Identification Using Encrypted Biometrics

Mohammad Haghighat, Saman Zonouz, and Mohamed Abdel-Mottaleb

Department of Electrical and Computer Engineering, University of Miami
haghighat@umiami.edu, {s.zonouz,mottaleb}@miami.edu

Abstract. Biometric identification is a challenging subject among computer vision scientists. The idea of substituting biometrics for passwords has become more attractive after powerful identification algorithms have emerged. However, in this regard, the confidentiality of the biometric data becomes of a serious concern. Biometric data needs to be securely stored and processed to guarantee that the user privacy and confidentiality is preserved. In this paper, a method for biometric identification using encrypted biometrics is presented, where a method of search over encrypted data is applied to manage the identification. Our experiments of facial identification demonstrate the effective performance of the system with a proven zero information leakage.

Keywords: face recognition, encrypted biometrics, search over encrypted data.

1 Introduction

In recent years, there has been a significant attention to substitute biometrics for passwords in authentication systems [7]. Biometric identifiers are distinctive, measurable characteristics used to label and describe individuals [6]. The well-known biometrics used for human identification are the fingerprints, face, iris, voice and DNA. Some of the advantages of biometrics over passwords are their precise identification, highest level of security, mobility, difficulty to forge, not being transferable, and user friendliness.

Besides the above-mentioned advantages, there are challenges that biometric systems face. One of the challenges is the changes in the biometric data over time. Biometric data of a person changes over time. For instance, facial features change due to the changes in illumination, head pose, facial expression, cosmetics, aging, and occlusions because of beard or glasses. Therefore, biometric systems usually identify subjects based on the nearest matches, rather than exact matches. Biometric-based identification is provided through a matching process between the biometric information of the querying subject and of the whole subjects available. The closest match will usually identify the subject. Therefore, these security systems have to store biometric information of all subjects in a database to be utilized at the time of query.

Another challenge faced in biometric systems relates to the possibility of identity theft. What if an attacker gains access to this database and steals the biometric information of an individual? The biometric information is unique and irrevocable, and unlike passwords you can never ask the users to change their biometrics. So, the system must guarantee the users' preservation of privacy, and the biometric information database has to be encrypted. However, the varying characteristic of biometrics brings about a serious

R. Wilson et al. (Eds.): CAIP 2013, Part II, LNCS 8048, pp. 440–448, 2013.

problem in encrypted domain since a little change in the plaintext results in big differences in the ciphertext. This difference misleads the classifier in the recognition process.

The problem of searching over encrypted biometrics has been considered for the first time by Bringer *et al.* in [3]. They have introduced an error tolerant searchable encryption system that makes it easy for the system to cope with the variations of the biometric data. In this method, the iris code system proposed in [4] is employed for biometric representation. This algorithm makes use of a locality sensitive hashing function that gives identical or very similar hash results to the biometric data that are close to each other. Using the locality sensitive hashing function, the (potentially malicious) database provider can cluster the data into groups based on the pattern that the data is being searched/processed for every received query. Therefore, he may eventually be able to sort the data records that could lead to a potential breach in the confidentiality of the data. To address the above-mentioned vulnerability, our proposed algorithm employs query over encrypted data with proven zero information leakage.

In this paper, we present a method for privacy-preserving identification using encrypted biometrics. The proposed algorithm applies a security approach to a face recognition system with a proven zero data disclosure possibility. While all the feature vectors are encrypted, it performs the classification using an effective method of search over encrypted data that can also consider the variations of the biometrics.

This paper is organized as follows. Section 2 demonstrates the structure of a typical pattern recognition system modified to be utilized in an encrypted domain and how the biometric data is stored in the database in an encrypted form. The structure of the search over encrypted data algorithm is presented in Section 3. Section 4 gives some experimental results to evaluate the performance of the system; and finally, Section 5 concludes the paper.

