Deep Face Image Retrieval for Cancelable Biometric Authentication

Young Kyun Jang and Nam Ik Cho
Department of Electrical and Computer Engineering INMC, Seoul National University
1, Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea
kyun0914@ispl.snu.ac.kr, nicho@snu.ac.kr

Abstract

This paper presents a cancelable biometric system for face authentication by exploiting the convolutional neural network (CNN)-based face image retrieval system. For the cancelable biometrics, we must build a template that achieves good performance while maintaining some essential conditions. First, the same template should not be used in different applications. Second, if the compromise event occurs, original biometric data should not be retrieved from the template. Last, the template should be easily discarded and recreated. Hence, we propose a Deep Table-based Hashing (DTH) framework that encodes CNN-based features into a binary code by utilizing the index of the hashing table. We employ noise embedding and intra-normalization that distorts biometric data, which enhances the non-invertibility. For training, we propose a new segment-clustering loss and pairwise Hamming loss with two classification losses. The final authentication results are obtained by voting on the outcome of the retrieval system. Experiments conducted on two large scale face image datasets demonstrate that the proposed method works as a proper cancelable biometric system.

1. Introduction

As the use of social media is increasing, an enormous amount of images including faces are uploaded on the web every day. The basic technologies required in many applications and services that use faces are face detection, facial feature extraction, and face analysis. In particular, face image retrieval [16, 8, 17], which finds similar faces on the Internet or database to the query face image, has received great interest in the research community.

In recent years, image features from deep convolutional neural networks (CNNs) demonstrate superior performance to the conventional methods in visual recognition and they have been widely used in many computer vision tasks. Moreover, deeper network architectures such as VGG [15] have shown that more enriched features can be extracted as the number of channels is increased. However, although the CNN-based features can be considered generalized features for image retrieval, they are not practical for large database systems due to their high-dimensionality.

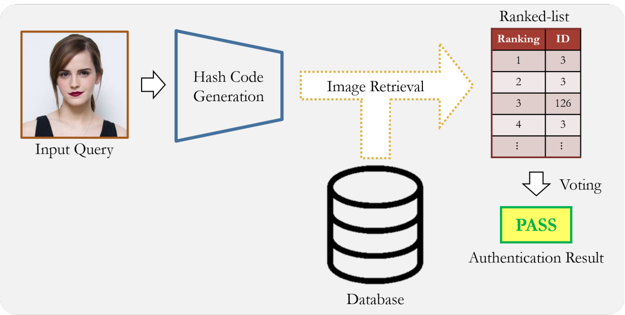

Figure 1. An illustration of the proposed cancelable biometric system. For a given authentication try, it generates a matched binary code to the input query and build a ranked-list by computing the Hamming distances from the database. Afterward, it applies voting to obtain a score for face authentication. If we are able to generate the proper hash code from an image and use it to authenticate, we can build a system that satisfies cancelable biometric conditions.

To resolve this problem, binary hashing has been widely applied as the approximate nearest neighbor (ANN) search. By calculating the Hamming distance between compact binary codes, images can be retrieved from a large database, reducing search time and memory usage. There were several CNN-based hashing methods that enabled efficient ANN search with Hamming ranking of binary codes [10, 21] and showed outstanding performance in image retrieval.

Likewise, in order to improve the computational efficiency of face image retrieval, binary code is generally selected as an image representative value. In particular for face authentication, binary code also has the advantage of removing most of the biometric data during the binary conversion process. Namely, we can take advantage of face image retrieval scheme to replace the biometric data with the binary code, which can prevent permanent biometric compromise, to build an efficient authentication system.

Cancelable biometrics [14, 13] represents the intentional and systematically repeatable distortion of biometric features to protect sensitive user-specific data. In this paper, we aim to construct a cancelable biometric system for face authentication, in which four principal conditions should be fulfilled:

(1) Unlinkability/Diversity: The same cancelable biometric template should not be used in two different applications/devices.

(2) Reusability: In the event of compromise, it should be revoked and reissued readily to generate a new template.

(3) Non-invertibility: The template should guarantee irreversibility to prevent the reconstruction of original biometric data.

(4) Performance: In the event of compromise, authentication performance between the old template and new one should not be severely different.

In order to satisfy the above conditions, BioHash [18, 19] approach has been proposed to generate a non-invertible binary hash code from biometric data. In recent years, CNN-based face template protection methods [12, 7] have been proposed for cancelable biometric authentication.

