Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 2404089, 9 pages
https://doi.org/10.1155/2018/2404089

Research Article

Modified Dynamic Time Warping Based on Direction Similarity for Fast Gesture Recognition

Hyo-Rim Choi and TaeYong Kim

Department of Advanced Imaging Science, Graduate School of Advanced Imaging Sciences, Film, and Multimedia, Chung-Ang University, Heukseok-dong, Dongjak-gu, Seoul 156-756, Republic of Korea

Correspondence should be addressed to TaeYong Kim; kimty@cau.ac.kr

Received 25 August 2017; Revised 12 December 2017; Accepted 21 December 2017; Published 22 January 2018

Academic Editor: Tae Choi

We propose a modified dynamic time warping (DTW) algorithm that compares gesture-position sequences based on the direction of the gestural movement. Standard DTW does not specifically consider the two-dimensional characteristic of the user's movement. Therefore, in gesture recognition, the sequence comparison by standard DTW needs to be improved. The proposed gesture-recognition system compares the sequences of the input gesture's position with gesture positions saved in the database and selects the most similar gesture by filtering out unrelated gestures. The suggested algorithm uses the cosine similarity of the movement direction at each moment to calculate the difference and reflects the characteristics of the gesture movement by using the ratio of the Euclidean distance and the proportional distance to the calculated difference. Selective spline interpolation assists in solving the issue of recognition-decline at instances of gestures. Through experiments with public databases (MSRC-12 and G3D), the suggested algorithm revealed an improved performance on both databases compared to other methods.

1. Introduction

In human-computer interaction (HCI), replacing the mouse and keyboard with the users' voice and movement as input mechanisms is a popular topic in present-day research. This has been a result of the substantial improvement in hardware performance and new sensor technology. An analysis of the movement obtained from the sensor in order to determine the user intention is an important process of these types of HCI. HCI gesture-recognition technology mostly consists of pattern-recognition technology. The recognition is mainly divided into two processes: the first process involves the extraction of the characteristics of a pattern, and the second process involves the categorization of the extracted features. A computer's acquisition of a user's characteristics typically occurs through a sensor or the processing of acquired data. The acquired characteristics from a gesture are sequential data, and pattern-recognition technologies are required to categorize them.

Methods such as dynamic time warping (DTW) and hidden Markov model (HMM) are used to analyze the sequential data. Research studies seek to improve these methods [1]. The DTW algorithm was developed to match sequence data that are of different lengths. The algorithm creates a cost table for each of the components of two sequence datasets, and it compares the two sequence datasets using dynamic programming that rotationally selects and saves the minimum cost. HMM is a probabilistic model that uses the transition probability of sequence data [2]. Neural network (NN) is a computer system modeled on the human brain and nervous system. Recently, deep learning methods (convolutional neural network, recurrent neural network) have provided reasonable results in computer vision research, while research is still on to improve their applicability in gesture recognition [3, 4]. The major challenge encountered in using deep learning during gesture recognition is the effective presentation of the gesture movements.

The use of deep learning-based methods requires remarkably large database for effectively training the inputs to obtain adequate results during the testing phase. Consequently and because all the processing is achieved within the hidden layers, it is challenging for the researcher to analyze the training

process. Therefore, the flexibility of DTW, its requirement of a small-sized database during the training phase, and its being viewable under this process make it a convenient tool for matching process and hence a flexible method for analyzing the extracted features.

The matching based on DTW algorithm involves lesser database-learning pressure and provides steadier results compared with a probability-based algorithm. Owing to these strengths, DTW could be applied to various areas that use sequence data such as gesture [5], voice [6], hand-written letters [7], and signatures [8]; moreover, favorable results can be achieved without the need for a large amount of learning data. DTW [1], edit distance with real penalty (ERP) [9], and edit distance on real sequence (EDR) [10] consider each datum rather than the shape of the sequence trajectory. Angular metric for shape similarity (AMSS) [11] and longest common subsequence (LCSS) [12] are less influenced by outliers; however, AMSS requires preprocessing and is more sensitive to short vibrations. With respect to methods that consider the sequence shape, Derivative DTW (DDTW) [13] compares the shape by using differential sequences. As DDTW considers only the shape, its performance exhibits significant deviation according to the characteristics of the database.

This study introduces a modified DTW algorithm that is based on direction similarity (DS), to calculate the similarity while considering both the movement and shape of gestures. The suggested gesture-recognition system considers the hand-position data acquired from the camera as the sequence data, and if the length of the sequence is insufficient, it interpolates the length. After the normalization of both the position and size of the acquired sequence, gesture recognition is achieved through the use of the DTW algorithm, which reflects the direction characteristics to detect fast gestures.

