Journal of Korea Multimedia Society Vol. 20, No. 2, February 2017(pp. 170-178)
https://doi.org/10.9717/kmms.2017.20.2.170

Multimodal Face Biometrics by Using Convolutional Neural Networks

Leslie Ching Ow Tiong†, Seong Tae Kim††, Yong Man Ro†††

ABSTRACT

Biometric recognition is one of the major challenging topics which needs high performance of recognition accuracy. Most of existing methods rely on a single source of biometric to achieve recognition. The recognition accuracy in biometrics is affected by the variability of effects, including illumination and appearance variations. In this paper, we propose a new multimodal biometrics recognition using convolutional neural network. We focus on multimodal biometrics from face and periocular regions. Through experiments, we have demonstrated that facial multimodal biometrics features deep learning framework is helpful for achieving high recognition performance.

Key words: Multimodal Biometrics Recognition, Face Recognition, Convolutional Neural Networks

1. INTRODUCTION

Majority of biometric system applications have been unimodal, which relies on a single source of biometric to achieve recognition. In [1], the authors introduced multimodal biometrics, which combined two or more biometric sources in order to prevent unacceptable error rates of unimodal biometric recognition and spoofing attack. In other words, the basic concept is to use multiple biometric sources to satisfy authentication and identification requirements.

Traditionally, the usage of different biometrics may increase recognition accuracy and precision. In particular, by obtaining dual features such as face and periocular regions from the same source, one can obtain more favorite results. This is a quick way of improving system performance, as this does not deploy multiple sensors alongside with additional feature extraction.

Convolutional Neural Network (CNN) is becoming mainstream in pattern and image recognition in recent years. Due to the uniqueness in its architecture, it performs segmentation, feature extraction and classification in one module, which has widely been used in a variety of areas, including face detection [2], face recognition [3], speech recognition [4], video analysis [5], and so forth. Therefore, using CNN typically leads to better performance for image and speech processing with the introduction of higher computation complexity requirements.

In this paper, we investigate face and periocular regions which are obtained from same source. Generally, the face is the most expressive biometric feature, and hence, it is a challenge to efficiently extract the identity information. Face recognition measures nodal points on the face, the distance between the eyes, and other distinguishable features by reorganizing an individual's face. The

※ Corresponding Author : Yong Man Ro, Address: (34141) KAIST 291, Daehak-ro, Yuseong-gu, Daejeon, Korea, TEL : +82-42-350-3494, FAX : +82-42-350-5494, E-mail : ymro@kaist.ac.kr
Receipt date : Jan. 12, 2017, Approval date : Jan. 20, 2017
† Dept. of Global Information and Telecommunication Technology, KAIST (E-mail : tiongleslie@kaist.ac.kr)
†† School of Electrical Engineering, KAIST (E-mail : stkim4978@kaist.ac.kr)
††† School of Electrical Engineering, KAIST
※ This work was partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No.2015R1A2A2A01005724).

identity challenge is often related to aging, camera pose, and illumination [6].

However, in recent years, the periocular region as a facial local region in the surrounding area of the eye, has been identified as an extended form of iris recognition, and become an active research area. The feature has made rapid strides in the feasibility as a biometric trait of an individual [7]. Several research studies have proven that the periocular is one of the several discriminating features in the face [8], [9]. The shape and texture of the periocular can also be extracted as individual biometric features, which differs from the entire face in order to provide an additional feature in recognition performance.

In this work, we propose facial multimodal biometrics recognition where multimodal biometrics are used as core components. The core components are used as inputs for CNN to take advantage of both features for feature learning. The contributions of this paper can be summarized as follows:

- The effectiveness of multimodal biometrics is investigated through latent features formed in CNN. Our experimental results show that latent features in CNN from multimodal biometrics could provide complementary information.
- Multimodal biometrics features are extracted by an effective learning with CNN, which significantly improves the recognition accuracy. We have observed that the latent feature from periocular is useful in providing additional information to achieve higher performance under variations including poses, illumination, and appearance.

This paper is structured in the following way. Section 2 reviews the background and related work in multimodal biometrics recognition. Section 3 provides the presentation of our approach of using multimodal biometrics with CNN architecture for recognition. Section 4 describes the presentation of experimental results and analysis. The conclusion and future work of the proposed approach conclude this paper.

