Performance Evaluation Model for the Face Recognition System

YongNyuo Shin¹, Jason Kim¹, YongJun Lee², Woochang Shin³, Jinyoung Choi⁴
Korea Information Security Agency¹, LG CNS², Seokyeong University³, Korea University⁴
ynshin@kisa.or.kr, bioman@lgcns.com, wcshin@skuniv.ac.kr, choi@formal.korea.ac.kr

Abstract

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 a performance evaluation model for biometric recognition evaluation tools, 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.

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 into contact with the recognition device intentionally, which results in fewer repercussions and less inconvenience when collecting the biometric information. An additional advantage exists, in that 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 U.S. These benefits, and the perceived necessity of increased security, have led to a rising social demand for face recognition systems, and certified performance evaluation has become important as a means of evaluating these face recognition systems.

This paper proposes a PEM (Performance Evaluation Model) 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.

Chapter 2 outlines existing studies related to performance evaluation of face recognition systems. Chapter 3 proposes a PEM to evaluate the performance of face recognition systems. Chapter 4 describes the design and implementation of the performance evaluation tool, based on the proposed PEM. Chapter 5 compares the performance evaluation method, utilizing the performance evaluation tool proposed in this paper, with existing performance evaluation programs. The last chapter provides a conclusion, and suggests some directions for future research.

2. Related Studies

FRVT (Face Recognition Vendor Testing) was a performance test for the face recognition system that was implemented using three FERET (Face Recognition Technology) 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 comparison, FRVT 2002 introduced an evaluation that made 15 billion comparisons in 264 hours[1][5][8].

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 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. In addition, 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.

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, and so forth) as well as performance evaluation measurement criteria such as FAR(False Acceptance Rate), FRR(False Rejection Rate), and EER(Equal Error Rate) should be selected. In addition, the face recognition system to be evaluated and a standardized interface for the performance evaluation system should be designed. International standards need to be applied at each stage of performance evaluation, in order to enhance fairness and reliability. To sum up, the 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.

Figure 1  Structure of the PEM

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 [7].

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 I/O file format, since the input data used for performance evaluation and the performance evaluation execution result data are generally transferred in a pre-defined 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) 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 BioAPI (Biometric Application Programming Interface) 1.1[3] 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[2].

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. The research face database that was developed by the KISA (Korea Information Security Agency) from 2002 to 2004 was used for performance evaluation[4].

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, as shown in Figure 2. 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.

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 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 1 shows the selected BioAPI for the 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. |

### 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 [2], were analyzed. Among these metrics, the criteria related with the technology evaluation of the face recognition system were chosen, as shown in Table 2.

| Metric type                          | Metric type                          |
|--------------------------------------|--------------------------------------|
| • Fail-To-Enroll rate                | • Enrollment Throughput (#case/sec.) |
| • Fail-To-Acquire rate               | • Enrollment Time (min, max, avg)    |
| • False Reject Rate (FTR = FTA + FNMR * (1-FTA)) | • Extraction Throughput (#case/sec.) |
| • False Accept Rate (FAR = FMR * (1-FTA))   | • Extraction Time (min, max, avg)   |
| • False Non-Match Rate               | • Matching Throughput (#case/sec.)   |
| • False Match Rate                   | • Matching Time (min, max, avg)      |
| • CMC Curves                         | • Transaction Throughput(#case/sec.)|
| • ROC Curves                         | • Transaction Time (min, max, avg)   |
| • Error Equal Rate                   | • Template Size (min, max, avg)      |

### 4.4. 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 DLL (Dynamic Link Library). 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. In the execution module, it is required to transform and calculate the similarity into numeric values, from 0(completely different), to 100(completely matching), which are commonly used in the performance evaluation. Assuming that the similarity returned by the selected face recognition module is $d$, the similarity of two images from completely different persons is $m$, and the similarity of the completely matching person is $M$. The function of $d$, $f(d)$, used to transform the similarity for performance evaluation, is as per the following:

$$f(d) = \begin{cases} 
\frac{d^*100}{M-m} & (M > m) \\
100 + \frac{d^*100}{M-m} & (M < m) 
\end{cases}$$

The face recognition module provided by the vendor was checked to verify that it provides the functions presented in Table 1, 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 3.

![Figure 3 Progress visualization when evaluating performance](image)

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

The following table 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.

Table 3 Comparison of evaluation methods (FERET, FRVT, and the proposed performance evaluation tool)

|                      | FERET                        | FRVT                          | Proposed evaluation tool |
|----------------------|------------------------------|-------------------------------|--------------------------|
| Evaluation target    | Face recognition system      | Face recognition system       | Face recognition system  |
|                      | - Face recognition system    | - Scenario                    |                          |
|                      | - Scenario                   | - Operation                   |                          |
| Face DB distribution method | 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                 | 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) | No meeting is required (low cost) |
|                      | - FRVT 2006: 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 |

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

- Disclosure of the face image database can be fundamentally prevented.
- Development of a face recognition module that complies with international standards will be encouraged.
- The performance evaluation target can be separated from the performance tester.
- 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:

- It presents a model and development method for the performance evaluation system.
- It applies the related international standards to the performance evaluation system.
- It enhances the consistency and reliability of the performance evaluation system.

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[4] 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.

7. References

[1] D.M. Blackburn, J.M. Bone, "FRVT 2000 Evaluation Report," FRVT 2000 documents, February 2001.
[2] ISO/IEC JTC 1/SC 37, Biometric Performance Testing and Reporting - Part 1: Test Principles and Framework, ISO Standards FCD, 2005.
[3] The BioAPI Consortium, BioAPI Specification Version 1.1, March 2001.
[4] Bon-Woo Hwang, Performance Evaluation of Face Recognition Algorithm on the Asian Face Database, KFDB, AVBPA, LNCS 2688, 2003
[5] P. J. Grother, R.J. Michaels and P. J. Phillips, "Face Recognition Vendor Test 2002 Performance Metrics," Proceedings 4th International Conference on Audio Visual Based Person Authentication, 2003.
[6] P. J. Phillips, P. Rauss, S. Der, "FERET (Face REcognition Technology) Recognition Algorithm Development and Test Report," ARL-TR-995, U.S. Army Research Laboratory, 1996.
[7] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki, "An Introduction to Evaluating Biometric Systems," IEEE Computer, February, pp. 56-63, 2000.
[8] P. J. Phillips, R.J Michaels, "FRVT 2002: Evaluation Report," FRVT 2002 documents, March 2003.