2. DTW and Classification of Gestures

2.1. Classic DTW. The DTW algorithm can be defined as a pattern-matching algorithm that permits nonlinear construction according to a time scale. To calculate the similarity of the lengths n and m of the two sequences $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_n\}$, respectively, a nonlinear adjustment course W is set, and the minimalized pathway of W's distance is determined. W is defined as follows:

$$W = \{w_1, w_2, w_3, \ldots, w_k\}, \tag{1}$$

and function d that calculates the distance of w is defined as follows:

$$d(w) = d(i, j) = \left\| a_i - b_j \right\|. \tag{2}$$

The minimum of the total distance of the adjustment course is calculated as follows:

$$\mathrm{DTW}(A, B) = \min \sum_{k=1}^{K} d(w_k). \tag{3}$$

Numerous calculations are required for (3) as all the feasible pathways must be calculated; meanwhile, dynamic programming could also be used to solve the equation. When DTW is applied to the recognition system, the accuracy and efficiency of the calculation increased with the following four limits: an end point limitation that quadrates the start and end points of the input pattern and reference pattern; a monotone-increasing limitation that requires the increase of the monotone for the optimized pathway; a global-path constraint that limits the permitted areas of the input pattern and reference pattern; and a local constraint that limits the pathway to a node to prevent overcontraction or overexpansion [14].

The application of the four constraints of DTW is depicted in Figure 1. The calculation of the optimum path D, in consideration of the local path limitation, is as follows:

$$\begin{aligned} &D(i, j) \\ &\quad = d(i, j) \\ &\qquad + \min\left[D(i-1, j-1), D(i-1, j), D(i, j-1)\right]. \end{aligned} \tag{4}$$

The main calculation cost is incurred during the calculation process of the optimum adjustment of (3), and although a few limitations and dynamic programming could alleviate such issues, the limitation cannot accurately determine the results in the event that an optimal result exists outside of the selected data.

2.2. Improvement of DTW. DTW is used in various areas including movement recognition [14], voice recognition [15], and data mining [1]. The previous research studies on DTW focused on the improvement of speed to solve the increasing complexity, followed by the increment of the sequence length. Sakoe and Chiba used several conditions to solve the problems of DTW [16], while Stan and Phillip achieved an improved speed through an approximation process [17]; Keogh and Ratanamahatana improved the speed by using the lowest-bound technology [18]. Figure 2 shows the Sakoe–Chiba and Itakura bands that are widely used to constrain the search area; limiting the search area tremendously aids the improvement of the speed; however, if the optimal matching pathway is outside of the search area, a favorable result is unlikely.

In addition to the purpose of speed improvement, a few researchers focused on improving the matching accuracy and recognition. Keogh and Pazzani recommended a Derivative DTW algorithm that uses the primary differential value [13], whereby the algorithm compares the shape of the sequence albeit not that of the sequence value, as shown in Figure 3.

2.3. Gesture Recognition by Using Classification Method. By comparing the operation sequence using DTW, only a simple distance can be obtained; this distance implies similarity. With a variety of target gestures, k-nearest neighbors (k-NN) is one of the common classification methods used with matching-based recognition algorithms. k-NN using one representative of each class is the 1-NN method, and 1-NN is a commonly used classification method to reduce the problem of computational complexity. In addition to the method of measuring the similarity, there is a method of classifying the operation sequence through SVM; however, it does not exhibit high performance.

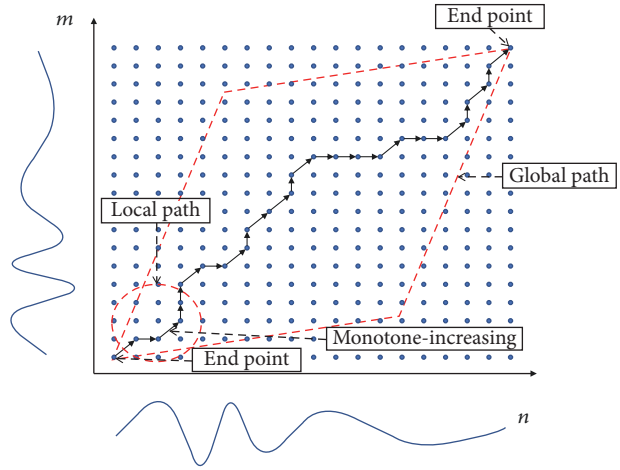


Figure 1: Four constraints of dynamic time warping.

