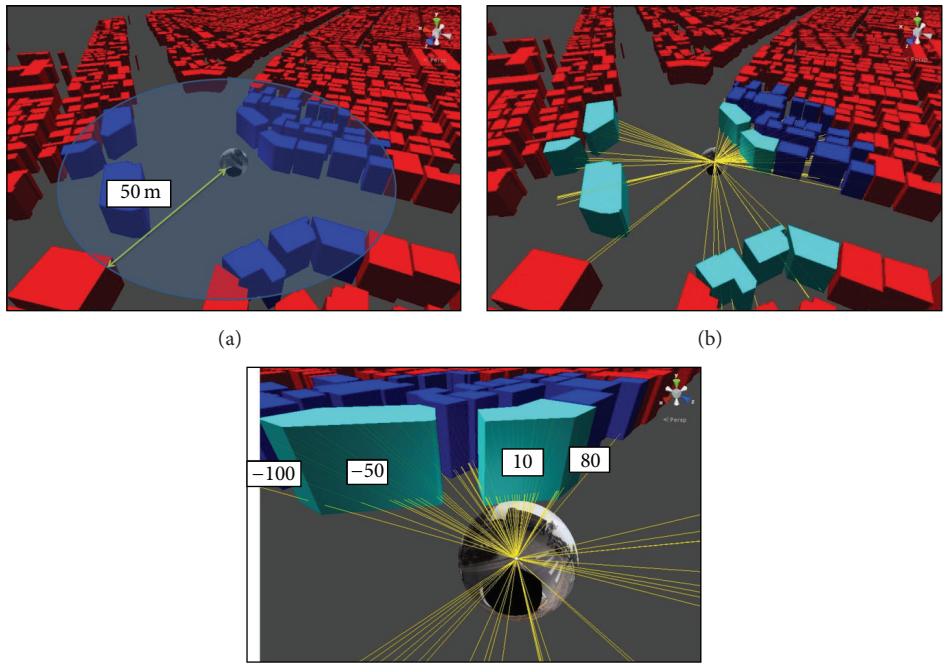


Figure 3: Criteria of the image/building analysis. (a) Distance between the image and buildings; (b) occlusion between buildings; and (c) facing viewing angle of the building façade, that is used to calculate the texture resolution.

surrounding environment [21], so user-input should be considered in order to more accurately estimate the location where the image was taken [22].

The error compensation utilizes this error bound to set the region of interest (ROI). Our objective in error compensation is to process the single 360-degree panoramic image into several normal-lens images to build each target, by partitioning and reprojecting. As we mentioned before, the error correction reduces the position and orientation error but there still exist mismatches, up to 2 m in position, between the base 3D model from the digital map and the panoramic image. So we calculate the FOV for each targeted building, which is ROI, with a 2-m margin. After that, the panoramic image is reprojected using rectilinear projection to generate an image which preserves the straight lines in 3D space in the projected image. Then the ROI contains the complete image of the target building and the complexity of the image-segmentation process is reduced.

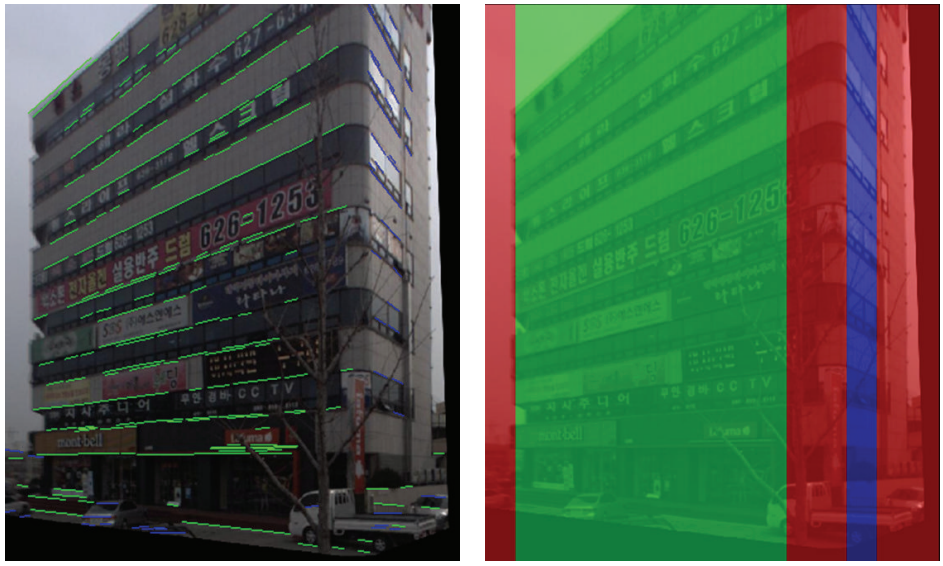
3.5. Segmentation and Validation. This step involves the segmentation and validation of individual buildings. As noted in earlier studies [23, 24], the most important information is the outer boundary in the image. Previous research has attempted to obtain the outer boundary of a building by estimating the vanishing point and corresponding line segments. Usually, the method proposed in [23] gives robust results, because horizontal line segments are less affected by occlusions, whereas [24], which relies on the vertical vanishing point, suffers from occlusions caused by pedestrians, trees, and cars. The outer boundary can be obtained by minimizing the 1D Markov random field energy, which is defined as

$$E(\mathscr{L}) = \sum_{i=1}^{n} D_i(l_i) + \sum_{i=1}^{n-1} V_{i,i+1}(l_i, l_{i+1}) \quad \text{(see [18])} \tag{2}$$

$$D_i(l_i) = \begin{cases} [T(x_i) > \beta] \cdot \lambda & \text{if } l_i = 0 \\ 100 - C_{l_i}(x_i) & \text{otherwise} \end{cases} \tag{3}$$

$$V_{i,j}(l_i, l_j) = [l_i \neq l_j] \cdot \mu\left(C_{l_i}(x_i) + C_{l_j}(x_j)\right), \tag{4}$$

where $\mathscr{L} = (l_1, l_2, \ldots, l_n)$ is the labeling of the entire column in the image so that n is the pixel width of the image and $l_i \in \{1, 2, \ldots, m\}$ is the label for each column x_i where m is the number of detected horizontal line orientations (e.g., $m = 2$ in Figure 4(a)). $C_{l_i}(x_i)$ is the number of line segments with the specific horizontal line orientation in the column x_i, and the total number of line segments of any orientation in x_i is $T(x_i) = \sum_j C_j(x_i)$. Therefore $D_i(l_i)$ is the unary potential, which has a lower energy when there are more horizontal line segments crossing x_i where β is the threshold and λ controls the cost for no-façade region. $V_{i,i+1}(l_i, l_{i+1})$ is the pairwise potential for the line segments corresponding to the vanishing points of the x_i and x_{i+1}, where μ is the weight factor. This describes the smoothness factor when labeling the different pixel values by providing higher energy when the labels differ between those pixels. The result is illustrated in Figure 4.



(a) (b)

Figure 4: (a) Detected horizontal line segments (green and blue); (b) segmentation results using [23].

In Figure 4(a), line segments corresponding to the two vanishing points from two major facades of the building are illustrated on top of the image with green and blue lines. The result labeling is illustrated in Figure 4(b), which minimizes the energy defined in (2). Each green and blue label indicates the same façade area in the image and the red label indicates no façade detected. To generate the rectified image according to the vanishing point, we applied metric rectification [25] and bilinear interpolation in image transformation, to fill in lost pixels. This is not the finished image, because the horizontal line segment method relies on the detected line segment. Hence, a novel image segmentation method that adopts the additional characteristics in the energy term is needed.

For validation, the segmented building boundary is evaluated for the presence of three possible defects in the segmented image.

(1) The existence of the building in the segmented region.

(2) The equality of the segmented building with a building in the 2D map.

(3) The completeness of the segmentation, which checks for pixel loss in the segmentation stage.

If the building consists of a set of planes, its existence can be confirmed by means of vanishing point estimation. The relationship between the normal directions of the planes can be calculated, and this can be compared to the relationship between the normal directions from the digital map to check the degree of sameness. This data can also be fed back to the second step to correct the position and orientation information.

Finally, pixel loss can occur even when the error correction/compensation process is complete. We expect that the continuity of the image segments used in the vanishing point estimation process can be used with the color information to guarantee subpixel accuracy in the segmented result. The resulting image is expected to resemble that shown in Figure 5. The leftmost figure shows the reprojected ROI image from the error correction/compensation process; through the segmentation, we obtain the outer boundary image of the target building illustrated in the middle. Note that the image is vertically rectified here. The rightmost figure illustrates an individually segmented image of major façades, which is classified from the vanishing point estimation and metric rectification.

3.6. Model Refinement and Texture Mapping. After segmentation, a process of model refinement and texture mapping is required. Model refinement involves modifying the geometrical shape of the buildings, with height being the most significant factor. The building height can be estimated from the vertical field of view in the original panoramic image, which was mapped inversely from the segmented vertical edge in the previous step. In addition, because we are referring to digital map data, building heights can be estimated using simple trigonometry by combining the distance from the camera to the building edge. Additional refinements can be done by adopting the methods in [26, 27], which extracts detailed shapes from images. The base 3D model is a quasi-Manhattan world in which buildings are composed of a set of planes. Thus, further refinement can improve the quality of the resulting 3D model both visually and physically.

The texture mapping stage is quite straightforward once the previous steps are complete. The resulting model is rendered using color data from the texture image and the UV-coordinate values of the vertices in the 3D model. To simplify the UV-coordinate modification, we consider the width of a single texture image to give the total perimeter of the entire building and assign to each façade image the ratio between this perimeter and the façade length. The resulting texture

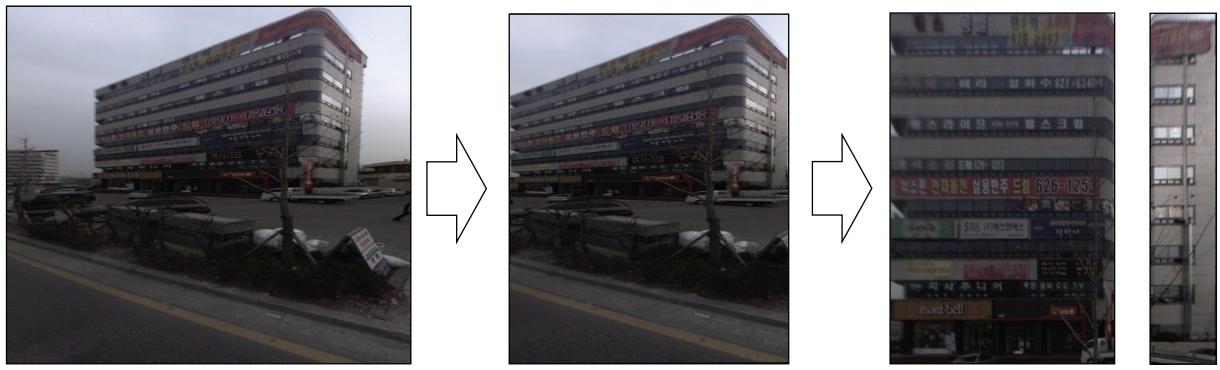


Figure 5: Image segmentation and validation. From the regional image of the building to the individual façade.

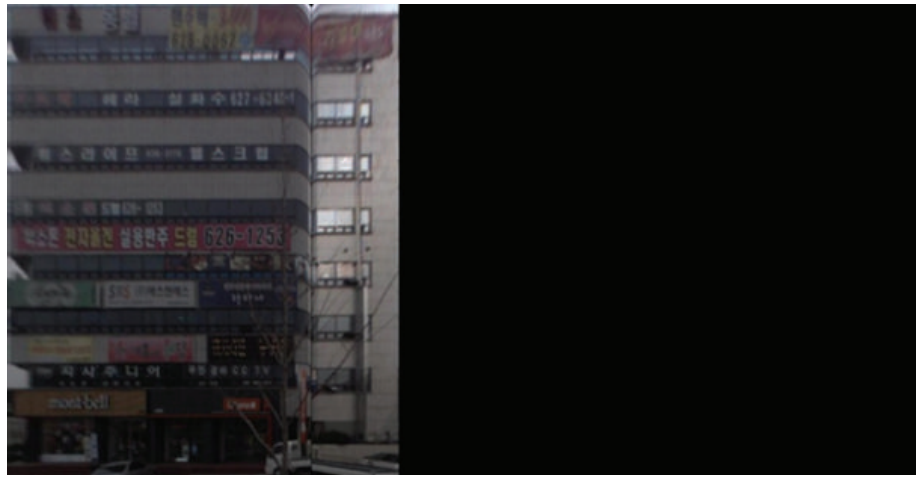


Figure 6: Resulting texture image.

is illustrated in Figure 6. Notice that the right portion of the texture image is empty since the panoramic image from the MMS only catches the façade facing the road. Therefore, only the image of the façade collected from panoramic image appears in the left portion of the texture image.

4. Experiments

4.1. Experimental Environment. For the experiments, we obtained panorama images from around Daejeon, South Korea, as well as coupled position and orientation data from a GPS/INS sensor. The source panorama images have a resolution of 5400 × 2700 pixels and a 360° horizontal/180° vertical field of view. Each panorama image was generated by stitching six perspective images using equidistant cylindrical mapping. A total of 233 images were gathered in the area bounded by [36.358933°, 36.339820°] latitude and [127.432165°, 127.436540°] longitude. The applied digital map was drawn at a scale of 1 : 5000, giving a horizontal accuracy of 1 m. By selecting digital map data through its classification code, only information related to the buildings was utilized.