Inspired from these, we propose a cancelable biometric scheme with a novel *Deep Table-based Hashing* (DTH) framework: CNN-based hashing method that extracts features from a face image and embeds them into the binary code. To train the DTH, we introduced several loss functions that leads to cancelable biometrics. In the middle of the network, we perform biometric salting by embedding noise into the descriptor (aggregated global feature vector) and apply intra-normalization. At the last part of the DTH, we generate the binary code from the descriptor by utilizing the matched indices of the look-up-table. The final authentication is done through the voting in retrieval results. The contributions of our proposed method mainly include:

- We propose the first cancelable biometric scheme designed for face authentication utilizing face image retrieval system based on CNN, where the whole process is trained end-to-end. Two cross-entropy loss functions are used to obtain a discriminative descriptor. Additionally, a new segment-clustering loss and pairwise Hamming loss is introduced to concentrate similar facial features and to reduce the performance difference between templates.

- A table-based hashing method (DTH) is introduced to fulfill the conditions of the cancelable biometrics, especially for the unlinkability and reusability. In addition, if the template is compromised, we can easily reissue the new one without an inference of the CNN. Further, the proposed approach eliminates the need to store original face images to create new templates.

- To evaluate the performance of our proposed method, we build experimental environments based on large-scale face image retrieval that includes authentication trials. Extensive experiments show that the proposed method is superior to the other methods in face image retrieval and outperforms other cancelable biometric authentication methods.

2. Related Work

There have been several cancelable biometric face authentication methods without CNNs. The most famous approach is BioHash [18, 19], which mixes a set of user-specific random vectors (tokens) with biometric features. With the genuine tokens, BioHash method shows extremely low error rates as compared to the sole biometric approach. However, there are two major limitations associated with tokens: 1) user should have to keep the token for the authentication, 2) in case of the stolen token scenario, the false acceptance ratio is extremely increasing. Therefore, instead of using tokens for authentication, we chose an image retrieval based approach which was better in terms of stability and performance.

Deep learning-based face template protection methods have been proposed and demonstrated outstanding performance. In [12], they utilized CNNs to learn a mapping from face images to maximum entropy binary codes. In [7], they employed CNNs with one-shot and multi-shot enrollment to learn a robust mapping from face images to the unique binary codes. Both methods have a limitation that a cryptographic hash methods such as SHA-3 512 is needed to remove the biometric information remaining in the binary code. However, the cancelable biometric system that we propose has the advantage of not requiring an additional hash function to generate the discriminative binary code. It also has the advantage of being robust to the intra-variations of each class because the ranked-list from image retrieval is exploited for authentication.

In case of face image retrieval, there are several deep learning-based hashing methods that focus on searching face images. Deep Hashing based on Classification and Quantization erros (DHCQ) [16] jointly learns the feature representation and binary hash codes by optimizing an objective function for classification and quantization errors. Discriminative Deep Hashing (DDH) [8] improved [16] by incorporating the divide-and-encode module to generate compact binary codes. Discriminative Deep Quantization Hashing (DDQH) [17] introduces a batch normalization quantization (BNQ) module and obtains the state-of-