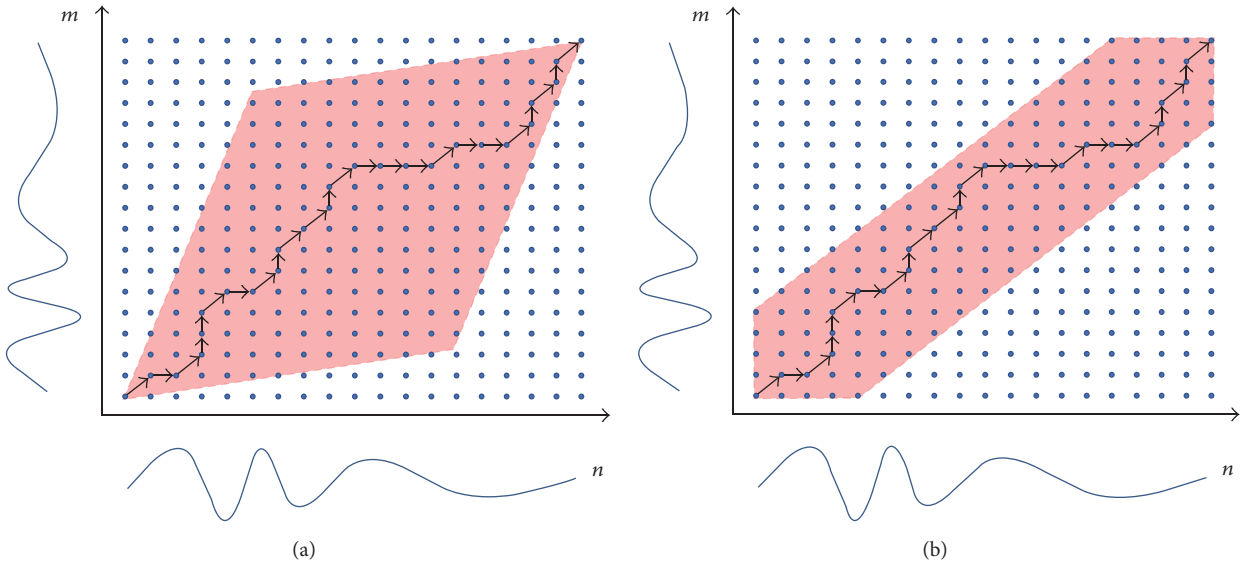


Figure 2: Sakoe–Chiba band (a) and Itakura band (b).

3. Modified DTW Based on DS

The DTW algorithm only compares the sequence value, and, as a result, it cannot consider a variety of gestural characteristics. The position and direction of a gesture are important factors for the expression of a gesture [19]. To compare two sequences while considering both the position and direction, we introduced a DTW with an added distance that is in proportion to the DS; Figure 4 shows the process flow of the proposed method. The recognition system that uses the proposed method involves the use of a depth camera (Kinect of MS was used in our experiment) to receive the position of the user's joint and also to determine the position; then, we interpolate the depth data based on sequence length. 1-NN matches the previously defined gestures that are in the database and selects the most identical gesture; furthermore, it uses the threshold value to filter out insignificant gestures.

3.1. Input Normalization. The sequences used in this section as examples demonstrate that the result of the normalization is a specially designed numeric action that is a convenient tool to analyze the performance of normalization.

Figure 4 shows the overall system flow. The data of hand positions acquired from the Kinect device is likely to exhibit a substantial difference regarding the distance of the hand position from the sensor. Hand-position data are normalized based on the overall size and the position of a gesture. First, the length of a longer axis is determined by measuring the horizontal and vertical lengths of the overall gesture, and the points are shifted based on the average of all of the values. Figure 5 shows the results of gesture normalization, where (a) shows the two gesture sequences before normalization and (b) shows the two sequences after normalization. It is evident from these figures that a similar size and position are attained through normalization.

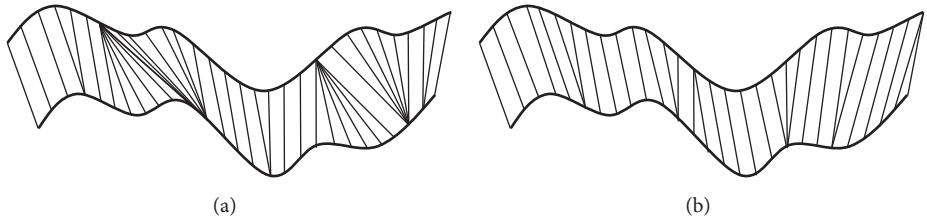


Figure 3: DTW sequence matching (a) and Derivative DTW sequence matching (b).

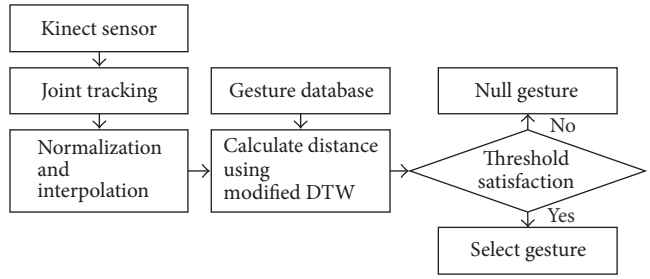


Figure 4: Gesture-recognition system using modified dynamic time warping based on direction similarity.

