IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 4, 2018, accepted October 1, 2018, date of publication October 5, 2018, date of current version October 29, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2874056

Convolutional Neural Network-Based Periocular Recognition in Surveillance Environments

MIN CHEOL KIM, JA HYUNG KOO, SE WOON CHO, NA RAE BAEK, AND KANG RYOUNG PARK
Division of Electronics and Electrical Engineering, Dongguk University, Seoul 100-715, South Korea

Corresponding author: Kang Ryoung Park (parkgr@dongguk.edu)

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041921), by the Bio & Medical Technology Development Program of the NRF funded by the Korean government, MSIT (NRF-2016M3A9E1915855), and by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1B03028417).

ABSTRACT Visible light surveillance cameras are currently deployed on a large scale to prevent crime and accidents in public urban environments. For this reason, various human identification studies using biometric data are underway in surveillance environments. The most active research area is face recognition, which generally shows excellent performance; however, aging, changes in facial expression, and occlusions by accessories cause a rapid decline in recognition performance. To resolve these problems, we propose a periocular recognition method in surveillance environments that is based on the convolutional neural network. In this paper, experiments were performed using the custom-made Dongguk periocular database and the open database of ChokePoint database. It was confirmed that the proposed method performs better than existing techniques used in periocular recognition. It was also found to perform better than conventional techniques in face recognition when an occlusion is present.

INDEX TERMS Visible light surveillance camera sensor, biometrics, periocular recognition, CNN.

I. INTRODUCTION

The periocular region is the area around eyes, and it includes the eyelids, eyebrows, etc. This region is useful for biometrics because it contains relatively stable patterns [1], and it has a stronger tolerance against changes in facial expression and aging [2]. Periocular recognition is a new biometric approach which is attracting a great deal of interest, and a lot of research is being performed to increase the accuracy of automated algorithms [2]. It is also receiving attention as a means of improving the performance of biometric techniques such as the combined method with face or iris recognition [3]. Also, the periocular region can provide better recognition accuracy than face recognition for images with a lot of deterioration (harsh illumination, occlusion, low resolution, etc.) [29]. However, the databases used in existing periocular recognition researches, such as the University of Beira interior periocular recognition (UBIPr) database [11], face recognition grand challenge (FRGC) database [42], and face and ocular challenges series (FOCS) database [43] contain high-resolution images which were not captured in the surveillance environment. Therefore, we examine periocular recognition using a convolutional neural network (CNN) in an unconstrained and surveillance environment. Our research is novel in the following five ways compared to previous works.

- This is the first study for periocular recognition using a CNN in a surveillance environment.
- We attempt to improve performance by applying a preprocessing technique that uses a focus assessment method to exclude images with severe blur when performing recognition. This helps to prevent the recognition errors in surveillance environments due to image blur caused by user movement.
- We compare the recognition accuracies based on feature values and feature normalization methods in the various layers of a CNN in order to find the feature values in most suitable layer and feature normalization. In addition, face recognition performance is compared even in circumstances where part of the face is occluded, which shows the usefulness of periocular recognition in a surveillance environment.
- Our method improves recognition performance by score fusion on loose region of interest (ROI) and tight ROI-based periocular recognition.
- In order to perform experiments with databases that are suitable for surveillance environments, this study uses the custom-made Dongguk periocular database (DP-DB1) and the ChokePoint database [4], [32] by the National ICT Australia Ltd. (NICTA). The DP-DB1 and

the trained periocular recognition CNN models with algorithms are shared with other researchers in [46] to enable fair performance evaluation.

II. RELATED WORKS

We have examined existing periocular recognition studies and compared their strengths and weaknesses. Existing studies have mainly used handcrafted features, and combined them with iris recognition [25], [27], [30], or face recognition [26], to improve recognition performance. Aside from these multimodality-based recognition, there have been previous researches using the single modality of periocular recognition. Lowe [5] used several methods of extracting handcrafted features, including scale invariant feature transform (SIFT), histograms of oriented gradients (HOG) [6], and local binary patterns (LBP) [7], in order to measure performance in the presence of various factors which degrade performance such as occlusions, different poses, and different sessions. They also confirmed that performing recognition via the periocular region is possible [8], [9]. Miller *et al.* [17] presented a method which uses LBP to extract the texture features of periocular skin and perform identification. The study [18] compared periocular recognition with face recognition in terms of blur, resolution change, illumination, and different color channels. The results showed that when the image quality is extremely bad due to blur and low resolution, the performance by face recognition decreased more than that by periocular recognition. Woodard *et al.* [19] combined texture and color information, and Adams *et al.* [10] used genetic and evolutionary feature extraction (GEFE) to optimize LBP features. Bharadwaj and Torralba [31] used circular LBP (CLBP) and gist descriptors as feature extraction methods, and fused the recognition scores of both eyes [13]. The LBP-based methods extract features from the whole area of ROI, so alignment is important. Most methods perform alignment via the iris center, and their recognition performance decreases with the images of unconstrained environment including various changes in pose and gaze. To resolve the problem of misalignment in an unconstrained environment, Padole and Proenca [11] performed experiments on factors which degrade performance in periocular recognition such as scale, pose, and occlusion. They proposed a method which does not set the iris center as a ROI but uses the eye corners to set the ROI for more accurate alignment in situations with different gazes and poses. Rather than using conventional LBP, Juefei-Xu and Savvides [23] used Walsh-Hadamard transform encoded LBP (WHT-LBP) as a feature extraction method. This method uses Walsh masks as a convolution filter to perform a Walsh-Hadamard transform, and then it applies LBP. The recognition accuracy is dramatically increased compared to conventional LBP [23]. Mahalingam and Ricanek [15] proposed a hierarchical three-patch LBP (H-3P-LBP) method. It has a hierarchical structure, so it can extract micro and macro textures more effectively than conventional LBP. Alonso-Fernandez and [24] used Gabor filters. Smereka and Kumar [29] performed experiments to find the periocular region that showed the best performance out of several regions. They divided the periocular region into four types, from the wide region (region 1) including the eyebrows and the skin to the tight region (region 4) where the iris is more magnified. They used probabilistic deformation models (PDMs) and modified SIFT (m-SIFT) on visible light (VL) and near-infrared (NIR) images. The experimental results showed that the recognition accuracies using regions 1 and 2 were high in VL whereas regions 2 and 3 showed better accuracies in NIR. Karahan *et al.* [28] measured the performance of various feature extractors based on the LBP combined with speeded up robust feature (SURF), as well as methods of SIFT combined with SURF. The method proposed by Bakshi *et al.* [16] uses a phase intensive global pattern (PIGP) to extract features. The method uses a 3×3 filter bank to perform convolution. Methods which use these kinds of handcrafted features have the advantage of being able to extract features without any separate training process. However, they have a disadvantage in that there are limits to improving the performance because it is difficult to extract the optimal features.

To resolve this problem, research has been performed on using CNN-based deep features. Zhao *et al.* proposed a semantics-assisted CNN (SCNN) method which not only learns identities but also uses semantic information such as left and right eye regions, gender etc. They combined the identities with semantic information to verify the test performance against completely different databases which were not used for training (open world setting) [2]. Proença *et al.* [49] proposed the periocular recognition without the regions of iris and sclera based on deep learning method. In another study that used a CNN, Luz *et al.* [34] proposed a periocular region recognition (PRR) architecture, which has an advantage in that it can modify the size of feature vector of output by replacing the last fully-connect layer in a visual geometry group (VGG) face-16 model with a new fully-connected layer; however, it was not tested in a surveillance environment consisting of low-resolution images.

Most of these existing studies were performed using databases such as FRGC, FOCS, UBIPr, university of Beira interior iris version 2 (UBIRIS v2), face recognition technology (FERET), and Chinese academy of sciences institute of automation (CASIA). These databases are organized according to factors which degrade recognition performance such as lighting changes, poses changes, and expression changes, etc; however, multiple factors do not occur on these databases at the same time. In a real surveillance environment, the factors degrading performance do not appear, one at a time but in complex ways (lighting and pose changes with image blurring), so these databases cannot be regarded as being collected in real surveillance environments. In addition, the camera is usually installed at the much higher position than user's height in surveillance environment, and it captures the moving people in the slanted direction at a distance. Therefore, the distorted eye images are frequently acquired,

TABLE 1. Comparisons of proposed and previous researches on periocular recognition.

Category		Method	Database	Strength	Weakness
Handcrafted features	Global features [46]	LBP with City-block distance [17, 18, 25]	FRGC, FERET, multiple biometric grand challenge (MBGC)	Can be used without training and features can be easily extracted from the whole area of ROI	Recognition accuracy in a surveillance environment is not verified
		Binarized statistical image features (BSIF) with sparse representation classification (SRC) [27]	Self-collected database		
		Gabor with Hamming distance [24]	CASIA-Iris v3, biometrics security (BioSec)		
		GIST + CLBP with x^2 distance [13]	UBIRIS v2		
		H-3P-LBP with Euclidean distance [15]	FRGC, MORPH album1, Georgia Tech face, WVU/ND twins		
		PIGP with Euclidean distance [16]	UBIRIS v2		
		LBP + GEFE with Manhattan distance [10]	FRGC, FERET		
		WHT-LBP with Cosine distance [23]	Compass		
		Red-green (RG) color histogram and LBP with Bhattacharya and City-block distances [19]	FRGC, MBGC		
	Local features [47]	SIFT + SURF with Brute-force (BF) matcher [28]	FERET	Can be used without training and features can be easily extracted by local key-points	
		m-SIFT and PDM with Euclidean distance [29]	FOCS, UBIPr		
	Global + Local features	LBP + HOG + SIFT with Euclidean distance and distance-ratio based scheme [8, 9]	FRGC	Combines features from the whole area of ROI and local key-points so that comprehensive features can be extracted	
		LBP + SIFT with Euclidean distance and distance-ratio based scheme [26, 30]	Plastic surgery database, FRGC, UBIRIS v2		
		LBP + HOG + SIFT with multilayer perceptron (MLP) [11]	UBIPr		
Deep features		SCNNs with Joint Bayesian scheme [2]	UBIPr, UBIRIS v2, FRGC, FOCS, CASIA v4-distance	Recognition is possible for completely different databases	- When semantic information is added, an additional CNN for matching should be designed - Recognition accuracy in a surveillance environment is not verified
		D-PRWIS [49]	UBIRIS v2, FRGC	High recognition accuracy by data augmentation	Experiments were performed only in closed world setting
		PRR CNN [34]	UBIRIS v2, mobile biometry (Mobbio)	The size of feature vector from final fully-connected layer can be adjusted	Recognition accuracy in a surveillance environment is not verified
		Proposed method	Self-collected database (DP-DB1), ChokePoint	Good recognition accuracy in an unconstrained environment	Requires adequate data for CNN training

and its resolution is very low. However, these databases used in [2], [34], and [49] do not include these kinds of images, and most images are captured by the camera in front of people (its height of installation is similar to that of people), and its resolution is fairly high. In addition, the research [49] performed the experiments only with the images where the classes in training data are same to those in testing data (closed world setting), which is different from the real scenarios of surveillance camera environments. To resolve these problems in the existing studies, this study proposes a CNN-based recognition method which uses low-resolution periocular images captured in a surveillance environment. Different from the researches [2], [34], [49], we used the custom-made Dongguk periocular database (DP-DB1) and the open database of ChokePoint database which were collected in real surveillance environments. Therefore, the various factors occur in complex way such as lighting and pose changes with severely (motion and optical) blurring in our experimental databases, and the image resolution is much lower than the databases used in [2], [34], and [49]. So, more challenging cases are dealt with in our research, and we performed the experiments based on open world setting considering the real scenarios of surveillance camera environments. Table 1 shows a comparison of the methods proposed in existing studies and this study.