For the 3D image/building analysis environment, we adopted Unity3D version 4.3 with ray casting and vector calculation functionality. Before the image/building analysis, the digital map data was parsed using the shapefile C Library in the Unity3D environment to generate the base 3D model. Error correction/compensation, image segmentation, and validation processes were implemented using MATLAB R2013a. A standard desktop PC (Intel Core i5 CPU 3.4 GHz, 4 GB RAM, Windows 7) was used as hardware.

4.2. Experimental Results. The resulting model, illustrated in Figure 7, has a photo-realistic appearance. This is because of the high-resolution texture obtained from the panoramic image. The following characteristics of the resulting model can be observed.

(1) Rich visual information from the photo-realistic appearance.

(2) Identical height values as for buildings in the physical world.

(3) Accurate geospatial information, identical to that of the digital map.

(4) Quasi-Manhattan world model that has a complete prismatic mesh structure.

The photo-realistic appearance provides an identical visual experience to the user, which is an important factor

Figure 7: Resulting textured building models.

(a) (b)

Figure 8: Comparison between our method (a) and aerial image-based method (b).

in immersive virtual environments. Additionally, the remaining text on buildings and signs provides additional visual information. The difference in visual quality is illustrated in Figure 8, which compares a 3D building model from an aerial image in V-World [28], a virtual earth serviced by the National Geographic Information Institute of Korea, and our result. The height values in our result are identical to those of the actual buildings; this information was not provided in the digital map. The height information directly influences the physical reliability. Moreover, because the position and direction of the resulting building models are coordinated with the information from the digital map, physical reliability is maintained for most applications. For instance, Figure 9 provides a comparison between the view from the center of the panoramic images mapped into the unit sphere and the resulting 3D world. The difference is less noticeable in terms of the buildings. Finally, as the geometry of the building is based on the footprint extruded from the digital map, the resulting 3D city model can be easily utilized in interactive virtual reality systems. This is because the complete prismatic mesh structure produces credible results for the algorithms that operate interactive content, for instance, in collision detection. In contrast to the results obtained using the stereo matching-based reconstruction, our approach does not require postprocessing.

Since the objective our research is offline generation of the model from the source data, the computation time is not the important target. Nevertheless, the image/building analysis takes 3.2 seconds for about 120 image locations with 7711 buildings in the digital map. The computation time of this process depends on the number of images and buildings in the targeted region. On the other hand, the computation time in the vanishing point estimation and reprojection in the error correction/compensation process as well as in segmentation depends on various factors, including the number of line segments detected, the resolution of the panoramic image, and the region occupied by the target building. For instance, the reprojection takes 36 seconds for a 30-degree horizontal and 120-degree vertical FOV which produces a 647×2977 resolution image from a 5400×2700 panoramic image. The processing time increases to 71 seconds for a 60-degree horizontal image with the same vertical FOV which produces a 1311×2977 resolution image. Vanishing point estimation takes 10 seconds on average by applying the line segments detector algorithm [29] and standard RANSAC technique. Then, the segmentation takes 2 seconds to solve the 1D minimization problem described in (2) using a dynamic programming method. The reprojection, vanishing point estimation, and segmentation are computed independently per building. Therefore the overall computation linearly increases

(a) (b)

Figure 9: Scene from the same geospatial position in resulting city model (a) and panoramic image in unit sphere (b).

according to the number of buildings. Although the overall results are positive, the proposed framework has several limitations. First, it is possible for small features to be truncated in the segmentation process, and in some cases, the segmentation was not successful. As we mentioned in Section 3.3, the current segmentation method is based purely on previous vanishing point estimation and vertical/horizontal labeling research [23]. Hence, the segmentation results are dependent on the existence of edge information, which can be disturbed by occlusions. As we can guarantee that the building exists in the ROI after error correction, color, and texture features [30] may be needed to complement the vanishing-point method and a novel segmentation algorithm [31] could be adopted. At this point, the error correction process is not complete, so in some cases buildings did not exist in the expected ROI. The error correction process can be improved using edge-based wide-baseline stereo matching. Edges would be identified completely after segmentation, because the validation process guarantees the completeness of the segmentation.

5. Conclusion

In this paper, we proposed a framework for the automatic generation of a textured high-resolution 3D city model for ground-level applications. Our approach was employed to produce a complete prismatic mesh based on a quasi-Manhattan world model. The cost-effectiveness of the framework is ensured by its use of MMS data, which allows access to a massive number of images and a large coverage area. To address the problems of existing MMS-based methods, our framework uses digital map data in four major steps. The proposed approach combines existing techniques with novel processes in each of these steps.

In future work, we will consider combining data from different sources. This is because data from the MMS system are restricted, as images of the façade facing away from the road cannot be gathered. The missing data can be supplemented by processing aerial images [32] or pedestrian-collected data. Aerial images can only provide low-resolution textures at ground level but can also be source data for the updated 2D digital map [33]. Pedestrian-collected data can generate high-resolution images that are similar to the panoramic images, although the data collection process cannot be made cost-effective. Thus, a compromise approach should be proposed according to the intended application. Moreover, the overall appearance of the building can be changed by reconstruction or alteration of its exterior over time, which could cause a mismatch between the 2D digital map and the panoramic image. Therefore, research should continue on detection of mismatches using automated, ground-level photography. The removal of occluding objects should focus on redundant images of the same building. Occluding objects can be severely disadvantageous to the visual experience, so occlusion removal will increase the visual quality of the resulting 3D city model.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publishing of this paper.

Acknowledgments

This work was supported by the Human Resources Development Program (no. 20134030200300) of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government Ministry of Trade, Industry, and Energy and Development of Integration and Automation Technology for Nuclear Plant Life-cycle Management grant funded by the Korea government Ministry of Knowledge Economy (2011T100200145).

References

[1] Google Earth, http://www.google.com/earth/.

[2] P. Musialski, P. Wonka, D. G. Aliaga, M. Wimmer, L. Gool, and W. Purgathofer, "A survey of urban reconstruction," *Computer Graphics Forum*, vol. 32, no. 6, pp. 146–177, 2013.

[3] J. Yao, H. Tawfik, and T. Fernando, "A GIS based virtual urban simulation environment," in *Proceedings of the 6th International Conference on Computational Science—Volume Part III*, pp. 60–68, 2006.

[4] B. Hagedorn and J. Döllner, "High-level web service for 3D building information visualization and analysis," in *Proceedings of the 15th ACM International Symposium on Advances in Geographic Information Systems (GIS '07)*, pp. 51–58, November 2007.

[5] R. B. Rusu, *Semantic 3D Object Maps for Everyday Robot Manipulation*, Springer, New York, NY, USA, 2013.

[6] N. J. Shih, C. Y. Lee, and T. Y. Chan, "3D scans of as-built street scenes for virtual environments," in *Proceedings of the Symposium on Simulation for Architecture and Urban Design*, pp. 46–51, 2011.

[7] D. Visintini, A. Spangher, and B. Fico, "The VRML model of Victoria Square in Gorizia (Italy) from laser scanning and photogrammetric 3D surveys," in *Proceedings of the 12th International Conference on 3D Web Technology (Web3D '07)*, pp. 165–168, Umbria, Italy, April 2007.

[8] G. Vosselman and H. G. Mass, *Airborne and Terrestrial Laser Scanning*, CRC Press, New York, NY, USA, 2010.

[9] X. Li, C. Wu, C. Zach, S. Lazebnik, and J. Frahm, "Modeling and recognition of landmark image collections using iconic scene graphs," in *Proceedings of the 10th European Conference on Computer Vision*, pp. 427–440, 2008.

[10] P. E. Debevec, C. J. Taylor, and J. Malik, "Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '96)*, pp. 11–20, New York, NY, USA, August 1996.

[11] X. Zheng, X. Zhang, and P. Guo, "Building modeling from a single image applied in urban reconstruction," in *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry (VRCAI '11)*, pp. 225–234, December 2011.

[12] O. Barinova, V. Konushin, A. Yakubenko, K. Lee, H. Lim, and A. Konushin, "Fast automatic single-view 3-d reconstruction of urban scenes," in *Proceedings of the 10th European Conference on Computer Vision*, pp. 100–113, 2008.

[13] H. Kim, Y. Kang, and S. Han, "A study on the photo-realistic 3D city modeling using the omnidirectional image and digital maps," Transactions of the Society of CAD/CAM Engineers, Accepted.

[14] Naver Maps, http://map.naver.com/.

[15] Google Street View, https://www.google.com/maps/.

[16] Baidu, http://www.map.baidu.com/.

[17] N. Haala and M. Kada, "An update on automatic 3D building reconstruction," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 6, pp. 570–580, 2010.

[18] A. Taneja, L. Ballan, and M. Pollefeys, "Registration of spherical panoramic images with cadastral 3D models," in *Proceedings of the 2nd Joint 3DIM/3DPVT Conference: 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT '12)*, pp. 479–486, October 2012.

[19] R. Cipolla, D. Robertson, and B. Tordoff, "Image-based localization," in *Proceedings of the International Conference on Virtual Systems and Multimedia*, pp. 22–29, 2004.

[20] P. David and S. Ho, "Orientation descriptors for localization in urban environments," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '11)*, pp. 494–501, San Francisco, Calif, USA, September 2011.

[21] Z. Kang, L. Zhang, S. Zlatanova, and J. Li, "An automatic mosaicking method for building facade texture mapping using a monocular close-range image sequence," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 3, pp. 282–293, 2010.

[22] R. Cipolla, D. Robertson, and B. Tordoff, "Image-based Localisation," in *Proceeding of International Conference on Virtual System and Multimedia (VSMM '04)*, pp. 22–29, 2004.

[23] G. Wan and S. Li, "Automatic facades segmentation using detected lines and vanishing points," in *Proceedings of the 4th International Congress on Image and Signal Processing (CISP '11)*, pp. 1214–1217, Shanghai, China, October 2011.

[24] D. Kim, H. Trinh, and K. Jo, "Object recognition of outdoor environment by segmented regions for robot navigation," in *Proceedings of the Intelligent Computing 3rd International Conference on Advanced Intelligent Computing Theories and Applications*, pp. 1192–1201, 2007.

[25] D. Liebowitz and A. Zisserman, "Metric rectification for perspective images of planes," in *Proceedings of the 11th IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '98)*, pp. 482–488, June 1998.

[26] N. Haala and C. Brenner, "Extraction of buildings and trees in urban environments," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 54, no. 2-3, pp. 130–137, 1999.

[27] K. Schindler and J. Bauer, "A model-based method for building reconstruction," in *Proceedings of the International Conference on Computer Vision Workshop on Higher-Level Knowledge in 3D Modeling and Motion (HLK '03)*, pp. 74–82, 2003.

[28] Vworld, http://map.vworld.kr/.

[29] R. Grompone von Gioi, J. Jakubowicz, J. Morel, and G. Randall, "LSD: a fast line segment detector with a false detection control," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 722–732, 2010.

[30] S. K. Vipparthi and S. K. Nagar, "Color directional local quinary patterns for content based indexing and retrieval," *Human-Centric Computing and Information Sciences*, vol. 4, no. 1, pp. 1–13, 2014.

[31] H. T. Manh and G. Lee, "Small object segmentation based on visual saliency in natural images," *Journal of Information Processing Systems*, vol. 9, no. 4, pp. 592–601, 2013.

[32] L. H. Liew, B. Y. Lee, Y. C. Wang, and W. Cheah, "Aerial images rectification using non-parametric approach," *Journal of Convergence*, vol. 4, no. 2, pp. 15–22, 2013.

[33] A. P. Nyaruhuma, M. Gerke, G. Vosselman, and E. G. Mtalo, "Verification of 2D building outlines using oblique airborne images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 71, pp. 62–75, 2012.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 350876, 9 pages
http://dx.doi.org/10.1155/2014/350876

Research Article

Secure eHealth-Care Service on Self-Organizing Software Platform

Im Y. Jung, Gil-Jin Jang, and Soon-Ju Kang

School of Electronics Engineering, Kyungpook National University, Daegu 702-701, Republic of Korea

Correspondence should be addressed to Gil-Jin Jang; gjang@knu.ac.kr and Soon-Ju Kang; sjkang@ee.knu.ac.kr

Received 8 April 2014; Accepted 30 May 2014; Published 17 July 2014

Academic Editor: Jong-Hyuk Park

There are several applications connected to IT health devices on the self-organizing software platform (SoSp) that allow patients or elderly users to be cared for remotely by their family doctors under normal circumstances or during emergencies. An evaluation of the SoSp applied through PAAR watch/self-organizing software platform router was conducted targeting a simple user interface for aging users, without the existence of extrasettings based on patient movement. On the other hand, like normal medical records, the access to, and transmission of, health information via PAAR watch/self-organizing software platform requires privacy protection. This paper proposes a security framework for health information management of the SoSp. The proposed framework was designed to ensure easy detection of identification information for typical users. In addition, it provides powerful protection of the user's health information.