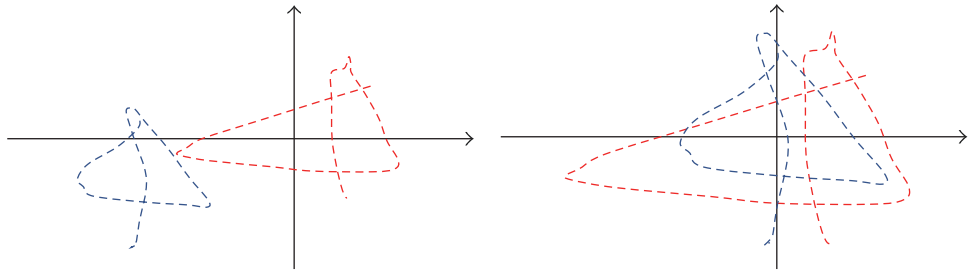


Figure 5: Before normalization (a) and after normalization (b).

3.2. Input Interpolation. An input sequence requires an interpolation process before the comparison of the gestures in the database can occur. The DTW algorithm allows for difference of sequence length; however, repeated matching pathways are introduced when it occurs. Furthermore, similar challenges are higher for repeated matching pathway with noise; therefore, sequence interpolation could be used to solve this problem. We used a spline interpolation method that provides an approximate value for the abruptly changing function. Figure 6 shows the sequence-interpolation result of a straightforward number gesture.

Spline interpolation was used in the experiment when the input sequence length was below the threshold value selected through experiments.

3.3. Modified DTW Based on DS. Previous DTW methods used the distance between elements to determine the optimal matching pathway between two sequences; however, this often results in matching pathways that do not reflect the gestural characteristics and also results in negative recognition result. Figure 7(a) shows the pathways of "3" and "3," and Figure 7(b) shows the pathways of "3" and "7"; when only the position data are used to compare the two, the right side exhibits a higher similarity. Even with the normalization, an inaccurate recognition result that is based on the person performing the gesture can appear; therefore, we propose a DTW algorithm that considers gestural-movement direction to supplement this problem.

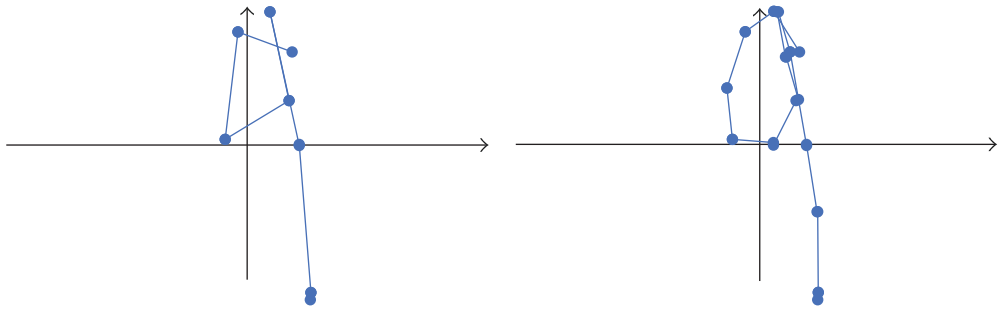
A modified DTW based on DS calculates the directional distance, which is based on the cosine similarity between two elements; then, this method calculates shape matching pathways by using the linear combination of the directional distance and Euclidean distance. The following new equation is used to calculate the distance between the two elements:

$$d(w) = (1 - \alpha)\, d(i, j) + \alpha s(i, j)\, d(i, j), \tag{5}$$

where function d calculates the Euclidean distance of the two sequence elements. Function S calculates the DS of the two elements and is defined as follows:

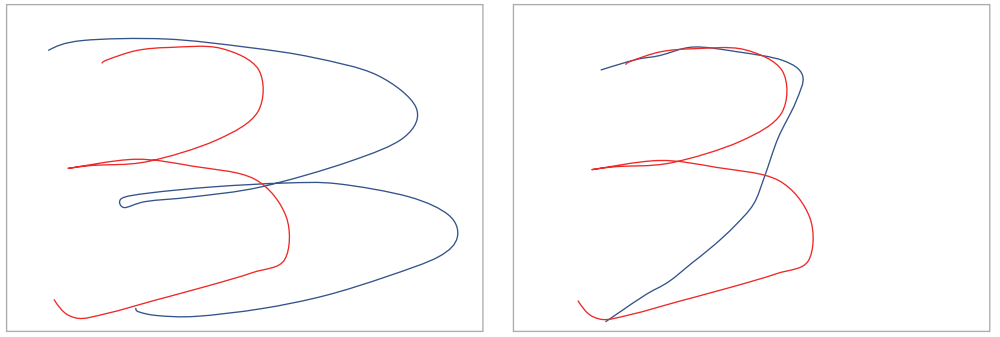
$$s(i, j) = 1 - \frac{\left(a_j' \times b_j'\right)}{\left\|a_i'\right\| \left\|b_i'\right\| + \varepsilon}, \tag{6}$$

which returns zero if the direction of the two vectors match and returns two if the directions are opposite to each other. ε



(a) (b)

Figure 6: Before interpolation (a) and after interpolation (b).