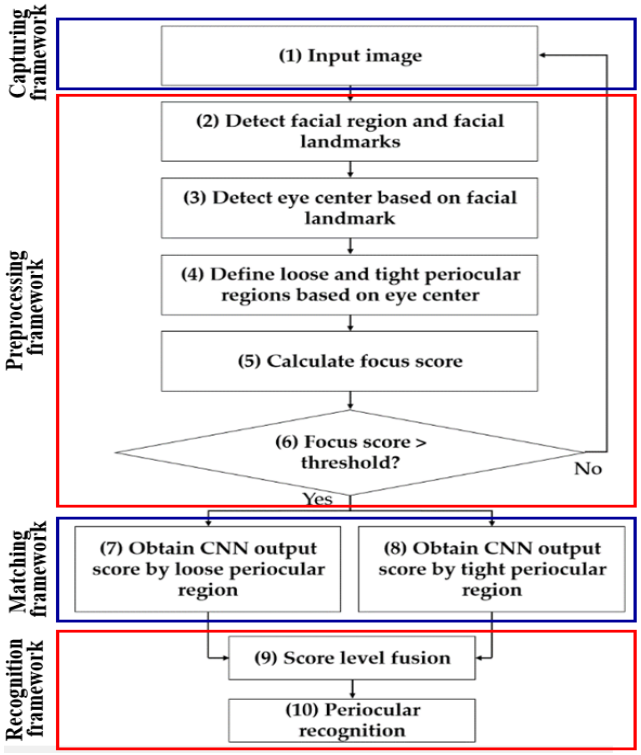

FIGURE 1. Overall framework representation of the proposed method.

III. PROPOSED METHOD

A. OVERALL PROCEDURE OF PROPOSED METHOD

Figure. 1 shows the overall framework representation for the proposed method. First an adaptive boosting (adaboost) algorithm is used on the input images to detect the face regions, and the dlib facial landmarks tracking method [33] is used on the faces to detect facial landmarks (step (2) of Fig. 1). Based on the 6 landmarks of the left and right eyes, the corners of each eye are used to calculate the center coordinates for the left eye and the right eye (step (3) of Fig. 1). The loose periocular ROI is set according to the ratio that was set based on the center coordinates whereas the tight periocular ROI is set based on the eye's center in the loose periocular ROI (step (4) of Fig. 1). Then, the 5×5 convolution kernel proposed in [12] is used to measure the focus score. The loose periocular ROI often contain background or hair which makes the recognition score not accurate, so this is also calculated based on the tight periocular ROI (step (5) of Fig. 1). If the calculated focus score is larger than the threshold, the CNN is used to extract features and then the Euclidean distances between the input and the gallery images (the enrolled images) are calculated (steps (6) $\sim$ (8) of Fig. 1). The two calculated distances based on loose and tight periocular ROIs are combined by score level fusion (step (9) of Fig. 1). The fused final score is used to determine whether the input image is the same class with the gallery one or not (step (10) of Fig. 1).

B. DETECTION OF FACE AND PERIOCULAR REGION

Referring to the results in [29], we defined the loose ROI and the tight ROI as shown in Figs. 2 (e) and (f), similar to regions 1 and 2 defined in [29]. Fig. 2 (a) is an image where facial regions have been detected using the adaboost algorithm to find facial landmarks before defining the periocular region. In the facial region detected by the adaboost algorithm, there are the cases where the jaw line is partially cut off. To consider these cases, the face width and height found by the adaboost algorithm is increased to an ROI of 1.8 times larger than the original size so that the jaw line is not cut off and an adequate ROI is selected, as shown in Fig. 2 (b). Next, the dlib facial landmarks tracking method is used to detect 68 coordinates on the face, as shown in Fig. 2 (c). Fig. 2 (d) is an image which shows the two coordinates in the corners of the eye and the center of the eye among the detected coordinates.

The reason why alignment is performed based on the center of the eye rather than the center of the iris or pupil is as follows. In unconstrained biometrics, user's poses or gazes can easily change, and accurate alignment is only possible if it is performed by referring to the eye corners [11]. In addition, it is difficult to detect the iris area in low resolution images captured by surveillance camera at a distance. Next, the distance between the blue points in the corners of the eye shown in Fig. 2 (d) is calculated and the loose ROI is set in proportion to the distance from the center of the eye, as in Fig. 2 (e). Here, the width and height of the loose ROI are set at 3.2 times the distance between the two corner coordinates in Fig. 2 (d), and it is important to note that the region should be set so the width and height of image are the same. If the width and height are different, distortion or stretching occurs when it

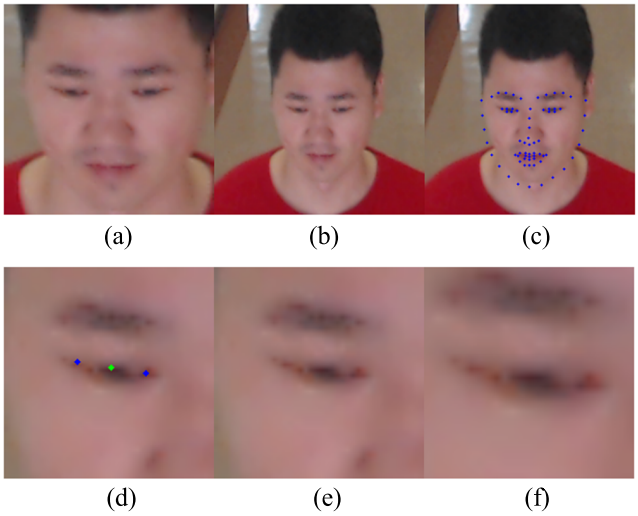

FIGURE 2. Detection of face and periocular region. (a) Face region detected by adaboost algorithm from input image, (b) 1.8 times sized-up face region, (c) facial landmarks found in face region, (d) corners and center coordinates for left eye used in alignment, (e) periocular image of loose ROI, and (f) periocular image of tight ROI.

is resized. Therefore, zero padding is not performed in this study and a skin region is added. The reason for this is that color information is also used when extracting features using CNN, so skin color can be good feature. The ROI is defined by a ratio because the size of eye region varies in each image depending on the distance to the camera. Then, the loose ROI of Fig. 2 (e) is resized to the size of 224×224 pixels, and it is used as input for the CNN. Within the loose ROI of 224×224 pixels, a tight ROI of 140×140 pixels is set based on the eye center coordinates extracted from Fig. 2 (d). It is resized to the size of 224×224 pixels, and it is used as input for the CNN, also. Here, the ratio based on the distance between the two eye corners is not used and a tight ROI is set to a specified numeric size (140×140 pixels) because the loose ROI was already captured according to the ratio based on the distance between the two eye corners, so even though the tight ROI is captured according to a fixed numerical size, the size of the periocular region remains regardless of the distance from the camera.

C. PERIOCULAR RECOGNITION

1) PREPROCESSING

When periocular recognition is attempted using blurred images as shown in Figs. 3 (b) and (d), the possibility of false rejection is high even though the images are from the same class, and image drop problems occur in which good quality images in good focusing condition cannot be captured from later sequential images due to the processing time required when recognition is attempted with these images.

To resolve these problems, this study used a focus score assessment method based on 5×5 convolution kernel proposed in study [12] as a preprocessing method to exclude blurred images like Figs. 3 (b) and (d). In detail, the

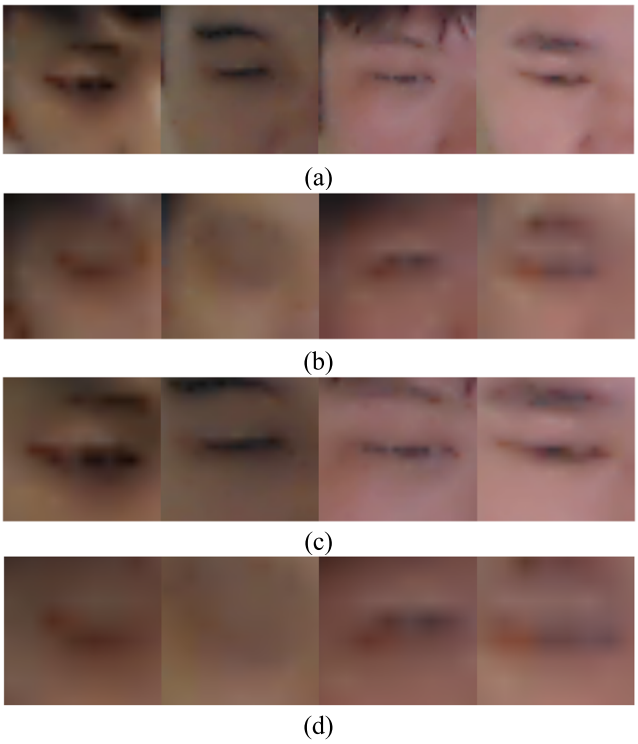

FIGURE 3. Images classified using focus score. The cases of (a) loose ROI having good focus score, (b) loose ROI having bad focus score, (c) tight ROI having good focus score, (d) tight ROI having bad focus score. The focus scores of images from left to right directions in (a) and (c) are 48, 48, 64, and 55, respectively. The focus scores of images from left to right directions in (b) and (d) are 22, 17, 24, and 26, respectively.

magnitude value calculated by the convolution operation with the 5×5 kernel in the image is used as the focus score [12]. The focus score has a value from 0 to 100 (the higher the score, the better the focusing status of image). Focus score is measured with the tight ROI, and the feature extraction using CNN and recognition was performed only if the score is above the threshold value. The reason why focus score is calculated only with the tight ROI is as follows. The part of background closed to the face can be included in the loose ROI. Therefore, if the score is measured with the loose ROI, an accurate score cannot be found. In DP-DB1, the images whose focus scores are higher than (or same to) 39 are determined as the images of good focus score.

In ChokePoint database, the images higher than (or same to) 30 are determined as those of good focus score. These optimal thresholds of 39 and 30 were determined experimentally based on the recognition accuracy of training data using each database. The focus scores are calculated separately for the left and right eye.

The images in Fig. 3 are from the same person in each column. Figs. 3 (a) and (c) show the loose and tight ROI images of the same periocular from the same image frame. Figs. 3 (b) and (d) are also loose and tight ROI images of the same periocular from the same frame. It can be observed that the images of bad score such as Figs. 3 (b) and (d) are severely blurred and difficult to be discernable.

2) FEATURE EXTRACTION USING CNNS

Currently, CNNs are gaining popularity due to their powerful ability to extract comprehensive features from visual patterns, and they are used in many successful applications. CNNs show better performance than traditional methods which use handcrafted features or other training approaches [2]. In this study, the pre-trained model of VGG face-16 [21] is used and fine-tuning is performed with our experimental databases. The VGG face-16 model is composed of 13 convolutional layers and 3 fully connected layers. In the 1st convolutional layer, 64 filters with the size of 3 × 3 and 3 channels are used. The size of feature map is 224 × 224 × 64 in the 1st convolutional layer, such that 224 and 224 are the output height and width, respectively, calculated based on (output height (or width) = (input height (or width) – filter height (width) + 2 × (the number of padding)) / (the number of stride) +1 [37]). For example, in the image input layer and 1st convolutional layer of Table 2, the input height is 224; the filter height is 3; the number of padding is 1; and the number of stride is 1. As a result, the output height is 224 (= (224 − 3 + 2 × 1)/1 + 1).