1. Introduction

Self-organization is a process in which the structure and functionality (pattern) of a system at the global level emerge solely from numerous interactions among the lower-level components without any external or centralized control. The system components interact in a local context, either by means of direct communication or through environmental observations, without reference to a global pattern [1].

The self-organizing software platform (SoSp) is a platform designed to actualize the self-organizing function of communication devices [2]. A SoSp is combined IT health devices. Such devices are designed to automatically interact with the physical environment and users based on the necessities and the surroundings of the users [3], for example, wearable healthcare, medical devices, watches, and bicycles, with ubiquitous computing allowing users to take care of their health needs easily regardless of their location or time and to cope with certain types of emergencies. The SoSp and its applications are practical products of IT convergence research for health improvement and are intended to be utilized by elderly users or patients. Ordinarily, people can check their health through periodic measurements and transmit their biosignals to their family doctors. During an emergency, they can call their doctors by pressing an emergency button and transmit their biosignals to them immediately regardless of their location or time [4, 5].

In this paper, an essential security framework to maintain, access, and transmit health information, such as an electrocardiogram (ECG) or other biosignals, is proposed. As sensitive private data, health information needs to be kept secure and should be accessed only by authorized persons. In addition, the users should be identified when their biosignals are acquired and accessed. The proposed framework was designed to ensure easy detection of verification information without the burden of using a smartcard or memorized password, particularly to patients or elderly users. However, it provides powerful protection of health information and a unique identity verification of biometric recognition [6].

The rest of this paper is organized as follows: Section 2 describes previous work related to this topic. Section 3 presents a brief review of the SoSp and its requirements of health information management. The proposed security framework for health information management applied on the SoSp is presented in Section 4. A security analysis and overhead estimation are then demonstrated in Section 5. Finally, some concluding remarks and areas of future work are provided in Section 6.

2. Related Works

Research on biometric authentication is actively being conducted as a means for password replacement [7]. Biometric features (fingerprints, faces, irises, hand geometries, palmprints, etc.) can be neither lost nor forgotten. They are neither copied nor shared easily [8]. In addition, they are extremely difficult to forge or distribute. Because they maintain their uniqueness, they are difficult to guess [9]. As powerful mobile devices and the internet are continuously developing, biometric authentication is being grafted into them [10]. Mobile biometric authentication using face and voice recognition simultaneously on a Nokia N900 mobile device was also introduced [11]. An authentication system using multiple biometrics such as face and hand features has also been studied [12, 13].

However, it is a significant issue how to preserve and to transmit the information for biometric authentication securely as well as the additional cost to adopt the device to acquire biometric data [14]. On the other hand, function creep, which refers to biometric data used outside of the original purpose, exists [15–17]. These days, personal information being sold or passed around without proper consent is a series issue. The unique and permanent nature of biometric information adds a more serious dimension to such a breach of confidentiality, that is, unlike passwords, fingerprints or retinal patterns cannot be changed when an identity theft is suspected [18]. A case involving Emilio Calatayud in the United States shows that systems containing an aggregation of identifiable personal information can be abused [19]. When patients use smartcards or barcodes to protect their privacy, managing such cards and barcodes against theft or loss becomes a double burden. Using cryptography to protect health information can generate complexity in the application, for example, the choice of whose information to encrypt (pharmacist, pharmacy, or group of pharmacies) and which public keys to use [20]. In [9], an efficient biometric-based authentication scheme for a telecare medicine information system (TMIS) with a nonce was proposed. When users are cared for using a TMIS, their mobile devices are connected to the TMIS through a public network, for example, the Internet. The goal is to overcome Internet security risks [20, 21].

On the other hand, the authors in [22] proposed a framework for security management in a self-organized mobile ad hoc network based on the assumption that individual nodes are themselves responsible for their own security level. However, in a network where health information is delivered, authentication and authorization should be further stressed because a node can be misused by another entity, that is to say, another person. The authors in [23] proposed a security protocol for self-organizing data storage through periodic verification. A self-organizing trust model was studied to trust the communications among nodes in P2P systems [24].

Researches into the security of the acquisition and transmission of health information have been conducted. However, there has been no research regarding health information management on internet of things (IoT) [25], especially on SoSp network.

3. Health Information Management on SoSp

3.1. Domain Description. Figure 1 shows a SoSp domain used to manage health information. The whole environment can be divided into several small spaces such as rooms or floors. Such spaces, called *unit spaces,* are the basic units for assessing location awareness. Communication devices are divided into stationary nodes and mobile nodes. A node means a communication device that is implemented in hardware. Mobile nodes are characterized through low-speed operations and can be attached to a person or physical mobile object in the form of small tags with limited H/W (e.g., 8-bit MCU, 4 K of SRAM, and a coin battery) and communication functionality. Unlike RFID tags, however, mobile nodes can communicate bidirectionally. In Figure 1, a mobile node is a mobile self-organizing software platform router (SoSpR) implemented on a smartphone. It provides a text message connection, emergent calls, and a fixed SoSpR agent. Stationary nodes attached to ceilings or walls of the unit spaces, for example, the fixed SoSpR shown in Figure 1, are characterized by high-speed operations with powerful H/W (e.g., an Arm Cortex A8 MCU with 512 MB of SDRAM, an 8 GB SDCard, and an IEEE 802.15.4 transceiver) compared with a mobile node. They are intended to function as location references and communication access for mobile nodes. A stationary node has a wired network for communication between stationary nodes and a wireless (sensor) network for communication between mobile nodes. The communication between stationary and mobile nodes is limited to a single hop. This approach can help minimize the network congestion and delay problems that take the form of a broadcast storm, packet replication, or routing table overflow in a multihop *ad hoc* sensor network [26, 27].

Figure 2 shows the SW stack of the SoSpR for health information management on the SoSp network [4]. The SW stack is composed of three parts: transportation for real-time streaming of health data, coordination for interaction, and user interface. The transportation part consists of messaging middleware, service broker and discovery, service routing, data processing, and simulation. Messaging middleware supports streaming transmission through publish and subscribe service. In Figure 2, Service Broker & Discovery creates or searches for new services. The coordination part provides cooperation among distributed services based on the consensus, group management, leader election, and presence protocol. This paper focuses on a security module of the SoSpR and distributed personal health record storage.

3.2. Healthcare via SoSp. Connected Patient in Home [4] states the service and its infrastructure, which can monitor patients or the aged at home in real-time and transmit the data monitored to their doctors using mobile technology [28]. One of the applications is the control of medicine dispersal. Not only the service and its infrastructure can reduce the user's medical expenses, but also they can improve the interaction between patients and doctors.

The SoSp as a self-organizing middleware platform for real-time biosignal transmission acquires various biosignals (e.g., ECG streaming data) from their measuring devices,

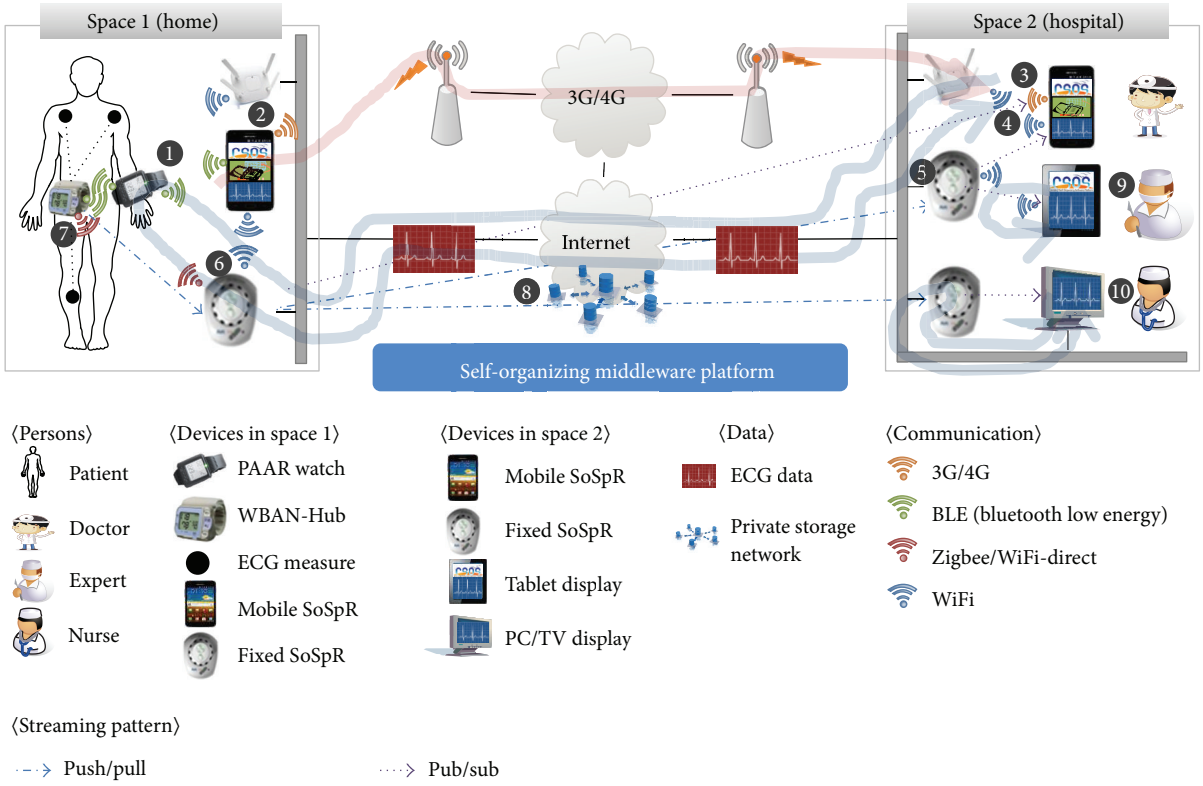


Figure 1: Domain description of SoSp for health information management [4].

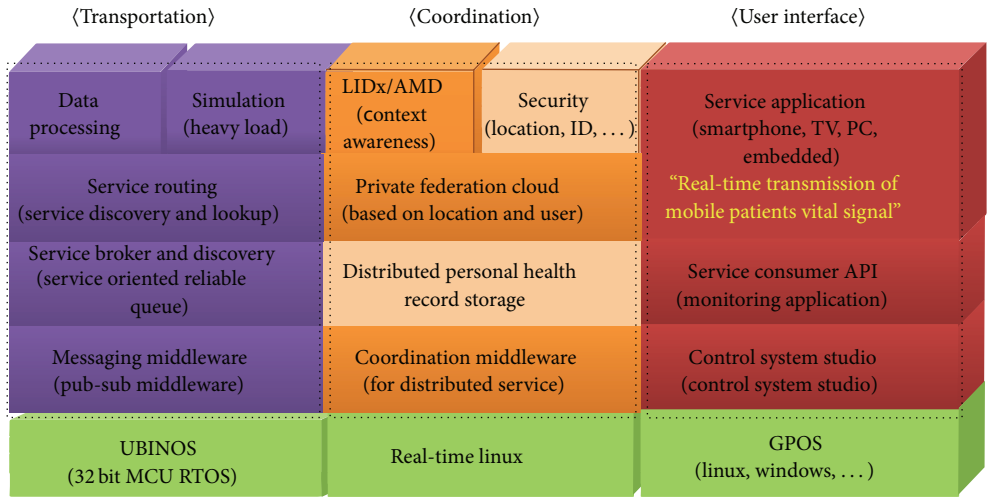


Figure 2: SW stack of SoSpR for health information management on SoSp [4].

transmits them to the SoSpR, and saves them in safe storage in real-time. Several devices can receive the signals from the SoSpR and display them. That is to say, the devices operated and managed by patients or the aged can interact with each other on the SoSp network and provide context-aware services for healthcare without any presetting. A flexible handover while moving and a predefined protocol can be used to handle emergencies.

Figure 1 shows the SoSp domain for health information management [4].

(1) When a patient or elderly user, wearing an ECG sensor, WBAN-Hub, and PAAR watch, presses the emergency button

in the PAAR watch, the PAAR watch notifies the mobile SoSpR (e.g., a smartphone) of the emergent state through bluetooth low energy (BLE) communication.

(2) The mobile SoSpR sends an SMS to a predefined person such as a doctor, medical team worker, or other family members.

(3, 4) When a doctor receives the emergency SMS, they request the biosignal of the patient or elderly user from a fixed SoSpR near in space 2, shown in Figure 1, if necessary.

(5) The fixed SoSpR in space 2 delivers a request to the fixed SoSpR, which can provide streaming service of the biosignal of the patient or elder user in space 1.

(6) The fixed SoSpR in space 1 sends a message to a WBAN-Hub allowing it to start measuring the ECG signal of the patient or elderly user.

(7) As WBAN-Hub measures the ECG signal, it sends the signal data to the fixed SoSpR in space 1 using a push-pull streaming pattern [29].

(8) The biosignal is saved in a private storage network and simultaneously sent to the client device that requested the signal. If the client device is a smartphone, the signal is transmitted through WiFi using a publish-subscribe pattern [30]. If the devices cannot be accessed through WiFi, the signal is sent to the fixed SoSpR near the devices using a push-pull streaming pattern.

(9, 10) Other authorized medical team workers or family members can monitor the biosignal of the patient or elderly user using a tablet PC, desktop PC, or TV located near the fixed SoSpR, which provides streaming service.