(a) (b)

Figure 7: Examples of misrecognition.

is a constant to prevent the denominator from becoming zero, and α is a constant coefficient that adjusts the size of the directional distance and Euclidean distance. The optimal value of α is changed based on the database characteristics; in this study, we select the value with the highest recognition by adjusting the constant from one to zero at the learning step. a' and b' are directional vectors that were calculated from the differences among the elements, and the calculation is as follows:

$$a'_n = \frac{(a_n - a_{n-1}) + ((a_{n+1} - a_{n-1})/2)}{2},$$
$$b'_n = \frac{(b_n - b_{n-1}) + ((b_{n+1} - b_{n-1})/2)}{2}. \quad (7)$$

Considering the near elements $(a-a_{n-1}, b-b_{n-1})$ as well as the neighbor elements $(a_{n+1}-a_{n-1}, b_{n+1}-b_{n-1})$, we can alleviate the noise and outliers.

4. Experiment

4.1. Gesture Set. Experiments were conducted on two public benchmark datasets (MSRC-12 Kinect Gesture Dataset, G3D) [20, 21]; the gestures in both databases can be applied as interfaces of various fields. MSRC-12 is a relatively large dataset for gesture/action recognition from 3D skeleton data captured by a Kinect sensor. The dataset has 594 sequences, containing 12 gestures by 30 subjects, 6244 gesture instances in total. The 12 gestures (as shown in Figure 8) are as follows: "lift outstretched arms," "duck," "push right," "goggles," "wind it up," "shoot," "bow," "throw," "had enough," "beat both," "change weapon," and "kick." For this dataset, cross subject protocol is adopted, that is, odd subjects for training and even subjects for testing.

The G3D database contains 20 gaming gestures (punch right, punch left, kick right, kick left, defend, golf swing, tennis swing forehand, tennis swing, backhand tennis serve, throw bowling-ball, aim and fire gun, walk, run, jump, climb, crouch, steer a car, wave, flap, and clap). Moreover, these gestures were captured through a Kinect sensor in an indoor environment as shown in Figure 9.

4.2. Interpolation Performance of Gesture Sequence. The DTW algorithm permits differences of sequence length; however, the recognition of fast-moving gestures is likely to cause problems. When the lengths of the compared sequences are different, repeated matching could occur; similar repetitive matching increases the incidence of error. To resolve this challenge, we used a spline sequence interpolation and tested its performance alongside other methods including linear-interpolation and low-pass interpolation methods. As the interpolation of all gesture sequences involves high computational expense, we interpolated only those sequences with lengths that are below average to the normal to compare the differences. We selected the sequence with the lowest distance among the similar gestures in the learning data and the measured recall factor using the 1-NN method. Figure 10

