Privacy Protection of Biometrics Evaluation Database – A Preliminary Study on Synthetic Biometric Database

Kazuhiko Sumi[†] Takashi Matsuyama[†]
[†]Graduate School of Informatics
Kyoto University
Kyoto 606-8501, Japan
sumi@vision.kuee.kyoto-u.ac.jp tm@i.kyoto-u.ac.jp

Abstract

Currently, evaluation of biometric authentication systems depends on biometric databases. Such databases usually consist of biometric data collected in cooperation with a lot of volunteers. If the number of volunteers involved in a database becomes very large and if the database is circulated from a organization to other, there arises a new threat that the privacy of the volunteers might be infringed in case of database abuse. To protect the privacy of volunteers and to build a freely circulated biometric database, we propose a new method to construct synthetic biometric examples from real examples. In this paper, first we analyze the requirement of biometric database for evaluation and show the guidelines to build a synthetic biometric database. Then we show the preliminary case study on face database. The case study includes the data distribution analysis in a PCA subspace of 330 volunteer faces. Finally, we show some examples synthesized by our method.

1. Introduction

Evaluation of biometric authentication systems, especially accuracy evaluation, requires a large-scale biometric database[5]. Conventionally, such databases are constructed with cooperation of many volunteers. As the biometric authentication systems being put on practical applications, the number of volunteers required for evaluation is becoming large. For example, Wayman et al. showed that to achieve 0.001-certainty factor of false rejection rate requires 3000 biometric examples to test. If we consider the performance changes caused by gender, age, and ethnos, we have to gather the same number of examples for each group. This kind of large biometric database also required training the algorithm because recent algorithms are based on learning.

Building such a large database requires great effort finding a large number of volunteers and keeping them cooperating along with the procedure to build the database. Demand for circulation from an organization to the other will increase to support multi-national evaluation, nation-wide evaluation, and to develop better authentication algorithms. Transferring a database to a different organization implies less responsibility of database protection and there will be more possibilities of leakage of the individual data. Currently, the privacy of the volunteers is guaranteed by the agreement between the individual and the database maintainer, and, if we strictly comply with the agreement, it will be impossible to transfer the database between different organizations or between countries.

Once, the individual data is leaked, there are several scenarios of database abuse and possible social threat. We analyze such social threat and propose a synthetic database as an alternative solution for privacy protection. An example of synthetic biometric database is SFINGE[1]. It is based on the model of fingerprint generation and has been applied in the public benchmarking FVC2004[3]. However, for biometrics whose generation processes are not modeled well, it is better to be based on from real database and keep the statistical characteristics of the real database.

In this paper, we propose a method to generate synthetic biometric database from a real database, keeping the same evaluation results form both database. In the following sections, first we analyze the threat of individual data leakage from biometric database in Section 2, then analyze the requirements of biometric database propose how the synthetic database should be built in Section 3. In addition to the general discussion, we perform a case study on face database in Section 4.

2. Threat analysis of biometric evaluation database

Various types of vulnerability have been alerted in biometric authentication systems. Figure shows points of

weakness in general enrollment/authentication procedure.

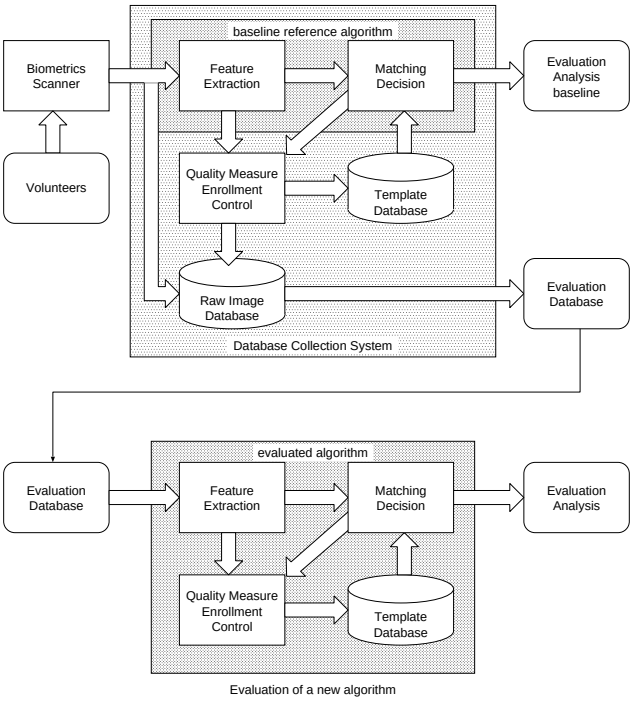

Figure 1. Schematic diagram of database collection and evaluation of a biometric authentication system