The output feature map for standard convolution considering padding and stride is usually acquired as [47]:

$$N_{k,l,n} = \Sigma_{i,j,m}(F_{i,j,m,n} \cdot M_{k+i-1,l+j-1,m}) \quad (1)$$

In Equation (1), $M_{k+i-1,l+j-1,m}$ is the input feature map having the size of $S_F \times S_F \times D$.S_F is the width and height of square input feature map, and D is the number of input channels (input depth). $N_{k,l,n}$ is the output feature map of the size of $T_F \times T_F \times E$.T_F is the spatial width and height of a square output feature map, and E is the number of output channels (output depth). In Eq. (1), $F_{i,j,m,n}$ is the convolution filter of size $S_K \times S_K \times D \times E$, and S_K is the spatial dimension of convolution kernel. Then, standard convolutions take the following computational cost of:

$$Cost = S_K \cdot S_K \cdot D \cdot E \cdot S_F \cdot S_F \quad (2)$$

Based on the Eq. (2), we can observe that the computational cost is dependent on multiplicatively on the kernel size of $S_K \times S_K$, the number of input channels of D, the number of output channels of E, and the input feature map size of $S_F \times S_F$ [47].

As shown in Table 2, there are 15 rectified linear unit (ReLU) layers and 5 max pooling layers. The ReLU is used as an activation function in the form shown in Eq. (3) [36].

$$o = max(0, i) \quad (3)$$

Here, o is the output, and i is the input. The ReLU function is linear, so it resolves the vanishing gradient problem [40], and has the advantage of a faster processing speed than non-linear functions. After passing through the convolutional layer and the ReLU layer, the feature map is passed through the max pooling layer, as shown in Table 2. Here, the 2nd convolutional layer is applied with a 3 × 3 filter, a padding of 1 × 1, and stride of 1 × 1, which is the same as the 1st convolutional layer. The feature map size of 224 × 224 × 64 is maintained. As shown in Table 2, the 13 convolutional layers use the same filter size of 3 × 3 and padding of 1 × 1, so the feature map size is maintained, and the number of filters is changed to 64, 128, 256, and 512. In addition, each ReLU layer has a connected structure behind each convolutional layer, and the feature map size that passes through the convolutional layer is maintained. After the 2nd, 4th, 7th, 10th, and 13th convolutional layer with ReLU layer, the max pooling layer is operated. The max pooling layer performs a kind of subsampling by selecting the largest value within the filter. After the 2nd convolutional layer with ReLU layer, when the max pooling layer is performed, the input feature map size is 224 × 224 × 64. The filter size is 2 × 2, and the number of strides is 2 × 2. Here, the number of strides is said to be 2 × 2, which means that there is a max pooling filter of 2 × 2, and it moves in horizontal and vertical directions, by 2 pixels at a time. As the filter moves, there is no overlapped area, so the feature map size is reduced to 1/4 (1/2 horizontally and 1/2 vertically).

Ultimately, the feature map size that passes through the max pooling layer is 112 × 112 × 64. As shown in Table 2, this max pooling layer is composed of the same filter of 2 × 2 and a stride of 2 × 2 in all cases, and through this, the feature map size is reduced to 1/4. If the 13 convolutional layers, 13 ReLU layers, and 5 max pooling layers are passed through, ultimately a feature map of 7 × 7 × 512 is obtained, and 3 additional fully connected layers (FCLs) are passed through. The number of output nodes of 1st, 2nd, and 3rd FCLs are 4096, 4096, and number of classes, respectively.

Normally, CNNs have an overfitting problem in which the network becomes too dependent on the training data, which can cause low recognition accuracy with testing data even when the accuracy with the training data is still high. To resolve this problem, this study used dropout methods [20, 48] which can reduce the effects of the overfitting problem. For the dropout method, we use the dropout probability of 50% to randomly disconnect the connections between the 1st and 2nd FCLs. After the 3rd FCL, a softmax layer is used to calculate the probability of output node as shown in Eq. (4) [38].

$$\sigma(k)_j = \frac{e^{kj}}{\sum_{n=1}^{R} e^{kn}} \quad (4)$$

Here, k is the output neuron array, and the probability of each class can be calculated by dividing the j^{th} element's value by the summation of all elements.

Table 2 and Fig. 4 shows the structure of the VGG face-16 model used in this study. In this study, the training images, which consist of the loose and tight ROI periocular images, are separately fine-tuned in the pretrained models of VGG face-16 from [35], and they are used to perform tests. Each trained model is used to extract 4096 features from the 1st FCL (Fc6 of Table 2) and the 2nd FCL (Fc7 of Table 2), respectively, and their performance are compared. We also make a performance comparison using the residual network

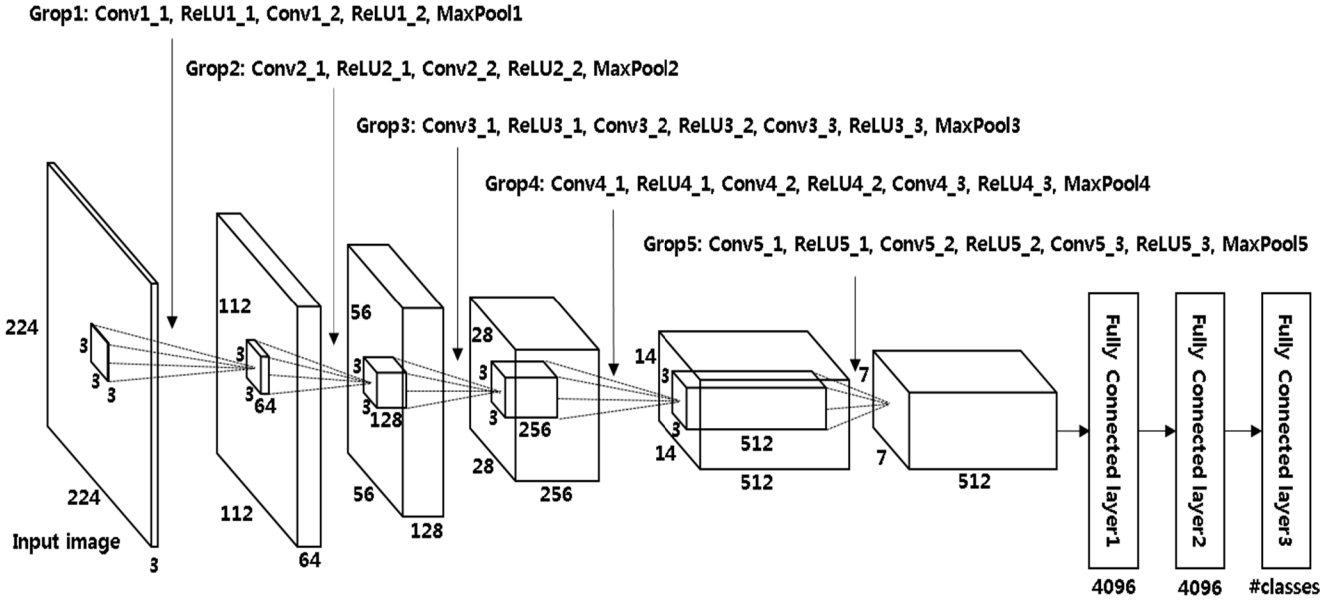

FIGURE 4. The architecture of VGG face-16 model.

(ResNet) of 50 layers [22], which is deeper than VGG face-16 in order to see how the depth of a deeper network affects performance. These experiments are described in detail in Sections IV.C.1 and IV.C.4.

3) FEATURE NORMALIZATION AND SCORE FUSION METHOD

We compare the weighted sum and weighted product methods as score level fusion. The 4096 features, which are the output of the Fc6 of the CNN which was trained with loose and tight ROI periocular images, are normalized each other as in Eq. (5) before they are used to calculate the Euclidean distance. This normalization can give a compensation effect for the case that feature values become too large or small due to factors of brightness change and noise in input image.

$$\boldsymbol{Ft}_{normalization} = \boldsymbol{Ft} \Big/ \sqrt{\sum_{j=1}^{n} Ft_j^2} \tag{5}$$

Here, $\boldsymbol{Ft}_{normalization}$ is the normalized feature vector, and $\boldsymbol{Ft}$ is the feature vector before normalization. n is the number of features.

To perform score level fusion, the two Euclidean distances calculated from each of the loose and tight ROI periocular images, are used as the two scores, and applied through the weighted sum and weighted product methods, as shown in Eqs. (6) and (7).

$$WS = \sum_{j=1}^{2} w_j s_j \tag{6}$$

$$WP = \prod_{j=1}^{2} s_j^{w_j} \tag{7}$$

WS and WP are scores calculated by applying the weighted sum and weighted product methods, respectively. S_j is the Euclidean distance calculated from the features, and w_j is the weight. The optimal weights (w_j) were found through recognition experiments using training data.

Periocular recognition is ultimately performed based on the combined score. In 1:1 matching (verification), genuine matching occurs when the combined score is smaller than the threshold whereas imposter matching happens when it is larger than the threshold. In this study, the threshold is set at the point where the false rejection rate (FRR) and the false acceptance rate (FAR) are the same. We call this point as the equal error rate (EER) point. In 1:n matching (identification), the combined scores are calculated for each of the images (gallery images) registered in the database. The image with the smallest score is ultimately determined to be the genuine matching class.

IV. EXPERIMENTAL RESULTS

A. DATABASE AND DATA AUGMENTATION

In this study, we used the custom-made DP-DB1 [45] and the open database of ChokePoint database [4], [32] provided by NICTA. The images in both databases were captured with visible light cameras. The UBIPr [11], FRGC [42], and FOCS [43] databases used in existing periocular recognition researches, are composed of high-resolution images which are difficult to be captured in an unconstrained and surveillance environment. However, the DP-DB1 and ChokePoint databases contain the severely blurred and low-resolution images that are frequently captured in the unconstrained and surveillance environment. The original sizes of periocular images detected in the DP-DB1 database and the ChokePoint databases are very small as 30×30 to 50×50 pixels. They are resized to 224×224 pixels to be used as input for the CNN.

TABLE 2. Detailed descriptions of VGG face-16 model.

	Layer type	Number of filter	Size of feature map	Size of kernel	Number of stride	Number of padding
	Image input layer		224 (height)× 224 (width)× 3 (channel)			
Group 1	Conv1_1 (1st convolutional layer)	64	224×224×64	3×3	1×1	1×1
	ReLU1_1		224×224×64			
	Conv1_2 (2nd convolutional layer)	64	224×224×64	3×3	1×1	1×1
	ReLU1_2		224×224×64			
	MaxPool1	1	112×112×64	2×2	2×2	0×0
Group 2	Conv2_1 (3rd convolutional layer)	128	112×112×128	3×3	1×1	1×1
	ReLU2_1		112×112×128			
	Conv2_2 (4th convolutional layer)	128	112×112×128	3×3	1×1	1×1
	ReLU2_2		112×112×128			
	MaxPool2	1	56×56×128	2×2	2×2	0×0
Group 3	Conv3_1 (5th convolutional layer)	256	56×56×256	3×3	1×1	1×1
	ReLU3_1		56×56×256			
	Conv3_2 (6th convolutional layer)	256	56×56×256	3×3	1×1	1×1
	ReLU3_2		56×56×256			
	Conv3_3 (7th convolutional layer)	256	56×56×256	3×3	1×1	1×1
	ReLU3_3		56×56×256			
	MaxPool3	1	56×56×256	2×2	2×2	0×0
Group 4	Conv4_1 (8th convolutional layer)	512	28×28×512	3×3	1×1	1×1
	ReLU4_1		28×28×512			
	Conv4_2 (9th convolutional layer)	512	28×28×512	3×3	1×1	1×1
	ReLU4_2		28×28×512			
	Conv4_3 (10th convolutional layer)	512	28×28×512	3×3	1×1	1×1
	ReLU4_3		28×28×512			
	MaxPool4	1	14×14×512	2×2	2×2	0×0
Group 5	Conv5_1 (11th convolutional layer)	512	14×14×512	3×3	1×1	1×1
	ReLU5_1		14×14×512			
	Conv5_2 (12th convolutional layer)	512	14×14×512	3×3	1×1	1×1
	ReLU5_2		14×14×512			
	Conv5_3 (13th convolutional layer)	512	14×14×512	3×3	1×1	1×1
	ReLU5_3		14×14×512			
	MaxPool5	1	7×7×512	2×2	2×2	0×0
	Fc6 (1st fully connected layer)		4096×1			
	ReLU6		4096×1			
	Dropout6		4096×1			
	Fc7 (2nd fully connected layer)		4096×1			
	ReLU7		4096×1			
	Dropout7		4096×1			
	Fc8(3rd fully connected layer)		#classes			
	Softmax layer		#classes			
	Output layer		#classes			