2. BACKGROUND AND RELATED WORK

Several approaches have been applied to improve the performance of biometrics recognition whenever it is affected by variability including illumination, appearance, and spoofing attacks. In [10], the authors addressed these limitations of unimodal biometric involving a large number of classes. This was motivated by the combination of multiple biometric traits to design new biometric systems for improving the performance of recognition and deterred spoofing, which is known as multimodal biometrics.

Multimodal biometrics consist of multiple units, sensors, biometrics, snapshots, and matchers. In this paper, we will focus only on the multiple units of the same biometric source and the performance of recognition. In [1], the authors conducted an experiment by combining multiple units of the same biometric characteristics, which was directly related to the improvement in performance of the combination. In addition, the performance depended on the decision level, which includes majority voting and behavior knowledge.

According to Chang et al. [11], Kang and Park [12], the limitations of unimodal biometric observation could be caused by several conditions, including the sensors and the individual's appearance. Thus, they proposed a new multimodal biometric approach at the score level, which increased the performance and efficiency in authenticating and recognizing an individual. They concluded that multimodal biometrics is more secure and robust when compared to unimodal biometric systems.

[13] studied the feasibility of periocular recognition under various circumstances. The authors highlighted the benefits of periocular biometric offered the eyes' shapes information to achieve high accuracy, especially compared to iris recognition. [8] also demonstrated the combination of face and

periocular features for authenticating individuals. The experimental results showed their proposed authentication method is more robust and more convenient than others.

Another research which conducted by [9], discovered that maximizing the correlation of face and periocular, which helped in increasing the performance and precision. Although face region comprises of periocular features, [8] and [9] justified that the periocular feature can be used independently as individual biometric due to its uniqueness in texture information, minimizing variations in aging, camera pose, and illumination. Another interesting work [14] demonstrated that periocular included the eyebrows, left and right eyes information, and local features effectiveness, which had accomplished high accuracy recognition. The high accuracy performance was achieved by fusion with face and periocular biometrics, but also minimized computational complexity.

In recent years, CNN is increasingly used in computer vision field with the great successes in image and video recognition. CNN is a feed-forward network with the ability to extract features from the input images, and then classifying the extracted features. [15] proposed using CNN to classify 1.2 million high-resolution images, thereby achieving impressive results to learn thousands of objects from millions of images. [16] also conducted experiments in image classifications. Their experimental results achieved the state-of-the-art on the classification tasks. We, therefore, conclude that CNN is outperforming in the recognition pipelines, which is built with visual representations for the classification.

In order to achieve better accuracy for reorganizing an individual, we proposed a new approach with the combination of face and periocular latent features via CNN. To the best of our knowledge, combining the face and the periocular latent feature through CNN in a multimodal way has never been attempted.

3. PROPOSED FRAMEWORK OF MULTI-MODAL BIOMETRICS RECOGNITION VIA CNN

This section describes our proposed approach by combining face and periocular as the core components of the CNN to recognize an individual as depicted in Fig. 1. We first denote the face as f and its periocular as p for the input features:

$$f \epsilon R^{(h_f \times w_f)} \tag{1}$$

and

$$p \epsilon R^{(h_p \times w_p)} \tag{2}$$

where h_f and w_f are the height and width are the height and width of face image and h_p and w_p are the height and width of periocular image, respectively. The inputs f and p were presented in

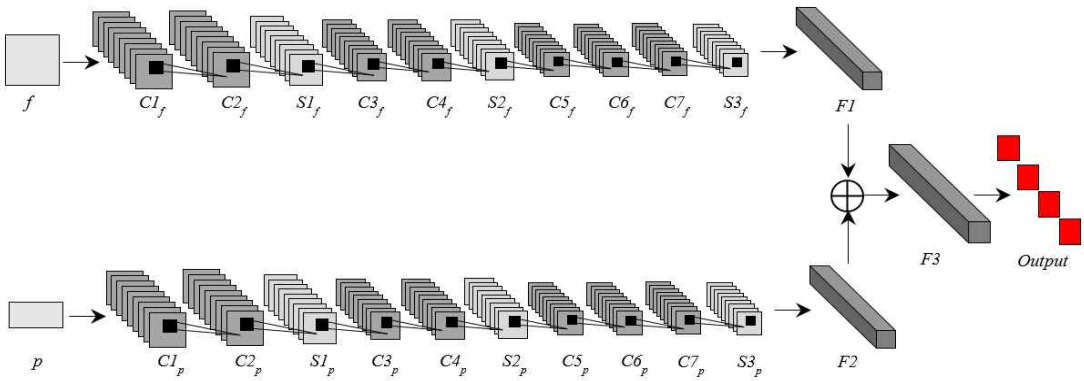


Fig. 1. The structure design of proposed CNN architecture.