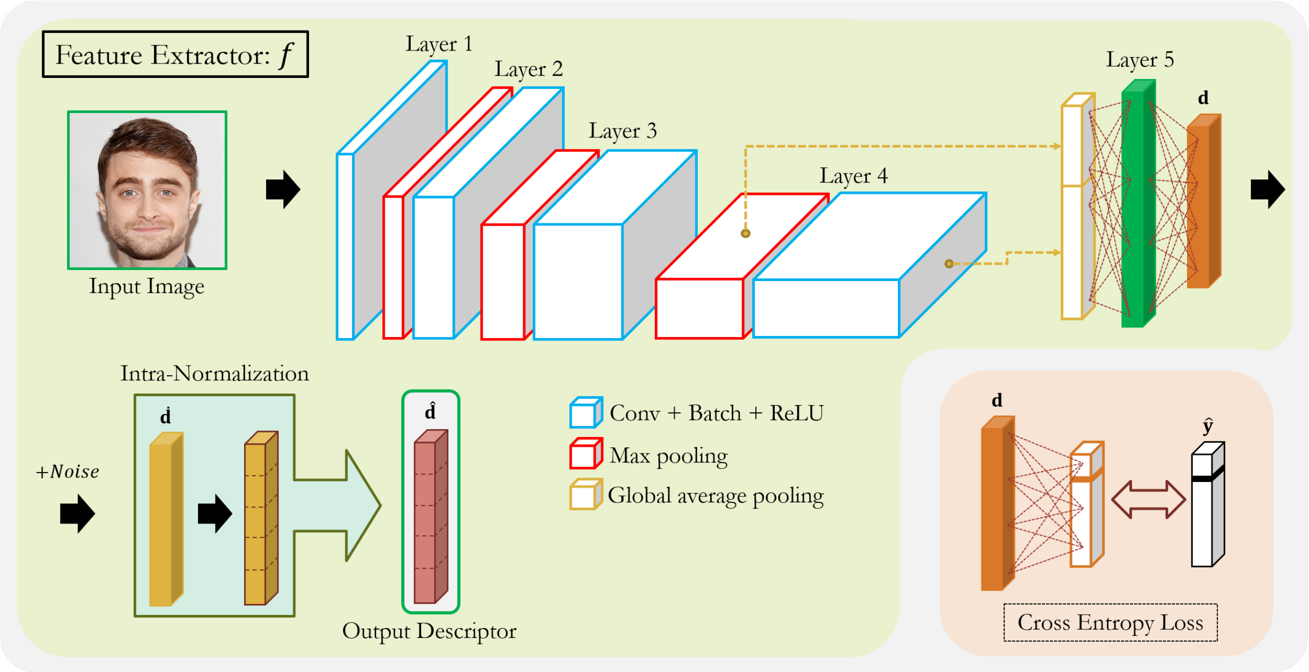

Figure 2. The illustration of the feature extractor in the proposed DTH framework. Input image is encoded into the descriptor $\hat{\mathbf{d}}$ while passing through the process of the noise embedding and the intra-normalization. Feature extractor is trained by the cross entropy loss with the one-hot encoded label $\hat{\mathbf{y}}$.

the-art performance on face image retrieval protocol. However, since it is difficult to satisfy the cancelable biometric conditions using the above methods, we propose a DTH framework for face authentication. This is not only suitable for cancelable biometrics, but also offers better performance in face image retrieval.

3. Deep Table-based Hashing

Like other similarity retrieval systems, goal of proposed DTH framework is to learn a mapping

$$h \circ f : \mathbf{I} \rightarrow \mathbf{b} \in \{-1, 1\}^K \tag{1}$$

that encodes an input image $\mathbf{I}$ into a K-bit binary hash code $\mathbf{b}$. To be specific, our proposed scheme consists of the CNN-based image feature extractor f and the table-based hash code generator h. For a given image label y, both feature extractor and hash code generator are trained to convey the similarity information of the same label and the discriminative characteristics between the different labels into the compact hash code. Specifically, we design (1) a classification loss for the image descriptor to learn the facial representations, (2) a noise embedding and intra-normalization to obtain the segment-wise normalized vectors, and (3) a segment-clustering loss and a trainable product quantization with pairwise Hamming loss to control the hashing quality. The proposed DTH framework can be trained through end-to-end pipeline by back-propagation.

3.1. Deep Feature Extraction

As illustrated in Fig. 2, the baseline of our proposed feature extractor is based on VGG [15]. We modify VGG-16 Network for the robust facial feature extraction by removing the last three convolutional layers that compute relatively small feature maps. Rather, we utilize ten convolutional layers with 3 × 3 kernels, which the number of channels doubles after the max pooling layer. The size of the feature maps is reduced after each max pooling layer by extracting the maximum value of the 2 × 2 kernel with stride 2. To accelerate the training speed, batch normalization [3] is employed between the each convolutional layer and the activation function. We replace the fully-connected layer with the global average pooling [9] from the bottom of the network to enhance the generalization and reduce the redundancy. To extract the robust and multi-scale facial representations, we concatenate the output of the global average pooling applied on the 3rd max pooling layer and the last convolutional layer. Finally, two fully-connected layers are adopted to obtain the face image descriptor $\mathbf{d}$. The detailed parameters are listed in Table 1.

To generate discriminative face image descriptors from a large number of classes, we train the feature extractor by a cross entropy loss:

Table 1. The detailed configuration of the feature extractor. 'Conv', 'Max' 'GAP', 'Concat' and 'FC' represent convolutional layer, max pooling layer, global average pooling, concatenation function and fully-connected layer, respectively. We employ batch normalization and ReLU [6] activation at the end of each convolutional layer.

Layer		Kernel Size	Output
input		-	$32 \times 32 \times 3$
Layer 1	$2\times$ Conv	3×3	$32 \times 32 \times 64$
	Max	2×2	$16 \times 16 \times 64$
Layer 2	$2\times$ Conv	3×3	$16 \times 16 \times 128$
	Max	2×2	$8 \times 8 \times 128$
Layer 3	$3\times$ Conv	3×3	$8 \times 8 \times 256$
	Max	2×2	$4 \times 4 \times 256$
Layer 4	$3\times$ Conv	3×3	$4 \times 4 \times 512$
	Max	2×2	$4 \times 4 \times 512$
Layer 5	GAP	-	$1 \times 1 \times 256$
	GAP	-	$1 \times 1 \times 512$
	Concat	-	$1 \times 1 \times 768$
	FC	-	$1 \times 1 \times 1024$
	FC	-	$1 \times 1 \times L$