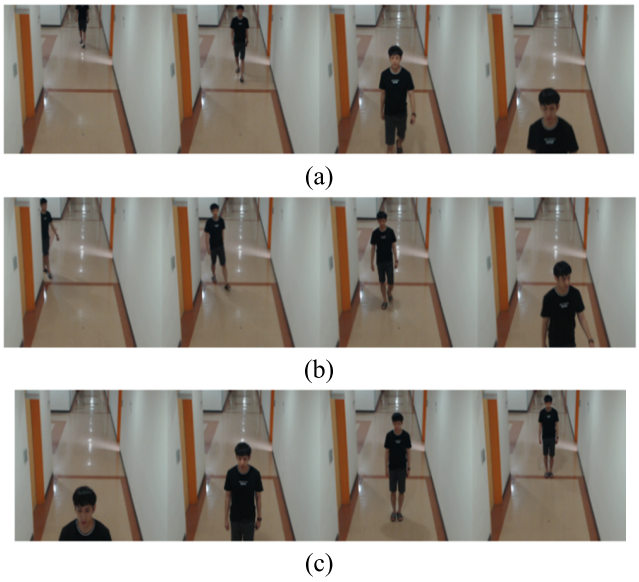

FIGURE 5. The examples in DP-DB1. Images captured based on the scenarios of (a) straight line movement, (b) corner movement, and (c) standing still.

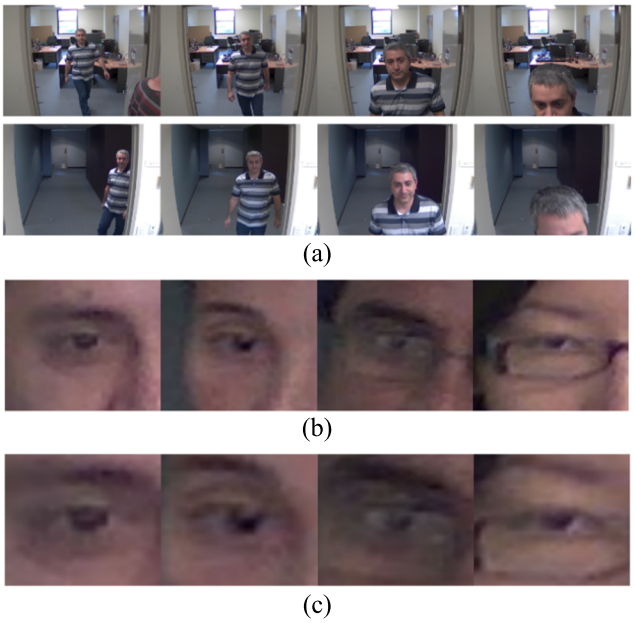

FIGURE 6. Examples of ChokePoint database. (a) Images based on Case study 1 scenario, and periocular images of (b) loose ROI, (c) tight ROI.

1) DP-DB1 DATABASE

The DP-DB1 database was created in this study for periocular recognition in an indoor surveillance environment. The camera used to capture the images was a Logitech BCC 950 [39], and the specifications of the camera are as follows; a camera viewing angle of 79 degrees, a maximum resolution of full high definition (Full HD) 1080p, and a frame rate of 30 fps with auto focusing. The camera was located in an indoor hallway (with indoor lights on) at a height of 2 m 40 cm.

This database can be used for both periocular and face recognition researches. It consists of 20 people captured in three scenarios: straight line movement, corner movement, and standing still, as shown in Fig. 5. In case of the scenario of standing still, the images were acquired at 4 different positions. The custom-built DP-DB1 and the trained periocular recognition CNN models with algorithms are to be shared with other researchers as shown in [45] so that they can perform impartial performance evaluations. Table 3 contains the detailed description of the DP-DB1 database. In this study, the experiments were performed by a two-fold cross validation scheme, so the DP-DB1 database was divided into the sub-datasets 1 and 2. As shown in Table 3, the people of training data was completely different from that of the testing data.

2) CHOKEPOINT DATABASE

The ChokePoint database is a real-world surveillance video database, and designed for person identification and verification experiments. It is provided by NICTA as an open database [4], [32]. It consists of portal 1, which has 25 people, and portal 2, which has 29 people. The portal 1 and portal 2 were captured in a one-month interval. In this paper, the images of case study 1 scenario presented in [32] were used in order to use images captured in the same indoor environment as the DP-DB1 database. The case studies 1, 2, and 3 from [32] are composed of 2 groups each, and case study 1 is only images captured over a short time interval indoors. Case study 2 includes the images captured over a short time interval indoors and outdoors, and case study 3 includes the images captured over a long time intervals indoors and outdoors.

We used the images of case study 1 for experiments and Fig. 6 shows the images and detected periocular images from ChokePoint database. Table 3 contains the detailed description of the databases used in the experiments. We performed the experiments using a two-fold cross validation. In the 1st fold cross validation, training was performed using the augmented data obtained from sub-dataset1, and sub-dataset2 was used to perform testing. In the 2nd fold cross validation, training was performed using the augmented data obtained from subdataset2, and sub-dataset1 was used to perform testing. As shown in Table 3, the same person's images were not included in the training data and the testing data at the same time.

3) DATA AUGMENTATION

When a CNN with a deep structure is trained, care should be taken to avoid overfitting. For this purpose, dropout and data augmentation methods are used [20]. In this paper, in order to train with an adequate amount of data and avoid overfitting, mirror images of the detected periocular images are made through horizontal flipping as shown in Fig. 7 to double the amount of data, and then these images are cropped at

TABLE 3. Description of experimental database (Loose: periocular images of loose ROI, Tight: periocular images of tight ROI).

Kinds of images, # of people, and # of images					DP-DB1		ChokePoint	
					Sub-dataset1	Sub-dataset2	Sub-dataset1	Sub-dataset2
Original images		# of people			10	10	13	12
		#of images	Loose	R	1,155	1,436	3,115	2,860
				L	1,155	1,436	3,115	2,860
			Tight	R	1,155	1,436	3,115	2,860
				L	1,155	1,436	3,115	2,860
Training	Augmented images	#of images	Loose		600,600	746,720	1,619,800	1,487,200
			Tight		600,600	746,720	1,619,800	1,487,200
Testing	Good focus images	#of images	Loose	R	539	765	2,235	2,106
				L	578	791	2,173	1,977
			Tight	R	539	765	2,235	2,106
				L	578	791	2,173	1,977
	Bad focus images	#of images	Loose	R	616	671	880	754
				L	577	645	942	883
			Tight	R	616	671	880	754
				L	577	645	942	883

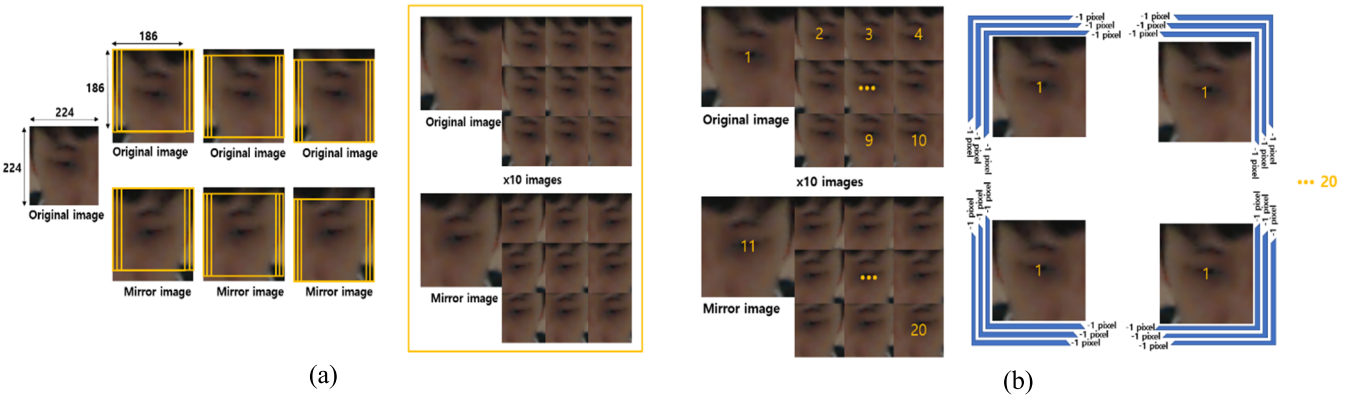

FIGURE 7. Data augmentation method. (a) Method of cropping into 9 regions, and (b) method of translation & cropping.

9 regions to create 10 times more data. The method for dividing the image into 9 regions involves the procedure of dividing the input image of 224 (height) × 224 (width) pixels into that of 186 × 186 pixels and moving it a fixed distance of 19 pixels at a time (Fig. 7 (a)). These images are translated and cropped by one pixel three times along the top left diagonal direction. This procedure is iterated along the top right diagonal direction, the bottom left diagonal direction, and the bottom right diagonal direction, also, in order to increase the amount of data 13 times. (Fig. 7 (b)). If this method is used, the number of images is increased two times by mirroring, 10 times by dividing images into 9 regions, and 13 times by pixel shifting, so that a total of 260 (2 × 10 × 13) times the number of images can be obtained as shown in Table 3. This kind of data augmentation has often been used in existing research [20]. The augmented data was used only in the training process; in the testing process, original images which were not augmented were used.

B. TRAINING PROCESS

All detected images were used for training, regardless of their focus score, and both left periocular and right periocular images were used without distinction. The detailed number of training images is shown in Table 3. The reason that we used all images without distinction to the left or right as well as blur images with incorrect focus for training is as follows. When we observe the experiment results according to periocular image augmentation in [1], the recognition accuracies are better when various augmentation methods are combined, including mirror, blur, noise, darken, and brighten methods, than those without augmentation. However, the augmentation effect is different according to the database, so it is important to find a suitable method [1].