RGB color space.

We also investigate the behavior and correlation between the face and periocular features for CNN in feature learning. The proposed approach (see Fig. 1) is designed with the same principles, as inspired by [16] and [17]. The baseline consisted of seven convolutional layers ($C1_f$, $C2_f$, $C3_f$, $C4_f$, $C5_f$, $C6_f$, and $C7_f$) for f and ($C1_p$, $C2_p$, $C3_p$, $C4_p$, $C5_p$, $C6_p$, and $C7_p$) for p; three subsampling layers for both f ($S1_f$, $S2_f$, and $S3_f$) and p ($S1_p$, $S2_p$, and $S3_p$), respectively; and four fully connected layers (F1 as f, F2 as p, a merged layer, and an output layer with softmax function).

The proposed CNN architecture was designed to learn multimodal biometrics, in order to achieve an increased accuracy in biometric recognition along with the variability of effects. Let the feature map in the convolution layer C be denoted as Y, which is computed as:

$$Y_a^{(l)} = B_a^{(l)} + \sum_{a=1}^{m} Y_a^{(a-l)} * K_{a,b}^{(l)} \tag{3}$$

where $Y_a^{(l)}$ denotes the ath feature map in the lth C layer with the input f or p consisting of RGB channels. $B_a^{(l)}$ is defined as bias matrix, $K_{a,b}^{(l)}$ as the filter, and m is the number of feature maps in the lth layer. Layer subsampling is used to reduce the resolution of features. The fully-connected layers (F1 and F2) are convolution layers with 1×1 convolution kernels and a full connection table. By concatenating all the features, the merged layer (F3) is formed as follows:

$$y = \{y_{F1}, y_{F2}\} \epsilon R^{(n \times d)} \tag{4}$$

where y denotes output of concatenated layer with $1 \times d$ dimension.

In the recognition stage, the output of F3 layer produced N class labels, which can be defined as:

$$\sigma(z) = \frac{\exp(z_c)}{\sum_{c=1}^{K} \exp(z_c)} \tag{5}$$

where z is the output of the predictions; $exp(.)$ is defined as exponential function; and K as the dimension of z. The function limits the range of predicted output to values as $0 \leq \sigma \leq 1$.

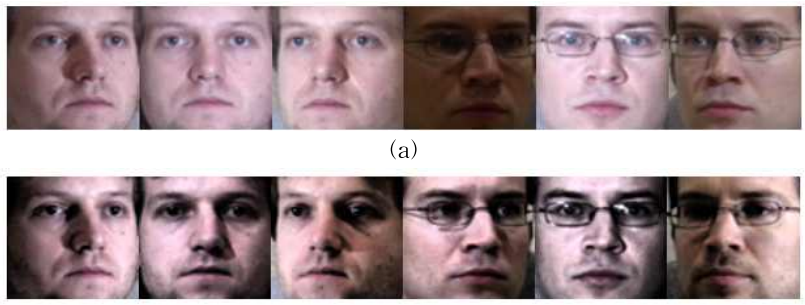
The feature representation processes of f and p are represented with and without image processing, which could lead to complementary effects in our experiment. An image processing technique, color-based Histogram Equalization (HE) was implemented based on the theory of [18]. The RGB image was converted into HSL color space in order to obtain the lightness (L) histogram channel. Then by applying HE in the L channel:

$$P(I_u^L) = \frac{n_u}{n} \tag{6}$$

and

$$CDF(I) = \sum_{v=0}^{l} P(I_v) \tag{7}$$

where $P(I_u^L)$ is defined as the histogram value of L channel; I as the probability of an occurrence of a pixel; and n_u as the number of occurrences of level u. $CDF(.)$ is defined as accumulated normal-



(a)

(b)

Fig. 2. Sample images of f (a) RGB color and (b) color-based HE.

ized histogram function to equalize the histogram value independently in the range $0 \leq v \leq 1$. Fig. 2 demonstrates several samples of the RGB images and the color-based HE images for input f, where both were used as the input features in our experiments.