$$\mathcal{L}_{cls:\mathbf{D}} = -\sum_{i=1}^{N} log \frac{\exp{(W_{D_{y_i}}^{T} \mathbf{d}_i)}}{\sum_{j=1}^{L} \exp{(W_{D_j}^{T} \mathbf{d}_i)}} \tag{2}$$

where $\mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_N]$ denotes the set of descriptors generated from N-number of training images and y_i denotes the i-th label. $W_{D_{y_i}}$ denotes the weight for the linear prediction of the L-class label by using the descriptor $\mathbf{d}_i$. Since the cross entropy loss has shown outstanding performances in learning-based classification tasks, the descriptor generated from the trained feature extractor can contain facial characteristics effectively.

For the cancelable biometrics, any personal information should be encrypted. The concept of biometric salting agrees with encryption in that independent input factors are blended with biometric data to derive a distorted version of the biometric template. In our case, we embed Gaussian random noise $n_i \sim \mathcal{N}(0, \sigma^2)$ into the descriptor for biometric salting as

$$\dot{\mathbf{D}} = [\dot{\mathbf{d}}_i \mid \dot{\mathbf{d}}_i = \mathbf{d}_i + n_i,\ i = 1, \dots, N]. \tag{3}$$

which enhances non-invertibility. Another advantage of noise embedding is related to the observation in [10]. Since excessively fitting the relaxed descriptor to a strict binary code results in the unsatisfactory performance, we embed noise to allow some room around the descriptor, which extends the selection range in the Hamming space.

Despite these advantages, noise can cause bursts in the descriptor that degrades the classification power. To handle this, we divide the descriptor $\dot{\mathbf{d}}_i$ into C segments ($\dot{\mathbf{d}}_i = [\dot{d}_{i,1}, \dots, \dot{d}_{i,C}]$) and apply the intra-normalization [1] to suppress the burst *i.e.*,

$$\hat{d}_{i,j} \leftarrow \dot{d}_{i,j} / ||\dot{d}_{i,j}||_2^2 \tag{4}$$

where $\hat{d}_{i,j}$ represents a normalized segment. In this way, the magnitude of the burst elements in the segment is reduced and the contribution of each segment is balanced. Besides, in terms of vector geometry, the similarity between the segments depends on the angle rather than the magnitude (e.g. cosine similarity), which is different from the CNN-based descriptor without intra-normalization. Finally, we can obtain the discriminative descriptor $\hat{\mathbf{d}}_i = [\hat{d}_{i,1}, \dots, \hat{d}_{i,C}]$.

3.2. Hash Code Generation

To build a cancelable face authentication system using image retrieval, we introduce a table-based hash code generator h as shown in Fig. 3. We utilize the index of the hash table element to embed image into the binary code. First of all, we configure a look-up-table which we call hashing table $\mathbf{T}$ consisting of C-codebooks with R-codewords

$$\mathbf{T} = \begin{bmatrix} t_1^1 & t_2^1 & \dots & t_C^1 \\ t_1^2 & t_2^2 & \dots & t_C^2 \\ \vdots & \vdots & \ddots & \vdots \\ t_1^R & t_2^R & \dots & t_C^R \end{bmatrix}. \tag{5}$$

Each column in the hashing table represents a codebook, and each element (a row in a hashing table) of the codebook represents a codeword. To aggregate the features of the image descriptors in each hashing table codeword, we introduce a segment-clustering loss:

$$\mathcal{L}_{cluster} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{C} \min_{k} ||\hat{d}_{i,j} - t_j^k||_2^2 \tag{6}$$

that minimizes the minimum value of the $l2$-distance between the segment of the descriptor $\hat{\mathbf{d}}_i$ and the codeword in the codebook, in order. This enables each codeword to collect general characteristics from descriptors. Second, inspired by [4, 21], we introduce a trainable product quantization on the descriptor $\hat{\mathbf{d}}_i$ that quantizes each segement separately, *i.e.*,

$$\mathbf{q}_i = [q_1(\hat{d}_{i,1}), \dots, q_j(\hat{d}_{i,j}), \dots, q_C(\hat{d}_{i,C})] \tag{7}$$