3.3. Requirements of Health Information Management on SoSp. At the domain handling the health information including the biosignal, privacy protection, authentication, and authorization are very important. This section describes the domain requirements for health information management on the SoSp including the security requirements.

The overlay network constructed by the SoSp is vulnerable to the same threats as any wireless network. However, the SoSp has the following additional threats and vulnerabilities owing to its basic nature [22].

(i) There is a lack of central administration, and neither central control nor prior contact is assumed.

(ii) The routing mechanisms are more vulnerable than in conventional networks because each node can act as a relay.

(iii) In terms of cooperation, if a node does not respect the cooperation rules, that is, it is selfish, the performance of the network can be severely affected.

(iv) There is a variation in memory and computation resources, in which many of the nodes are expected to be low-priced consumer electronics with cheap and slow computation capability and a limited storage size.

(v) Finally, there is an energy constraint during operation, in which many of the nodes are expected to operate on battery power. Sleep or standby modes are used to conserve energy, during which they may not be reachable. A sleep deprivation torture (exhausting the battery power) attack may be implemented by attackers.

These threats place great demand for flexibility of the security system because one must be capable of tailoring the range of supplied security services toward a wide variety of network topologies and application requirements, rather than fixing them. Further, each application imposes cost, performance, complexity, flexibility, and ease-of-use constraints, which affect the feasibility of a particular security solution [31].

In addition, node security, secure information handling, authentication, and authorization are needed. In particular, easy human computer interaction (HCI) and an efficient and simple sequence of security mechanisms should be seriously considered for elderly users and patients.

A security framework for health information management on the SoSp was proposed that satisfies the requirements mentioned in this section.

4. Security Framework for Health Information Management on SoSp

4.1. System Architecture. Figure 3 shows the proposed system architecture of security framework for health information management on the SoSp. The health information of the patients, authorization information of the doctors who can access the information, and the biometric recognition information of the patients and doctors are saved in secure storage. Biometric recognition devices can be equipped in a PAAR watch [4] worn by the patients or doctors. Multiple biometric recognition devices can be equipped in one PAAR watch. Through such devices, biometric recognition information, such as fingerprint, face, voice, and iris information, can be easily acquired. Several types of biometric recognition information can be combined to improve the exactness of verification.

4.2. Service Management. The fixed SoSpR, which loads the streaming service of a biosignal, can provide streaming service for mobile nodes or devices that have registered for a service subscription. However, data leakage during transmission is one weak point in an overlay network because there is no central server that manages all of the nodes providing streaming service, or the paths that the data are delivered upon. Any security scheme should be lightweight so that it does not cause a delay in streaming.

Figure 4 shows the security scheme used to protect health information delivered on the SoSp.

When a doctor requests the streaming of a patient's biosignal, he/she issues one ID of the patient, $\mathrm{id}_{\mathrm{temporary}}$, and send $\mathrm{ID}_{\mathrm{patient}}$, which is composed of $\mathrm{id}_{\mathrm{temporary}}$ and the issuing timestamp, T, to the mobile SoSpR of the patient directly. Consider

$$\mathrm{ID}_{\mathrm{patient}} = \mathrm{id}_{\mathrm{temporary}} \,||\, T. \tag{1}$$

If the issuing time is 2014-01-20 15:30, T will be a string of 201401201530.

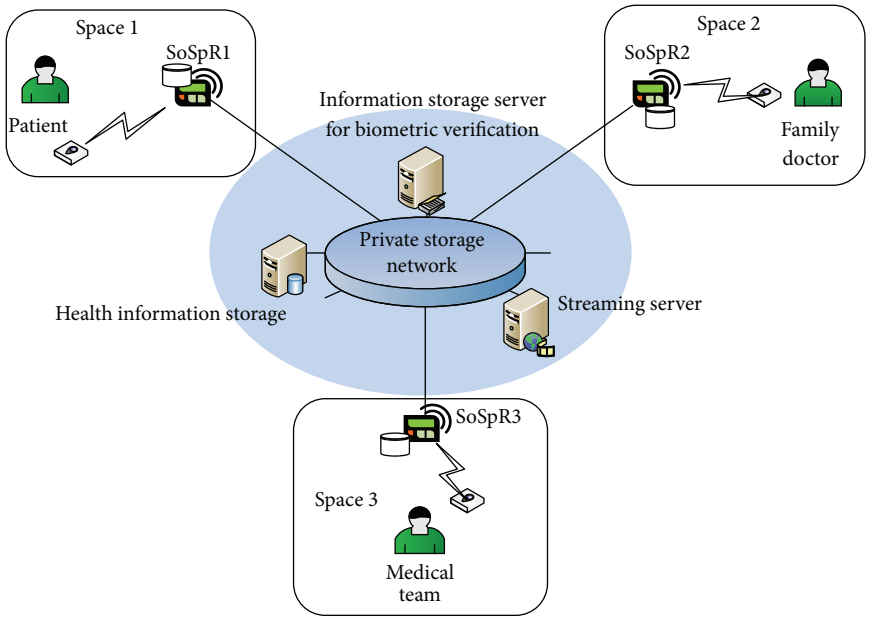


FIGURE 3: System architecture of security framework for health information management on SoSp.

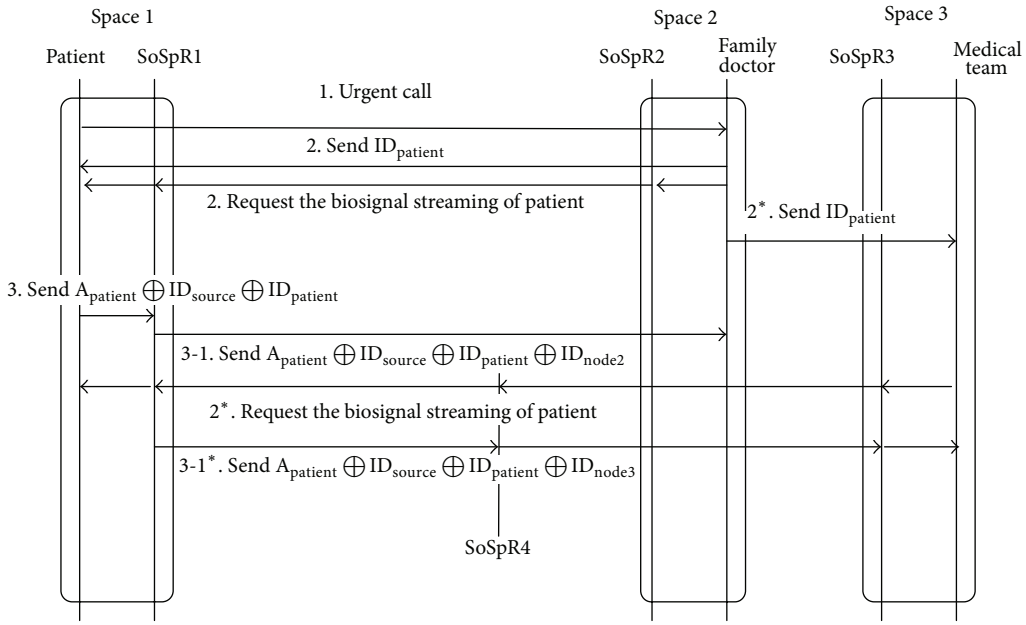


FIGURE 4: Health information protection on SoSp.

When the biosignal of the patient is A_{patient}, and the ID of the equipment used to measure the biosignal is $\text{ID}_{\text{source}}$, the streaming data includes the XOR results of A_{patient}, $\text{ID}_{\text{source}}$, and $\text{ID}_{\text{patient}}$. Consider

$$A_{\text{patient}} \oplus \text{ID}_{\text{source}} \oplus \text{ID}_{\text{patient}}. \tag{2}$$

An XOR operation is very fast and lightweight.

If the destination identity, ID_{node}, is added,

$$A_{\text{patient}} \oplus \text{ID}_{\text{source}} \oplus \text{ID}_{\text{patient}} \oplus \text{ID}_{\text{node}}. \tag{3}$$

Only the node whose identity is ID_{node} can know $A_{\text{patient}} \oplus \text{ID}_{\text{source}} \oplus \text{ID}_{\text{patient}}$. This prevents eavesdropping during transmission in an overlay network and ensures exact delivery to the destination.

In addition, only the doctor who knows $\text{ID}_{\text{source}}$ and $\text{ID}_{\text{patient}}$ can see A_{patient}. Therefore, when cooperative medical treatment is needed, the doctor notifies the medical team members of $\text{ID}_{\text{source}}$ and $\text{ID}_{\text{patient}}$ through another communication channel, not the SoSp network. $\text{ID}_{\text{source}}$ is permanent and $\text{ID}_{\text{patient}}$ is temporary. Because $\text{ID}_{\text{patient}}$ is a temporary identity issued by a doctor, anonymity is guaranteed and the

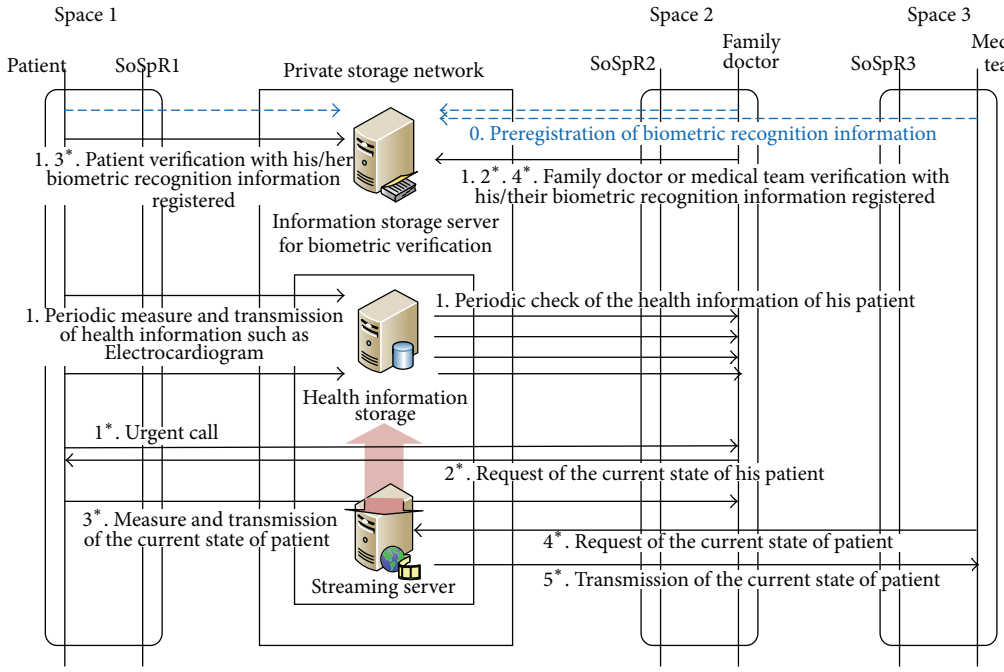


Figure 5: Secure health information management on SoSp.

patient's privacy is protected. The issuing time, T, prohibits a relay attack.

4.3. Authentication and Authorization. Figure 5 shows a sequence diagram of health information management on the SoSp.

Verification information on the biometric recognition for the patient, family doctor, and medical team should be registered in advance. Verification of the family doctor or medical team is done every time they are requested to access the patient's health information. They should be verified periodically while accessing and monitoring the biosignal of the patient because of the limited effective verification time. The verification of the patient is done each time their biosignal is measured, and the information is kept in private storage.

During an emergency, the patient can call his/her family doctor by pushing the emergency button on his/her PAAR watch. The family doctor requests the biosignal of his/her patient in real-time, who is verified simultaneously. The biosignal is measured and transmitted to both the family doctor and the private storage. When the family doctor requests cooperation of the medical team, the medical team can obtain the patient's state from the streaming server through the verification procedure.

The strong points of the proposed security framework for health information management on the SoSp are its exact verification capability using biometric recognition, secure protection of data and its simple and convenient user interface.

5. Evaluation

5.1. Framework Analysis. This section analyzes whether the proposed framework satisfies the requirements of health information handling on the SoSp.

5.1.1. Lack of Central Administration. All nodes on the SoSp should route and search for services for users without a central administration through a central server. However, it overcomes a single point of failure of the central server, the centralized burden of processing, and the nonscalability. The performance of the SoSp with mobile nodes is better than a centralized environment [32]. The proposed framework manages important data such as the biometric authentication and the biosignal of the patients or elderly users for periodic check-ups on the private storage network. This framework provides a safe deposit of private information.

5.1.2. Routing Mechanisms. In the framework, fixed SoSpRs are connected through a wired network. Mobile nodes have a one-hop connection to fixed SoSpRs. This simple scheme lessens the side effects of routing on the SoSp.

5.1.3. Cooperation. Cooperation among the nodes on the SoSp is under implementation. There are many nodes providing the same services in the SoSp network. The SoSp network will be implemented in such a way that the service group is reorganized by isolating any selfish nodes [33–36].

5.1.4. Performance Variation in Memory and Computation Resources. Based on an XOR operation, the security framework does not cause a delay in real-time streaming. We limit the operations to an XOR because it is a lightweight operation [37]. Therefore, there is no problem stemming from performance variations in the node resources in the SoSp network.