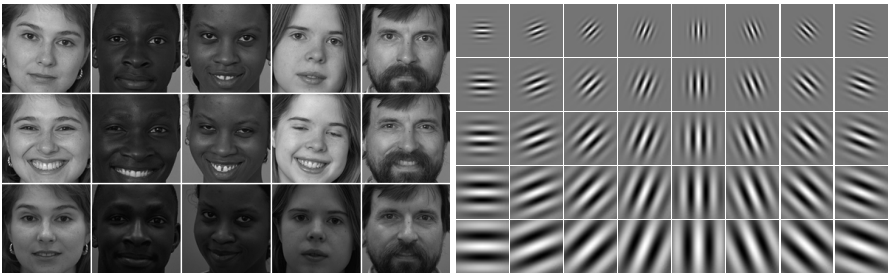
2 Populating the Biometric Database

Although this work can be applied to any biometric data, we have used face images as the biometric identifiers. In order to evaluate our algorithm, the Facial Recognition Technology (FERET) database is used in our approach [10]. We selected six hundred frontal face images for 200 subjects, where all the subjects have three different images. In FERET database, these images are letter coded as *ba*, *bj*, and *bk*. For preprocessing, Viola and Jones face detection algorithm [12] was applied to the images and the detected facial images were resized to 120×120 pixels. Fig. 1(a) illustrates some samples of the database after this preprocessing.

We can divide the feature vector generation into two major steps: feature extraction and dimensionality reduction. However, since we need to encrypt the feature vectors, there is also an additional quantization stage. Note that, feature vectors consist of real numbers; however, encryption algorithms are applied on integer values. Therefore, there is a need to consider an effective quantization method.

2.1 Feature Extraction

The frequency and orientation representations of Gabor wavelets (filters) are similar to those of the human visual system and they have been found to be particularly



(a) (b)

Fig. 1. (a) Face samples from FERET database after face detection. First row: *ba*. Second row: *bj*. Third row: *bk*. (b) Gabor wavelets in five scales and eight orientations.

appropriate for texture representation and discrimination [11]. Gabor filters have been widely used in pattern analysis applications [8, 9, 11]. The most important advantage of Gabor filters is their invariance to illumination, rotation, scale, and translation. Furthermore, they can withstand photometric disturbances, such as illumination changes and image noise.

In the spatial domain, a two-dimensional Gabor filter is a Gaussian kernel function modulated by a complex sinusoidal plane wave, defined as:

$$G(x,y) = \frac{f^2}{\pi\gamma\eta} exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right) exp\big(j2\pi f x' + \phi\big) \qquad (1)$$
$$x' = \quad xcos\theta + ysin\theta$$
$$y' = -xsin\theta + ycos\theta$$

where f is the frequency of the sinusoidal factor, θ represents the orientation of the normal to the parallel stripes of a Gabor function, ϕ is the phase offset, σ is the standard deviation of the Gaussian envelope and γ is the spatial aspect ratio which specifies the ellipticity of the support of the Gabor function.

Our proposed algorithm, employs forty Gabor filters in five scales and eight orientations as shown in Fig. 1(b). Fig. 2 illustrates the face detection step along with the real parts of the result images after applying Gabor filters on the face image. Given the fact that the adjacent pixels in the image are highly correlated, we can remove the information redundancy by downsampling the feature images resulting from Gabor filters [8, 11].

Gabor filters are extrac the variations in different frequencies and orientations in the face. Here the size of the output feature vector is the size of the image (120×120) multiplied by the number of scales and orientations (5×8) divided by the row and column downsampling factors (4×4) which is $120 \times 120 \times 5 \times 8/(4 \times 4) = 36000$ in total. The feature vector is still very large even after downsampling. Therefore, we will need to use dimensionality reduction methods [5].

For dimensionality reduction, we use general discriminant analysis (GDA). The basic idea is close to the support vector machines (SVM) in a way that the GDA method provides a mapping of the input vectors into high-dimensional feature space. In the transformed space, linear properties make it easy to extend and generalize the classical

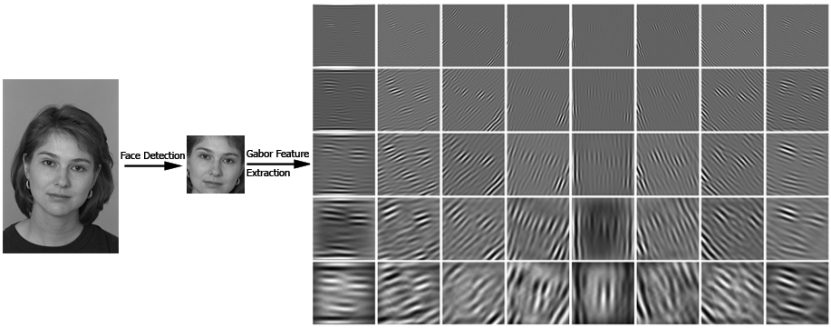


Fig. 2. Face detection and feature extraction, results of applying filters shown in Fig. 1(b) on a face image

linear discriminant analysis (LDA) to non-linear cases. More details for this approach are presented in [1, 11].