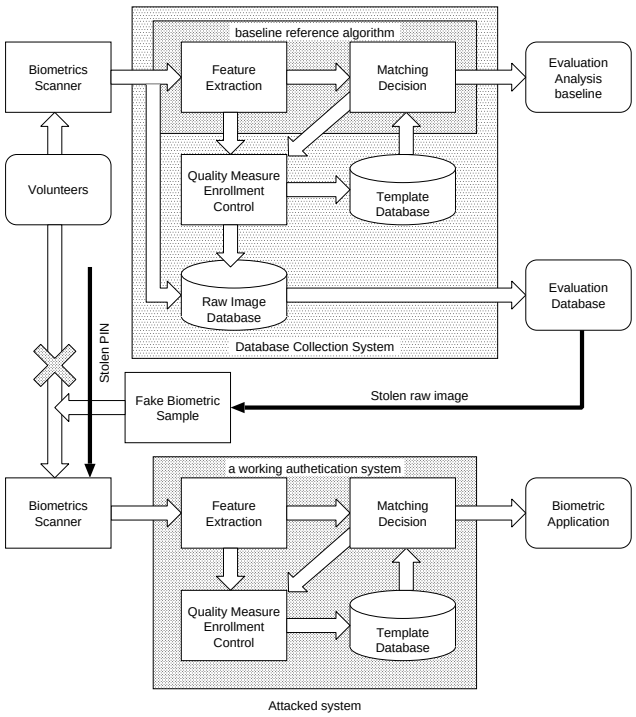

Figure 2. Threat of attack against a working biometric authentication system using a fake biometric sample made from evaluation database

Figure. 1 shows the schematic diagram of database collection and evaluation procedure. Generally, threat of database leakage is less recognized than individual template leakage from operating biometric authentication system. However, it is much easier to steal evaluation database than to steal individual templates, because of the following reason:

1. Individual templates are usually stored in a system are and secured. But evaluation database are exposed outside of the system and are not secured.
2. Many of individual templates are not raw images but feature vectors. But most of the evaluation databases have raw images.

The first scenario is to produce a fake biometric example from stolen database. A fake biometric sample, such as a fake fingerprint, an iris and a fake facemask, can be produced from the raw image. This fake example can be used to attack a biometric authentication system under operation shown in Figure 2. Suppose the attacker steal the template database DB_B of biometric evaluation system B. If the attacker knows the PIN N_j of the person $P_j, (P_j \in DB_B)$, a fake biometric, which produces the same impression as T_j, is produced from T_j, then it can be used to attack a 1-to-1 authentication system A and obtain the access permit of the owner j. Even if the PIN is not known, but it is certain to be enrolled in the specific system A, the fake example can be used to obtain access permit to the system, if the system allows 1-to-N authentication scheme.

The second scenario is to find a real-existing close example from the whole database shown in Figure3. Suppose the attacker is eager to access a certain system A with the PIN i, and obtained the template database of the system DB_A. Then the attacker will obtain very large evaluation database DB_B and find a person P_j, whose biometric template $T_j, (T_j \in DB_B)$ is very close to the one $T_i (T_i \in DB_A)$ enrolled in the target system T_i. If the attacker track the PIN of DB_B and possibly find the owner of T_j, the attacker force or persuade the person P_j obtain access to the system A. This can fake the mechanism of liveness detection of the biometric system A.