5.1.5. Energy-Constrained Operation. Most types of wireless networks have limited computational power such that they

cannot perform operations over large finite fields [37]. An XOR operation has been evaluated as having good operations and can be used for wireless network coding.

5.2. Security Analysis

5.2.1. Node Security. SoSpRs do not maintain the health information of users. All of the data are processed at the RAM of the SoSpRs. It is infeasible to acquire data that passes the SoSpRs. In the proposed security framework, the node security is less focused. Instead, the security scheme covers the overall security for the health information handling on the SoSp.

5.2.2. Secure Information Handling. A health information handling service is only served to the clients who are authenticated. In addition, only the person who knows the ID of a patient can see his/her health data.

According to the streaming patterns in the SoSp network, health information is protected in the following manner.

(1) Push-pull transmission between the WBAN-Hub and a fixed SoSpR.
Because the fixed SoSpR receives $A_{\text{patient}} \oplus \text{ID}_{\text{source}} \oplus \text{ID}_{\text{patient}}$ and transmits it, the SoSpR cannot know A_{patient}.

(2) Push-pull transmission between fixed SoSpRs.
When there is no SoSpR that provides streaming service, the biosignal is relayed to the SoSpR, which can provide the service. In this case, the SoSpR that relays the signal cannot know A_{patient} because it does not know $\text{ID}_{\text{source}} \oplus \text{ID}_{\text{patient}}$.

(3) Publish-subscribe transmission between a fixed SoSpR and various devices such as a tablet PC, desktop PC, or TV.

The proposed security framework only allows a one-hop connection between the fixed SoSpR and the end nodes. Of course, the end nodes should subscribe to the service. This lessens the security threat caused by a multihop connection.

5.2.3. Easy HCI. Identification by a password is not suitable because the process memorizing the password and its input is somewhat burden for a patient or elderly user, particularly during an emergency [38]. Because biometric authentication is easy, it is a good HCI instead of password identification.

5.2.4. Efficient and Simple Security Mechanism. An efficient and simple sequence is another important security mechanism for a patient or elderly user. The security mechanism should be fast and cooperate simply with the HCI. A self-organizing platform is characterized by no presettings or central management. The proposed security framework does not require any complex sequences that the users should follow or participate in.

5.2.5. Exact Identification and Authentication. Biometric authentication is secure because the biometric data are unique and are hard to copy or falsify. A password can be easily leaked or cracked [38].

5.2.6. Authentication and Authorization. The proposed framework provides a strict authorization for the health data. Only the doctors or medical team members who are authenticated and authorized can access the data of the patient or elderly user.

Today, many security incidents from loose authentication procedures have occurred [39]. In the proposed framework, all users are identified through their biometric authentication. Because biometric data are unique, they are difficult to copy or modify. The biometric method is also easy for patients and elderly users to use.

5.3. Estimation of Performance and Overhead. The performance of the proposed framework and the overhead imposed on the SoSp by the framework are estimated as follows.

(i) The performance of the biometric authentication added to the real-time streaming service of biosignals
One type of biometric authentication, fingerprint verification, is loaded onto smartphones. High-performance SW products for fingerprint verification of smartphone users have been developed [40]. Under the SoSpR prototype with Exynos5 Octa Cortex-A15 1.6 Ghz quad core and Cortex-A7 quad core CPU, 2 GB LPDDR3 RAM, the XOR overhead of 4 B and 8 B were negligible (~0 sec); the unit of data streaming is 3 B at current state.
It may therefore not be a problem to use such biometric verification on the SoSp in near future. In the proposed framework, a storage server is used to keep the biometric data for authentication safe and conduct a heavy authentication procedure. Therefore, biometric authentication does not cause a delay in real-time service.

(ii) The performance of the XOR operation when streaming data are created and streaming data are restored at the destination.
Because an XOR operation is sufficiently lightweight to be used as a wireless network packet [37], real-time streaming is not affected seriously by an XOR operation at the source or destination of the streaming data.
Because biometric authentication and XOR operations are extra modules for the SoSp, some overhead exists when adopting them. However, they overcome the overhead in the following way.

(iii) Biometric authentication is exact, not copy-prone, and unmodifiable. It is easy for a patient or elderly user to adopt, especially during an emergency.

(iv) An XOR operation is simple and lightweight. When such an operation is applied to health information, the information is hidden without the need for complex cryptography.

In addition, even though the streaming data are revealed, it is difficult to know whose data belongs to which $\text{ID}_{\text{patient}}$. Anonymity guarantee is another strong point of the proposed security framework.

6. Conclusion and Future Works

This paper proposed an essential security framework of the SoSp to maintain, access, and transmit health information such as a biosignal. As sensitive and private data, health information needs to be kept securely, and should be strictly accessible only by the authorized persons.

The proposed security framework requires mutual authentication between the patient or elderly user and the medical team through biometric authentication. In addition, it protects the privacy of the patient or elderly worker with a temporary anonymous ID issued by the family doctor or doctor in charge. Through a simple and lightweight operation, that is, an XOR, the biosignal is hidden. The signal can be recovered only by those persons who know the temporary ID, $ID_{patient}$, and the ID of the device measuring the biosignal, ID_{source}.

The proposed security framework is under implementation. After implementation, the framework will be tuned according to the upgraded SoSp to improve its performance and reduce the overhead. In addition, node security, particularly the SoSpR, will be intensively considered.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the IT R&D program of MSIP/KEIT (10041145, Self-Organizing Software platform (SoSp) for Welfare Devices).

References

[1] F. Dressler, "A study of self-organization mechanisms in ad hoc and sensor networks," *Computer Communications*, vol. 31, no. 13, pp. 3018–3029, 2008.

[2] Center of Self-Organizing Software-Platform, http://www.csosp.org/.

[3] S. Oh, "Using an adaptive search tree to predict user location," *Journal of Information Processing Systems*, vol. 8, no. 3, pp. 437–444, 2012.

[4] H.-Y. Kang, S.-Y. Jeong, C.-S. Ahn, Y.-J. Park, and S.-J. Kang, "Self-organizing middleware platform based on overlay network for real-time transmission of mobile patients vital signal stream," *The Journal of Korea Information and Communications Society* , vol. 38, no. 7, pp. 630–642, 2013.

[5] J. Ahn and R. Han, "An indoor augmented-reality evacuation system for the smartphone using personalized pedometry," *Human-Centric Computing and Information Sciences*, vol. 2, article 18, 2012.

[6] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[7] S. Sonkamble, R. Thool, and B. Sonkamble, "Survey of biometric recognition systems and their applications," *Journal of Theoretical and Applied Information Technology*, vol. 11, no. 1, pp. 45–51, 2010.

[8] A. Chaudhary, K. Vatwani, T. Agrawal, and J. L. Raheja, "A vision-based method to find fingertips in a closed hand," *Journal of Information Processing Systems*, vol. 8, no. 3, pp. 399–408, 2012.

[9] D. Mishra, S. Mukhopadhyay, S. Kumari, M. K. Khan, and A. Chaturvedi, "Security enhancement of a biometric based authentication scheme for telecare medicine information systems with nonce," *Journal of Medical Systems*, vol. 38, no. 5, article 41, 2014.

[10] C.-L. Tsai, C-J. Chen, and D.-J. Zhuang, "Trusted M-banking verification scheme based on a combination of OTP and biometrics," *Journal of Convergence*, vol. 3, no. 3, pp. 23–30, 2012.

[11] S. Marcel, C. Cool, C. Atanasoaei et al., "MOBIO: mobile biometric face and speaker authentication," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR '10)*, 2010.

[12] J. Rokita, A. Krzyzak, and C. Y. Suen, "Cell phones personal authentication systems using multimodal biometrics," in *Proceedings of the International Conference on Image Analysis and Recognition (ICIAR '08)*, pp. 1013–1022, 2008.

[13] S. Basak, M. I. Islam, and M. R. Amin, "A new approach to fingerprint detection using a combination of minutiae points and invariant moments parameters," *Journal of Information Processing Systems*, vol. 8, no. 3, pp. 421–436, 2012.

[14] A. P. Pons and P. Polak, "Understanding user perspectives on biometric technology," *Communications of the ACM*, vol. 51, no. 9, pp. 115–118, 2008.

[15] A. Chandra and T. Calderon, "Challenges and constraints to the diffusion of biometrics in information systems," *Communications of the ACM*, vol. 48, no. 12, pp. 101–106, 2005.

[16] A. Sprokkereef and P. de Hert, "Ethical practice in the use of biometric indentifiers within the EU," *Law, Science & Policy*, vol. 3, no. 2, pp. 177–201, 2007.

[17] L. A. Jones, A. I. Antón, and J. B. Earp, "Towards understanding user perceptions of authentication technologies," in *Proceedings of the 6th ACM Workshop on Privacy in the Electronic Society (WPES '07)*, pp. 91–98, October 2007.

[18] S. Khan and P. Gurkas, "Identification using biometric technology: issues and attitudes," in *Proceedings of the IADIS International Conference ICT, Society and Human Beings*, 2010.

[19] K. Poulsen, "DEA Data Thief Sentenced to 27 Months, Security Focus," 2002, http://online.securityfocus.com/news/1847.

[20] E. Ball, D. W. Chadwick, and D. Mundy, "Patient privacy in electronic prescription transfer," *IEEE Security & Privacy*, vol. 1, no. 2, pp. 77–80, 2003.

[21] J. L. Fernández-Alemán, I. C. Señor, P. Á. O. Lozoya, and A. Toval, "Security and privacy in electronic health records: a systematic literature review," *Journal of Biomedical Informatics*, vol. 46, no. 3, pp. 541–562, 2013.

[22] R. Savola and I. Uusitalo, "Towards node-level security management in self-organizing mobile ad hoc networks," in *Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services (AICT/ICIW '06)*, 2006.

[23] N. Oualha, M. Önen, and Y. Roudier, "A security protocol for self-organizing data storage," Research Report RR-08-208, Institut Eurecom, 2008.

[24] A. B. Can and B. Bhargava, "SORT: a self-organizing trust model for peer-to-peer systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 10, no. 1, pp. 14–27, 2013.

[25] J. Couturier, D. Sola, G. Scarso Borioli, and C. Raiciu, "How can the internet of things help to overcome current healthcare challenges," *Digiworld Economic Journal*, vol. 87, pp. 67–81, 2012.

[26] D. K. Lee, T. H. Kim, S. Y. Jeong, and S. J. Kang, "A three-tier middleware architecture supporting bidirectional location tracking of numerous mobile nodes under legacy WSN environment," *Journal of Systems Architecture*, vol. 57, no. 8, pp. 735–748, 2011.

[27] A. Shikfa, "Security issues in opportunistic networks," in *Proceedings of the 2nd International Workshop on Mobile Opportunistic Networking (MobiOpp '10)*, pp. 215–216, Pisa, Italy, February 2010.

[28] J. Kee-Yin Ng, "Ubiquitous healthcare: healthcare systems and applications enabled by mobile and wireless technologies," *Journal of Convergence*, vol. 3, no. 2, 2012.

[29] M. Fouzia and J. Subrun, "Push pull services offering SMS based m-banking system in context of Bangladesh," *International Arab Journal of e-Technology*, vol. 1, no. 3, 2010.

[30] D. Lagutin, K. Visala, A. Zahemszky, T. Burbridge, and G. F. Marias, "Roles and security in a publish/subscribe network architecture," in *Proceedings of the 15th IEEE Symposium on Computers and Communications (ISCC '10)*, pp. 68–74, June 2010.

[31] T. S. Messerges, J. Cukier, T. A. M. Kevenaar, L. Puhl, R. Struik, and E. Callaway, "A security design for a general purpose, self-organizing, multihop ad hoc wireless network," in *Proceedings of the 1st ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN '03)*, pp. 1–11, October 2003.

[32] S. Y. Jeong, H. G. Jo, and S. J. Kang, "Remote service discovery and binding architecture for soft real-time QoS in indoor location-based service," *Journal of Systems Architecture*, 2014.

[33] S. Silas, K. Ezra, and E. B. Rajsingh, "A novel fault tolerant service selection framework for pervasive computing," *Human-Centric Computing and Information Sciences*, vol. 2, p. 5, 2012.

[34] X. Zhou, Y. Ge, X. Chen, Y. Jing, and W. Sun, "A distributed cache based reliable service execution and recovery approach in MANETs," *Journal of Convergence*, vol. 3, no. 1, pp. 5–12, 2012.

[35] V. Viswanathan and I. Krishnamurthi, "Finding relevant semantic association paths through user-specific intermediate entities," *Human-Centric Computing and Information Sciences*, vol. 2, article 9, 2012.

[36] D. Werth, A. Emrich, and A. Chapko, "An ecosystem for user-generated mobile services," *Journal of Convergence*, vol. 3, no. 4, pp. 10–15, 2012.

[37] A. Khreishah, I. M. Khalil, P. Ostovari, and J. Wu, "Flow-based XOR network coding for lossy wireless networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 2321–2329, 2012.

[38] J. Catone, "Bad Form: 61% Use Same Password for Everything," January 2008.

[39] APWG, "Phishing Activity Trends Report," 2013.

[40] Precise BioMatch Mobile, http://www.precisebiometrics.com.