When the augmented images are fine tuned in the VGG face-16 pretrained model, the mean value should be set, so we use the value provided by [35] without modification, and

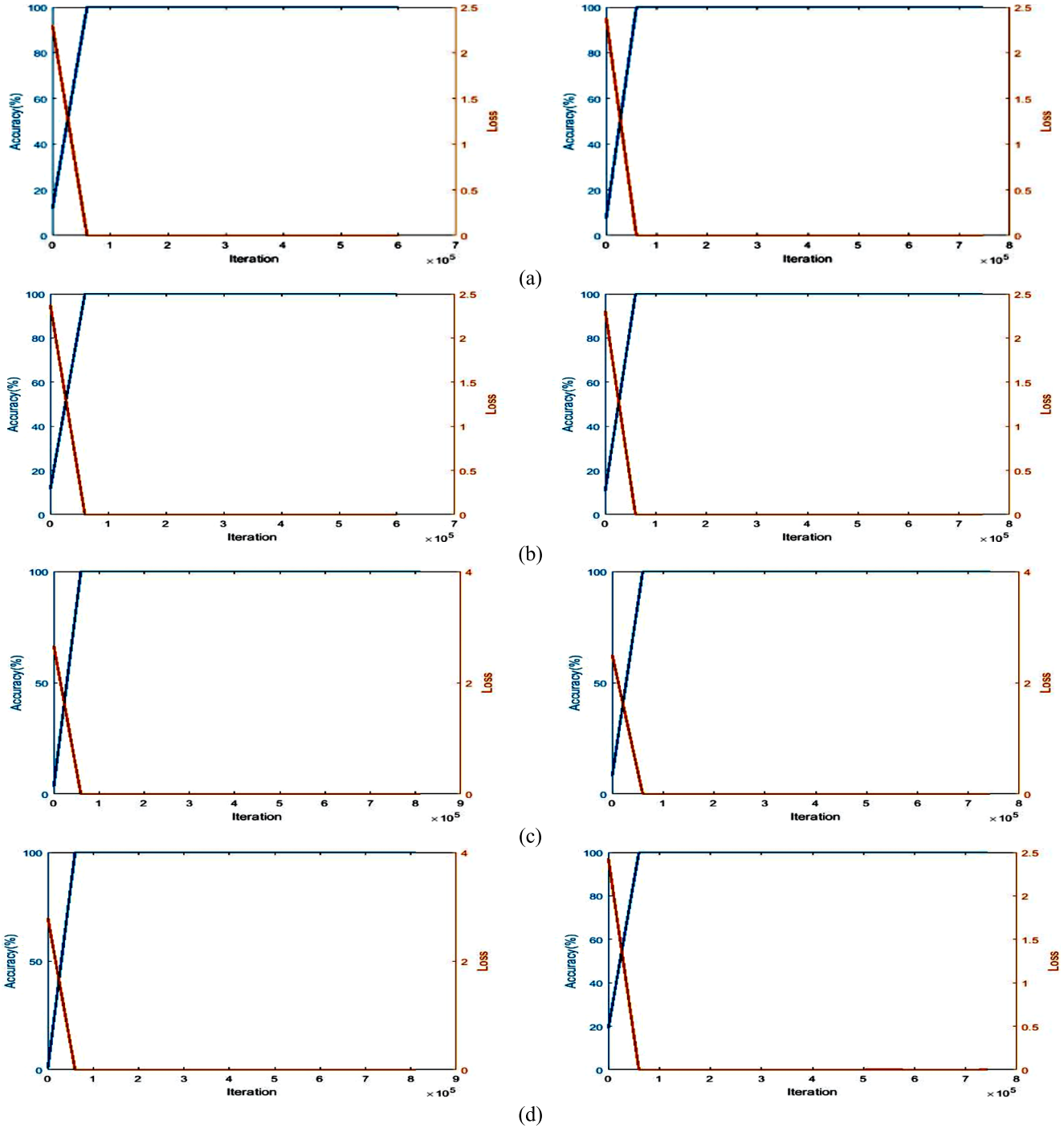

FIGURE 8. Loss and accuracy curves with training data of two-fold cross validation according to databases. (a) Loose ROI of DP-DB1 database, (b) tight ROI of DP-DB1 database, (c) loose ROI of ChokePoint database, (d) tight ROI of ChokePoint database. In (a) ~ (d), left and right figures show the results by the 1^{st} fold (sub-dataset1) and 2^{nd} fold (sub-dataset2) cross validation, respectively.

training parameters for the periocular images of loose and tight ROIs in the same way. For the training, the stochastic gradient descent (SGD) method as in [35] is used as shown in Eqs. (8) and (9) [20]. Unlike the gradient descent (GD) method, in the SGD method, the number of training sets divided by mini-batch size is defined as iteration, and one epoch is set when training is performed for all the number of iterations.

$$h_{j+1} := mh_j - dlrw_j - lr < \frac{\partial Q_j(w)}{\partial w} | w_j >_{D_j} \quad (8)$$

$$w_{j+1} := w_j + h_{j+1} \quad (9)$$

Where w_j is the weight to be learnt at the j^{th} iteration. m is momentum, h_j is the momentum variable, d is the weight

TABLE 4. Comparisons of EER according to loose ROI and tight ROI before score fusion.

Method		EER (%)									
		DP-DB1					ChokePoint				
		L-L		R-R		Average	L-L		R-R		Average
		S1	S2	S1	S2		S1	S2	S1	S2	
VGG face-16 model with fine tuning **(proposed method)**	Loose ROI	4.49	9.38	4.65	6.20	6.18	5.90	4.40	4.89	5.00	5.05
	Tight ROI	11.22	15.41	7.60	8.78	10.75	7.22	6.32	5.82	8.27	6.91

TABLE 5. Comparisons of EER according to feature normalization and fully-connected layer features (S1 and S2 represents Sub-dataset1 and Sub-dataset2, respectively).

Method			EER (%)									
			DP-DB1					ChokePoint				
			L-L		R-R		Average	L-L		R-R		Average
			S1	S2	S1	S2		S1	S2	S1	S2	
Fc6	Original features	Weighted sum	3.91	9.73	2.40	7.24	5.82	6.09	5.30	5.09	4.92	5.35
		Weighted product	4.14	9.71	2.40	7.41	5.92	6.01	5.30	5.14	4.95	5.35
	Normalization features **(proposed method)**	Weighted sum	4.56	8.76	3.89	4.47	**5.42**	4.85	3.89	4.17	4.94	4.46
		Weighted product	4.59	8.88	3.97	4.41	5.46	4.90	3.82	4.18	4.86	**4.44**
Fc7	Original features	Weighted sum	4.85	9.59	2.75	6.36	5.89	5.40	4.96	5.70	5.10	5.29
		Weighted product	5.90	9.63	2.79	7.76	6.52	5.40	5.01	5.78	5.14	5.33
	Normalization features	Weighted sum	5.11	9.70	3.82	4.97	5.90	5.46	5.07	4.64	5.09	5.06
		Weighted product	5.21	10.21	3.78	5.03	6.06	5.50	5.04	4.70	5.07	5.08

decay, and lr is the learning rate. $< \frac{\partial Q_j(w)}{\partial w} | w_j >_{D_j}$ is the average over the j^{th} batch D_j of the derivative of the object with respective to w, evaluated at w_j. We set m, d, and lr at 0.9, 5×10^{-4}, and 5×10^{-4}, respectively, and training was performed with a batch size of 20. For the DP-DB1 and ChokePoint databases, we used 20 epochs and 10 epochs for training, respectively. The reason why the numbers of epochs are different is that in a comparison of the number of augmented images used in training for the DP-DB1, the ChokePoint database has over two times as many augmented images as shown in Table 3, so the number of epochs was reduced.

Fig. 8 shows the training loss and accuracy of the sub-datasets 1 and 2 of DP-DB1 and the ChokePoint databases. The x axis means the number of iterations whereas the right and left y axes show the loss value and training accuracy, respectively. As shown in Fig. 8, in all cases, the training loss was close to 0% and the training accuracy was close to 100%.

C. TESTING PROCESS AND RESULT

To perform the performance evaluation, we set the gallery features (the features of the enrolled images) by calculating the geometric center for each class. For that, we use 4096 features extracted from the fully-connected layer. Based on the gallery features, the Euclidean distance with the probe features (the features of the input images) was calculated to measure the EER through authentic and imposter matching. The left and right periocular images were tested separately from each other. In Table 4, the performance measured among left-periocular images is labeled L-L; the performance measured among right-periocular images is labeled R-R; and the ultimate performance was set as the mean of the EER for left- and right-periocular recognition. As explained in Section III.C.3, EER is the error rate at the point where the FAR and FRR are the same.

1) COMPARISONS OF ACCURACIES ACCORDING TO LOOSE OR TIGHT ROI, FEATURE TYPE, AND NORMALIZATION

As the first test, we compared the recognition accuracies only by loose ROI or tight ROI. As shown in Table 4, average EERs with loose ROI are lower than those by tight ROI on both DP-DB1 and ChokePoint databases. As the next experiment, we compare the accuracies using the features in Fc6 of Table 2 with and without normalization. In addition, the accuracies using the features in Fc7 of Table 2 with and without normalization were compared as shown in Table 5. The experiment results showed that the normalized 4096 features in Fc6 had the best performance. The weighted sum method showed the best performance in the DP-DB1 database

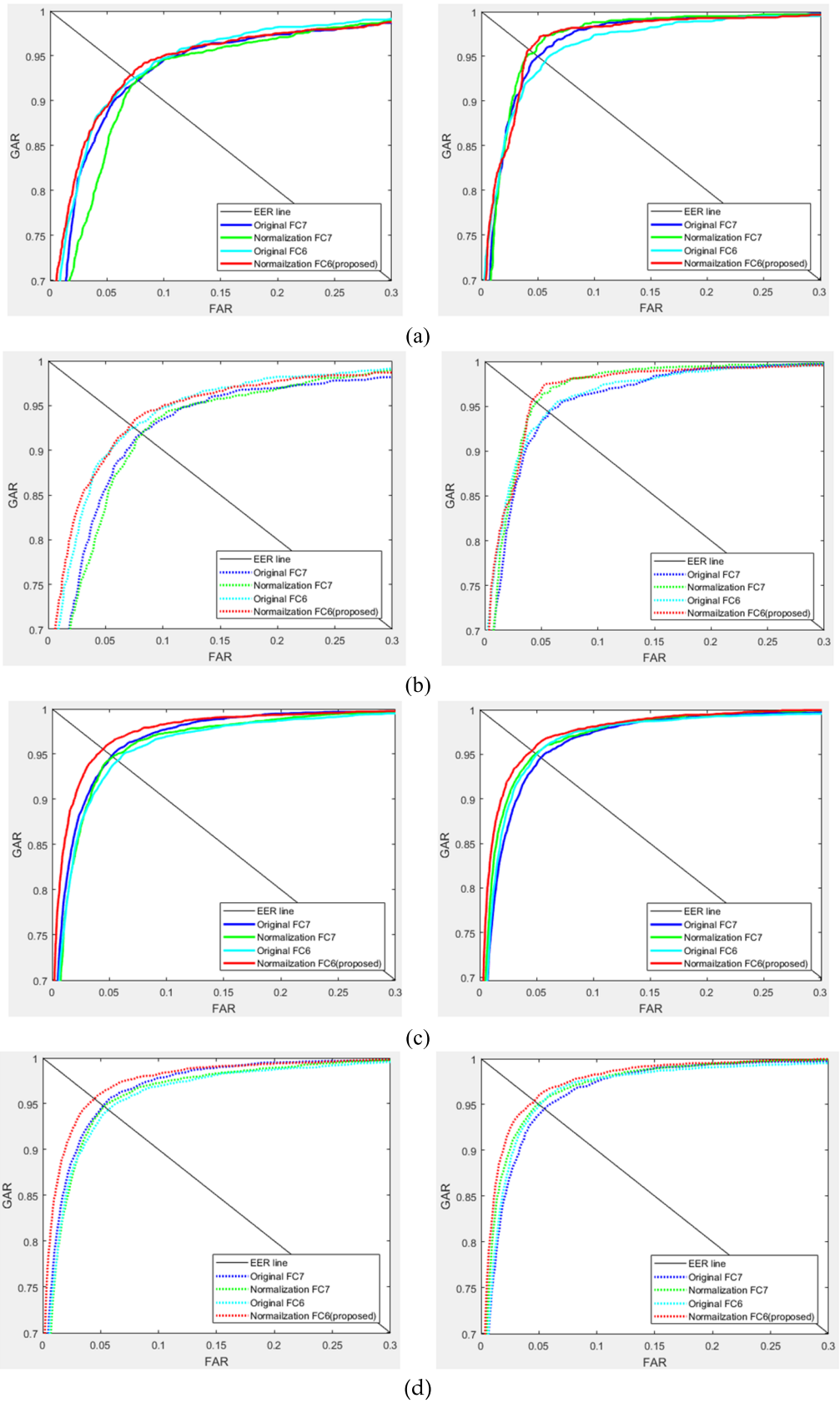

FIGURE 9. ROC curves of recognition using normalization and fully-connected layer features. (a) Weighted sum method with the DP-DB1 database, (b) weighted product method with the DP-DB1 database, (c) weighted sum method with the ChokePoint database, (d) weighted product method with the ChokePoint database. In (a) ∼ (d), left and right figures show the results by the L-L and R-R periocular recognitions, respectively.