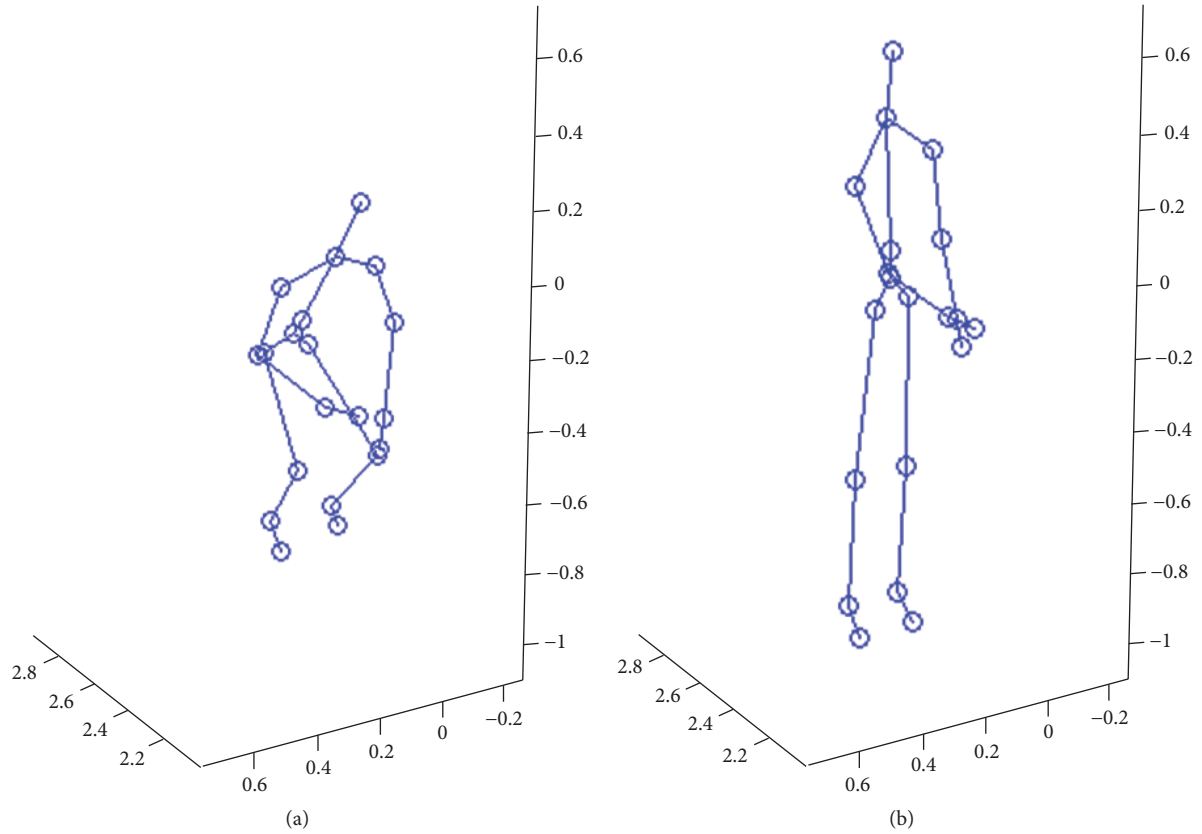


Figure 8: Example of MSRC-12.

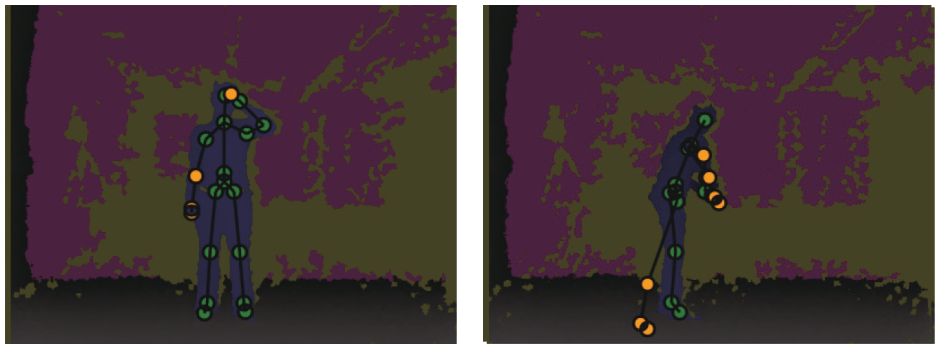


Figure 9: Examples of G3D.

shows the mean recall value that was measured for both the datasets for various interpolation methods.

All the interpolation methods improved the recognition rate. The spline and low-pass interpolation methods exhibit similar results, and their performances are superior to that of the linear-interpolation method. We eventually selected the spline method with the most favorable performance result. We did not interpolate all the sequences that were subsequently used in the experiments and used only the spline interpolation method on those sequences with a below-average length.

4.3. Determination of Parameter α. The modified DTW based on DS uses the directional distance and Euclidean distance, and the ratio is adjusted through α. The optimal value of α could differ depending on the characteristics of the gesture set. To determine an optimal α value, we changed α from "0" to "1" at intervals of "0.1" during the learning stage and measured the recognition rate using the 1-NN method. We selected the representative sequence prior to measuring the recognition rate. Furthermore, we selected the sequence that exhibited the lowest average distance by calculating the DTW distances of all the sequences for similar gestures and used all

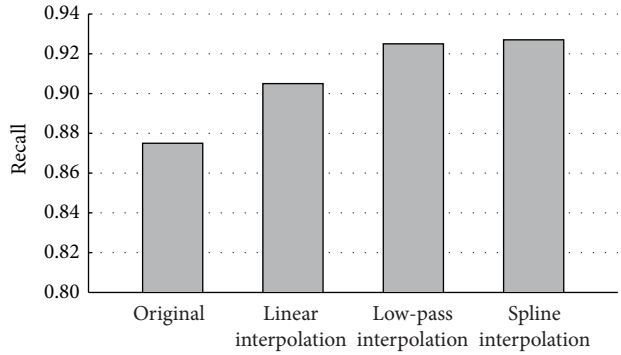


Figure 10: Recognition rate based on interpolation methods.

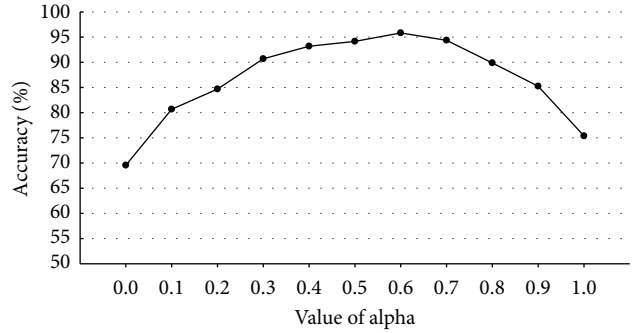


Figure 11: Recognition rate of varying α using 1-NN method (G3D).

the sequences excluding the representative sequence as inputs to measure the recognition rate; thereby, parameter α was changed with the use of a modified DTW-based DS. Figures 11 and 12 show the recognition rates of various values of α. The G3D gestures were set to $\alpha = 0.6$, while the MSRC-12 gestures were set to $\alpha = 0.8$.