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 814593, 9 pages
http://dx.doi.org/10.1155/2014/814593

Research Article

Reliable Fault Classification of Induction Motors Using Texture Feature Extraction and a Multiclass Support Vector Machine

Jia Uddin,[1] Myeongsu Kang,[1] Dinh V. Nguyen,[2] and Jong-Myon Kim[1]

[1] *School of Electrical Engineering, University of Ulsan, Building No. 7, Room No. 308, 93 Daehak-ro, Nam-gu, Ulsan 680-749, Republic of Korea*
[2] *School of Computer Engineering, Hanoi University of Science and Technology, Hanoi 010000, Vietnam*

Correspondence should be addressed to Jong-Myon Kim; jongmyon.kim@gmail.com

Received 18 May 2014; Accepted 15 June 2014; Published 29 June 2014

Academic Editor: Jong-Hyuk Park

This paper proposes a method for the reliable fault detection and classification of induction motors using two-dimensional (2D) texture features and a multiclass support vector machine (MCSVM). The proposed model first converts time-domain vibration signals to 2D gray images, resulting in texture patterns (or repetitive patterns), and extracts these texture features by generating the dominant neighborhood structure (DNS) map. The principal component analysis (PCA) is then used for the purpose of dimensionality reduction of the high-dimensional feature vector including the extracted texture features due to the fact that the high-dimensional feature vector can degrade classification performance, and this paper configures an effective feature vector including discriminative fault features for diagnosis. Finally, the proposed approach utilizes the one-against-all (OAA) multiclass support vector machines (MCSVMs) to identify induction motor failures. In this study, the Gaussian radial basis function kernel cooperates with OAA MCSVMs to deal with nonlinear fault features. Experimental results demonstrate that the proposed approach outperforms three state-of-the-art fault diagnosis algorithms in terms of fault classification accuracy, yielding an average classification accuracy of 100% even in noisy environments.

1. Introduction

Induction motors are widely used in rotary machinery systems, including both heavy-and light-duty machinery [1], and play an important role in the industry due to their hardiness, low cost, and low maintenance requirements. An induction motor usually falls out of service due to the following reasons: the application of an unexpected heavy load, unsuitable or inadequate lubrication, and ineffective sealing. With the increase in the production capabilities of current manufacturing systems, machines are expected to run continuously, making unexpected interruptions due to machine failure being more costly than ever. Fault diagnosis for industrial induction motors is therefore a significant issue. Reliable, fast, automated, and state-monitoring schemes have been widely used to identify specific failures in various induction motor components.

In general, feature extraction can be performed by matching similarities between different signals [2, 3]. In order to extract the features of numerous faults, current, voltage, vibration, infrared thermography, and acoustic emission signals were utilized as sources to monitor fault states [4]. Among the signals, vibration analysis has been the most frequently employed methodology for identifying induction motor faults due to its ability to represent intrinsic information of them [5].

These vibration signals are analyzed in time domain, frequency domain, and time-frequency domain [6, 7]. The time-domain analysis determines numerous characteristics from a signal waveform, such as the root mean square (RMS), skewness, kurtosis, and crest factor, and applies these parameters in the diagnosis of inductions motors [8]. Although the scalar descriptors extracted from time signals are robust in terms of fluctuating load, they are not consistently effective, particularly in the presence of increased defects [9]. In contrast, frequency analysis considers only the frequency components, neglecting amplitude, where the frequency patterns are closed to the rotation properties of

the machine. A number of time-frequency techniques have been proposed as a trade-off between the time and frequency domain techniques.

In the traditional approaches, signals are processed as a one-dimensional (1D) representation, where the relationship information between time and frequency coefficients may be easily lost. In addition, the features collected from either the raw or processed signals have large dimensions that usually increase the computational burden of the subsequent classifier and degrade the generalization capability of a classifier. A significant number of well-known classifiers have been used for fault diagnosis, including artificial neural network (ANN), support vector machine (SVM), simplified fuzzy art map (SFAM), and fuzzy art map (FAM). Conventional feature extraction techniques that collect feature vectors from raw signals are not functional in practical scenarios, because load and rotational speed can vary at any instant. Therefore, an efficient approach for feature extraction with an effective dimension reduction technique is essential.

In this paper, a multifault detection and classification approach is proposed that first converts vibration signals to two-dimensional (2D) gray images by transforming the amplitude of the signals into the intensity of the pixels in an image. Due to the fact that the converted gray-level images show texture patterns (or repetitive patterns), the dominant neighborhood structure (DNS) map is generated to extract the texture features of the converted images in this paper. In general, a feature vector including the extracted texture features is high-dimensional and its high dimensionality can degrade classification performance. Hence, this paper exploits a principal component analysis (PCA) for dimensionality reduction of the feature vector. Finally, the reduced feature vector with discriminative fault features is used as input for the multiclass support vector machines (MCSVMs) to identify each fault in an induction motor. The proposed approach is compared with three conventional techniques in terms of classification accuracy in both noiseless and noisy environments.

The rest of the paper is organized as follows. Section 2 introduces related works, and Section 3 presents the proposed fault diagnosis method. Section 4 describes a detailed implementation of the proposed approach. Section 5 analyzes the experimental results of the proposed approach and compares the classification accuracy of the proposed approach with other conventional techniques. Finally, Section 6 concludes this paper.

2. Related Works

Several fault-state monitoring techniques have been proposed to improve robustness, reliability, and production capacity and reduce maintenance costs [4, 8, 10–19]. The proposed techniques include short-time Fourier transform (STFT), scale-invariant feature transform (SIFT), wavelet packet transform (WPT), Wigner-Ville distribution (WVD), Gabor filter, Hilbert Huang transform (HHT), empirical mode decomposition (EMD), local mean decomposition (LMD), and local characteristic-scale decomposition (LCD).

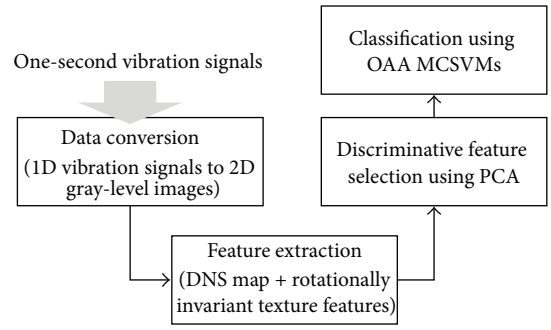


Figure 1: A block diagram of the proposed fault diagnosis methodology.

Lei et al. proposed an intelligent classification method using WPT and EMD [10]. Dimensionless time-domain feature vectors were extracted from each of the original vibration signals and preprocessed to form a combined feature set. In [11], a systematic procedure using an adaptive learning system such as ANNs [20] was proposed for fault diagnosis in induction motors, where the time-domain features were calculated from the input signals. William and Hoffman proposed an early fault detection approach using the zero-crossing time-domain features [12]. These features were collected from the input during the zero-crossing intervals. As a result, estimation of the rotational frequency was not required. Rodriguez-Donate et al. proposed a quantitative general methodology for online monitoring of an induction motor and an automatic identification of multiple faults based on the startup vibration transient analysis [4]. Yang and Kim proposed a fault diagnosis model using the Dempster-Shafer theory, a combinational rule using current and vibration signals concurrently to extract feature vectors [13]. Bafroui and Ohadi proposed a fault diagnosis model to process nonstationary vibration signals, where a resample technique at a constant angle increment is combined with the continuous wavelet transform (CWT) and measured the numerous statistical parameters from the wavelet coefficients [19]. An improved HHT-based detection methodology was proposed in [21] for removing undesirable intrinsic mode functions (IMFs), where noisy IMFs were selected based on a threshold value. A new self-adaptive time-frequency domain method, LMD, was proposed that produces a number of product functions, and a multiscale entropy of each product function was utilized to calculate the feature vectors [22]. Zheng et al. proposed a new nonstationary fault diagnosis method, LCD, which decomposes the vibration signal adaptively in a series of intrinsic scale components (ISC) for different scales [23].

3. Proposed Model

Figure 1 presents a block diagram of the proposed model consisting of the following four blocks: data conversion, texture feature extraction, texture feature reduction, and classification.

First, data conversion process is carried out in order to produce 2D gray-level images from 1D vibration signals.

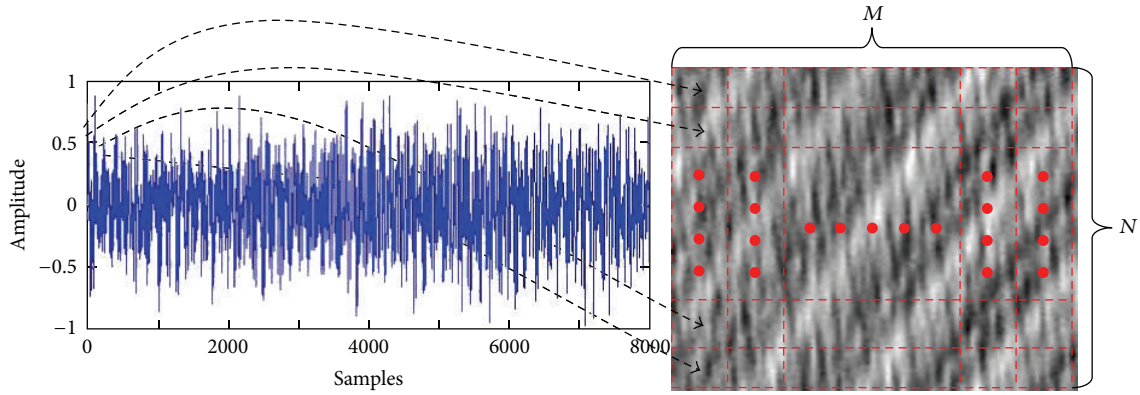


Figure 2: Data conversion process to represent a two-dimensional gray-level image from a one-dimensional vibration signal.

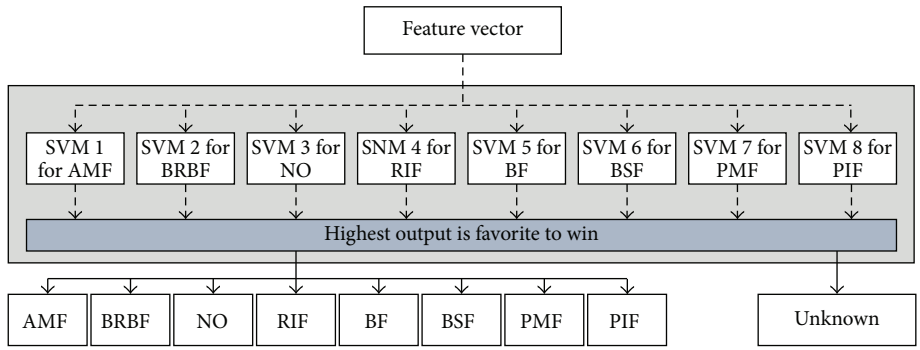


Figure 3: OAA MCSVMs for identifying induction motor failures.

In the data conversion process, the amplitude of each sample of the vibration signal is normalized to range from 0 to 255, and the normalized amplitude of each sample becomes the intensity of the corresponding pixel, as shown in Figure 2. The coordinate of the corresponding pixel for the ith sample of the vibration signal is pixel (j, k), where j = modulo (i/M), k = floor $(i/M) + 1$, M is the column length of an $M \times N$ image, and N is the row length of an $M \times N$ image.

We then generate a DNS map to extract texture features in the converted image [24] and the overall process of the DNS map generation can be summarized as follows.

Step 1. The center pixel is located at the center of the search window and the other pixels in the search window are neighboring pixels. To generate a DNS map, vectors $v(N_i)$ and $v(N_j)$ are first created, where $v(N_i)$ is the vector of the neighborhood pixel values around the pixel i surrounded by the $n \times n$ neighborhood window, $v(N_j)$ is the vector of neighboring pixels within the neighborhood window, the pixel i refers to the pixel intensity at the center of the $m \times m$ search window, and the pixel j indicates each pixel value of the search window.

Step 2. Euclidean distances are then computed between the two generated vectors and all of the pixels in the search window are replaced with Euclidean distances. In this study, the center pixel is also considered to be a neighborhood pixel and the Euclidean distance can be 0 if $v(N_i) = v(N_j)$.

Step 3. The spacing interval is the distance between pixels used to build the DNS map, which affects the resolution at which the global neighborhood similarity is captured. To produce the DNS map at a given scale of resolution, it is important to have a sufficient number of neighborhood structure maps. This can be achieved by moving the predefined $m \times m$ search window at a certain spacing interval between the previous search window and the current search window. The spacing interval was experimentally determined in this study, which is set to 5.

Since the converted texture image can be rotated by a certain rotation angle, it is necessary to extract rotationally invariant features. To deal with this problem, the DNS map values are sampled on circumferences of concentric circles of various radii centered at the center pixel of the map. However, the dimensionality of a feature vector including these texture features is high, which can degrade classification performance. Thus, selection of the most discriminative feature is vital for reliable fault diagnosis. For the purpose of dimensionality reduction in this study, we utilize PCA [25] to select the most distinctive features between classes. The PCA algorithm reduces the dimension of the feature vectors by selecting only the principal components, where a principal component is defined by a linear transformation of the original variables into a new set of uncorrelated variables. PCA finds a new set of feature vectors with principal directions (PDs) and insignificant directions (IDs), where PDs have

Table 1: Description of induction motor faults.