Note that the number of features in a LDA-based method (like Fisherfaces or GDA) depends on the number of classes in the classification problem and can be at most $C-1$, where C is the number of classes. Here, the maximum size of the projected vectors is 199 which has a significant reduction in comparison with 36000.

2.2 Feature Quantization

The proposed method combines the feature quantization and classification in range queries over a few parallel k-d trees. The feature vectors are fed into a k-d tree which partitions the feature space along each feature using hyperplanes. These boundary lines are treated as the quantization thresholds. A k-d tree of the depth L that can quantize $L-1$ features has $2^{L-1}-1$ nodes and 2^{L-1} leaves, therefore, it needs 2^L-1 samples to be built. The maximum number of quantized features can be calculated using the equation below. In our experiments, since the number of classes (C) is 200, the number of quantized features is restricted to seven. In order to make use of more features, we employ several k-d trees in parallel using different sets of features.

$$L_{max} = \lceil \log_2 (C+1) \rceil - 1 \tag{2}$$

GDA sorts the features according to their discriminative power. Therefore, it is important to select the first features of the GDA to construct the k-d trees. On the other hand, the deeper you go towards the leaves of the tree, the number of boundary lines (i.e., 2^L) increases. In order to have a reasonable quantization we are just using the four deepest levels of each tree for quantization. For this reason, it is also better to use the most discriminative feature in a higher level of the tree where the boundary lines are more dense. Therefore, the features will be used in reverse order, i.e., the least discriminative feature for the root and the most discriminative feature for the leaves.

In our experiments, we make use of 77 features to construct 11 k-d trees. Each of the first 11 features of the GDA, i.e., the most discriminative ones, are used for the leaves of a tree where we have the highest number of boundary lines. Consequently, the last 11 features (67 to 77) are used for the roots where there is only one boundary line defined.

Then, the first three levels of each tree, which have the coarsest quantizations, are disregarded such that each tree is quantizes four features. Therefore, using 11 trees quantizes the first 44 features of the GDA output. Note that in real systems the number of subjects will be higher, which, based on equation (2), increases the number of features used in each tree.

2.3 Biometric Record Encryption

Most of the security systems maintain a single biometrics database that their agents can have access to. Increasing the number of subjects makes the database huge such that storing all these records locally is often not feasible. Therefore, we assume that the database is stored in a public cloud. Face features as the biometric key attribute are encrypted to prevent attackers from unauthorized access to these confidentiality-sensitive records.

Let us assume that the trusted gateway intends to encrypt and send a user's biometric data record to the database and that the encryption should be such that it preserves the capability to be searchable by specific queries. Our solution makes use of search-aware encryption [2] and delicately applies it on a biometric identification system. C_i denotes the encrypted feature vectors of the i-th person. In the first step, all the cryptographic parameters, i.e., public keys, P, and private (secret) keys, S, are generated:

$$P \leftarrow (P_1, P_2, \cdots, P_t), \tag{3}$$

$$S \leftarrow (S_1, S_2, \cdots, S_t), \tag{4}$$

where t denotes the possible number of predicates (queries) that the customer can ask for in the future. t directly depends on the number of features and quantization boundary lines. The Encryption is carried as using the following assignments for $0 < j < t+1$

$$C_j \leftarrow \begin{cases} \text{Encrypt}(P_j, M) \ \textit{if Predicate}_j(F) = 1, \\ \text{Encrypt}(P_j, \perp) \ \text{Otherwise.} \end{cases} \tag{5}$$

where F is the feature vector of the subject's biometric and M is his/her ID or any relevant data. M is encrypted using the public key corresponding to the predicate which is only true by his biometric. Once completed, this step will give us a ciphertext vector

$$C \leftarrow (C_1, C_2, \cdots, C_t), \tag{6}$$

whose length is the number of possible predicates t. Note that, only one of the ciphertexts includes the data of the subject. The encryption will be completed locally by the trusted gateways so that the (untrusted) cloud provider will only have access to ciphertexts, and not the encryption keys.

