Methodology Report

Formal Implementation of a Performance Evaluation Model for the Face Recognition System

Yong-Nyuo Shin,¹ Jason Kim,¹ Yong-Jun Lee,² Woochang Shin,³ and Jin-Young Choi⁴

¹ Korea Information Security Agency, 78 Karak-dong, Songpa-gu, Seoul 138-160, South Korea

² LG CNS, Hoehyun-dong, 2-ga, Jung-gu, Seoul 100-630, South Korea

³ Department of Internet Information, College of Natural Science & Engineering, Seokyeong University, 16-1 Jungneung-dong, Sungbuk-ku, Seoul 136-704, South Korea

⁴ Division of Computer Science and Engineering, College of Information and Communications, Korea University, Anam-dong, Sungbuk-ku, Seoul 136-701, South Korea

Correspondence should be addressed to Woochang Shin, wcshin@imail.skuniv.ac.kr

Received 5 October 2007; Accepted 25 October 2007

Recommended by Daniel Howard

Due to usability features, practical applications, and its lack of intrusiveness, face recognition technology, based on information, derived from individuals' facial features, has been attracting considerable attention recently. Reported recognition rates of commercialized face recognition systems cannot be admitted as official recognition rates, as they are based on assumptions that are beneficial to the specific system and face database. Therefore, performance evaluation methods and tools are necessary to objectively measure the accuracy and performance of any face recognition system. In this paper, we propose and formalize a performance evaluation model for the biometric recognition system, implementing an evaluation tool for face recognition systems based on the proposed model. Furthermore, we performed evaluations objectively by providing guidelines for the design and implementation of a performance evaluation system, formalizing the performance test process.

Copyright © 2008 Yong-Nyuo Shin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Face recognition systems provide the benefit of collecting a large amount of biometric information in a relatively easy and cost-effective manner, because they do not require subjects to bring any part of their body in contact with the recognition device intentionally, which results in fewer repercussions and less inconvenience when collecting the biometric information. An additional advantage exists, that is, the widely deployed image acquisition equipment can be used without modification. In particular, various face recognition algorithms and commercial systems have been developed and proposed, and the marketability of face recognition systems has increased, with many immigration-related facilities such as air and sea ports in many countries anxious to introduce face recognition systems after the 9.11 terror attacks in the US. These benefits, and the perceived necessity of increased security, have led to a rising social demand for face recognition systems; and certified performance evalua-

tion has become important as a means of evaluating these face recognition systems.

This paper proposes a performance evaluation model (PEM) to evaluate the performance of biometric recognition systems; and designs and implements a performance evaluation tool that enables comparison and evaluation of face recognition systems, based on the proposed PEM. The PEM is designed to be compatible with related international standards, and contributes to the consistency and enhanced reliability of the performance evaluation tool that is developed with reference to the model.

Section 2 outlines existing studies related to performance evaluation of face recognition systems. Section 3 proposes a PEM to evaluate the performance of face recognition systems. Section 4 describes the design and implementation of the performance evaluation tool, based on the proposed PEM. Section 5 compares the performance evaluation method, utilizing the performance evaluation tool proposed in this paper, with existing performance evaluation

programs. The last section provides a conclusion, and suggests some remarks for future research.

2. RELATED STUDIES

The representative certified performance evaluation programs for facial recognition systems are Face Recognition Technology (FERET) and Face Recognition Vendor Testing (FRVT). As has been used since 1993, FERET includes not only performance evaluation of facial recognition systems but also the development of algorithms and the collection of a face recognition database. Headed by the U.S. Department of Defense (USDoD), FERET is an evaluation tool that has been systematically executed from 1993 through 1997 by testing changes in the environment (e.g., size, posture, background, etc.), differences in the time when pictures are taken, and the performance of algorithms in processing a mass-volume database. In particular, the FERET performance evaluations have been general evaluations designed to measure algorithms at the level of research centers. The major purpose of the FERET performance evaluation tool has been to implement adaptations to the latest facial recognition technologies and their flexibility. Therefore, the FERET test is neither used to clearly measure the influence of algorithms on the performance of individual components nor to assess performance in fully organized scientific manners under all operating conditions of a system [1–3].

