DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-470

Two-step Learning of Deep Convolutional Neural Network for Discriminative Face Recognition under Varying Illumination

Yeoreum Choi, Hyung-Il Kim, and Yong Man Ro [1]*; IVY Lab, School of Electrical Engineering, KAIST; Republic of Korea*

Abstract

In real-world face recognition (FR) scenario, illumination variation has been known to be a challenging problem because face appearance dramatically changes depending on the illumination conditions. In order to deal with this illumination variation effectively, an illumination-reduced feature learning method using deep convolutional neural network (DCNN) is proposed in this paper. It is motivated by the capability of deep learning that represents highly complicated nonlinear structures. Our learning method is mainly comprised of following two-steps: 1) learning illumination patterns for eliminating illumination effect and 2) learning for maximizing discriminative power of feature representation. Experimental results on CMU Multi-PIE database have demonstrated that the proposed method outperforms the previous works in terms of FR accuracy.

Introduction

Face recognition (FR) has received a great deal of attention for the wide range of applications (e.g., video surveillance security and biometric identification) [1], [2]. In spite of the recent progress, precise FR under uncontrolled environments is still a challenging problem. In particular, varying illumination condition including low light intensity and shading effect has been known to be a great challenging factor in FR [3]. Face appearance could be changed as illumination changes. So face appearance variation caused by illumination could surpass face identification variation between persons [3].

To deal with the varying illumination problem in FR, one of the typical approaches is photometric normalization-based method [4]. This approach is basically to classify a face into a specific identity after normalizing illumination effect. There have been several works with similar approach [5], [7]-[9]. Traditional image processing method such as histogram equalization [5] tries to compensate illumination changes by adjusting gray level distribution. Some researchers have suggested Lambertian reflectance model-based methods [7]-[9]. They have been motivated by the fact that facial parts (e.g., eyes, tips, etc.) are considered as relatively higher spatial frequencies, while the illumination parts are considered as low spatial frequencies [6]. The authors in [7] have proposed a multi-scale retinex algorithm that reduces the effect of illumination by dividing facial image with a smoothed version of facial image. In [8], the authors have proposed the GradientFace, which shows the gradient direction of face images. In [9], Weber-face extracts locally salient patterns using relative terms (i.e., relative intensity difference of a pixel against its neighbors and the intensity of current pixel) inspired by Webers law [10].

However, previous methods abovementioned have mainly two limitations. First, a varying illumination condition encountered in an uncontrolled environment could not be specifically modelled. Moreover, [11] have pointed out that, for an object with Lambertian reflectance, there are no discriminative functions which are invariant to illumination. Thus the existing methods could merely determine functions which are insensitive to illumination changes. Second, by normalizing face images (i.e., reducing illumination effect), facial appearance information (e.g., texture information) could be reduced. In particular, in [9], only edge information remains. So detailed facial texture information could disappear, which is important feature for discriminating persons. Consequently, it causes poor recognition performance due to the reduction of identity information.

A learning-based approach to minimize the illumination effect instead of photometric normalization models could be promising. Face images with varying illumination conditions are known to be distributed on highly complex nonlinear manifold [12]. A learning approach could handle the complicated nonlinearity and enhance discriminative power for the purpose of FR.

In this paper, we propose a robust FR method which addresses both learning illumination patterns and enhancing discriminative power of feature representations under varying illumination conditions. More specifically, motivated by the property of learning hierarchical nonlinear representation of a deep learning, we propose the two-step learning of deep convolutional neural networks (DCNN): 1) learning of illumination patterns and 2) enhancement of discriminative power. By learning DCNN in the two steps, the proposed method can learn a latent facial feature representation robust to illumination variation with enhanced discriminative power. The contributions of the paper are summarized as two folds:

- Through the first step learning, illumination effect on a face image is learned by a DCNN. By eliminating the learned illumination effect, face image could be robust to shadow and varying light intensity while texture information of face image could be preserved.
- For illumination-reduced face image from the first step, subsequent DCNN layers are trained to maximize variations between persons in the second step. Then DCNN is learnt to minimize the intra-class variation between features. Consequently, the proposed method make DCNN maximize discriminative power in terms of feature representation.

Experimental results with CMU Multi-PIE database [13] show that the proposed method successfully improves FR performance compared to the previous works under varying illumination conditions.