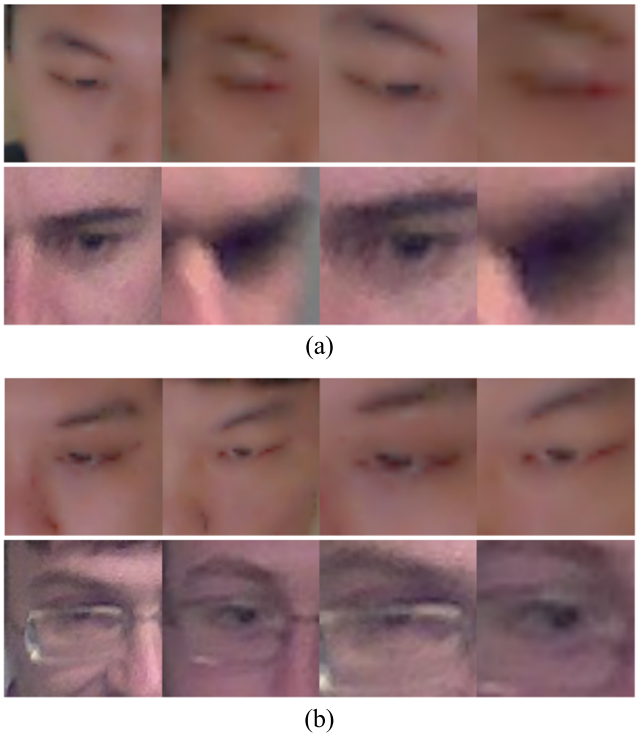

FIGURE 10. True acceptance and true rejection cases. (a) True acceptance cases, and (b) true rejection cases. In (a) and (b), left two images are respectively the gallery and probe images of loose ROI whereas right two images are the gallery and probe images of tight ROI, respectively.

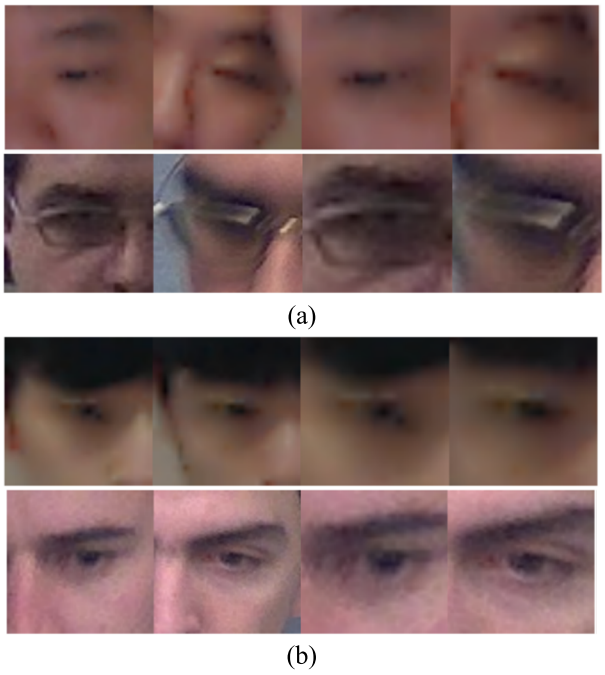

FIGURE 11. False rejection and false acceptance cases. (a) False rejection cases, (b) false acceptance cases. In (a) and (b), left two images are respectively the gallery and probe images of loose ROI whereas right two images are the gallery and probe images of tight ROI, respectively.

whereas the weighted product method showed the best performance in the ChokePoint database.

Fig. 9 shows the receiver operation characteristic (ROC) curve of the experiment results in Table 5. The graphs on the left show the average one of 1st and 2nd fold results of L-L matching whereas the graphs on the right show the average one of 1st and 2nd fold results of R-R matching. In addition, the result by weighted sum method is shown as a solid line, and that by the weighted product method is shown as a dotted line. The horizontal and vertical axes show FAR and genuine acceptance rate (GAR), respectively, and the GAR was calculated via 1-FRR. In Fig. 9, the normalized features in Fc6 still show the best performance, and the weighted sum method showed the best performance in the DP-DB1 database whereas the weighted product method showed the best performance in the ChokePoint database.

2) ANALYSIS OF GENUINE AND IMPOSTER MATCHING RESULT

To confirm how well the proposed method recognizes periocular images, we examined genuine matching and imposter matching results as shown in Figs. 10 and 11. As shown in Fig. 10 (a), it can be seen that even when there was blurring, when there was reflected light from glasses, or when the face was greatly rotated, genuine matching was done correctly, and these were cases where it would be difficult for an actual person to distinguish a genuine match. The images in Fig. 10 (b) are imposter matching cases which look like the same person but were properly distinguished as imposters. It can be seen that the images contained similar-looking skin color, eye shape, glasses, etc., but they were correctly distinguished as imposters. In Figs. 10 (a) and (b), the two images on the left are gallery and probe images with loose ROI, respectively whereas the two images on the right are gallery and probe images with tight ROI.

Fig. 11 shows a false rejection case and a false acceptance case, which were error cases found when measuring the performance of the proposed method. Fig. 11 (a) is an error case that the genuine person was incorrectly rejected. As shown in the first row of images of Fig. 11 (a), it can be seen that there is a completely different pose, a slight occlusion occurring due to hand movement, and the eye is closed. This is a complex case with three different performance degrading factors.

As shown in the second row of images of Fig. 11 (a), there is severe blur, a pose change, and an occlusion due to eyeglasses. Like the first row, it is a complex case with three different performance degrading factors. Looking at these kinds of cases, we decided that it is necessary to add an algorithm for pose compensation and to perform preprocessing for open and closed eyes. In addition, it was decided that there is a need for an additional algorithm to deal with various cases of occlusion, as well as an algorithm which can reduce performance degrading factors which appear in complex ways. Fig. 11 (b) is an error case in which an imposter was incorrectly determined to be the genuine person.

TABLE 6. Comparisons of periocular recognition accuracies by our method with those by previous methods (S1 and S2 represents Sub-dataset1 and Sub-dataset2, respectively).

Method	EER (%)									
	DP-DB1					ChokePoint				
	L-L		R-R		Average	L-L		R-R		Average
	S1	S2	S1	S2		S1	S2	S1	S2	
SIFT [9]	32.65	29.5	31.06	28.46	30.42	36.96	43.85	33.74	33.58	37.03
HOG [9]	22.06	19.31	17.56	20.31	19.81	26.86	24.79	21.71	27.41	25.19
LBP [9]	21.91	18.35	17.25	16.44	18.49	21.38	19.23	12.33	19.38	18.08
SIFT+HOG+ LBP weighted sum [9]	21.41	17.95	17.22	15.80	18.095	21.52	19.55	12.46	19.76	18.32
SCNN [2]	20.03	23.25	15.49	19.26	19.51	28.31	28.20	28.26	28.84	28.40
D-PRWIS [49]	23.87	28.91	22.83	25.92	25.38	30.19	21.60	25.65	26.80	26.06
Proposed method	4.56	8.76	3.89	4.47	**5.42**	4.90	3.82	4.18	4.86	**4.44**

As can be observed, the eyebrow region is completely covered by the hair, and eye shape is very similar. Even the eye color is similar, so it is difficult to resolve the person's identity using periocular information alone. To overcome errors in such cases, there are the method combining face information or other additional information, and we can consider the methods which combine VL, NIR, or thermal images.

3) COMPARISONS WITH EXISTING METHODS

Experiments were performed to compare the proposed method with the existing studies [2], [9], [49], and the results are shown in Table 6. There were many blur images in the DP-DB1 database and the Chokepoint database as shown in Figs. 10 and 11, and periocular images had low resolution. Therefore, it is impossible to use the iris detection-based ROI setting methods used in [9], so the periocular region was set according to the method described in Section III.B. In addition, the tight ROI was chosen among the tight and loose ROIs to be used in the experiments because it is the most similar to the test images used in [2], [9], and [49].

As shown in Table 6, the method using the SIFT features showed the highest EER. This is because periocular images detected in our databases have low resolution so small key-points are not detected properly and only relatively large key-points are detected. The HOG and LBP methods, which showed a lower EER than SIFT, extracted features globally from the entire image, so they were able to extract good features more than SIFT even though the images had low resolution. In addition, when score level fusion is performed on SIFT, LBP, and HOG via the weighted sum method, the EER is slightly higher than when only LBP is used on the ChokePoint database. The reason for this is that the EER by SIFT is too high. In addition, the study [9] used the images of high quality where the periocular region can be defined through iris detection. However, irises cannot be detected in the images in the DP-DB1 database and the ChokePoint database. When the experiments were performed, the periocular region was is using the center of the eye corner coordinates detected through the dlib facial landmarks tracking method described in Section III.B, so there is a difference with the performance presented in [9].

Next, the SCNN method presented in [2], which uses deep features rather than handcrafted features, was used to perform a performance evaluation. For the evaluation method, the source code obtained from [41] was used. All-to-all matching method used in [2] performs a larger number of matching tests than the authentic and imposter matching method that we used. Therefore, in order to perform the fair evaluation, we modified the method to use authentic and imposter matching so that there would be the same number of matching tests as the method that we performed. However, the SCNN method showed a higher EER than the method using LBP. By analyzing this, we can see that the use of deep features does not mean that good features are always extracted, and excellent performance can be achieved if training is performed properly according to the database environment and the biometric category (face, periocular, iris, etc.). The reason that the EER by SCNN was high is that the database used in the tests had a complex factors that included illumination changes, pose changes, and image blurring, etc.; however, it contains high resolution images that can be used for iris recognition as in the study in [9], and surveillance environment images were not used at all in training. On the other hand, our system was fine-tuned using periocular images obtained from a surveillance environment in the VGG face-16 pretrained model, which has been proven to have excellent performance, so it was more robust in