4. EXPERIMENTAL RESULT AND DISCUSSION

4.1 Experiment Design

In order to verify our proposed methods, several experiments were conducted using unimodal and multimodal biometric features via CNN. The input features of biometrics were RGB color space and color-based HE. Table 1 describes the architecture design of CNN with unimodal and multimodal biometrics features. For instance, 64 feature maps were defined with size 64×64, which were denoted as 64@64×64. Each subsampling layer was applied to reduce the size of the previous output. For instance, subsampling layer $S1_f$ is denoted as 64@32×32.

We also applied unimodal biometrics via the Sparse Representation Classifier (SRC) algorithm since it had received a lot of attentions in the past few years with great successes in biometric recognition. In our experiment, the algorithm was implemented based on [19]. Therefore, this paper would not describe this algorithm in detail as it was covered in [19].

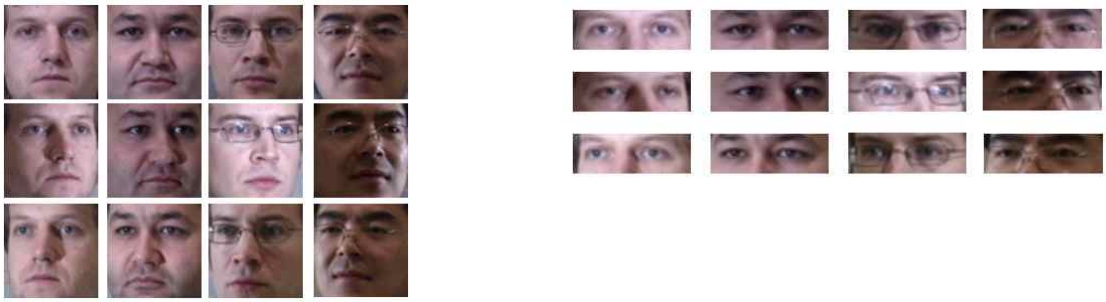
4.2 Result Analysis and Discussion

In our experiments, we first selected the Multi-PIE dataset [20] to train and validate our proposed approach. The input images of the f and p, which is shown in Fig. 3, contained 246 classes. Each class had 60 images for training, which contained frontal, left, and right poses; 10 images for each pose as validation. All the images were selected randomly with 'neutral' form under several variations in illumination and appearances. The periocular images were extracted directly from the same source as the face images.

We presented a multimodal biometrics recognition (see Fig. 1) with CNN, which achieved very competitive accuracy using the Multi-PIE dataset. In the majority of our experiments, our proposed approach has outperformed the baseline as shown in Fig. 4. The comparison indicated that our proposed approach with RGB color space, multimodal biometrics recognition via CNN outperformed the other approaches by achieving 98.5% (front pose), 98.09% (left pose), and 97.97% (right pose), respectively. Furthermore, our proposed approach

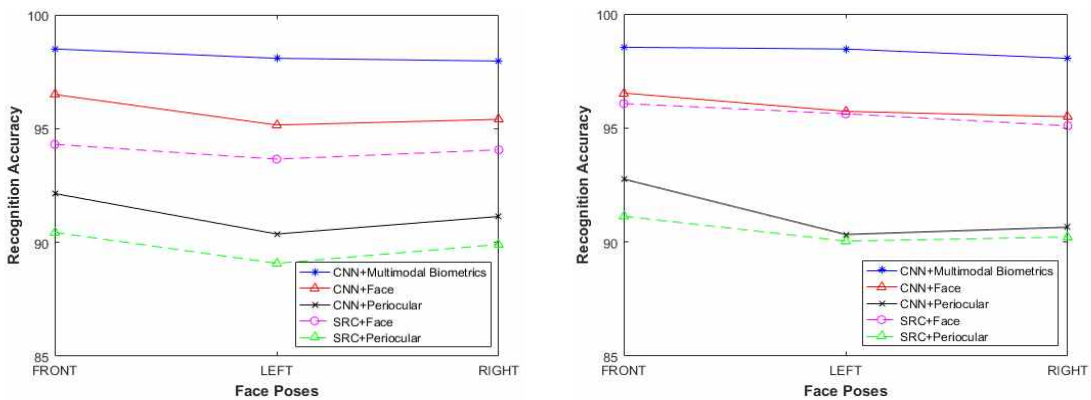
Table 1. Configurations of the proposed CNN architecture with unimodal and multimodal biometrics features.