where $q_j(\cdot)$ is a quantizer for $\hat{d}_{i,j}$ defined as

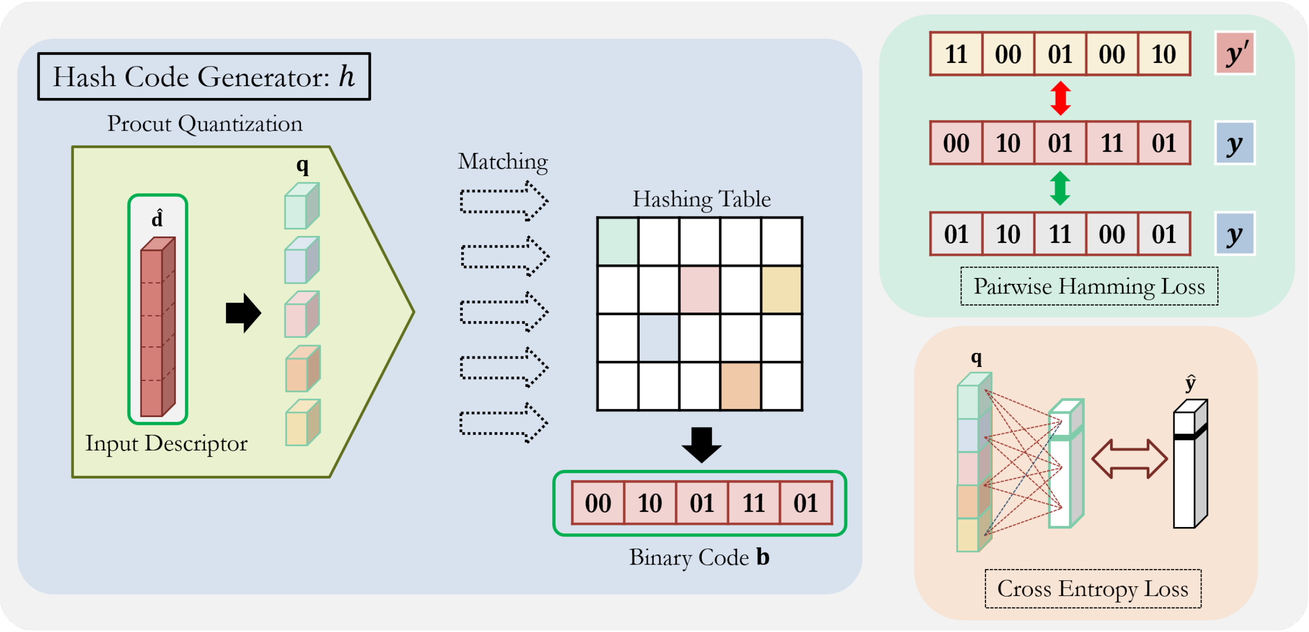

Figure 3. The illustration of the hash code generator in DTH framework. First, input descriptor turns into the quantized descriptor $\mathbf{q}$. Second, the closest hashing table codeword is found by calculating the distance. Last, each index of the hashing table codeword is converted to binary. Hash code generator is trained by the pairwise Hamming loss of binary codes with corresponding labels (y, y') and the cross entropy loss with one-hot encoded label $\hat{\mathbf{y}}$.

$$q_j(\hat{d}_{i,j}) = \sum_{a=1}^{R} \frac{\exp\left(-\alpha_j^a ||\hat{d}_{i,j} - t_j^a||_2^2\right)}{\sum_{b=1}^{R} \exp\left(-\alpha_j^b ||\hat{d}_{i,j} - t_j^b||_2^2\right)} t_j^a \tag{8}$$

that contains trainable parameter $\mathbf{a}_j = [\alpha_j^1, \ldots, \alpha_j^R]$ which belongs to each codebook. Applying this method makes it possible to share gradients generated from matched codeword with other codewords during training stage. In addition, by learning the $\mathbf{a}$ properly, the selection of codewords can be evenly distributed.

Unlike other product quantization methods that compute Euclidean distance between two descriptors, we try to exploit the Hamming distance as a similarity measure to fulfill the cancelable biometric principals with high accuracy. Therefore, we introduce two additional loss functions. First, to make codebooks discriminative among each other, we calculate another cross entropy loss:

$$\mathcal{L}_{cls:\mathbf{Q}} = -\sum_{i=1}^{N} log \frac{\exp\left(W_{Q_{y_i}}^{T} \mathbf{q}_i\right)}{\sum_{j=1}^{L} \exp\left(W_{Q_j}^{T} \mathbf{q}_i\right)} \tag{9}$$

where $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_N]$ and $W_{Q_{y_i}}$ denotes the weight for the linear prediction of the L-class label by using the quantized descriptor $\mathbf{q}_i$. In this way, combination of the quantized segments can represent identifiable information for each class. Furthermore, we can obtain the discriminative binary code $\mathbf{b}_i = [b_{i,1}, \ldots, b_{i,C}]$ by computing

$$b_{i,j} = \mathrm{Bin}(\underset{k}{\mathrm{argmin}}(||q_j(\hat{d}_{i,j}) - t_j^k||_2^2)) \tag{10}$$