The rest of this paper is organized as follows: Section 2 describes the proposed two-step learning of DCNN with their training procedure in detail. In Section 3, we present and discuss

[1]Corresponding author (ymro@kaist.ac.kr)

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-470

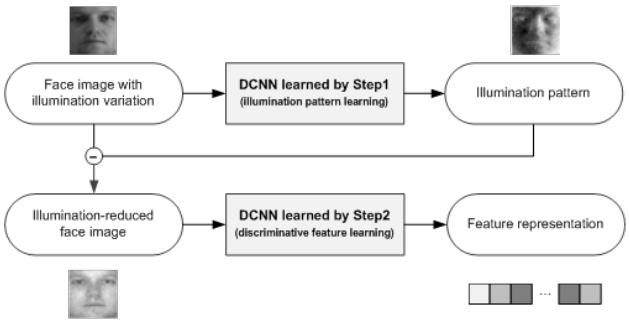


Figure 1. *FR method robust to illumination variations learned with the proposed two-step learning method.*

experimental results. Finally, Section 4 provides concluding remarks.

Proposed Two-step Learning of DCNN

As seen in Fig. 1, the proposed FR method is based on the learned DCNN model by two-step learning, which is comprised of: 1) learning of illumination patterns (Step 1) and 2) enhancement of discriminative power with the illumination-reduced face image (Step 2). In Step 1, for face images with a variety of illumination variations, latent illumination patterns are learned by the DCNN, i.e., mapping from input face images with illumination variations to illumination patterns. Then, by subtracting input face images with the learned illumination patterns, illumination-reduced face images are generated. Based on the illumination-reduced face images, subsequent DCNN is learned for enhancing discriminative power by considering both inter-class variation and intra-class variation. Note that even if some facial appearance information might be affected by reducing illumination variation, Step 2 could provide the enhancement of discriminative power in the feature subspace for FR. More details of the proposed method are described in the next subsections.

Step 1: Learning illumination patterns

The objective of Step 1 is to learn latent illumination patterns for face images with various illumination conditions in DCNN structure. In order to achieve the objective, the DCNN of Step 1 is learned by minimizing the similarity between input face image which has an illumination variation and expected illumination. The expected illumination for the input face image is estimated as difference from corresponding neutral illumination face image. An effective learning can be performed by mapping from the input face image to the expected illumination which the input face image has. Furthermore, since the proposed learning method in Step 1 takes only illumination effect, facial structural information could be preserved. More details of the learning in Step 1 are as follows.

Given a training face image $\mathbf{x}_{c,n}$ from the n-th image of the c-th class, the difference between $\mathbf{x}_{c,n}$ and the neutral illumination face image $\overline{\mathbf{x}}_c$ for the c-th class is computed as the expected illumination, i.e., $\mathbf{i}_{c,n} = \mathbf{x}_{c,n} - \overline{\mathbf{x}}_c$. Slight mis-alignment between $\mathbf{x}_{c,n}$ and $\overline{\mathbf{x}}_c$ could be negligible due to translational and rotational invariance property [14] of DCNN. Note that since the difference is taken for each subject, the identity information could be preserved. Then, based on the similarity between the feature $\mathbf{y}_{c,n}$ (i.e., the output of the last fully-connected layer in the DCNN) and the expected illumination of $\mathbf{i}_{c,n}$, the DCNN is learned in Step 1 by minimizing the following equation:

$$E_{\text{Step1}} = \frac{1}{2}\sum_{c,n} g\left(\|\mathbf{y}_{c,n} - \mathbf{i}_{c,n}\|^2\right), \tag{1}$$

where the function $g(z)$ is defined as $g(z) = (1/\beta)\log(1 + \exp(\beta z))$ with a sharpness parameter β, which function is known as the smoothed approximation of $[z]_+ = \max(0,z)$ [15]. After learning the DCNN based on Eq. (1), illumination pattern (denoted as $\Lambda(\mathbf{y}_{c,n})$) could be obtained by reshaping the last fully-connected layer (F5). By subtracting the input face image with the learned illumination pattern, we can obtain illumination-reduced face image $\hat{\mathbf{x}}_{c,n} = \mathbf{x}_{c,n} - \Lambda(\mathbf{y}_{c,n})$, which will be the input for Step 2 learning. The illumination-reduced face image obtained by Step 1 can eliminate illumination effect while preserving face structural information for face recognition.

Step 2: Maximizing discriminative power of feature representation

From Step 1, we obtain illumination-reduced face images. Using these images, we learn subsequent DCNN in Step 2 for maximizing the discriminative power of feature representation. In this paper, in order to utilize the learned illumination patterns information in Step 1, DCNN parameters (e.g., weights and bias) to be learned in Step 2 is initialized by the parameters learned in Step 1. These learned weights have already plentiful information about face characteristics, thus they are useful to converge faster. In order to increase the discriminative power, Step 2 learns DCNN based on two objective functions for a class label and feature representation (see below Eq. (2) and (3)). By Eq. (2), we learn DCNN in terms of classification accuracy for class label. In addition, by considering Eq. (3) in terms of feature representation, we can significantly enhance the discriminative power.