Face	Periocular	Multimodal Biometrics (Face+Periocular)	
$C1_f$: 64@64×64 $C2_f$: 64@64×64	$C1_p$: 64@20×60 $C2_p$: 64@20×60	$C1_f$: 64@64×64 $C2_f$: 64@64×64	$C1_p$: 64@20×60 $C2_p$: 64@20×60
$S1_f$: 64@32×32	$S1_p$: 32@10×30	$S1_f$: 64@32×32	$S1_p$: 32@10×30
$C3_f$: 128@32×32 $C4_f$: 128@32×32	$C3_p$: 64@10×30 $C4_p$: 64@10×30	$C3_f$: 128@32×32 $C4_f$: 128@32×32	$C3_p$: 64@10×30 $C4_p$: 64@10×30
$S2_f$: 128@16×16	$S2_p$: 128@5×15	$S2_f$: 128@16×16	$S2_p$: 128@5×15
$C5_f$: 256@16×16 $C6_f$: 256@16×16 $C7_f$: 256@16×16	$C5_p$: 64@5×15 $C6_p$: 64@5×15 $C7_p$: 64@25×15	$C5_f$: 256@16×16 $C6_f$: 256@16×16 $C7_f$: 256@16×16	$C5_p$: 64@5×15 $C6_p$: 64@5×15 $C7_p$: 64@25×15
$S3_f$: 256@8×8	$S3_p$: 256@2×7	$S3_f$: 256@8×8	$S3_p$: 256@2×7
$F1_f$	$F2_p$	$F1_f$	$F2_p$
		Merged- F3	
Output	Output	Output	



(a) (b)

Fig. 3. Samples of biometrics from Multi-PIE dataset (a) face features and (b) periocular features.



(a) (b)

Fig. 4. Experimental results of unimodal and multimodal biometrics recognition via SRC and proposed approach using (a) RGB color space, and (b) color-based HE.

using color-based HE also outperformed the other approaches by achieving 98.54% (front pose), 98.46 % (left pose), and 98.05% (right pose), respectively.

Furthermore, Receiver Operating Characteristics (ROC) curves computed from the test data which implemented the multimodal biometrics features, are shown in Fig. 5. These curves demonstrate that the effectiveness of our proposed approach is better than other methods. Fig. 5 shows that multimodal biometric features perform better than using unimodal biometric, and our proposed approach significantly outperforms in Fig. 5a and 5b, achieving the best results.

These results proved that the combination of both features further boosted the recognition performance. Our approach also addresses that the different kinds of poses recognition between the genuine and imposter, which can, therefore, be enhanced by periocular. This is because periocular contained hidden texture information such as the eyes' shapes and eyebrows information. Thus, the proposed approach has achieved effective recognition performance based on features learning.

In order to verify the generalization ability of our approach, Table 2 presents a comparison of methods used by our proposed approach with other state-of-the-art approaches. [21] implemented unimodal biometric via CNN, and the approach achieved 95.7%. Both [22] and [23] applied the SRC algorithm with facial biometrics and achieved a recognition accuracy of 96.3% and 92.2%, respectively. In [24], multiple features based on facial images were used as input parameters to design the dictionary, and the verification performance