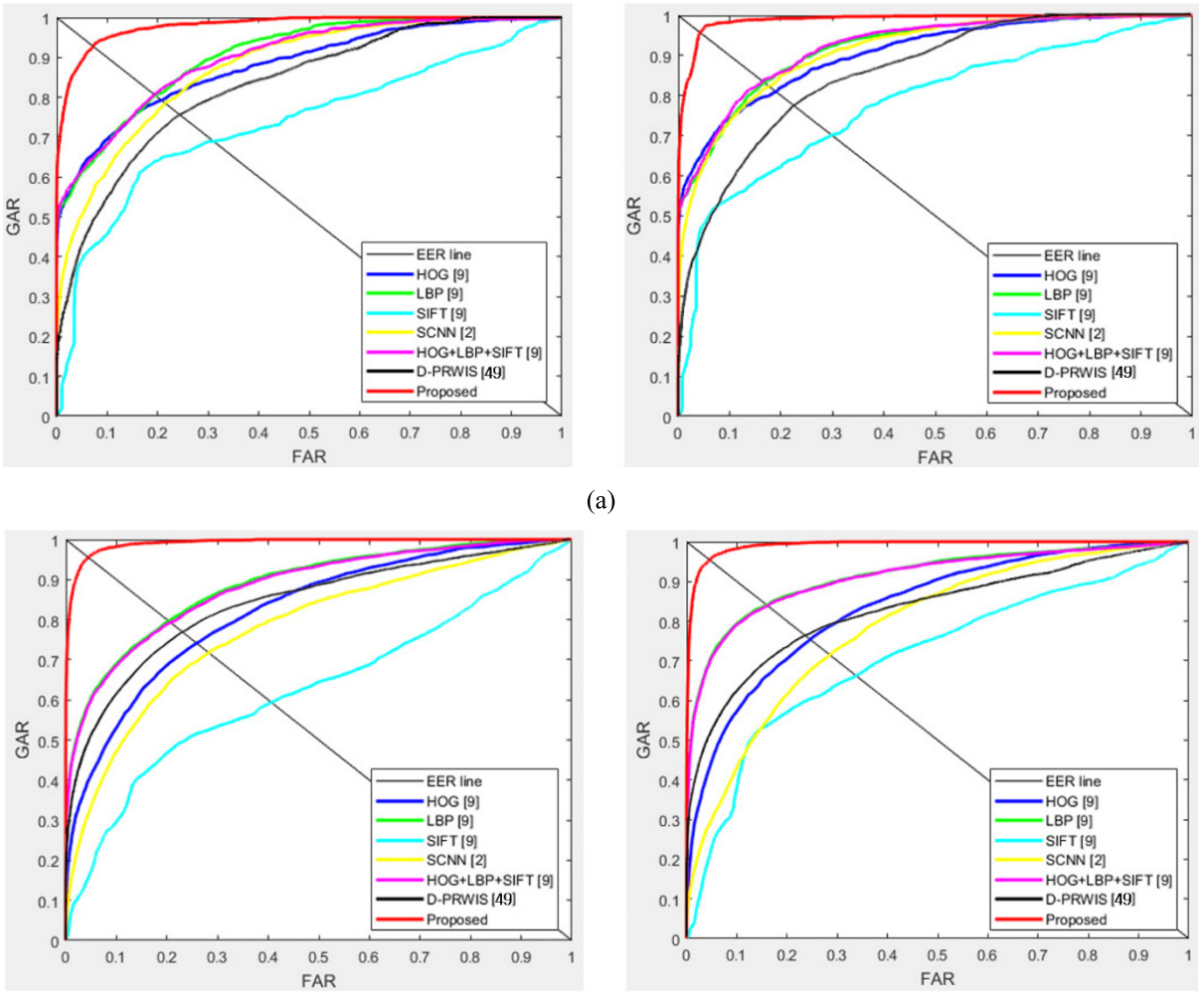

FIGURE 12. ROC curves of recognition. (a) DP-DB1 database, and (b) ChokePoint database. In (a) ∼ (b), left and right figures show the results by the L-L and R-R periocular recognition, respectively.

performing biometrics using periocular data in a surveillance environment. Therefore, it showed excellent performance compared to other methods, as seen in Table 6. In details, the EERs of periocular recognition by our method are 5.42% and 4.44% which are lower than those by other methods as shown in Table 6.

Fig. 12 contains graphs which show the experiment results from Table 6 as ROC curves. Like Fig. 9, the average result of the 1st fold and 2nd fold validations are shown. As can be seen in Fig. 12, the VGG face-16 model fine tuning method that we proposed showed the best performance. Fig. 13 contains graphs which show the experiment results from Table 6 as cumulative match characteristic (CMC) curves. The horizontal axis is the rank whereas the vertical axis is the accuracy by rank. Like the ROC curves, the CMC curves show the average result of the 1st fold and 2nd fold validations for the L-L matching and R-R matching, respectively. As can be seen in Fig. 13, the accuracy of the method proposed in this study was the highest in the CMC curves as well.

4) COMPARISONS WITH OTHER CNN NETWORK ARCHITECTURE

To understand how a deeper network architecture causes performance changes, a comparison of recognition accuracy for the DP-DB1 database was made using ResNet-50 [22], which has more layers than the proposed VGG face-16 model, after fine tuning the models with the same method. To tune ResNet-50, fine tuning was performed on the images used in the fine tuning of VGG face-16 without modification. The test methods were also the same, except that the ResNet-50 has only one fully connected layer, so 2048 features were used in the last pooling layer. The feature normalization methods which were described in Section III.C.3 were applied, and the scores obtained from the images of loose and tight ROIs were fused. The results are shown in Table 7. As shown in the results, the performance did not improve as the number of network layers increased, so it is not necessary to choose a deep structure. In addition, ResNet-50 does not simply have an increased number of layers, it also uses

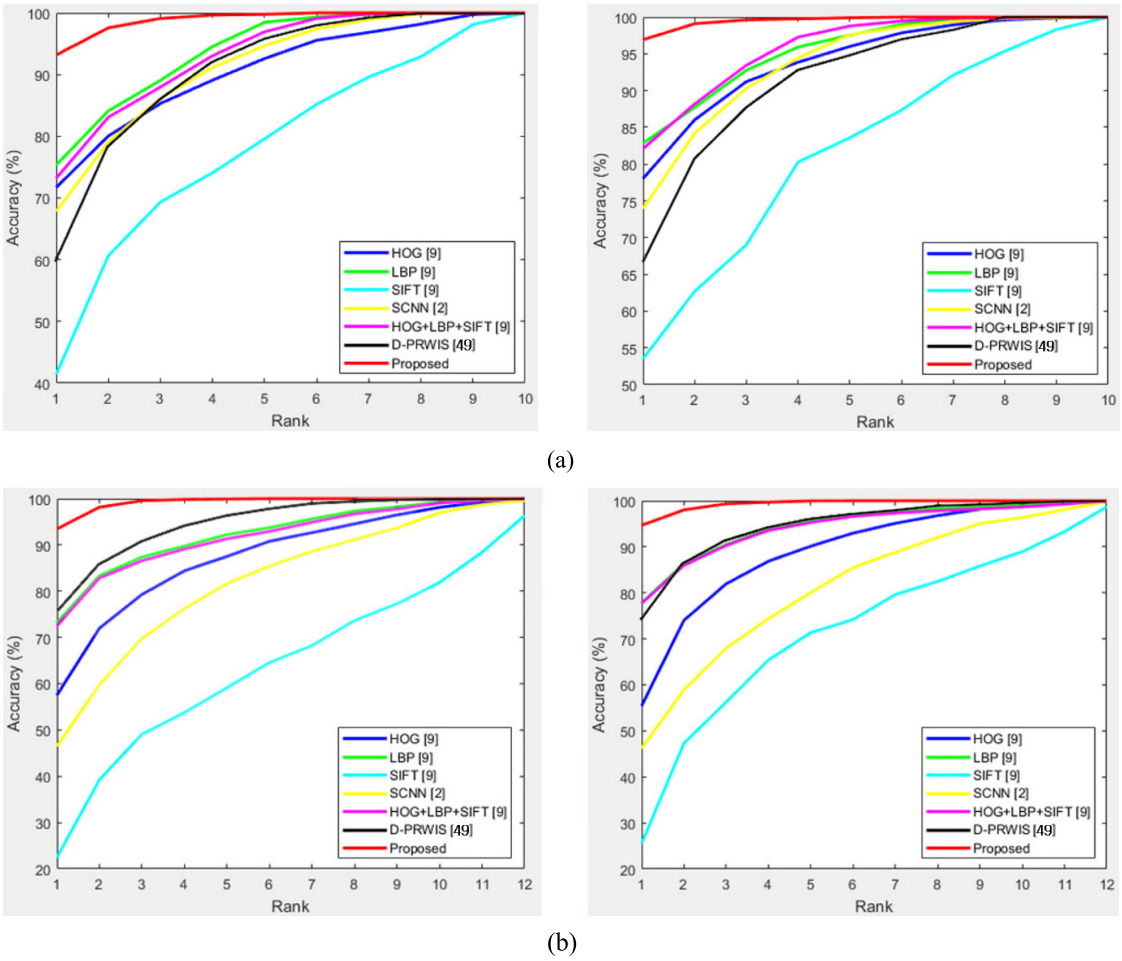

FIGURE 13. CMC curves of recognition. (a) DP-DB1 database, and (b) ChokePoint database. In (a) ∼ (d), left and right figures show the results by the L-L and R-R periocular recognition, respectively.

TABLE 7. Comparison of periocular recognition accuracies by proposed method with those by ResNet-50.

Method	EER (%)				Average
	L-L		R-R		
	Sub-dataset1	Sub-dataset2	Sub-dataset1	Sub-dataset2	
ResNet-50 model fine tuning [22]	6.88	11.96	3.68	6.38	7.23
VGG face-16 model fine tuning **(proposed method)**	4.56	8.76	3.89	4.47	**5.42**

shortcuts to preserve the high frequency component of the feature map of the previous layer as much as possible when performing training [22]. This is a good method when the focus of the input images is good enough to make clear distinctions, but it can be seen that the shortcuts can have negative effects when there are obstructive factors such as blur, etc. in low resolution images captured in a surveillance environment. In [44], several kinds of CNN architectures were compared and tested using blur images, and the results of an experiment comparing VGG-16 fine tuning and ResNet-50 fine tuning showed that VGG-16 fine tuning was better, also.

5) COMPARISONS WITH OCCLUDED FACE IMAGES

In the last experiment, periocular recognition was compared to face recognition when there were occlusions present in order to understand what benefits periocular recognition has in a surveillance environment. In the face recognition, the DP-DB1 database was used, and a two-fold cross validation method was performed. An occlusion at the bottom part of the face was created, as in Fig. 14 (b), and a focus score was applied to the face, just as in the proposed method. Face images which had a score higher than the threshold were used for recognition. The VGG face-16 pretrained model provided by [35] was used on these images without fine tuning to

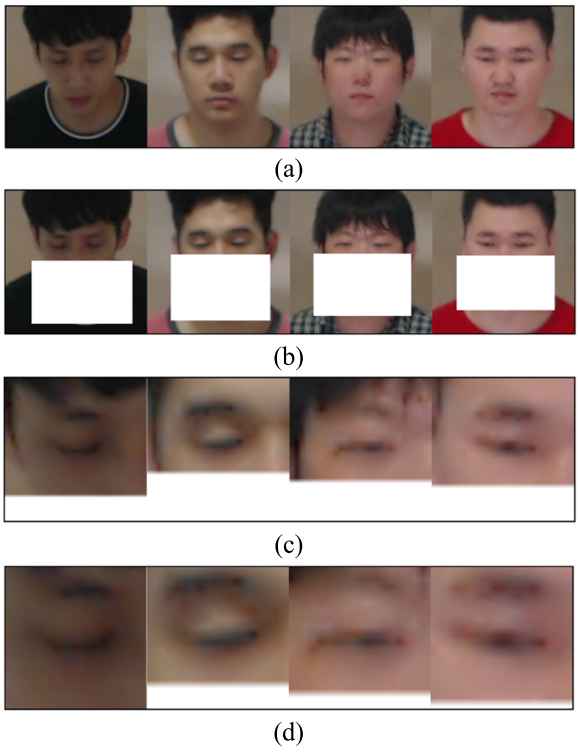

FIGURE 14. Occluded face and periocular images by mask. (a) No occluded face, (b) face occluded by mask, (c) loose ROI periocular occluded by mask, and (d) tight ROI periocular occluded by mask.

TABLE 8. Comparisons of accuracies by face and periocular recognition.

Type		EER (%)		
		Sub-dataset1	Sub-dataset2	Average
Face	No occlusion	0.94	1.35	1.15
	Occlusion by mask	10.52	12.68	11.60
Periocular	No occlusion	4.23	6.62	5.43
	Occlusion by mask	5.36	8.05	6.71

extract 4096 Fc6 features. The feature normalization method presented in Section III.C.3 was applied. The geometry center for each class was calculated, and gallery features were selected. Then, Euclidean distance was calculated between the gallery features and probe features, and the EER was measured through authentic and imposter matching.

Fig. 14 (a) shows face images without occlusions, and Fig. 14 (b) shows face images with occlusions. In addition, Fig. 14 (c) shows the periocular images of loose ROI, which are more affected by occlusions than those of tight ROI shown in Fig. 14 (d). However, relatively little information is lost by occlusion compared to the face images. As can be seen in Fig. 14 (d), occlusions had the smallest effect on the tight ROI.