for every codebooks, where $\mathrm{Bin}(\cdot)$ is a function that converts decimal to binary. This means that each segment of the descriptor matches with the nearest codeword in the codebook and the index of codeword transformed to the binary code. Second, to make the obtained binary code be independent and evenly distributed, we compute pairwise Hamming loss for every training batch $\mathbf{N} = [N_1, \ldots, N_M]$ as:

$$\mathcal{L}_{pair} = \sum_{m=1}^{M} \sum_{i,j}^{N_m} \left[\Omega_{i,j} \times s_{i,j} + (1 - \Omega_{i,j}) \times (1 - s_{i,j})\right] \tag{11}$$

where $s_{i,j}$ denotes a pairwise similarity relationship which is $s_{i,j} = 1$ when the input image pair $\mathbf{I}_i$ and $\mathbf{I}_j$ have an identical label and $s_{i,j} = 0$ when they are not. $\Omega_{i,j}$ denotes the relaxed Hamming distance computed by inner product between two binary codes.

$$\Omega_{i,j} = \frac{1}{2} \tanh\left(b_i \cdot b_j\right) \tag{12}$$

where $\mathbf{b}_i$ and $\mathbf{b}_j$ represents binary codes generated from $\mathbf{I}_i$, $\mathbf{I}_j$, respectively. In addition, we employ $l2$-regularization $\mathcal{L}_{reg}$ [2] with a non-negative hyper-parameter β to avoid overfitting. From Eqn. 1 to 12, the loss function of our network is summarized as

$$\begin{aligned}\mathcal{L} = \quad & \mathcal{L}_{cls:\mathbf{D}} + \mathcal{L}_{cluster} + \mathcal{L}_{cls:\mathbf{Q}} \\ & + \mathcal{L}_{pair} + \mathcal{L}_{reg}\end{aligned} \tag{13}$$

4. Experiments

4.1. Datasets

To validate the face image retrieval based cancelable biometric scheme, we adopt two well-organized large scale face image datasets for both model training and evaluation. The first dataset is *YouTube Faces* ($Y.T.F$) [20] which contains average 181.3 frames of 3,425 videos involved with 1,595 people. We take 63,800 images with randomly selected 40 face images per person as the training set and 7,975 images with 5 images per person as the test set. The second is *FaceScrub* ($F.S$) [11] which contains 106,863 face images of 530 people retrieved from the Internet. We select 63,000 images with about 120 face images per person for the training and 2,650 images with 5 images per person for the test. All face images are resized to 32 × 32 × 3.

4.2. Evaluation Metric

Like other cancelable biometric systems [13], the performance of our proposed system is expressed in terms of the equal error rate (EER), which means that the false accept rate (FAR) and the false reject rate (FRR) are same. We measured the FAR and FRR by varying the threshold of the authentication score from top-A retrieval results. To compare the EER performance of CNN-based face image retrieval methods, we create two experimental environments. *Environment 1*: for the model trained by $Y.T.F$, we enroll 63,800 training images form $Y.T.F$ into the database. For the authentication, we randomly select 2,500 for the positive samples and 500 images for the negative from the test set of $Y.T.F$ and $F.S$, respectively. *Environment 2*: similar to the first environment, for the model trained by $F.S$, 63,000 training images from $F.S$ are enrolled and 2,500 positive samples from $F.S$ and 500 negative samples from $Y.T$ are used for the test.

4.3. Ablation Study and Training Details

At first, we performed experiments to find the best hyper-parameters by evaluating the classification performance of the quantized descriptor **q**. We fix β as 0.0002 for the $l2$-regularization. We vary the number of codeword R for $\{2^2, 2^4, 2^6, 2^8\}$ and adjust C to fit the length of the binary code. The length of each codeword is fixed as 8.

From the experimental results, we could observe that the best performance is obtained when $R = 2^4$. Also, we could find that the classification performance is best when the initial value of α is 40. In experiments to investigate the effect of noise embedding (biometric salting), prominent degradation of performance was observed at noise levels (σ) above 0.6. As a result, we set the hyper-parameters as $\{R, \alpha, \sigma\} = \{2^4, 40, 0.6\}$.

Table 2. Face authentication result (EER, %) according to the noise level (σ) and the number of returned images (A).

Env.	σ			
	0.0	0.3	0.6	0.9
1	.0092	.0083	.0384	5.984
2	.0806	.0774	.1881	9.051

Env.	A			
	5	10	20	40
1	.0102	.0094	.0112	.0120
2	.0932	.0827	.0950	.1086

Table 3. Face authentication result (GAR@0.1%FAR, %) according to the noise level (σ) and the number of returned images (A).