The third scenario is to get the database for fun. Some

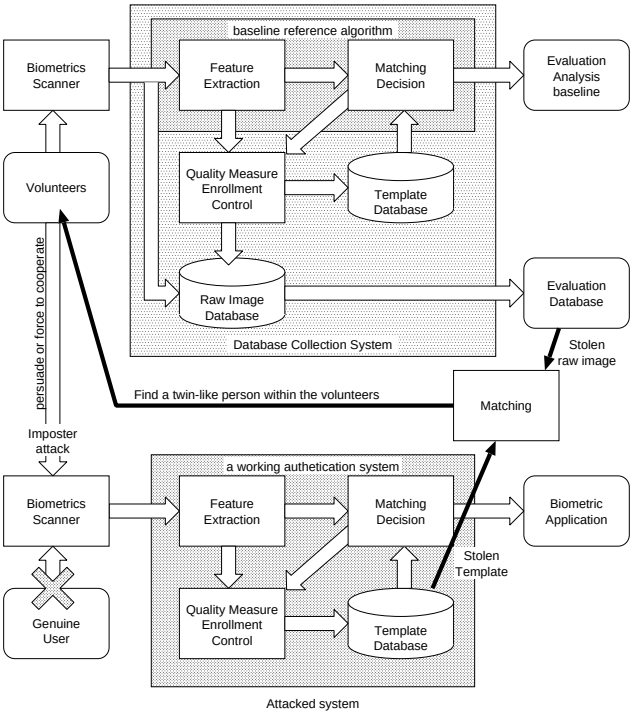

Figure 3. Threat of attack against a working biometric authentication system using a twin-like impostor found in evaluation database

might try to identify an subject enrolled in a biometric database for evaluation. If the biometrics data is face, it is not easy but possible to find the person from a community. Then a story of the personal facts of the subject might be revealed. The privacy of the volunteer will be violated.

To prevent the above scenarios, protecting the database is desirable. However, an evaluation database has necessity of circulation among different organizations and it is impossible to monitor all data access. Also, hiding raw image is impossible, because the evaluation usually includes feature extraction algorithm as well as matching and decision algorithms and the input of the algorithm must be a raw image. So, it is inevitable to expose individual raw data. One solution is to evaluate the system in a network isolated white room, in which only limited number operators can come in. However, there is no confidence that every evaluator operates such a secure evaluation environment. The database builder, which is responsible to the volunteers' privacy, should take measures to protection. Otherwise, the database builder will have risk of being accused or difficulties gathering wide range of volunteers.

3. Requirement of synthetic biometric database

A synthetic biometric database consists of virtual individual biometric examples is one of the solution. However, to satisfy the demand on precise evaluation of biometric authentication systems, the database should have the following characteristics:

1. (precision requirement) The evaluation results derived from a synthetic biometric database should be the same with the real database.
2. (universality requirement) The precision requirement should be satisfied for all the authentication algorithms to be evaluated.
3. (privacy requirement) Each biometric data in the synthetic database should not represent any real person.

The precision requirement can be resolved in the following way.

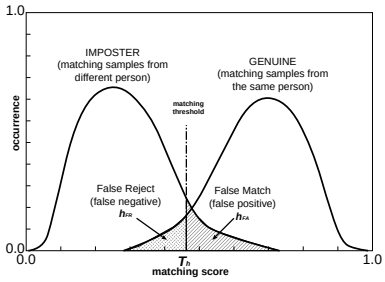

Figure 4. Typical similarity distribution of biometrics authentication

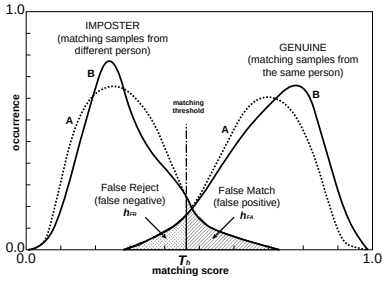

Figure 5. An ideal similarity distribution of synthetic biometrics database B comparison with real database A

Suppose group A is a real database corrected from existing individuals consist of M_A examples. Using algorithm Θ, an biometric raw example $a_i, (a_i \in A, 1 \leq i \leq M_A)$ is projected to $\theta(a_i)$ in a future space. If we obtain a similarity distribution like Figure.4, it means that the distribution of h_{FA} at threshold T_h is the number of impostor samples closer than T_h in the feature space Θ. Another group B is a synthetic database derived from A consists of M_B examples. ($M_B = M_A$ in this case) Using algorithm θ, a biometric raw example $b_i, (b_i \in B, 1 \leq i \leq B_A)$ is projected to $\theta(b_i)$ in a future space. If we like to have the same false rejection rate (FRR) and false accept rate (FAR) at threshold T_h, the number of pairs, which are closer than T_h in the feature space Θ should be the same with the case of A. This suggests that we should be careful not to change the distance of samples, whose distance is less than T_h, but we don't have to be careful about the distance of samples, whose distance is larger than T_h.