For the class label, we learn DCNN based on a cross entropy error function [16] between target probability distribution (target class label) and the predicted probability distribution (predicted class label) for the softmax layer (see Fig. 3(b)). The error function is defined as

$$E^1_{\text{Step2}} = -\sum_{c,n}\sum_i t^i_{c,n}\log\hat{t}^i_{c,n} = -\sum_{c,n}\log\hat{t}^c_{c,n}, \tag{2}$$

where $\mathbf{t}^m_{c,n} = [t^1_{c,n}, t^2_{c,n}, \cdots, t^C_{c,n}]^T$ is target probability distribution about $\hat{\mathbf{x}}_{c,n}$ which is the n-th illumination-reduced training sample of the c-th class, $\mathbf{t}^l_{c,n}$ is predicted probability distribution, and C denotes the number of classes in training data. Then, in order to enhance discriminative power more, we consider the intra-class similarity [17] in terms of feature representation. In other words, feature vector extracted by the DCNN is enforced to minimize the following function:

$$E^2_{\text{Step2}} = \frac{1}{2}\sum_{c,n} g\left(\|\mathbf{q}_{c,n} - \mathbf{m}_c\|^2 - \left(d^c_{\min}\right)^2\right), \tag{3}$$

where $\mathbf{q}_{c,n}$ is feature vector to be learned for the illumination-reduced training samples $\hat{\mathbf{x}}_{c,n}$, $\mathbf{m}_c$ denotes the mean vector of feature vectors for the c-th class with N_c training samples, i.e., $\mathbf{m}_c = (1/N_c)\sum_n \mathbf{q}_{c,n}$ and $d^c_{\min}$ is the half of the minimum distance between $\mathbf{m}_c$ and $\mathbf{m}_k$ for all k except the c-th class. Through

DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-470

the minimization of the function, Eq. (3) makes the distance between feature vectors within a class be smaller than $d^c_{\min}$ in multi-dimensional space.

Classification

With the learned DCNN in Step 1 of the proposed method, a test face image $\mathbf{x}_{\mathrm{Tst}}$ is feed-forwarded to obtain the illumination-reduced face image $\hat{\mathbf{x}}_{\mathrm{Tst}}$. Then, the illumination-reduced face image is forwarded into the DCNN learned in Step 2 of the proposed method. Finally, a feature vector $\mathbf{q}_{\mathrm{Tst}}$ is obtained for the purpose of FR. For the classification for face recognition, we compute gallery feature vectors $\mathbf{q}^1_{\mathrm{Gal}}, \cdots, \mathbf{q}^C_{\mathrm{Gal}}$ corresponding to gallery face images $\mathbf{x}^1_{\mathrm{Gal}}, \cdots, \mathbf{x}^C_{\mathrm{Gal}}$ of C classes. For the classification, 1-nearest neighborhood classifier is used based on Euclidean distance.

Experimental Results

Figure 2. *Example face images from CMU Multi-PIE dataset under 20 different illumination conditions.*

Experimental Setups

In order to verify the proposed method, the subset of the publicly available CMU Multi-PIE database [13] was used. For considering the effectiveness of the proposed method under varying illumination, frontal face images with all illumination conditions (i.e., 20 conditions [13]) were selected as shown in Fig. 2. The number of subjects is 337 persons across the 4 sessions. All face images used in our experiments were cropped by using the facial landmarks in [17]. Each cropped facial image was resized to 32 32 pixels. To evaluate the proposed method, we measured FR accuracy, where the training set (for learning the proposed DCNN structure) and the test set were mutually exclusive. The first 200 subjects were used for the training set and remaining 137 subjects were chosen for the test set. When testing, the gallery images were set to face images with neutral illumination and other images with varying illuminations were used as the probe images. The number of training, gallery, and probe images is 4,000, 137, and 2,603, respectively.