Env.	σ			
	0.0	0.3	0.6	0.9
1	96.04	96.71	91.22	53.79
2	83.45	82.91	65.86	11.67

Env.	A			
	5	10	20	40
1	93.88	97.04	96.34	94.42
2	80.17	82.62	81.03	78.96

In the training procedure, we adopt Adam algorithm [5] to optimize our deep neural network from the loss functions. Especially for the pairwise Hamming loss term, we need to be careful in organizing the training batch, since we conduct experiments with a large number of classes (more than 500). Therefore, we selected 5 images from the 100 randomly extracted classes to build a batch.

4.4. Using the DTH as Cancelable Biometrics

In this section, we describes how to use the face image retrieval results for face authentication. Like other face image retrieval methods [16, 8, 17], we are able to build a ranked-list by calculating the Hamming distance between binary codes generated from the query image and database. From this, we can calculate the authentication score by computing weighted voting on top-A ranked list: A points to the highest one and 1 point to the lowest one in descending order. We set the threshold for authentication score as half of A.

In terms of reusability, we randomly shuffle the position of codeword in each codebook to reissue the template as

$$\dot{\mathbf{T}} = [\dot{t}^i_j \mid \dot{t}^i_j = t^k_j,\ k = \mathrm{Rand}(M)] \tag{14}$$

Table 4. The mean Average Precision of different hashing approaches on both face images datasets and the EER on both experimental environments.

Method	Face Image Retrieval (mAP)						Face Authentication (EER, %)					
	$YoutubeFaces$			$FaceScrub$			$Environment$ 1			$Environment$ 2		
	24-bit	48-bit	72-bit	24-bit	48-bit	72-bit	24-bit	48-bit	72-bit	24-bit	48-bit	72-bit
DHCQ	.7004	.7667	.8421	.4522	.5579	.6393	.1567	.1345	.0884	.2852	.2046	.1079
DDH	.8655	.9412	.9620	.6074	.6585	.6885	.0854	.0141	.0089	.1844	.1560	.1052
DDQH	.8754	.9524	.9725	.6345	.6594	.6987	.0873	.0139	.0078	.1115	.1098	.1009
DTH	.9880	.9928	.9934	.7416	.7985	.8242	.0112	.0094	.0048	.0961	.0827	.0653
BioHash	-	-	-	-	-	-	13.50	10.24	7.624	13.67	11.16	8.135
Pandey's [12]	-	-	-	-	-	-	.1854	.0542	.0116	.3004	.1442	.1112
Jindal's [7]	-	-	-	-	-	-	.1632	.0408	.0096	.2851	.1473	.1057

where $\mathrm{Rand}(\cdot)$ is a function that generates non-overlapping random non-negative integers less than M, and $\dot{\mathbf{T}}$ is a shuffled hash table which has different positions of codewords than T.

To minimize the performance difference between the reissued template and original one, we shuffle the hash table for every 10th epoch in the training process. In addition, trainable parameter **a** which is influenced by the pairwise Hamming loss, is also shuffled along to the hash table. Furthermore, the Hamming distances between binary codes in a batch are revealed directly to the production of quantized descriptor **q** through **a** as shown in Eqn. 8. Therefore, **a** guides to have a uniform hamming distance regardless of the shuffle, reducing the performance difference originated from the template reissue. Moreover, since quantized descriptor **q** is also influenced by the classification error, we can keep the codebook-wise discriminativity for the authentication. Furthermore, with the proposed template generation method, we can recreate a template by using only a set of stored descriptors which has the advantage that there is no need to infer the heavy deep learning model.

The performance related to the cancelable biometrics of our proposed DTH method is evaluated by the mean of EER and the mean Genuine Accept Rate (GAR) at different FAR for $R = 2^4, C = 12$, i.e. for code length 48 (48-bit binary codes). The mean of EER and GAR were computed by making 100 new templates with shuffle. We conducted experiments on our experimental environments in 4.2 by varying the number of returned images (A) and the noise level (σ) as shown in Table 2 and 3. In addition, we calculated the standard deviation of EER for each case in order to investigate the performance change according to the reissue, and it is confirmed that there is almost no difference in performance since the standard deviation is about 0.3% on both experimental environments. To sum up, we can observe that our proposed method fulfills the principal conditions of cancelable biometrics (unlinkability, non-invertibility, reusability and performance), by introducing the face image retrieval based face authentication approach.

4.5. Results and Security Analysis

Since our cancelable biometric scheme is based on the results from face image retrieval, we conducted experiments to investigate the retrieval performance in terms of mAP (mean Average Precision) at top 50 images on both face image datasets. We perform the experiments by varying the code length for $\{24, 48, 72\}$ to observe the results according to the length of the binary codes. Experiments are also conducted on DHCQ [16], DDH [8] and DDQH [17]. For a fair comparison, the feature extractor of DTH is applied to other hashing methods that use relatively shallow network architectures.