Type of faults	Fault description
Angular misalignment fault (AMF)	Angular misalignment is the effective angle between the two shaft centerlines, the angle between the shaft centerlines is 0.48°
Broken rotor bar fault (BRBF)	Among 34 rotor bars, 12 rotor bars are involved in the plastic deformation of the grinding furrow: 5 mm in diameter and 15 mm in depth
Parallel misalignment fault (PMF)	The offset between two centerlines of the motor and load has been changed 0.1 mm
Rotor imbalance fault (RIF)	Unbalance mass of 15.64 g cm is added at the right end of the rotor
Bearing fault (BF)	A spalling on the outer race of the bearing is replicated
Bowed shaft fault (BSF)	The shaft is slack in the middle (0.075 mm), which causes dynamic air-gap eccentricity
Phase imbalance fault (PIF)	4.3 Ω resistance is connected to one of the three-phase wires of the induction motor

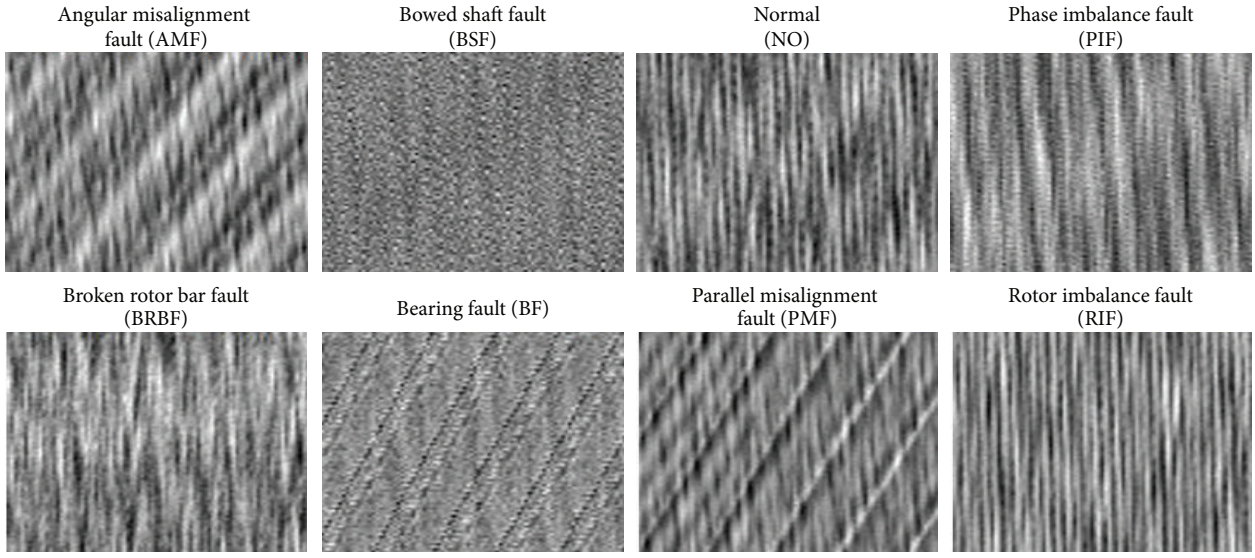


Figure 4: Examples of 2D representation of vibration signals.

large covariance and IDs have small covariance. Among these new feature vectors, only the most significant features are utilized as inputs to the classifier.

Finally, we utilize multiclass SVM (MCSVM) as a classifier, where the Gaussian radial basis function is used as a function kernel to ensure efficient nonlinear classification between faults. In order to design MCSVMs, the following three approaches can be considered: one-against-all (OAA), one-against-one (OAO), and one-acyclic-graph (OAG). Among them, the OAA method is employed, which is one of the most popular and simplest techniques for multiclass classifiers. In the OAA approach, each SVM structure discriminates one class from the others, and the final decision can be made by selecting the SVM structure yielding the highest output value.

In order to design OAA MCSVMs, each SVM structure is separately evaluated to achieve the maximum classification accuracy for its own class. Then, all SVM structures cooperate together to make a final decision. If none of the classes can recognize what type of fault the input signal is, the final decision is assigned to *unknown*. OAA MCSVMs for classifying faults are illustrated in Figure 3. To train and test OAA MCSVMs, it is necessary to build a training dataset using 50% of the 105 vibration signals for each fault, including normal signals and a test dataset with the remaining signals in this study (e.g., 53 one-second vibration signals are used as the training data set, and 52 one-second vibration signals are used as the test data set).

Table 2: Parameter setups for generating a DNS map.

Parameters	Values
Searching window size	21×21
Neighbor window	13×13
Number of central pixels	144
Size of DNS map	21×21
Gap between two central pixels	5 pixels

4. Implementation of the Proposed Model

In our experiment, we utilized eight types of signals: angular misalignment fault (AMF), bearing fault (BF), bowed shaft fault (BSF), phase imbalance fault (PMF), broken rotor bar fault (BRBF), rotor imbalance fault (RIF), phase imbalance

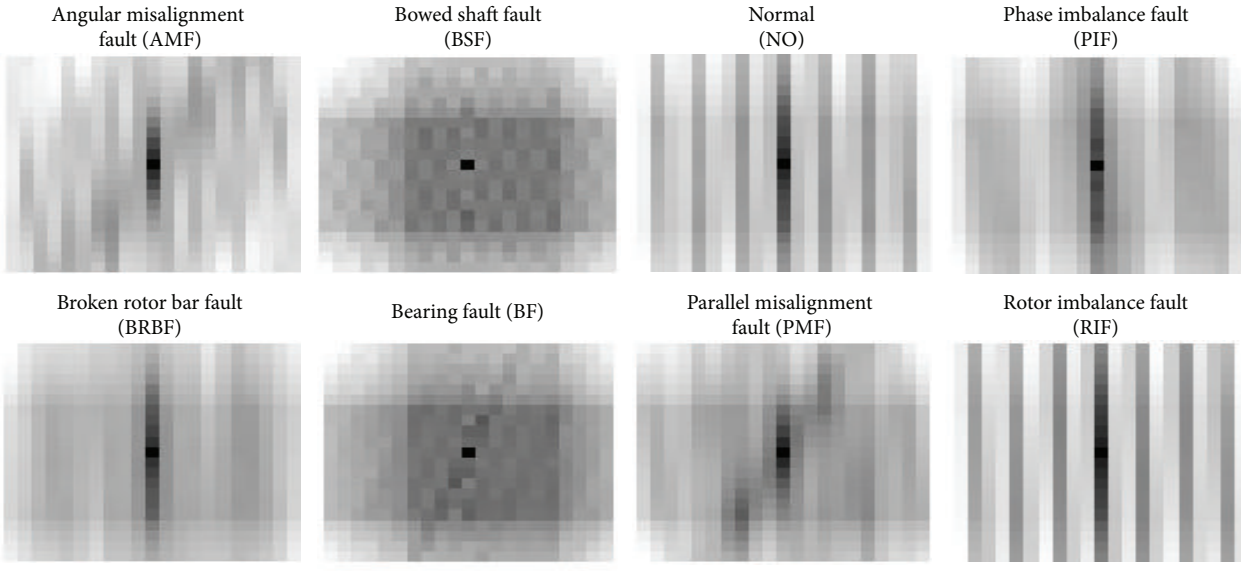


Figure 5: DNS maps obtained from the converted texture images.

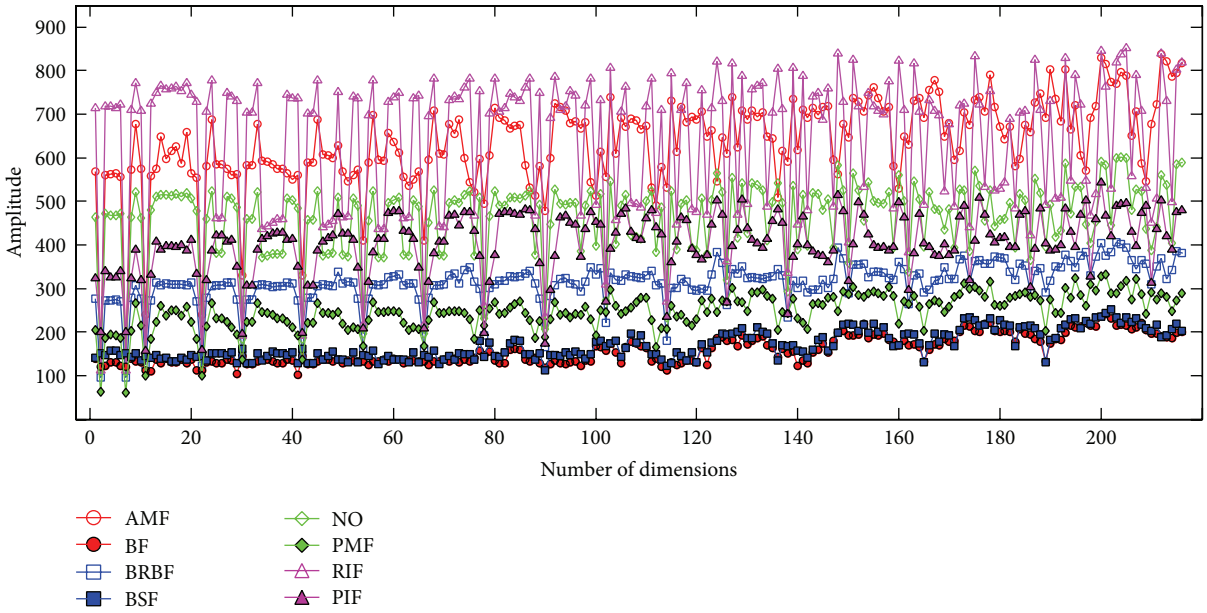


Figure 6: 1 × 216 Feature vectors including rotationally invariant features for various induction motor faults.

fault (PIF), and healthy (NO). We acquired one-second 105 vibration signals for each fault condition from the accelerometer located at the axial direction of the induction motor and Table 1 presents a brief description of each fault used in this study.

To efficiently observe the relationship of samples, this paper converts one-second 1D time-domain vibration signal into a 2D gray-level image with a size of 89 × 89 and Figure 4 shows a 2D representation of vibration signals for each fault condition. Likewise, Table 2 shows parameters (e.g., search window size, neighborhood window size, and so on) for generating a DNS map and Figure 5 depicts the DNS map for each fault condition.

As mentioned in previous section, DNS map values are sampled on circumferences of concentric circles of various radii centered at the center pixel of the map. For a DNS map of size 21 × 21, ten circles are used in order to cover all of the DNS map regions. For circles of radii greater than two pixels, 24 uniform angular measurements are extracted. Only 8 and 16 measurements are extracted from the first two innermost circles. Therefore, the number of dimensions of the feature vector is 216 (= 8 + 16 + 8 × 24) and Figure 6 illustrates

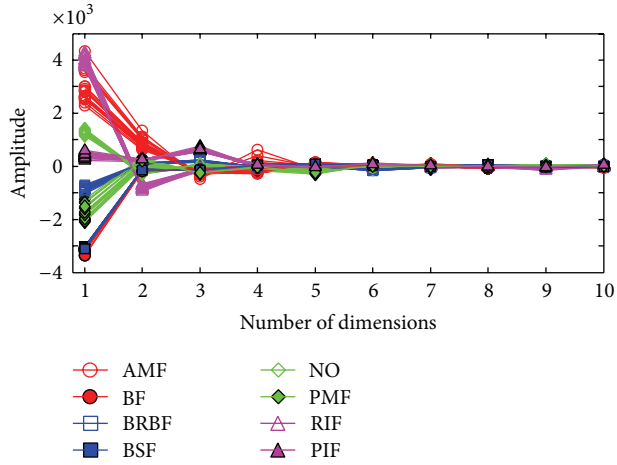


Figure 7: Patterns of principal components of each induction motor fault.

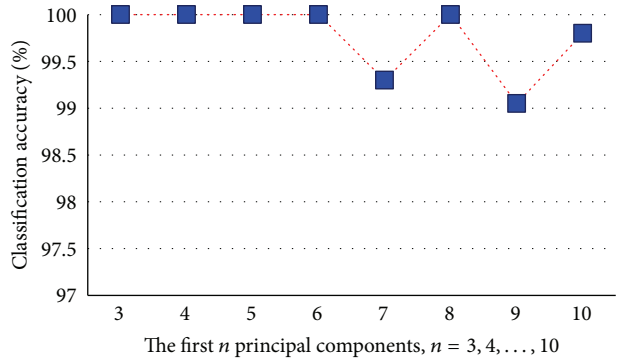


Figure 8: Classification accuracy with different numbers of principal components.

1×216 feature vectors including rotationally invariant features for each fault condition.

It was found that computational complexity typically increases and overall classification accuracy decreases with the high-dimensional feature vector. Thus, we employed the PCA to reduce computational complexity and improve the overall classification accuracy by selecting the most discriminative features from the high-dimensional feature vector. Figure 7 demonstrates magnitudes of several principal components after applying the PCA to the high-dimensional feature vectors. For the PCA, it is necessary to determine the most influential component for fault diagnosis and Figure 8 shows the classification results with different numbers of principal components. As depicted in Figure 8, the proposed fault diagnosis methodology achieves maximum classification accuracies when we utilize first three to six principal components as inputs of OAA MCSVMs and consequently this paper utilizes the first three principal components as fault features in this study providing both the highest classification performance and the lowest computational time.