Face Recognition Vendor Testing (FRVT) was a performance test for the face recognition system that was implemented using three Face Recognition Technology (FERET) performance evaluations (1994, 1995, and 1996). The FERET program introduced the evaluation technique in the face recognition area, and developed the face recognition area at the earliest level (system prototype development). However, as face recognition technology matured from the prototype level to the commercial system level, FRVT 2000 measured the performance of these commercial systems, and evaluated how far the technology had evolved through comparison with the last FERET evaluation. The public began to pay more attention to face recognition technology in 2002. As a consequence, FRVT 2002 measured the degree of technical development since 2000, evaluated the large-size databases that were in use, and introduced new experiments to better understand the performance of face recognition. Size, difficulty, and complexity of this performance evaluation were on the rise as the evaluation theory as well as the face recognition technology grew. For example, FERET SEP96 performed just 14.5 million comparisons over a period of 72 hours, while FRVT 2000 carried out 192 million comparisons in 72 hours. In contrast, FRVT 2002 introduced an evaluation that made 15 billion comparisons in 264 hours [4–6].

Certified performance evaluation programs like FERET and FRVT were designed to measure the algorithm accuracy of face recognition systems. For these projects, a common face image database was provided for the test, face recognition was performed for a certain period of time according to the respective method, and the results were evaluated. However, this method provides evaluation only for face recognition technology vendors that participated in the program

TABLE 1: Classification of factors affecting the performance of biometric system.

Classification	Particular factors
Population demographics	Age, ethnic origin, gender, occupation, etc.
Application	Template aging, shooting time, User friendliness, user motivation, etc.
User physiology	Hair, baldness, illnesses and diseases, eye-lashes, nail length, skin tone, etc.
User behavior	Facial expression, dialect, motion, posture, angle, distance, stress, nervousness, etc.
Environmental influences	Background features (color, shadow, sound, etc.), lighting strength and direction, weather, (temperature, humidity, rain, snow, etc.)
User appearance	Outfit (hat, sleeves, pants, dress shoes, etc.), band (bandage), contact lenses, makeup, glasses, fake nails, hairstyle, rings, tattoos, etc.
Sensor and hardware	Dirt on camera lenses, focus, sensor quality, sensor change, transmission channel, etc.
User interface	Feedback, instructions, supervising director

during the evaluation period. In particular, database items were limited to image size, target posture, image acquisition environment, and time, which left the problem that various conditions of algorithm evaluation were not satisfied dynamically. Moreover, the algorithm evaluation environment was commissioned to each of the face recognition system developers, creating the problem of inconsistency in establishing a performance evaluation system environment. Furthermore, additional tasks were required in order to determine the accuracy of each algorithm, and to analyze the algorithm implementation result again. Therefore, it is necessary to design an algorithm evaluation method that can resolve these problems, to build a standardized evaluation environment, and to automatically figure out the evaluation result of the algorithm whose performance is measured in this environment.

2.1. Factors affecting performance evaluation

Results from the performance evaluation of facial recognition systems change in accordance with varying factors, such as lighting, posture, facial expression, and elapsed time. The JTC 1/SC 37/WG 5 International Standard [7] classifies those factors affecting the performance of biometric systems.

As outlined in Table 1, there are a number of factors affecting the performance of facial recognition systems, and such factors must be prudently taken into consideration when the probe and gallery are selected during the performance evaluation of these systems. Algorithm performance evaluation of biometric recognition technology is conducted in such a way that the standard gallery is trained or registered, and is then compared with the test biometric information (probe) to be recognized, after which the similarities between the two sets of information are measured.

TABLE 2: Classification of the KISA's database.