4.4. Comparison and Analysis of DTW Performance Based on DS. To analyze the DTW performance based on DS (DTWDS) in our study, we used the 1-NN and conventional gesture-recognition methods. To verify the differences from the other DTW algorithms, we compared DTW and modified DTW. In comparison with the general methods, we used discrete HMM, SVM (for MSRC-12 datasets only) [22], and two CNN methods (SOS, JTM) [23, 24]. The 1-NN method was used to find the most similar movement in the gesture-recognition system, and the threshold-value setting of the gesture-recognition system was used to filter out insignificant gestures. We used the 1-NN method to measure the recall ratio by comparing the input gestures that were not used in the learning and the representative gestures of each movement to select the gesture with the lowest distance. In the event of a similarity, we recorded it as "true"; otherwise, we recorded it as "false." Moreover, all the sequences in the database were used as inputs. Then, every single neighbor of each class in 1-NN was adjusted when the input was changed.

Figures 13 and 14 show the experimental result with MSRC-12 datasets and G3D, respectively.

For the HMM results, the empirical numbers of states are used for each database. It is challenging for HMM with static number of states to model continuous sequenced gestures compared to other methods. SVM performed better than standard DTW, DDTW, and HMM; however, CNN was more effective for MSRC-12 datasets.

The proposed method (DTW with DS) also exhibits favorable performances in both the gesture sets. Neither the position nor the shape of the gesture was selected by the modified DTW based on DS, ensuring a more favorable performance than that of DTW and DDTW. The proposed method outperformed the other methods for MSRC-12 datasets, with an average value of 3.89%; however, for the G3D datasets, the proposed method exhibited no significant difference with the two CNN methods. The modified DTW based on DS adjusts the portion of the Euclidean distance and expands the difference between the true and false gestures; therefore, it exhibits a higher degree of accuracy compared to the other methods. Moreover, we applied normalization and interpolation methods for the matching process in the modified DTW based on DS (proposed method).

5. Conclusion

DTW is a pattern-matching algorithm that is used in different areas and has a number of advantages including a simple calculation process and a lower learning pressure when compared to HMM and deep learning-based methods. This study alleviated the matching errors that result from the

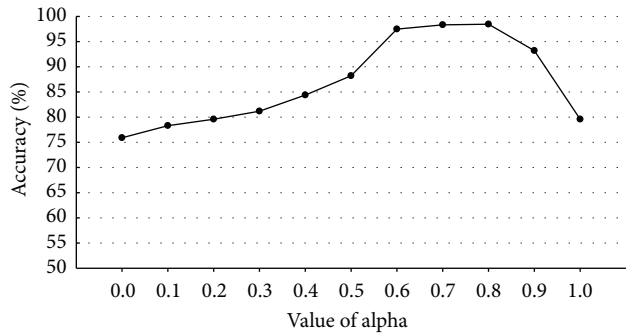


Figure 12: Recognition rate of varying α using 1-NN method (MSRC-12).

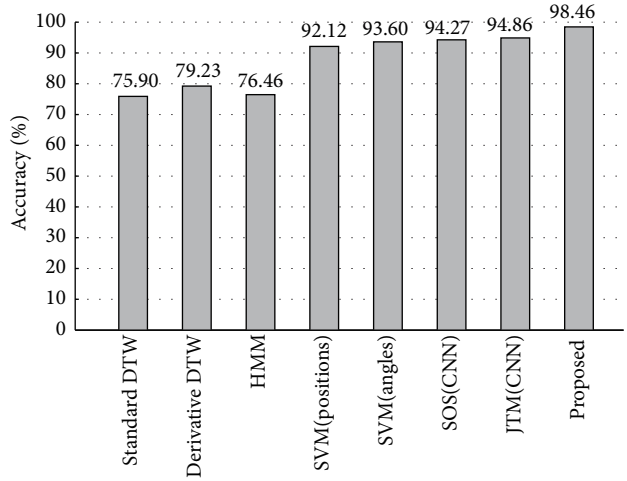


Figure 13: Experimental result with MSRC-12 datasets.