In the experiment, a DCNN structure was set up (refer to Fig. 3), which consisted of three convolutional layers, a max-pooling layer, and two fully-connected layers. For the first convolutional layer (C1), 32 different 55 filters were used for the convolution. The pooling layer (P1) took the maximum value in each 22 region. For C2 and C3, the number of filters was equally 64 and the size of a filter was 55 and 33, respectively. The fully-connected layer F4 consisted of 1,024 units, and the last fully-connected layer F5 only used in Step2 had 200 units which was the number of classes in the training set. In order to train the proposed DCNN architecture,

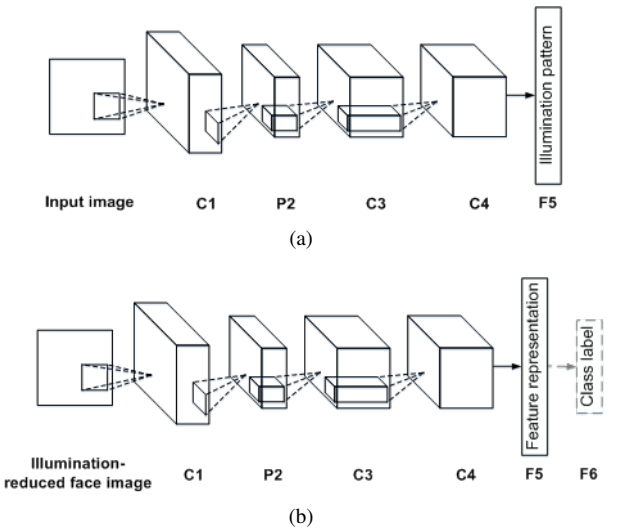


Figure 3. *DCNN architectures adopted in the proposed learning method. (a) DCNN for learning illumination patterns. (b) DCNN for maximizing the discriminative power of feature representation.*

the initial learning rate was set to 0.002 in Step 1. Otherwise, in Step 2, initial learning rate was set differently according to the equation. was set to 0.01 for Eq. (2), and 0.002 for Eq. (3). The learning rate decayed exponentially as [20], where was the learning rate of previous epoch and was that of current epoch.

Visualization of Feature Subspace

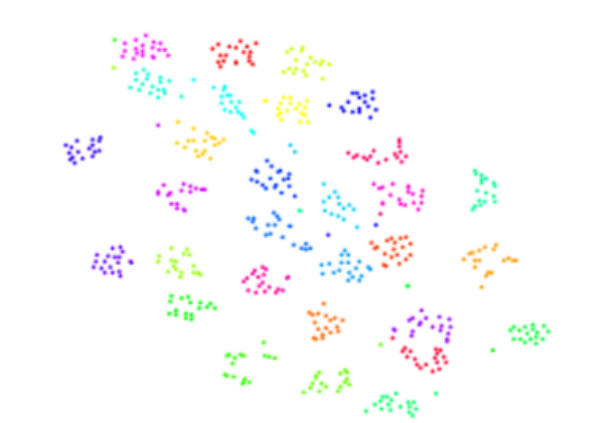
Fig. 4 visualizes a 2D feature spaces from the DCNN learned with the proposed method for the probe face images under varying illuminations, where each dot represents a feature. To visualize the feature subspace in 2D space, we adopted t-SNE [21]. And, the features under 20 different illumination conditions are plotted in different colors from 30 different classes in Fig. 4(a), (b), and (c). As seen in Fig. 4(a), feature distribution under different illumination variations is severely overlapped. So, it is difficult to discriminate between persons. After Step 1, feature becomes more separable by eliminating illumination effect on face images. Furthermore, after Step 2, the proposed method shows higher separability, i.e., minimizing intra-class variations as well as maximizing intra-class variations.

Classification Results

Table 1 shows FR accuracy and comparison with the existing FR methods robust to illumination variations. For the comparison, histogram equalization [5], multi-scale retinex [7], GradientFace [8], Weber-face [9], and Conventional CNN [17] were used, where the first four methods were photometric normalization-based approaches and the last one was a conventional deep learning-based approach. In [8], after applying GradientFace method, they used modified gradient-based L1 distance for the classification. In other methods, 1-nearest neighbor classifier was used based on Euclidean distance. Among photometric normalization-based methods, Weber-face showed the best FR accuracy (90.47%). Conventional CNN showed lower performance than some illumination normalizing methods. The conventional CNN was directly trained by face images without considering il-

DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-470

(a) Original 2D feature space under illumination variations.



(b) 2D feature space learned with the proposed method (after Step 1).

(c) 2D feature space learned with the proposed method (after Step 2).

Figure 4. *Visualization of 2D feature spaces. Each dot represents a feature from 30 different classes under 20 illumination variations. (Best viewed in color.)*

lumination variations, it failed to extract discriminative facial features. On the other hand, in our proposed method, varying illumination effects were eliminated in Step 1 and the latent feature extraction for face recognition could be performed well by Step 2. The proposed method showed the best performance which was 96.24%.