3 Query over Encrypted Biometrics

After storing the biometric samples of the individuals in the database, the system is ready to respond to queries. The query over encrypted biometrics approach is briefly

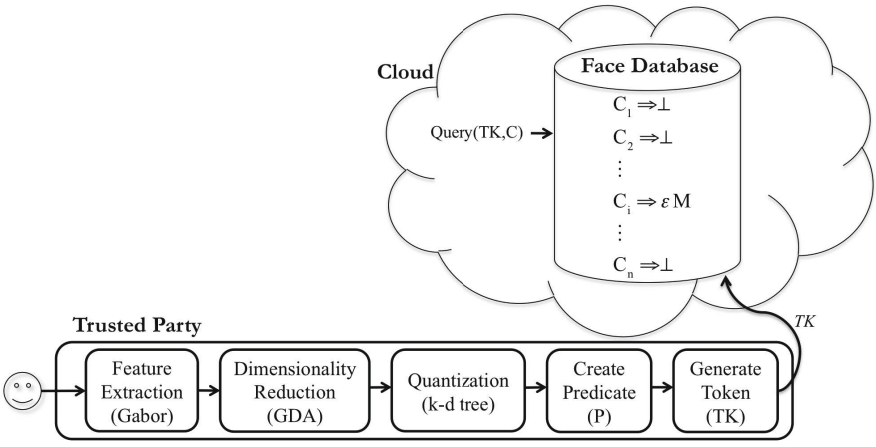


Fig. 3. Framework of the system

described in this section. Different components of the proposed system and their interconnection are illustrated in Fig. 3. At the time of the query, a face image of the client is fed into the system. Features of the face image are extracted and quantized using the k-d tree method described in Section 2. Because of the changing characteristic of biometrics, an upper boundary and a lower boundary are derived from the query feature vector to be compared to the existing feature vectors stored in database. The query is answered by records whose feature vectors fall in between these two boundaries. Therefore, a conjunctive predicate is created for the comparison query.

$$\mathbf{P} = \begin{bmatrix} l_1 < p_1 < u_1 \\ l_2 < p_2 < u_2 \\ \vdots \\ l_n < p_n < u_n \end{bmatrix} \tag{7}$$

Conjunctive query over encrypted data makes it possible to utilize a number of queries altogether with minimum amount of information leakage. For example, in the above predicate, $2n$ comparisons must be verified conjunctively $((l_1 < p_1 \wedge p_1 < u_1) \wedge (l_2 < p_2 \wedge p_2 < u_2) \wedge \ldots \wedge (l_n < p_n \wedge p_n < u_n))$. In order to guarantee the privacy-preserving ability of the system, such a query should carry no more information than the truth value of the conjunctive predicate. That is, if all the comparisons are true, then the query result is εM; however, if any of the comparisons is false, then the query not only will give no output, but even the cloud provider will be unable to identify which comparison is false. For example, if a conjunction $(P_1 \wedge P_2)$ is false, an eavesdropper should not be able to identify which one of the P_1 or P_2 or both were false.

3.1 Cryptographic Token Generation

As mentioned in the last section, for any query, we will have an accepted range. The range is defined by a lower and upper boundary where the features of the query fall in

between these two boundaries. Using the range query predicate, our algorithm defines a token including the encrypted private key corresponding to that query predicate[1]. The comparisons are applied in a conjunctive manner such that the ciphertext can be decrypted only if the result of all the propositions are true in union.

3.2 Biometric Database Query

The final step of the searchable encryption scheme is processing the received query on the encrypted biometric records using the single private key that the trusted gateway has sent. It is easy to follow that the system will retrieve (correctly decrypt) only the records which match the received query, and will get $\perp$ as the result of all other decryptions (see equation (5)). The retrieved records are then sent back to the trusted gateway. It is worthy of mention that for absolute data disclosure prevention, the trusted party should encrypt the records using a symmetric key such that M is already encrypted (εM) and the database provider is not allowed to see the actual record.

4 Experimental Results and Evaluations

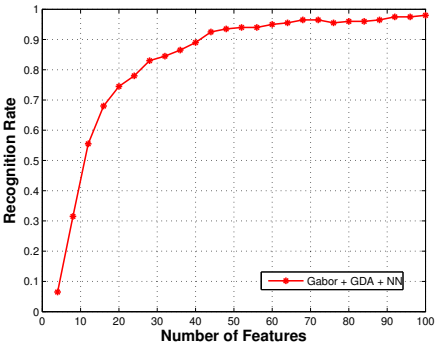
Accuracy of a biometric recognition system depends on different factors in the system, e.g., the informativity of features, number of features, and classifier performance. There is a huge variety of face recognition systems in the literature. Depending of what features and which dimensionality reduction method they use and the applied classifier, these methods have different recognition accuracies.