Criterion	Content
Gender	Male or female
Age	Age
Birthplace	Gyeonggi-do, Seoul
Lighting color	Fluorescent lamp, Incandescent lamp
Lighting direction	Front, right/left 45°, right/left 90°, right/left 135°, 180°
Glasses direction	Front, Right/Left 45°, Right/Left 90°
Facial expression	No Expression, Smiling, Angry, Closed Eyes, Surprised
Posture	Front, Right/Left 15°, Right/Left 30°, Right/Left 45°
Hair condition	Natural Hairstyle, Hair Band/s, Glasses

Generally, basic factors such as posture, angle, facial expression, lighting brightness, and gender are considered in the construction of a face image database. However, records of the face image database need to be further subdivided in order to process the test under conditions similar to those prevalent in the real world.

The research facial database that was developed by Korea Information Security Agency (KISA) from 2002 to 2004 was used for performance evaluation [8]. Table 2 shows a classification of KISA's database.

3. PERFORMANCE EVALUATION MODEL (PEM)

Many factors must be considered when building a fair performance evaluation system for biometric recognition systems. For example, to evaluate the performance of face recognition systems, a database of facial information for use in face recognition should be collected, and performance evaluation items (changes in facial expression, lighting, etc.) as well as performance evaluation measurement criteria such as false acceptance rate (FAR), false rejection rate (FRR), and equal error rate (EER) should be selected. The face recognition system to be evaluated and a standardized interface for the performance evaluation system should be designed and international standards need to be applied at each stage of performance evaluation, in order to enhance fairness and reliability.

Thus, the performance evaluation model (PEM) is created to analyze and arrange the criteria to be considered in building up the performance evaluation system, and to support the development of the performance evaluation system. The PEM presents the basic system structure, guidelines, and development process used to build a system for performance evaluation.

The PEM proposed in this paper is designed to (1) evaluate the performance of the biometric recognition algorithm, and (2) to build a system that automatically evaluates performance and outputs the results in tandem with the biometric recognition system.

3.1. Structure of the PEM

The PEM structure for the system that evaluates the biometric recognition algorithm is composed of a data preparation module, an execution model, and a result analysis module, as shown in Figure 1.

3.1.1. Data preparation module

The data preparation module prepares the biometric information used for performance evaluation, for which the development of a biometric information database and the design of the test criteria are the major elements to consider. As the biometric information database used affects the evaluation reliability of the performance evaluation system to a large extent, it should be considered a priority at the initial stage of system development. In addition, the biometric information used for performance evaluation should never be exposed, so that evaluation reliability can be improved [9].

Generally, algorithm performance evaluation of biometric recognition technology is conducted in such a way that the standard gallery is trained or registered, and is then compared with the test biometric information (probe) to be recognized, after which the similarities between the two sets of information are measured. At this time, algorithm performance varies according to the information generation environment or conditions. For example, if an expressionless front-view facial image is registered in the gallery, and a smiling facial image photographed at a 15-degree angle from the left is used as the probe, we can compare the strength of the different algorithm technologies in terms of facial expression and angle. In this paper, the "test criteria" refers to an item that could affect the performance evaluation result, and the performance evaluation system developer should design test criteria that are suitable for the objective of the evaluation. The test criteria selected by the PEM are limited to the classification criteria of the biometric information database.

3.1.2. Execution module

The execution module activates the biometric recognition system to be evaluated, and executes a performance evaluation. There are two methods for establishing an interface between the performance evaluation and the biometric recognition system. The first one consists in developing two systems as independently applied programs, while the second one consists in creating the biometric recognition algorithm as a component or library and then inserting it into the performance evaluation system. The former requires advance agreement between two systems with regards to the input/output file format, since the input data used for performance evaluation and the performance evaluation execution result data are generally transferred in a predefined form (generally, XML). Even though FRVT 2006 did not use the performance evaluation tool, participating companies submitted the biometric recognition system as an execution file, and the name of the input file used for evaluation and the output file that records the evaluation result were transferred as the program argument. For the component (or library)

FIGURE 1: PEM Framework.

method, the standardized interface should be agreed upon in advance. The agreed interface should be as simple as possible, and compatibility with international standards is desirable. The related international standards include biometric application programming interface (BioAPI) 1.1 [10] and BioAPI 2.0.