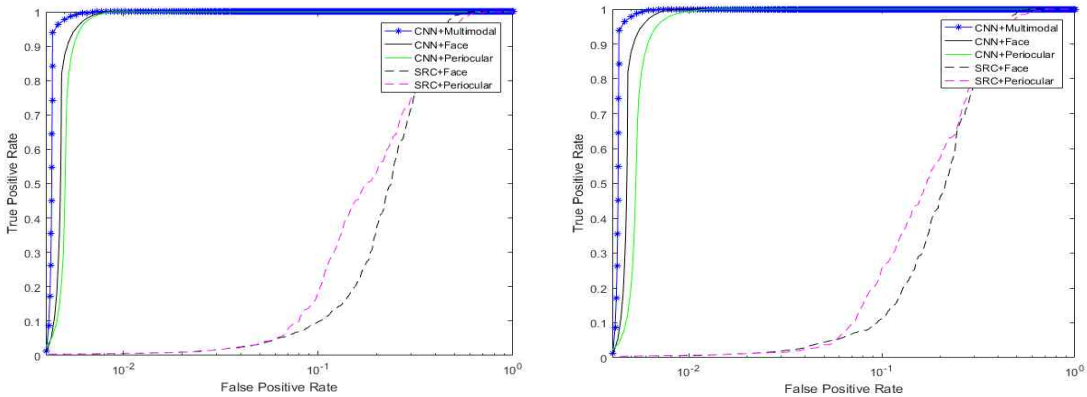


Fig. 5. Performance of ROC curves (a) using colour-based HE to process input images for CNN and HE to process images for SRC, and (b) using RGB images for CNN and greyscale images for SRC.

achieved 95.19% accuracy.

Since the experimental conditions were different, it is, therefore, impossible to draw absolute conclusions through the comparisons. As shown in Table 2, our proposed approach using CNN with multimodal biometrics is competitive with respect to [21] and [22], also significantly outperforming than [23] and [24] by achieving 3% to 4% improvements in recognition accuracy. As a result, this is a remarkable achievement considering our best result is achieved by combining multimodal biometrics features. In addition, our experiments demonstrated that our proposed multimodal biometric approach is beneficial for the improved classification accuracy. Moreover, the state-of-the-art performance on the Multi-PIE dataset can be achieved using multimodal biometrics with CNN.

Our experimental results showed that unimodal biometric is unreliable due to the inconsistencies in the image features extraction and human behavior such as poses and variation in appearances. Thus, the periocular feature can maintain or even improve recognition accuracy. This implies that the periocular can be used independently as a standalone biometric. Also, it is capable of providing unique textural information to minimize variations due to aging, appearance differences, and illumination.

5. CONCLUSION

Unimodal biometric had a limitations due to unreliable features which could be caused from single biometric. In order to improve the recognition accuracy and precision, we proposed multimodal biometrics recognition using face and periocular regions. Comparative experiments showed that the proposed approach achieved better results compared to unimodal biometric. We validated our proposed model on the CNN architecture using multi-

Table 2. Comparison among proposed approach and previous studies

Approaches	Biometrics	Accuracy [%]
Proposed Approach	Multimodal (Face+Periocular)	98.35
[21]	Face	95.70
[22]	Face	96.30
[23]	Face	92.20
[24]	Face (Multiple LE + Comp)	95.20

modal biometric features and demonstrated its excellent generalization ability. In addition, our experiments also verified that the periocular feature is one of the discriminative features obtainable from the same facial image source, thereby increasing the recognition performance. This means that the periocular region could be used independently as a standalone biometric feature due to its uniqueness in its hidden texture information. Moreover, the periocular region could be useful at facial pose variation. For our future work, we intend to add different kinds of biometric features to enhance the performance of our proposed multimodal biometrics approach.

REFERENCE

[1] A. Ross and A.K. Jain, "Multimodal Biometrics: an Overview," *Proceeding of European Signal Processing Conference*, pp. 1221-1224, 2004.

[2] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A Convolutional Neural Network Approach for Face Identification," *Proceeding of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5325-5334, 2015.

[3] M.M. Kasar, D. Bhattacharyya, and T.H. Kim, "Face Recognition Using Neural Network: a Review," *International Journal of Security and Its Application*, Vol. 10, No. 3, pp. 81-100, 2016.

[4] O. Abdelhamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," *IEEE Transactions Audio, Speech, Language Processing*, Vol. 22, No. 10, pp. 1533-1545, 2014.

[5] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F.F. Li, "Large-Scale Video Classification with Convolutional Neural Networks," *Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1725-1732, 2014.

[6] Y Choi, H. Kim, and Y.M. Ro, "Local Feature Learning using Deep Canonical Correlation for Heterogeneous Face Recognition," *Journal of Korea Multimedia Society,* Vol. 19, No. 5, pp. 848-855, 2016.

[7] C.N. Padole and H. Proenca, "Periocular Recognition: Analysis of Performance Degradation Factors," *Proceeding of International Association of Pattern Recognition International Conference on Biometrics*, pp. 439-445, 2012.