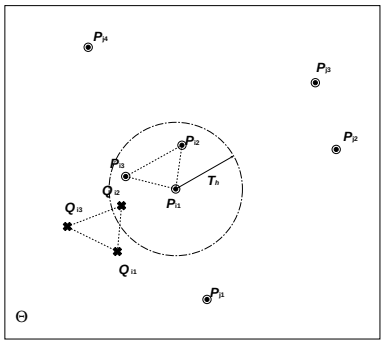

Figure 6. Relationships of critical samples in a real database and the corresponding synthetic database

Figure.6 shows an example of such a deformation. Suppose P are the biometric samples. For a arbitrary index i, select samples closer than the threshold T_h in the feature space Θ. In this figure, they are P_{i1}, P_{i2}, and P_{i3}. If we generate synthetic examples Q_{i1}, Q_{i2}, and Q_{i3}, and the distance between Q_{i1} and Q_{i2}, Q_{i1} and Q_{i3}, and Q_{i2} and Q_{i3} are equal to the original distance between P_{i1} and P_{i2}, P_{i1} and P_{i3}, and P_{i2} and P_{i3}, respectively, the synthetic samples satisfy with the three requirements explained in this section.

In the above deformation, we should consider isolated samples which have only one neighbor or no neighbors within the threshold T_h. In case of doubles, we rotate the pair of samples around its center. In case of standalone, we move the sample along a certain displacement, which has a fixed length and a random direction.

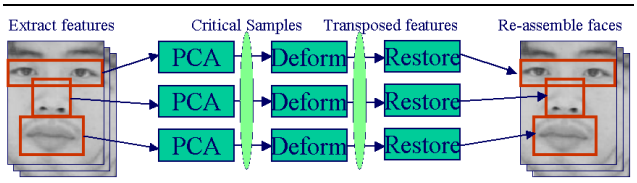

Figure 8. Schematic diagram of face image deformation based on facial parts regions

4. A case study using image based PCA

According to the idea in Section. 3, we have synthesized a face database from a real face database. The real faces are from HOIP face database contains 300 subjects of various age (from 20 to 60) and gender (150 males and 150 females), in a illumination controlled environment. It is available under usage agreement with the database builder. Figure. 7 shows the examples of the face images in the database. We used only the frontal faces.

In this study, we deform the faces in image based PCA subspaces, and then reconstruct face images. Because of the nature of image based PCA, location of each pixel or spatial relationships between pixels is not considered at all. When there exists a contour with sharp intensity gap and the shape of the contour is changing more than few pixels between images, PCA cannot interpolate the intermediate changes between two samples. This will result in artifacts on a synthesized image reconstructed from PCA subspace. To avoid this problem, we perform the deformation in a following way as shown in Figure. 8. First, we select the three major parts, eyes, nose and mouse, from a face. Those parts, defined as fixed sized rectangular regions, are cropped from a face image and aligned at their center. Then PCA is applied to each region. Deformation is performed in the PCA subspaces, then synthetic parts images are reconstructed. Finally, those synthetic parts are superimposed on the original image.

Regards to the details of deformation, triples are detected in each PCA subspace Θ of facial parts region. Then the center of the triples P_{i1}, P_{i2}, and P_{i3} are calculates as C. The synthetic face samples Q_{i1}, Q_{i2}, and Q_{i3} are placed at the symmetrical position of P_{i1}, P_{i2}, and P_{i3}, respectively. The relationships of P_{i1}, P_{i2}, P_{i3}, Q_{i1}, Q_{i2}, Q_{i3}, and C are shown in Figure. 9.

Examples of synthetic faces are shown in Figure. 10. In the figure, upper row is the real faces and the lower row is the synthesized images. As we can see, the pair (left and right) of images is apparently different person but has similar impression. The distances between three samples are same within the PCA feature space.