3.1.3. Result analysis module

The result analysis module performs a final analysis of recognition algorithm performance, using the result value obtained from the execution module. The performance of the specific algorithm can be expressed using several measurement criteria, and the appropriate measurement factor is decided upon depending on the objectives of the performance evaluation. Measurement factors can be broadly grouped into error rates and throughput rates. Error rates are basically derived from matching errors and sample acquisition errors, and the focus is on whether the algorithm is working properly and accurately. The throughput rate shows the number of users that the face recognition system can process in a given unit time. This throughput rate has significant meaning when performing the verification in a large image database [7].

3.2. Formalization of PEM

As biometric products are being used in establishing national infrastructure, a need for more effective and objective biometric performance test is on the increase. As examples are being announced such as FRVT, however, no objective and proved methodology is reported yet. In this paper, by presenting and formalizing biometric system performance test models, firstly, securing efficiency by eliminating unnecessary processes or factors that can occur when evaluating the performance of a biometric recognition system, and secondly, guaranteeing objectivity may be accomplished

by generating credible test factors and processes for the performance test, which is completely dependent on heuristic methodology.

The performance test was executed by elevating the usability of presented models in this paper, and by formulating model-based tools for validation.

PEM is defined as a structure $\Gamma = (DB, ATTR, INT, MET)$ with the following meanings.

- (a) DB is a set of all images of the test database.
- (b) ATTR is a set of pair $\langle \text{pattr}, \text{gattr} \rangle$ representing classification factors for probe and gallery set.
 - (i) $\text{pattr}, \text{gattr} \subseteq \text{fact}$ where fact is a set of factors influencing performance.
 - (ii) $\text{fact} = \{\text{age, sex, background, expression, pose, illumination, } \dots\}$.
- (c) INT is an interface of biometric recognition system being tested.
- (d) MET is a set of performance metrics.

DB represents all facial images in the image database to be used for the purpose of face recognition performance evaluation. ATTR represents the factors that affect the performance test. The factors include age, sex, photographing time, expression, background, posture, lighting, and costume; and these factors are used when selecting probe and gallery image set. Probe image set is $\text{PSET} = \sigma_{\text{pattr}}(\text{DB})$, and gallery image set is $\text{GSET} = \sigma_{\text{gattr}}(\text{DB})$. $\sigma_{\text{condition}}$ is a function selecting elements that satisfy the condition. For example, in case of $\text{ATTR} = \{\langle \{\text{Man, ExprSmile}\}, \{\text{Front, Normal}\} \rangle\}$ to perform the performance test for laughing men's faces, the probe image set used in the performance test is $\sigma_{\text{Man, ExprSmile}}(\text{DB})$. Gallery image set is $\sigma_{\text{Front, Normal}}(\text{DB})$.

When executing the performance test, all images of GSET will be matched to each image of PSET. That is, all image matching set MSET is a Cartesian product of PSET and GSET.

$$(a) \text{ MSET} = \{(x, y) \mid x \in \text{PSET and } y \in \text{GSET}\}$$

INT is the interface of face recognition module, an object of the performance test. Through this interface, performance-testing tools call the function of face recognition module. The interface should be as simple as possible, and it is desirable to be compatible with the international standards. INT should include the matching function F_{matching} , and this function outputs the matching results (accept or reject) by accepting one factor of image matching set.

$$(b) \exists F_{\text{matching}} \in \text{INT}, \text{ such that } F_{\text{matching}}(a) \text{ returns "accept" or "reject," where } a \in \text{MSET}$$

The execution result of face recognition function for the performance evaluation is expressed as the results from calling F_{matching} function using all element of MSET as factors, and this can be expressed as a two dimensional matrix, RMATRIX. That is, $\text{RMATRIX}_{i,j}$ element value is $F_{\text{matching}}(\text{PSET}_i, \text{GSET}_j)$. When the result value is accept, the outcome is 1, when the resulting value is reject, it is 0. For example, $\text{PSET} = \{p1, p2, p3\}$, and $\text{GSET} = \{g1, g2, g3\}$, MSET becomes $\{\langle p1, g1 \rangle, \langle p1, g2 \rangle, \langle p1, g3 \rangle, \langle p2, g1 \rangle, \langle p2, g2 \rangle, \langle p2, g3 \rangle, \langle p3, g1 \rangle, \langle p3, g2 \rangle, \langle p3, g3 \rangle\}$. If the resulting values of applying each element of MSET to F_{matching} is 1, 0, 0, 0, 1, 1, 0, 0, 1, consecutively, this can be expressed as the following 2-dimensional matrix:

$$\text{RMATRIX} = \begin{matrix} & \begin{matrix} g1 & g2 & g3 \end{matrix} \\ \begin{matrix} p1 \\ p2 \\ p3 \end{matrix} & \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{matrix}. \quad (1)$$

MET is a set of performance test measures such as fail-to-enroll rate (FTER), fail-to-acquire rate (FTAR), and false nonmatch rate (FNMR), false match rate (FMT). Such matching error-related metrics as FNMR, FMT can be calculated using RMATRIX:

FNMR

$$= \frac{\sum_{i=1}^n \sum_{j=1}^m (\text{Same}(\text{PSET}_i, \text{GSET}_j) \wedge (\text{RMATRIX}_{i,j} = 0))}{\sum_{i=1}^n \sum_{j=1}^m \text{Same}(\text{PSET}_i, \text{GSET}_j)},$$

$$\text{FMR} = \frac{\sum_{i=1}^n \sum_{j=1}^m (\text{Diff}(\text{PSET}_i, \text{GSET}_j) \wedge \text{RMATRIX}_{i,j})}{\sum_{i=1}^n \sum_{j=1}^m \text{Diff}(\text{PSET}_i, \text{GSET}_j)}, \quad (2)$$

where the following hold:

- (i) n is the size of PSET,
- (ii) m is the size of GSET,
- (iii) $\text{Same}(a, b)$ is 1, if a and b are images of same person, otherwise $\text{Same}(a, b)$ is 0,
- (iv) $\text{Diff}(a, b)$ is 0 if a and b are images of same person, otherwise, $\text{Diff}(a, b)$ is 1,
- (v) $a \wedge b$ is 1, if a and b are 1, otherwise it is 0,
- (vi) $a = b$ is 1, if values of a and b are same, otherwise it is 0.

3.3. Evaluation system development process

The following section describes how to build a performance evaluation system according to the PEM.

- (1) Describe the objectives of developing and evaluating a performance evaluation system.
- (2) Develop or select the biometric information database that will be used for performance evaluation.
- (3) Design test criteria that fit into the evaluation objectives.
- (4) Determine the type of interface to be used between the performance evaluation system and the biometric recognition system. If it runs as a standalone program, select the input/output file format; or, if it is linked by the component method, design the component interface.
- (5) Select the measurement criteria that fit into the evaluation objectives.
- (6) Implement the "data preparation module" which reads the biometric information according to the test criteria through the interface with the biometric information database.
- (7) Implement the "execution module" which executes the recognition algorithm with the gallery/probe biometric information provided by the data preparation module. The execution result (degree of similarity) should be saved in the database containing the results of performance evaluation.
- (8) Calculate the value of the measurement criteria by analyzing the similarity saved in the performance evaluation result database, and implement the "result analysis module" to generate a report on the results of the performance evaluation.

The test items and measurement criteria can be decided when performing the actual performance evaluation instead of building the performance evaluation system. In this case, select test items and measurement criteria that can be selected at steps 3 and 5 as described above, the test should be able to select the necessary items from among these when running the performance evaluation tool.

4. DESIGNING AND IMPLEMENTING THE PERFORMANCE EVALUATION TOOL

The performance evaluation tool was designed and implemented using the PEM proposed by this paper. The following section describes the contents and results by step, according to the evaluation system development process. The purpose of performance evaluation is to identify the technology level of the face recognition system through objective performance evaluation and certification, so as to encourage public trust in face recognition products and enhance their competitiveness.