Table 8 shows the results of the experiments comparing the EER of the face and periocular methods when an occlusion is present. Without the occlusion, the face recognition showed better performance than periocular recognition. However, when an occlusion occurred at the lower part of the face, the performance of face recognition declined rapidly whereas periocular recognition experienced a slight decline. The reason for this is that the occlusion on the face caused the loss of a fairly large portion of information in the face region, including the nose, mouth, jaw line, etc. However, in the case of the periocular images, a slight loss of the skin region occurred in some pixels at the bottom part, but outside of this, most of the important information such as the eyebrows, eye shape, eye color, etc. remained. Therefore, the performance decline was very small, and the EER was lower than that by face recognition. Through this experiment, it was found that periocular recognition can be used in a surveillance environment not just as a recognition method added to a multimodal method, but as an alternative to facial recognition in case of occlusions.

V. CONCLUSION

In this study, CNN-based periocular recognition was proposed using the images captured in a surveillance environment. Experiments were performed using not only a custom-made database but also an open database, and the proposed method showed better performance compared to methods from existing researches. In addition, when there were occlusions on the bottom part of the face, the performance degradation by periocular recognition is lower than that by face recognition. Through this study, it was confirmed that periocular recognition can be used in a surveillance environment not just as a recognition method added to a multimodal method, but as the main biometric information, and it can be used as an alternative to face recognition when occlusions occur on the face. In addition, an error analysis showed that recognition errors occurred when several complex factors degrading performance (posture changes, occlusions, closed eyes, etc.) occur at the same time in images. It is expected that there would be a need to verify this method in a surveillance environment with more extreme illumination changes such as an outdoor environment.

As such, future research directions include using an additional pose compensation CNN to deal with the error cases, using a restoration algorithm to deal with blur images more completely, and using a method which combines NIR or thermal images.

VI. ACKNOWLEDGMENT

Portions of the research in this paper use the ChokePoint collected by the National ICT Australia Ltd. (NICTA).

REFERENCES

[1] R. Dellana and K. Roy, "Data augmentation in CNN-based periocular authentication," in *Proc. 6th Int. Conf. Inf. Commun. Manage.*, Hatfield, U.K., Oct. 2016, pp. 141–145.

[2] Z. Zhao and A. Kumar, "Accurate periocular recognition under less constrained environment using semantics-assisted convolutional neural network," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 5, pp. 1017–1030, May 2017.

[3] J. R. Lyle, P. E. Miller, S. J. Pundlik, and D. L. Woodard, "Soft biometric classification using periocular region features," in *Proc. 4th IEEE Int. Conf. Biometrics, Theory, Appl., Syst.*, Washington, DC, USA, Sep. 2010, pp. 1–7.

[4] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Colorado Springs, CO, USA, Jun. 2011, pp. 74–81.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, USA, Jun. 2005, pp. 886–893.

[7] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[8] U. Park, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum: A feasibility study," in *Proc. 3rd IEEE Int. Conf. Biometrics, Theory, Appl., Syst.*, Washington, DC, USA, Sep. 2009, pp. 1–6.

[9] U. Park, R. R. Jillela, A. Ross, and A. K. Jain, "Periocular biometrics in the visible spectrum," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 1, pp. 96–106, Mar. 2011.

[10] J. Adams, D. L. Woodard, G. Dozier, P. Miller, K. Bryant, and G. Glenn, "Genetic-based type II feature extraction for periocular biometric recognition: Less is more," in *Proc. 20th IEEE Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 2010, pp. 205–208.

[11] C. N. Padole and H. Proenca, "Periocular recognition: Analysis of performance degradation factors," in *Proc. 5th Int. Conf. Biometrics*, New Delhi, India, Mar./Apr. 2012, pp. 439–445.

[12] B. J. Kang and K. R. Park, "A robust eyelash detection based on iris focus assessment," *Pattern Recognit. Lett.*, vol. 28, no. 3, pp. 1630–1639, 2007.

[13] S. Bharadwaj, H. S. Bhatt, M. Vatsa, and R. Singh, "Periocular biometrics: When iris recognition fails," in *Proc. 4th IEEE Int. Conf. Biometrics, Theory, Appl., Syst.*, Washington, DC, USA, Sep. 2010, pp. 1–6.

[14] A. Joshi, A. K. Gangwar, and Z. Saquib, "Person recognition based on fusion of iris and periocular biometrics," in *Proc. 12th Int. Conf. Hybrid Intell. Syst.*, Pune, India, Dec. 2012, pp. 57–62.

[15] G. Mahalingam and K. Ricanek, Jr., "LBP-based periocular recognition on challenging face datasets," *EURASIP J. Image Video Process.*, vol. 2013, no. 1, pp. 1–13, Dec. 2013.

[16] S. Bakshi, P. K. Sa, and B. Majhi, "Phase intensive global pattern for periocular recognition," in *Proc. Annu. IEEE India Conf.*, Pune, India, Dec. 2014, pp. 1–5.

[17] P. E. Miller, A. W. Rawls, S. J. Pundlik, and D. L. Woodard, "Personal identification using periocular skin texture," in *Proc. ACM Symp. Appl. Comput.*, Sierre, Switzerland, Mar. 2010, pp. 1496–1500.

[18] P. E. Miller, J. R. Lyle, S. J. Pundlik, and D. L. Woodard, "Performance evaluation of local appearance based periocular recognition," in *Proc. 4th IEEE Int. Conf. Biometrics, Theory, Appl., Syst.*, Washington, DC, USA, Sep. 2010, pp. 1–6.

[19] D. L. Woodard, S. J. Pundlik, J. R. Lyle, and P. E. Miller, "Periocular region appearance cues for biometric identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 162–169.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1097–1105.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, May 2015, pp. 1–14.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[23] F. Juefei-Xu and M. Savvides, "Unconstrained periocular biometric acquisition and recognition using COTS PTZ camera for uncooperative and non-cooperative subjects," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Breckenridge, CO, USA, Jan. 2012, pp. 201–208.

[24] F. Alonso-Fernandez and J. Bigun, "Periocular recognition using retinotopic sampling and Gabor decomposition," in *Proc. Eur. Conf. Comput. Vis. Workshops Demonstrations*, Florence, Italy, Oct. 2012, pp. 309–318.

[25] D. L. Woodard, S. Pundlik, P. Miller, R. Jillela, and A. Ross, "On the fusion of periocular and iris biometrics in non-ideal imagery," in *Proc. 20th IEEE Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 2010, pp. 201–204.

[26] R. Jillela and A. Ross, "Mitigating effects of plastic surgery: Fusing face and ocular biometrics," in *Proc. 5th IEEE Int. Conf. Biometrics, Theory, Appl., Syst.*, Arlington, VA, USA, Sep. 2012, pp. 402–411.

[27] K. B. Raja, R. Raghavendra, and C. Busch, "Binarized statistical features for improved iris and periocular recognition in visible spectrum," in *Proc. 2nd IEEE Int. Workshop Biometrics Forensics*, Valletta, Malta, Mar. 2014, pp. 1–6.

[28] Ş. Karahan, A. Karaöz, Ö. F. Özdemir, A. G. Gü, and U. Uludağ, "On identification from periocular region utilizing SIFT and SURF," in *Proc. 22nd Eur. Signal Process. Conf.*, Lisbon, Portugal, Sep. 2014, pp. 1392–1396.

[29] J. M. Smereka and B. V. K. V. Kumar, "What is a 'good' periocular region for recognition?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Portland, OR, USA, Jun. 2013, pp. 117–124.

[30] G. Santos and E. Hoyle, "A fusion approach to unconstrained iris recognition," *Pattern Recognit. Lett.*, vol. 33, pp. 984–990, Jun. 2012.

[31] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[32] *ChokePoint Database*. Accessed: Feb. 21, 2018. [Online]. Available: http://arma.sourceforge.net/chokepoint/

[33] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1867–1874.

[34] E. Luz, G. Moreira, L. A. Z. Junior, Jr., and D. Menotti, "Deep periocular representation aiming video surveillance," *Pattern Recognit. Lett.*, to be published.

[35] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, Swansea, U.K., Sep. 2015, pp. 1–12.

[36] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, Jun. 2010, pp. 807–814.

[37] *CS231N Convolutional Neural Networks for Visual Recognition*. Accessed: Dec. 26, 2017. [Online]. Available: http://cs231n.github.io/convolutional-networks/#overview

[38] H. G. Hong, M. B. Lee, and K. R. Park, "Convolutional neural network-based finger-vein recognition using NIR image sensors," *Sensors*, vol. 17, no. 6, p. 1297, 2017.

[39] *Logitech BCC 950 Camera*. Accessed: Dec. 26, 2017. [Online]. Available: https://www.logitech.com/en-roeu/product/conferencecam-bcc950

[40] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, Fort Lauderdale, FL, USA, Apr. 2011, pp. 315–323.

[41] *SCNN Codes*. Accessed: Dec. 30, 2017. [Online]. Available: http://www.comp.polyu.edu.hk/~csajaykr/scnn.rar

[42] P. J. Phillips *et al.*, "Overview of the face recognition grand challenge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, pp. 947–954.

[43] *Face and Ocular Challenge Series (FOCS) Database*. Accessed: Dec. 30, 2017. [Online]. Available: http://www.nist.gov/itl/iad/ig/focs.cfm

[44] S. Dodge and L. Karam, "A study and comparison of human and deep learning recognition performance under visual distortions," in *Proc. 26th Int. Conf. Comput. Commun. Netw.*, Vancouver, BC, Canada, Jul./Aug. 2017, pp. 1–7.

[45] *Dongguk Periocular Database (DP-DB1) With CNN Models and Algorithms*. Accessed: Mar. 29, 2018. [Online]. Available: http://dm.dgu.edu/link.html

[46] F. Alonso-Fernandez and J. Bigun, "A survey on periocular biometrics research," *Pattern Recognit. Lett.*, vol. 82, pp. 92–105, Oct. 2016.

[47] A. G. Howard *et al.* (Apr. 2017). "MobileNets: Efficient convolutional neural networks for mobile vision applications." [Online]. Available: https://arxiv.org/abs/1704.04861

[48] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jun. 2014.

[49] H. Proença and J. C. Neves, "Deep-PRWIS: Periocular recognition without the iris and sclera using deep learning frameworks," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 4, pp. 888–896, Apr. 2018.

MIN CHEOL KIM received the B.S. degree in avionics from Hanseo University, Seosan, South Korea, in 2015. He has been pursuing the M.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, since 2016. He designed the periocular recognition system based on CNN and analyzed results of experiments. His research interests include biometrics and pattern recognition.

JA HYUNG KOO received the B.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2016, where he has been pursuing the combined course of M.S. and Ph.D. degree in electronics and electrical engineering since 2016. He helped to experiments and collecting databases. His research interests include biometrics and pattern recognition.

SE WOON CHO received the B.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2017, where he has been pursuing the combined course of M.S. and Ph.D. degrees in electronics and electrical engineering since 2017. He helped to experiments and collecting databases. His research interests include biometrics and pattern recognition.

NA RAE BAEK received the B.S. degree in electronics and electrical engineering from Dongguk University, Seoul, South Korea, in 2017, where she has been pursuing the combined course of M.S. and Ph.D. degrees in electronics and electrical engineering since 2017. She helped to experiments and collecting databases. Her research interests include biometrics and pattern recognition.

KANG RYOUNG PARK received the B.S. and M.S. degrees in electronic engineering and the Ph.D. degree in electrical and computer engineering from Yonsei University, Seoul, South Korea, in 1994, 1996, and 2000, respectively. He has been a Professor with the Division of Electronics and Electrical Engineering, Dongguk University, since 2013. His research interests include image processing and biometrics.