Additionally, experiments were conducted on our DTH method and other face template protection methods depending on the experimental environments that we have set up for face authentication. We set the number of returned images (A) as 10. To make a fair comparison, we use the same feature extractor to the other methods (BioHash [18, 19], Pandey's [12] and Jindal's [7]) and replace the scoring method with retrieval based scheme as we proposed. For BioHash, we calculate the inner product between descriptor and user-specific random basis. For Pandey's and Jindal's, we train the network from the scratch and employ SHA-3 512 algorithm to encrypt assigned binary codes.

From the results, we can observe that the DTH outperforms other face image retrieval hashing methods and has the lowest EER in all code lengths. Furthermore, since our template is based on the index of the codeword matched to the hashing table, attackers cannot extract the original biometric from embedded binary codes because it does not contain any user information. The original descriptor cannot be restored even if the hashing table is compromised.

For the security analysis, assume a situation where an attack on authentication. Without CNN parameters, the search space for brute force attacks would be 2 to the power of code length, making the brute force attacks computationally infeasible ($2^{24}, 2^{48}, 2^{72}$). Furthermore, since our proposed method is based on the retrieval results, the attacker should have to find perfectly matched binary code to exceed the threshold.

In the scenario with CNN parameters, the attacker would attempt to generate the template using the face images. However, even for the same face image, the same descriptor cannot be generated because a random noise is inserted in the process of generating the descriptor. Even if the similar image descriptors are generated, it is hardly possible to obtain the same template because the possibility of building the same hash table is one over $R \times C$.

5. Conclusion

We have proposed a CNN-based face image retrieval scheme for the face authentication which can especially work as cancelable biometrics. To get the discriminative descriptor, we trained a feature extractor by introducing a loss term that shows high classification performance. Also, we embedded noise into the descriptor and applied normalization to guarantee the non-invertibility. Furthermore, we have proposed a new hashing table-based binary encoding method with a segment-clustering loss to learn the table and a pairwise Hamming loss to satisfy the unlinkability and reusability while maintaining the authentication performance. Experimental results show that the proposed method fulfills principal cancelable biometric conditions and achieves the lowest equal error rate on face image datasets with a large number of classes.

6. Acknowledgments

This research was supported by Samsung Electronics Co., Ltd.

References

[1] R. Arandjelovic and A. Zisserman. All about vlad. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1578–1585, 2013.

[2] S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.

[3] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shif. *In ICML*, 2015.

[4] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2011.

[5] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *In ICLR*, 2015.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[7] A. Kumar Jindal, S. Chalamala, and S. Kumar Jami. Face template protection using deep convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 462–470, 2018.

[8] J. Lin, Z. Li, and J. Tang. Discriminative deep hashing for scalable face image retrieval. In *Proceedings of International Joint Conference on Artificial Intelligence*, 2017.

[9] M. Lin, Q. Chen, and S. Yan. Network in network. *In ICLR*, 2014.

[10] H. Liu, R. Wang, S. Shan, and X. Chen. Deep supervised hashing for fast image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2064–2072, 2016.

[11] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 343–347. IEEE, 2014.

[12] R. K. Pandey, Y. Zhou, B. U. Kota, and V. Govindaraju. Deep secure encoding for face template protection. In *CVPR Workshops*, pages 77–83, 2016.

[13] V. M. Patel, N. K. Ratha, and R. Chellappa. Cancelable biometrics: A review. *IEEE Signal Processing Magazine*, 32(5):54–65, 2015.

[14] N. K. Ratha, J. H. Connell, and R. M. Bolle. Enhancing security and privacy in biometrics-based authentication systems. *IBM systems Journal*, 40(3):614–634, 2001.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *In ICLR*, 2015.

[16] J. Tang, Z. Li, and X. Zhu. Supervised deep hashing for scalable face image retrieval. *Pattern Recognition*, 75:25–32, 2018.

[17] J. Tang, J. Lin, Z. Li, and J. Yang. Discriminative deep quantization hashing for face image retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

[18] A. B. Teoh, A. Goh, and D. C. Ngo. Random multispace quantization as an analytic mechanism for biohashing of biometric and random identity inputs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1892–1901, 2006.

[19] A. B. Teoh, Y. W. Kuan, and S. Lee. Cancellable biometrics and annotations on biohash. *Pattern Recognition*, 41(6):2034–2044, 2008.

[20] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 529–534. IEEE, 2011.

[21] T. Yu, J. Yuan, C. Fang, and H. Jin. Product quantization network for fast image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 186–201, 2018.