4.1. Test criteria

The test criteria are designed in such way that those do not have to be selected when developing the evaluation system,

enabling the tester to select it in the course of performing the actual performance evaluation. Basically, the test item lists all the classification criteria so that the tester can select from them separately, based on the condition of the gallery image set and the probe image set. The gallery image set and the probe image set, each of which is composed of several items, are referred to as the “test set.” One performance evaluation project can generate several test sets, and each test set can generate a different result report. For instance, even though the collection of facial images to be registered in the face recognition system is white-front (normal) or purple-front (illum. yellow), the actual image acquired by the image acquisition device to identify the user can be normal, eye-closed, or tilted slightly to the left or right. The test set can be configured as in Figure 2. by assuming this kind of face recognition system.

If we assume that 10 face images exist per category in the above example, there are 20 facial images in the gallery set, and 40 facial images in the probe set. Therefore, the template generation function will be invoked 20 times for the gallery image, and image processing will be performed 40 times for the probe image, which means that 800 (20×40) comparison operations will be executed if performance evaluation is performed on the above test set. Among these comparison operations, image comparison will be performed 80 times for all the persons involved, because the image for a specific person appears twice in the gallery set and 4 times in the probe set, which results in an image comparison being performed 8 times for the same person, with 10 persons in total being compared. Therefore, 720 different image comparisons are made for the different persons, based on this system. Using this method, image comparison times can be estimated in advance, and the tester can estimate the time required for performance evaluation in advance, because the comparison times are calculated beforehand.

4.2. Selected BioAPI

The performance evaluation tool provides a standard interface for the face recognition system, and the examinee provides the face recognition module that satisfies this interface as the dynamic library. The performance evaluation tool is designed to enable the tester to change the face recognition module during the run time so that the tester can perform evaluation by changing the face recognition module without modifying the performance evaluation tool.

BioAPI 1.1 was applied for the compatibility with international standards, and only the minimum number of functions required for algorithm performance evaluation was selected, in order to reduce the examinee’s development burden. Table 3 shows the selected BioAPI for the face recognition module.

4.3. Measurement criteria

The performance evaluation metrics used by FERET and FRVT, as well as the metrics proposed by JTC 1/SC 37/WG 5 Standard [7], were analyzed. Among these metrics, the cri-

FIGURE 2: Example of setting the facial image test set.

teria related with the technology evaluation of the face recognition system were chosen, as shown in Table 4.

4.4. Class diagram of metadata

Within the performance evaluation tool implemented in this paper, individual projects created for performance evaluation internally generate metadata in a specific structure in order to save the settings related to performance evaluation and performance evaluation results. The structure of these metadata is as illustrated in Figure 3.

4.4.1. CMetaProject class

Whenever a new project is created, this class creates an instance in connection with the project. It maintains the path used to save the project name and the project itself, and the path to access a database in character strings. This information is related to the project configurations, and contains values established by the tester upon the project’s creation. In addition, the list of face image data designed by the tester for performance evaluation receives the lists of CMetaTestSet class. A single project may have at least one group of face image data for its performance evaluation, and independently create a report based on these individual evaluation results. Therefore, the information exists in the extendible list format. Moreover, this class allows users to ascertain the number of total images to test (compare), and that of probes and gallery images with regards to such information.

4.4.2. CMetaCategory class

This class contains data related to the subcategories of face images. Face images are influenced by the direction of the picture taken, the location of lighting, and posture, and so forth. According to these conditions, they are classified, and the tester may choose some of the classified images by using the performance test tool as the test subject group.

FIGURE 3: Example of setting the facial image test set.

TABLE 3: Selected BioAPI for face recognition module.

Function name	Meaning
BioAPI_Init	initializes the face recognition module
BioAPI_Terminate	terminates the face recognition module
BioAPI_FreeBIRHandle	releases the BIR data memory
BioAPI_GetBIRFromHandle	returns the BIR data from the data handle
BioAPI_CreateTemplate	generates the data template from the gallery
BioAPI_Process	generates the BIR data by processing the probe image
BioAPI_VerifyMatch	returns the verification result by comparing two processed images

TABLE 4: Selected performance evaluation metric.