In the observation and testing process, we utilized SVM with a Gaussian radial basis function (RBF) kernel [17] because the RBF kernel shows an exceptional ability to handle the nonlinear problem inherent in the time-varying vibration signals caused by motor slippage and variable rotational speed. The Gaussian radial basis function kernel is represented as follows [26]:

$$k\left(sv_i, sv_j\right) = \exp\left(-\frac{\left\|sv_i - sv_j\right\|^2}{2\sigma^2}\right), \quad (1)$$

where $k(sv_i, sv_j)$ is the RBF kernel, sv_i and sv_j are the input feature vectors, and σ is a parameter set by users to determine the effective width of the RBF kernel. The selection of an optimal σ value of the RBF kernel is an important issue. Therefore, this paper explores the impact of σ values on the classification performance in the range from 0.1 to 2 at intervals of 0.1. Table 3 presents the optimal range of σ values for each SVM in order to classify faults of the induction motor.

Table 3: Optimal sigma values of each SVM strutcure.

Faults	Optimal range	Selected values
AMF (SVM1: class 1)	$0.2 < \sigma < 2.0$	1.0
BRBF (SVM2: class 2)	$0.1 < \sigma < 2.0$	0.9
NO (SVM3: class 3)	$0.1 < \sigma < 1.4$	0.6
RIF (SVM4: class 4)	$0.1 < \sigma < 1.9$	1.0
BF (SVM5: class 5)	$0.1 < \sigma < 0.3$	0.2
BSF (SVM6: class 6)	$0.1 < \sigma < 0.4$	0.2
PMF (SVM7: class 7)	$0.1 < \sigma < 1.2$	0.6
PIF (SVM8: class 8)	$0.1 < \sigma < 1.1$	0.5

5. Experimental Results

The performance of the proposed model is evaluated in terms of true positive (TP) and false positive (FP) classification accuracy [27], where TP represents the number of faults in class i that are correctly classified into class i, and FP is the number of faults in other classes that are incorrectly classified into class i. Furthermore, since the vibration signals

Table 4: True positive and false positive of classification accuracy without noise.

Algorithm	Average classification accuracy in terms of true positive and false positive								Average
	AMF	BRBF	NO	RIF	BF	BSF	PMF	PIF	
Algorithm 1 [11]									
TP	91.35	91.54	84.23	90.38	96.73	98.46	97.11	80.19	91.24875
FP	1.54	1.24	2.17	0.80	0.22	0.47	0.63	3.24	1.28875
Algorithm 2 [10]									
TP	98.85	93.85	90.00	100.00	100.00	100.00	96.73	99.04	97.30875
FP	0.00	1.59	1.02	0.00	0.00	0.00	0.76	0.00	0.42125
Algorithm 3 [4]									
TP	100.0	100.0	100.0	100.00	100.00	100.00	100.00	100.0	100
FP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
Proposed approach									
TP	100.0	100.00	100.00	100.00	100.00	100.00	100.00	100.0	100.00
FP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5: True positive and false positive of classification accuracy with noise (SNR = 15 dB).

Algorithm	Average classification accuracy in terms of true positive and false positive								Average
	AMF	BRBF	NO	RIF	BF	BSF	PMF	PIF	
Algorithm 1 [11]									
TP	91.45	88.00	84.00	90.91	96.00	98.91	95.19	79.82	90.535
FP	1.95	0.49	2.57	0.99	0.16	0.57	0.25	3.40	1.2975
Algorithm 2 [10]									
TP	96.15	85.57	48.26	78.46	100.00	100.00	96.92	100.0	88.17
FP	0.00	10.83	0.73	0.03	0.00	0.000	3.18	3.65	2.3025
Algorithm 3 [4]									
TP	34.42	0.00	7.50	100.00	100.00	100.00	91.73	91.92	65.69625
FP	0.00	0.00	7.45	23.74	2.31	4.53	0.00	0.00	4.75375
Proposed approach									
TP	100.0	94.61	100.0	100.00	100.00	100.00	100.00	100.0	100.00
FP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

can be influenced by various noise levels in real industrial environment, we evaluate the robustness of the proposed fault diagnosis approach in noisy environments by artificially adding white Gaussian noise to the acquired vibration signals and setting the signal-to-noise ratio (SNR) between the originally acquired vibration signals and the noise-inserted vibration signals at 15 dB and 20 dB, respectively. To guarantee the reliability of the proposed fault diagnosis approach in terms of classification accuracy, we evaluate it ten times with randomly generated training data and test data at each iteration. The final classification accuracy can be obtained by averaging the total classification accuracies. Tables 4, 5, and 6 show the average classification accuracies of the proposed and conventional algorithms in noiseless and noisy environments.

Experimental results show that the proposed approach achieves 100% classification accuracy even in noisy environments. We observe that the number of significant pixels does not show abrupt change even if noise is included in the vibration signals to a limited value. In contrast, Algorithm 1 [11], which utilizes the time-domain statistical features (RMS, variance, skewness, and kurtosis) from the vibration signal, provides relatively lower classification accuracy for BRBF, NO, RIF, and PIF than those for AMF, BF, BSF, and PMF in noisy environments. This is due to the characteristics of BRBF, NO, RIF, and PIF which are not distinctive enough each other. Algorithm 2 [10], utilizing a combination of wavelet packet transform (WPT) and empirical decomposition, also fails to provide high classification accuracy in noisy environments due to its limitation in selecting the correct number of significant IMFs from the EMD algorithm. Algorithm 3 [4] uses the reconstruction of discrete wavelet packet nodes with information entropy and also exhibits low classification accuracies in noisy environments due to the nonadaptive selection of the decomposition signal level. Overall, the proposed fault diagnosis model outperforms other conventional algorithms in classification accuracy both with and without a noisy environment.

Table 6: True positive and false positive of classification accuracy with noise (SNR = 20 dB).

Algorithm	Average classification accuracy in terms of true positive and false positive								Average
	AMF	BRBF	NO	RIF	BF	BSF	PMF	PIF	
Algorithm 1 [11]									
TP	91.54	88.08	84.04	90.96	95.96	99.04	95.96	79.81	90.67375
FP	1.92	0.49	2.58	0.99	0.14	0.58	0.41	3.38	1.31125
Algorithm 2 [10]									
TP	98.27	86.92	72.31	99.23	99.42	100.00	93.84	100.0	93.74875
FP	0.00	3.65	1.32	0.14	0.00	0.14	0.91	1.02	0.8975
Algorithm 3 [4]									
TP	97.31	87.88	84.42	100.00	100.00	100.00	100.00	100.0	96.20125
FP	0.00	0.00	1.43	2.61	0.30	0.00	0.00	0.00	0.5425
Proposed approach									
TP	100.0	100.0	100.0	100.00	100.00	100.00	100.00	100.0	100.00
FP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

6. Conclusions

This paper proposed a robust fault detection and classification approach for induction motors using texture feature extraction and OAA MCSVMs. In the proposed approach, time-domain vibration signals were first converted into gray images to exploit texture features from the converted images of each faulty vibration signal. Feature vectors were then calculated from the DNS maps of the images. The PCA was utilized to select the most discriminative features by eliminating the trivial ones. Finally, using the distinctive features as inputs, OAA MCSVMs identify each fault of the induction motor. Experimental results showed that the proposed approach achieves 100% classification accuracy even in noisy environments.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science, and Technology (NRF-2012R1A1A2043644), and this work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean government (MEST) (no. NRF-2013R1A2A2A05004566).

References

[1] M. Zhao, X. Jin, Z. Zhang, and B. Li, "Fault diagnosis of rolling element bearings via discriminative subspace learning: visualization and classification," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3391–3401, 2014.

[2] E. Namsrai, T. Munkhdalai, M. Li, J. Shin, O. Namsrai, and K. H. Ryu, "A feature selection-based ensemble method for arrhythmia classification," *Journal of Information Processing Systems*, vol. 9, no. 1, pp. 31–40, 2013.

[3] K. Goswami, G. S. Hong, and B. G. Kim, "A novel mesh-based moving object detection technique in video sequence," *Journal of Convergence*, vol. 4, no. 3, pp. 20–24, 2013.

[4] C. Rodriguez-Donate, R. J. Romero-Troncoso, E. Cabal-Yepez, A. Garcia-Perez, and R. A. Osornio-Rios, "Wavelet-based general methodology for multiple fault detection on induction motors at the startup vibration transient," *Journal of Vibration and Control*, vol. 17, no. 9, pp. 1299–1309, 2011.

[5] L. Batista, B. Badri, R. Sabourin, and M. Thomas, "A classifier fusion system for bearing fault diagnosis," *Expert Systems with Applications*, vol. 40, no. 17, pp. 6788–6797, 2013.

[6] S. Bansal, S. Sahoo, R. Tiwari, and D. J. Bordoloi, "Multiclass fault diagnosis in gears using support vector machine algorithms based on frequency domain data," *Measurement*, vol. 46, no. 9, pp. 3469–3481, 2013.

[7] M. Kang and J. M. Kim, "Singular value decomposition based feature extraction approaches for classifying faults of induction motors," *Mechanical Systems and Signal Processing*, vol. 41, pp. 348–356, 2013.

[8] S. Silas, K. Ezra, and E. B. Rajsingh, "A novel fault tolerant service selection framework for pervasive computing," *Journal of Human-Centric Computing and Information Sciences*, vol. 2, no. 5, pp. 1–14, 2012.

[9] M. Kedadouche, M. Thomas, and A. Tahan, "Empirical mode decomposition of acoustic emission for early detection of bearing defects," in *Advances in Condition Monitoring of Machinery in Non-Stationary Operations*, Lecture Notes in Mechanical Engineering, pp. 367–377, Springer, Berlin, Germany, 2014.

[10] Y. Lei, Z. He, and Y. Zi, "Application of an intelligent classification method to mechanical fault diagnosis," *Expert Systems with Applications*, vol. 36, no. 6, pp. 9941–9948, 2009.

[11] J. Zarei, "Induction motors bearing fault detection using pattern recognition techniques," *Expert Systems with Applications*, vol. 39, no. 1, pp. 68–73, 2012.

[12] P. E. William and M. W. Hoffman, "Identification of bearing faults using time domain zero-crossings," *Mechanical Systems and Signal Processing*, vol. 25, no. 8, pp. 3078–3088, 2011.

[13] B.-S. Yang and K. J. Kim, "Application of Dempster-Shafer theory in fault diagnosis of induction motors using vibration

and current signals," *Mechanical Systems and Signal Processing*, vol. 20, no. 2, pp. 403–420, 2006.

[14] Z. T. Li and H. Li, "EMD and envelope spectrum based bearing fault detection," *Advanced Materials Research*, vol. 459, pp. 233–237, 2012.

[15] Z. Xu, J. Xuan, T. Shi, B. Wu, and Y. Hu, "Application of a modified fuzzy ARTMAP with feature-weight learning for the fault diagnosis of bearing," *Expert Systems with Applications*, vol. 36, no. 6, pp. 9961–9968, 2009.

[16] H. Keskes, A. Braham, and Z. Lachiri, "Broken rotor bar diagnosis in induction machines through stationary wavelet packet transform and multiclass wavelet SVM," *Electric Power Systems Research*, vol. 97, pp. 151–157, 2013.

[17] S. M. Chang, H. H. Chang, S. H. Yen, and T. K. Shih, "Panoramic human structure maintenance based on invariant features of video frames," *Journal of Human-Centric Computing and Information Sciences*, vol. 3, article 14, 2013.

[18] S. G. Kim and Y. G. Seo, "A TRUS prostate segmentation using Gabor texture features and snake-like contour," *Journal of Information Processing Systems*, vol. 9, no. 1, pp. 103–116, 2013.

[19] H. H. Bafroui and A. Ohadi, "Application of wavelet energy and Shannon entropy for feature extraction in gearbox fault detection under varying speed conditions," *Neurocomputing*, vol. 133, pp. 437–445, 2014.

[20] M. Malkawi and O. Murad, "Artificial neuro fuzzy logic system for detecting human emotions," *Journal of Human-Centric Computing and Information Sciences*, vol. 3, no. 3, pp. 1–13, 2013.

[21] J. Yan and L. Lu, "Improved Hilber-Huang transform based weak signal detection methodology and its application on incipient fault diagnosis and ECG signal analysis," *Journal of Signal Processing*, vol. 98, pp. 74–87, 2014.

[22] H. Liu and M. Han, "A fault diagnosis method based on local mean decomposition and multi-scale entropy for rolling bearing," *Journal of Mechanism and Machine Theory*, vol. 75, pp. 67–78, 2014.

[23] J. Zheng, J. Cheng, and Y. Yang, "A rolling bearing fault diagnosis approach based on LCD and fuzzy entropy," *Journal of Mechanism and Machine Theory*, vol. 70, pp. 441–453, 2013.

[24] F. M. Khellah, "Texture classification using dominant neighborhood structure," *IEEE Transactions on Image Processing*, vol. 20, no. 11, pp. 3270–3279, 2011.

[25] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[26] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 533–544, 2001.

[27] D. Ghimire and J. Lee, "A robust face detection method based on skin color and edges," *Journal of Information Processing Systems*, vol. 9, no. 1, pp. 141–156, 2013.