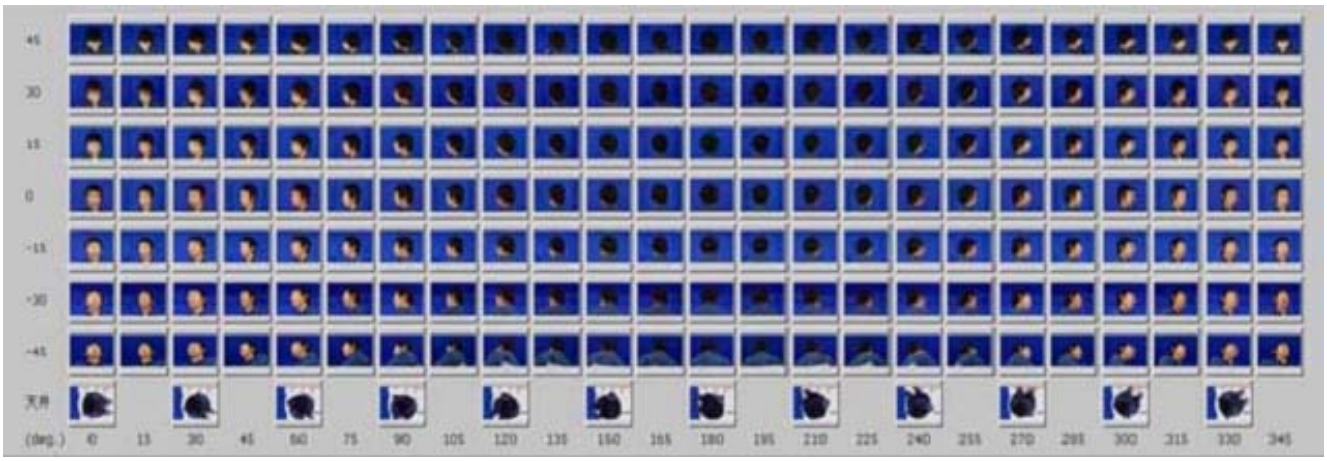

Figure 7. Example images in HOIP face database

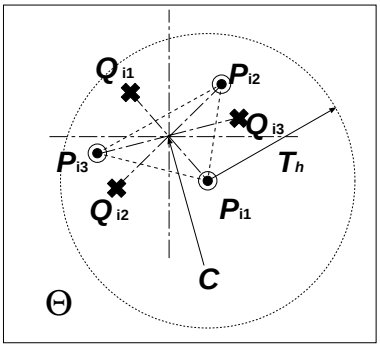

Figure 9. Generating synthetic face images from the most close triples

5. Discussion and future direction

At this moment, we have not completed the way to satisfy the universality requirement yet. Those synthetic images, whose distances are same with the real images in PCA sub-space, are not equal-distant in other feature space, such as graph feature space derived by Gabor filter banks.

To manage more algorithms proposed so far[6], our feature work will be exploration of the better feature space, which is more universal than PCA subspace. For example, to satisfy both image-based algorithms and feature-based algorithms, we should take shape parameters into account. One direction is to describe a face with its shape and texture normalized with the shape. Figure. 11 shows such a face modeling[2]. In the figure, faces corresponding to the same person appeared in Figure. 10, are modeled by shape (left column) and texture (right column) respectively. Using this scheme, we will be able to describe facial feature location by the shape parameter. Texture of a face (right column) is normalized by shape and contours of facial parts are aligned precisely. This will achieve better facial texture synthesis without artifacts around the contour.

Another discussion is how to deal with intra-personal variations. Since, some of the face recognition algorithms require multiple images with different appearances, we have to simulate intra-personal changes. One way is to analyze intra-personal variation space, which is introduced by Moghaddam[4], and generate synthetic images with intra-personal changes. The other idea is to take multiple images at registration and apply same deformation to the series of images. The former approach requires only a single image but the variation is a statistically calculated one and is not real variations. The later approach contains the real variations captured in real situation. However, which way is more realistic is not examined yet.

6. Summary

In this paper, we have analyzed the vulnerability and threat of the biometric evaluation database and proposed the method to generate synthetic database based on real examples. The reason we generate the database from real example is that we can keep the statistical distribution of the original database, thus, we are expecting that the evaluation result will be the same as original real database. The proposed database, which does not have privacy problem, can be circulated freely among biometric vendors and testers. We hope that this technique will accelerate the development practical biometric authentication systems.