Metric type	
Fail-to-enroll rate	enrollment throughput (no. of cases/s)
Fail-to-acquire rate	enrollment time (min, max, avg)
False reject rate (= $f_{ta} + f_{nm} * (1 - f_{ta})$)	extraction throughput (no. of cases/s)
false accept rate (= $f_{mr} * (1 - f_{ta})$)	extraction time (min, max, avg)
false nonmatch rate	matching throughput (no. of cases/s)
False match rate	matching time (min, max, avg)
cmc curves	transaction throughput (no. of cases/s)
roc curves	transaction time (min, max, avg)
error equal rate	template size (min, max, avg)

FIGURE 4: Selection window for face image probe set and gallery set.

FIGURE 5: Progress visualization when evaluating performance.

The chosen information is individually saved in the *Cmeta-category* class. These metadata contain variables, such as the category name of each criterion, the total number of images in the category, and the number of images failed to enroll (for gallery items) or acquire (for probe items).

4.4.3. *CmetaTestItem* class

This class contains data pertaining to one face image. Each face image has a unique ID for the photograph target, its location, and the face image items. Additionally, it contains

Boolean variables in order to save whether each item failed to enroll (for gallery items) and acquire (for probe items) or not.

4.4.4. *CMetatTestResult class*

This class stores the verification results of the comparison of one probe item with one gallery item in order to determine whether or not they come from an identical person. It contains each item's criteria, location, and ID information, along with variables, such as the similarity value and the comparison time created as a result of a comparison between two images.

4.4.5. *CMetaEnrolled class*

This class saves the template data created so as to recognize a face from the image data for the system to execute the enrollment of a gallery item. It holds not only the item's criteria and location information but also a binary space to store template data, as well as data used to save the results of the template creation and the time required to create a template.

4.4.6. *CMetaTestset class*

This class includes information about the face image probe group and gallery group in order to conduct the test. One project may have several test groups, and each test group individually saves the performance evaluation results. Metadata of the test groups bear the following information. Firstly, the class contains CMetaCategory which is the information holding the face image criteria as the list information for each of the probe and gallery groups. Namely, each of the probe and gallery groups may include a face image group with several different criteria. In addition, the class maintains the list of CMetaTestItem for each of the probe and gallery groups. This is not the criteria information but the metadata with individual image item information used for the test. It also contains the list of CMetaTestResult, containing the test results between a single probe item and a single gallery item.

Furthermore, where a system enrolls a gallery item, this class will contain the list of CMetaEnrolled classes in order to store template data that are created when each face recognition module generates its own template data. This is accomplished by using the face image data. As general data of such list data, the class contains variables to contain the test start time, end time, number of total probe items, number of total gallery items, number of total gallery items that failed to enroll, and number of total probe items that the system failed to acquire. We developed the six major modules and metadata classes that we examined above with the program for Windows, under Microsoft's Visual Studio development configuration. The face recognition modules, which are the subject of the performance evaluation, work in connection with the performance evaluation tool in the format of a dynamic link library.

4.5. *Implementing data preparation/execution/result analysis module*

The performance evaluation tool was developed as an application program running on Windows OS, and the face recognition module to be evaluated was implemented as a dynamic link library (DLL). The data preparation module that has the function of connecting with the biometric database and of setting the gallery and probe image set was implemented for use in the performance evaluation, as shown in Figure 4.

The face recognition module provided by the vendor was checked to verify that it provides the functions presented in Table 3, and the execution module that performs evaluation was implemented, using the functions of the selected face recognition module. A function that visually displays whether performance evaluation is progressing properly or not was included in the execution module, as shown in Figure 5.

Finally, the value of evaluation criteria is calculated by analyzing the similarity saved in the performance evaluation result database, and the result analysis module that generates the performance evaluation result report is implemented. This performance evaluation tool is equipped with a function that generates the evaluation result, as well as a function that issues the certificate for the face recognition module, depending on the evaluation result.

5. COMPARISON OF PERFORMANCE EVALUATION METHODS

Table 5 shows a comparison made between FERET and FRVT, which are the representative face recognition evaluation cases, and the evaluation method that uses the performance evaluation tool proposed by this paper.