[8] K.B. Raja, R. Raghavendra, R.R. Martin, and S. Christoph, "Fusion of Face and Periocular Information for Improved Authentication on Smartphones," *Proceeding of International Conference on Information Fusion*, pp. 2115-2120, 2015.

[9] N.S. Lakshmiprabha, "Fusing Face and Periocular Biometrics Using Canonical Correlation Analysis," *Proceeding of Computer Vision and Pattern Recognition*, pp. 1-5, 2016.

[10] S. Prabhakar and A.K. Jain, "Decision-Level Fusion in Fingerprint Verification," *Pattern Recognition*, Vol. 35, No. 4, pp. 861-874, 2002.

[11] K. Chang, K.W. Bowyer, S. Sarkar, and B. Victor, "Comparison and Combination of Ear and Face Images in Appearance-Based Biometrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 9, pp. 1160-1165, 2003.

[12] B.J. Kang and K.R. Park, "Multimodal Biometric Authentication Based on the Fusion of Finger Vein and Finger Geometry," *IET Computer Vision*, Vol. 4, No. 3, pp. 209-217, 2009.

[13] U. Park, R.R. Jillela, A. Ross, and A.K. Jain, "Periocular Biometrics in the Visible Spectrum: a Feasibility Study," *Proceeding of IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pp. 1-6, 2009.

[14] U. Park, R.R. Jillela, A. Ross, and A.K. Jain, "Periocular Biometrics in the Visible Spectrum," *IEEE Transactions on Information Forensics Security*, Vol. 6, No. 1, pp. 96-106,

2011.

[15] Y. Lecun and Y. Bengio, *Convolutional Networks for Images, Speech, and Time-Series*, The Handbook of Brain Theory and Neural Networks, MIT Press, Cambridge, MA, 1995.

[16] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Proceeding of International Conference on Learning Representations*, pp. 1-14, 2015.

[17] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S.Z. Li, et al., "When Face Recognition Meets with Deep Learning: an Evaluation of Convolutional Neural Networks for Face Recognition," *Proceeding of IEEE International Conference on Computer Vision*, pp. 384-392, 2015.

[18] Histogram Equalization, http://cache. freescale. com/files/dsp/doc/app_note/AN4318.pdf. (accessed Oct., 11, 2016).

[19] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust Face Recognition Via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 31, No. 2, pp. 210-227, 2009.

[20] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Visual and Computer Vision*, Vol. 28, No. 5, pp. 807-813, 2010.

[21] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-View Perceptron : a Deep Model for Learning Face Identity and View Representations," *Advanced in Neural Information Processing Systems*, pp. 1-9, 2014.

[22] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a Practical Face Recognition System: Robust Alignment and Illumination by Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 2, pp. 372-386, 2012.

[23] L. Zhang, M. Yang, X. Feng, and H. Kong, "Sparse Representation or Collaborative Representation: which Helps Face Recognition?," *Proceedings of International Conference on Computer Vision*, pp. 471-478, 2011.

[24] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face Recognition with Learning-Based Descriptor," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2707-2714, 2010.

Leslie Ching Ow Tiong

He is currently doing his Ph.D. in KAIST. He received his BSc (Hons) in Computer Science and MSc in Computer Science (by research) from Sunway University and Lancaster University. His research interests include biometrics recognition, biometric template protection, image processing, and deep learning.

Seong Tae Kim

He received the B.S. degree from Korea University in 2012 and the M.S. degrees from KAIST in 2014. He is currently pursuing the Ph.D. degree at KAIST. His research interest include machine learning, medical image analysis, and biometrics.

Yong Man Ro

He received the B.S. degree from Yonsei University and the M.S. and Ph.D. degrees from KAIST. He was a Researcher with Columbia University in 1987, and with the University of California at Irvine and KAIST from 1992 to 1995. He was a Research Fellow with the University of California at Berkeley and a Visiting Professor with the University of Toronto in 1996 and 2007, respectively. He is currently a Professor and the Chair of signals and systems group in the School of Electrical Engineering, KAIST, where he is directing the Image and Video Systems Laboratory. His research interests include image/video processing, 3D video, visual recognition, and visual data deep learning.