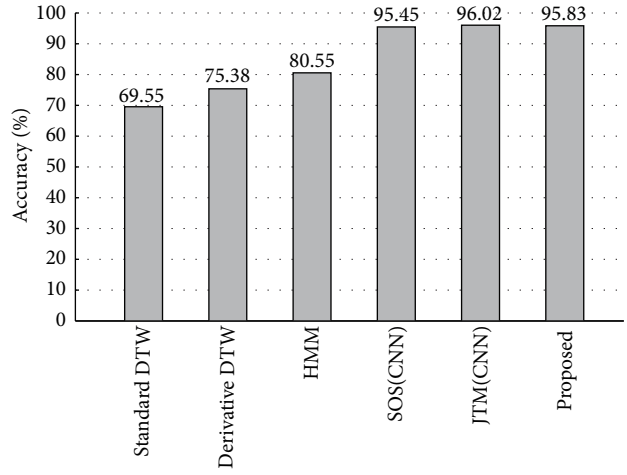


Figure 14: Experimental result with G3D datasets.

differences of sequence length through the use of spline interpolation when such differences arose. The proposed modified DTW considers position and shape of gestures, which are important characteristics in gesture expression, and incorporates the ratio of the directional distance and the Euclidean distance. The value of α was obtained empirically throughout the experiment to determine the optimal ratio of the directional distance and Euclidean distance based on the targeted gesture. Even when the gesture set was changed, we verified its gesture-recognition performance by adjusting the value at the learning stage. In comparison with the method that only considers one characteristic or a linear summation of two characteristics, the results revealed that the proposed method achieved remarkable performance.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2015R1D1A1A01058394).

References

[1] S. Laxman and P. S. Shanti, "A survey of temporal data mining," *Sadhana*, vol. 31, no. 2, pp. 173–198, 2006.

[2] L. R. Rabiner and B.-H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[3] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 4, pp. 498–509, 2016.

[4] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 1110–1118, June 2015.

[5] K. Jung, J.-L. Park, and B.-U. Choi, "Interactive auto-stereoscopic display with efficient and flexible interleaving," *Optical Engineering*, vol. 51, no. 2, Article ID 027402, 2012.

[6] K. Inthavisas and D. Lopresti, "Speech biometric mapping for key binding cryptosystem," in *Proceedings of the Sensing Technologies for Global Health, Military Medicine, Disaster Response, and Environmental Monitoring; and Biometric Technology for Human Identification VIII*, April 2011.

[7] R. M. Saabni and J. A. El-Sana, "Word spotting for handwritten documents using Chamfer distance and dynamic time warping," *IST/SPIE Electronic Imaging*, 2011.

[8] M. Faundez-Zanuy, "On-line signature recognition based on VQ-DTW," *Pattern Recognition*, vol. 40, no. 3, pp. 981–992, 2007.

[9] L. Chen and R. Ng, "On the marriage of lp-norms and edit distance," in *Proceedings of the Thirtieth international conference on Very large data bases*, distance. and On. the marriage of lp-norms and, Eds., vol. 30, 2004.

[10] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '05)*, pp. 491–502, ACM, June 2005.

[11] T. Nakamura, K. Taki, H. Nomiya, K. Seki, and K. Uehara, "A shape-based similarity measure for time series data with ensemble learning," *PAA. Pattern Analysis and Applications*, vol. 16, no. 4, pp. 535–548, 2013.

[12] G. Das, D. Gunopulos, and H. Mannila, "Finding similar time series," *Principles of Data Mining and Knowledge Discovery*, vol. 1263, pp. 88–100, 1997.

[13] E. J. Keogh and M. J. Pazzan, *Derivative Dynamic Time Warping*, vol. 1, SDM, 2001.

[14] T. Darrell and A. Pentland, "Space-time gestures," in *Proceedings of the 1993 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 335–340, June 1993.

[15] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," 2010, https://arxiv.org/abs/1003.4083.

[16] H. Sakoe and S. Chiba, "Dynamic programming algorit hm optimization for spoken word recognition," *IEEE Transactions on Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[17] S. Salvador and P. Chan, "Fastdtw: toward accurate dynamic time warping in linear time and space," *KDD workshop on mining temporal and sequential data*, 2004.

[18] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," *Knowledge and Information Systems*, vol. 7, no. 3, pp. 358–386, 2005.

[19] H.-S. Yoon, J. Soh, Y. J. Bae, and H. Seung Yang, "Hand gesture recognition using combined features of location, angle and velocity," *Pattern Recognition*, vol. 34, no. 7, pp. 1491–1501, 2001.

[20] V. Bloom, D. Makris, and V. Argyriou, "G3D: A gaming action dataset and real time action recognition evaluation framework," in *Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2012*, pp. 7–12, June 2012.

[21] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proceedings of the 30th ACM Conference on Human Factors in Computing Systems, CHI 2012*, pp. 1737–1746, May 2012.

[22] D.-D. Nguyen and H.-S. Le, "Kinect Gesture Recognition: SVM vs. RVM," in *Proceedings of the 7th IEEE International Conference on Knowledge and Systems Engineering, KSE 2015*, pp. 395–400, October 2015.

[23] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1-1, 2016.

[24] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using Convolutional Neural Networks," in *Proceedings of the 24th ACM Multimedia Conference, MM 2016*, pp. 102–106, October 2016.