Compared with performance evaluation programs such as FERET and FRVT, the performance evaluation tool proposed by this paper provides the following benefits.

- (i) Disclosure of the face image database can be fundamentally prevented.
- (ii) Development of a face recognition module that complies with international standards will be encouraged.
- (iii) The performance evaluation target can be separated from the performance tester.
- (iv) The evaluation cost can be reduced significantly, and individual evaluations can be performed for each vendor.

6. CONCLUSION

This paper proposed a PEM to evaluate the performance of biometric recognition systems. The proposed PEM is designed for compatibility with the related international standards, thereby contributing to the enhanced consistency and reliability of the performance evaluation tool that is developed according to this design. The proposed PEM is essential for the following reasons.

TABLE 5: Comparison between evaluation methods (FERET, FRVT, and the proposed performance evaluation tool).

	FERET	FRVT	Proposed evaluation tool
Evaluation target	Face recognition system	Face recognition system Scenario Operation	Face recognition system
Face DB distribution method	DB distribution onsite (possibility of face DB disclosure)	DB distribution on-site (possibility of face DB disclosure) FRVT 2006: Vendors provide the face recognition module. (No risk of disclosing face DB)	Vendors provide the face recognition module. (No risk of disclosing face DB)
Performance evaluator	Vendor staff	Vendor staff FRVT 2006: Evaluator	Evaluator
Result analysis tool	Performance result analysis tool is required	Performance result analysis tool is required	No analysis tool is required
Result report generation	Manual work	Manual work	Automatically generated by the tool
Individual evaluation	Impossible	Impossible	Individual valuation by vendor
Evaluation costs	All related vendor staff should congregate in one place (expensive)	All related vendor staff should congregate in one place (expensive) FRVT 2006: No meeting is required (low cost)	No meeting is required (low cost)
Responsibility of the vendor	I/O file format should be complied with	I/O file format should be complied with	Complies with recognition module interface

- (i) It represents a model and development method for the performance evaluation system.
- (ii) It applies the related international standards to the performance evaluation system.
- (iii) It enhances the consistency and reliability of the performance evaluation system.
- (iv) It provides guidelines for the design and implementation of the performance evaluation system by formalizing the performance test process.

In addition, a performance evaluation tool capable of comparing and evaluating the performance of the commercialized facial recognition systems was designed and implemented, and an evaluation that executed 800 billion comparisons in 596 hours using the KFDB [8] was conducted. The certificate issuance criteria regarding the performance of the face recognition systems should be presented systematically, and a method should be prepared that can promote certification.

REFERENCES

- [1] P. J. Phillips, P. Rauss, and S. Der, "FERET (Face REcognition Technology) Recognition Algorithm Development and Test Report," ARL-TR-995, U.S. Army Research Laboratory, 1996.
- [2] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, 1998.
- [3] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.
- [4] D. M. Blackburn, J. M. Bone, and P. J. Phillips, "FRVT 2000 Evaluation Report," FRVT 2002 documents, February 2001.
- [5] P. Grother, R. J. Micheals, and P. J. Phillips, "Face recognition vendor test 2002 performance metrics," in *Proceedings of the 4th International Conference on Audio- and Video-Based Person Authentication (AVBPA '03)*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 937–945, Guildford, UK, 2003.
- [6] P. J. Phillips and R. J. Michaels, "FRVT 2000: Evaluation Report," FRVT 2002 documents, March 2003.
- [7] ISO/IEC JTC 1/SC 37, "Biometric performance testing and reporting—part 1: test principles and framework," ISO IS, 2006.
- [8] B.-W. Hwang, H. Byun, M.-C. Roh, and S.-W. Lee, "Performance evaluation of face recognition algorithms on the Asian face database, KFDB," in *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA '03)*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 557–565, Guildford, UK, June 2003.
- [9] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki, "An introduction to evaluating biometric systems," *Computer*, vol. 33, no. 2, pp. 56–63, 2000.
- [10] "The BioAPI Consortium," BioAPI Specification Version 1.1, March 2001.