As mentioned before, our test face recognition algorithm utilizes Gabor features and general discriminant analysis dimensionality reduction. Fig. 4(a) demonstrates the performance of this approach using a simple nearest neighborhood classifier. Note that a leave-one-out cross-validation is considered in our experiments with three samples for each subject. The previous methods mentioned in Section 1 have not evaluated the biometric recognition rate or false positive rate after applying their security algorithms. Table 4(b) illustrates the recognition rate and false positive rate of the proposed security algorithm using different number of trees and features.

Note that we accept a subject as a potential match whenever more than half of the trees gives a true response for its predicate. As it can be seen in Table 4(b), with an increase in the number of features, the recognition rate increases and we have better results that prove the discriminative power of our employed features. On the other hand, using more trees makes the classifier more accurate. Note that since we are using conjunctive query, more features bring about more probability to have a negative response which on the other hand leads to a reduction in false positive rate.

As mentioned before, we apply a range query over the encrypted database. Here, the recognition rate is directly related to the width of the accepted region. As the number of quantization levels increases exponentially with the level of the tree ($2^{L-1}+1$), the intervals get smaller. Therefore, in order to have a consistent range in all levels, the number of accepted intervals should also increase. In the above experiment we have

[1] Recall from equation 4 that S includes a private key for each possible predicate (query).



(a)

No. of Trees	No. of Features	Recognition Rate	False Positive Rate
1	4	77.50	24.62
3	12	82.50	15.58
5	20	83.00	10.93
7	28	87.50	8.38
9	36	91.50	5.38
11	44	93.50	3.32

(b)

Fig. 4. (a) Accuracy of the face recognition using Gabor features, GDA dimensionality reduction, and nearest neighborhood classifier. (b) Accuracy of the secure face recognition system.

chosen the width of the acceptance region to be $2^{L-2}+2$. In addition to the number of features, the accepted region width will also be another degree of freedom for the designer to consider a trade-off between the recognition rate and false positive rate.

5 Conclusions and Future Work

In this paper, we presented a privacy-preserving biometric identification solution that stores and processes individual biometric records while they are encrypted. The proposed algorithm applies a security approach to a face recognition system with a proven zero data disclosure possibility. Applying a secure approach will allow the users of biometric-based systems to store their information in public clouds which will give them the opportunity of having more effective storage along with the computational power of the cloud infrastructures as well as the controlled and flexible data accessibility by multiple agents. A working prototype of the proposed system is implemented and evaluated using a real-world biometric database. The experimental results show that proposed system can be used in practice for trustworthy storage of sensitive records as well as precise identification of clients with proven zero data disclosure possibility.

References

1. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. Neural Computation 12(10), 2385–2404 (2000)
2. Boneh, D., Waters, B.: Conjunctive, subset, and range queries on encrypted data. In: Vadhan, S.P. (ed.) TCC 2007. LNCS, vol. 4392, pp. 535–554. Springer, Heidelberg (2007)
3. Bringer, J., Chabanne, H., Kindarji, B.: Error-tolerant searchable encryption. In: IEEE International Conference on Communications, ICC 2009, pp. 1–6 (2009)
4. Daugman, J.G.: High confidence visual recognition of persons by a test of statistical independence. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(11), 1148–1161 (1993)

5. Haghighat, M.B.A., Namjoo, E.: Evaluating the informativity of features in dimensionality reduction methods. In: 2011 5th International Conference on Application of Information and Communication Technologies, AICT, pp. 1–5 (October 2011)
6. Jain, A., Hong, L., Pankanti, S.: Biometric identification. Communications of the ACM 43(2), 90–98 (2000)
7. Jain, A., Ross, A., Pankanti, S.: Biometrics: a tool for information security. IEEE Transactions on Information Forensics and Security 1(2), 125–143 (2006)
8. Liu, C., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing 11, 467–476 (2002)
9. Meshgini, S., Aghagolzadeh, A., Seyedarabi, H.: Face recognition using gabor-based direct linear discriminant analysis and support vector machine. Computers & Electrical Engineering 39(3), 727–745 (2013)
10. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10), 1090–1104 (2000)
11. Shen, L.L., Bai, L., Fairhurst, M.: Gabor wavelets and general discriminant analysis for face identification and verification. Image and Vision Computing 25(5), 553–563 (2007)
12. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)