Table 1. FR accuracy comparisons with the proposed method and the previous works.

Method	Recognition rate
Histogram equalization [5]	43.10%
Multi-scale retinex [7]	60.55%
GradientFace [8]	84.75%
Weber-Face [9]	90.47%
Conventional CNN [18]	72.22%
Proposed method	**96.24%**

Conclusion

In this paper, we proposed a two-step learning of deep convolutional neural networks for a new illumination-reduced feature representation. For that purpose, the proposed method consisted of the following two steps: 1) the DCNN in Step 1 is learned to extract illumination pattern from face images. 2) The DCNN in Step 2 is learned to enhance a discriminative power. Through the comparative experiments with CMU Multi-PIE dataset under varying illumination conditions, we showed that the proposed method outperformed photometric normalization-based methods and the conventional CNN.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No.2015R1A2A2A01005724).

References

[1] K. W. Bowyer, "Face recognition technology: Security versus privacy," *IEEE Technol. Soc. Mag.*, vol. 23, no. 1, pp. 9-19, 2004.

[2] A. K. Jain, A. Ross, and S. Prabhaker, "An introduction to biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4-20, 2004.

[3] Y. Adini, Y. Moses, and S. Ullman, "Face recognition: The problem of compensating for changes in illumination direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 721-732, 1997.

[4] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635-1650, 2010.

[5] S. M. Pizer and E. P. Amburn, "Adaptive histogram equalization and its variations," *Comput. Vis. Graph. Image Process.*, vol. 39, no. 3, pp. 355-368, 1987.

[6] R. Basri, and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.* 25, 2, 218-233.

[7] D. Jobson, Z. Rahman, and G. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Process.*, vol. 6, no. 7, pp. 965-976, 1997.

[8] T. Zhang, Y. Tang, B. Fang, Z. Shang, and X. Liu, "Face recognition

DOI: 10.2352/ISSN.2470-1173.2016.11.IMAWM-470

under varying illumination using gradient faces," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2599-2606, 2009.

[9] B. Wang, W. Li, W. Yang, and Q. Liao, "Illumination normalization based on Weber's law with application to face recognition," *IEEE Signal Process. Lett.*, vol. 18, no. 8, pp. 462-465, 2011.

[10] A. K. Jain, *Fundamentals of Digital Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1989.

[11] H. F. Chen, P. N. Belhumeur, and D. W. Jacobs, "In search of illumination invariants," in *Proc. CVPR*, 2000, pp. 254-261.

[12] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *arXiv preprint arXiv:1502.04383*, 2015.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, no. 5, pp. 807-813, 2010.

[14] [Online] Available:. `https://en.wikipedia.org/wiki/Convolutional_neural_network.`

[15] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *Proc. CVPR*, 2012.

[16] P. Golik, P. Doetsch, and H. Ney, "Cross-entropy vs. squared error training: a theoretical and experimental comparison," in *Proc. Interspeech*, 2013.

[17] J.-J. Seo, H.-I. Kim, and Y. M. Ro, "Pose-robust and discriminative feature representation by multi-task deep learning for multi-view face recognition," in *Proc. ISM*, 2015.

[18] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *Proc. ICCV*, 2013.

[19] A. Krizhevsky, I. Sutskever, G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012.

[20] A. Senior, G. Heigold, M.-A. Ranzato, and K. Yang, "An empirical study of learning rates in deep neural networks for speech recognition," in *Proc. ICASSP*, 2013.

[21] L.J.P. van der Maaten and G.E. Hinton, "Visualizing highdimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.

Author Biography

Yeoreum Choi received the B.S. degree from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2015. He is currently working toward the M.S. degree at KAIST. His research interests include pattern recognition, object recognition, and deep feature modeling.

Hyung-Il Kim received the B.S. degree (Summa Cum Laude) from Dongguk University, Seoul, Korea, in 2011 and the M.S. degree from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2013. He is currently pursuing the Ph.D. degree in electrical engineering at KAIST. His research interests include image/video processing, machine learning, and face recognition.

Yong Man Ro received Ph.D. degrees from KAIST. He was a researcher at Columbia University and a research fellow at the UC, Berkeley. He is currently a professor and the chair of signals and systems group of the school of electrical engineering in KAIST. His research interests are image processing, 3-D video processing, computer vision, visual recognition. Dr. Ro received the young investigator finalist award of ISMRM. He served as an associate editor for IEEE SPL.