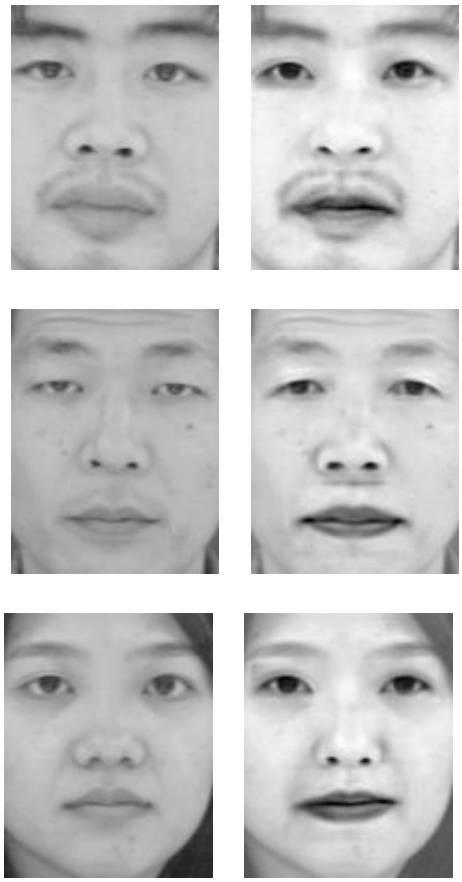

Figure 10. The real face image (left) and the synthetic face image (right) using our proposed method

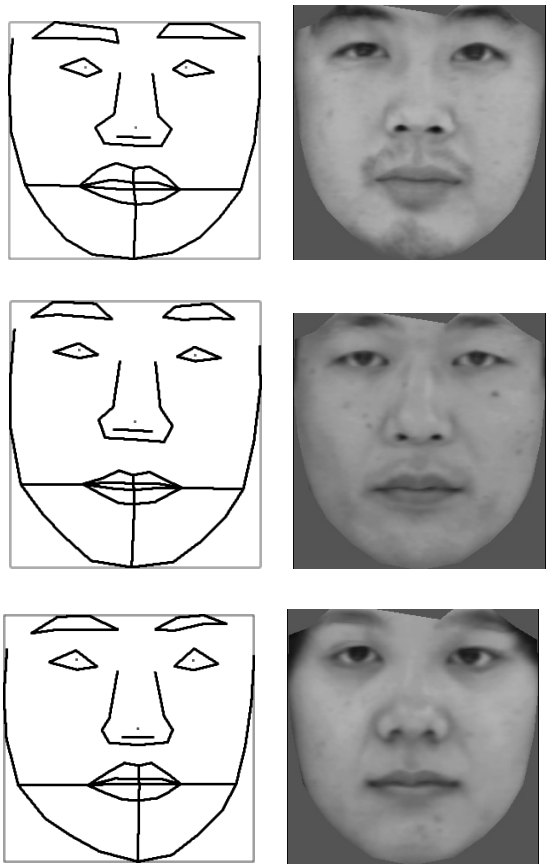

Figure 11. Shape representation (left) and shape normalized texture representation (right) of faces using active appearance model. Each row contains the same person as shown in Figure.10

Acknowledgments

This research is supported in part by the Informatics Research Center for Development of Knowledge Society Infrastructure, 21st. Century COE program and by contracts 13224051 and 14380161 of the Ministry of Education, Culture, Sports, Science and Technology, Japan. This research is also supported in part by the research contracts with Japan Automatic Identification Systems Association. Figure. 11 contains the latest results of Chang Liu.

References

[1] R. Cappelli, A. Erol, D. Maio, and D. Maltoni. Synthetic fingerprint-image generation. In *Proc. International Conference on Pattern Recognition*, September 2000.

[2] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. In *AFGR00*, pages 227–232, 2000.

[3] D. Maio, D. Maltoni, R. Cappelli, J. Wayman, and A. K. Jain. Fvc2004: Third fingerprint verificatin competition. In *Proc. International Conference on Biometric Authentication*, pages 1–7, July 2000.

[4] B. Moghaddam, T. Jebara, and A. S. Pentland. Baysian face recognition. *PR*, 33:1171–1782, 2000.

[5] C. L. Wilson. Large scale usa patriot act biometric testing. In *Proc. International Meeting of Biometrics Expert*, March 2004. http://www.biometricscatalog.org/document_area/view_document.asp?pk={5E0CA69A-B4AC-4FE9-9624-6ED3450E9CCF}.

[6] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.