2015 IEEE International Symposium on Multimedia

Pose-robust and Discriminative Feature Representation by Multi-task Deep Learning for Multi-view Face Recognition

Jeong-Jik Seo[†], Hyung-Il Kim[†], and Yong Man Ro[‡]
School of Electrical Engineering, KAIST
Daejeon, Republic of Korea
{jj.seo, hyungil.kim, ymro}@kaist.ac.kr

***Abstract*—Automatic face recognition (FR) under uncontrolled environments has attracted considerable research attention. In the uncontrolled environments, pose variation is known as one of the crucial factors that influences FR performance. In this paper, we propose a discriminative and pose-robust feature representation using the multi-task learning in deep convolutional neural networks (ConvNet). We introduce four tasks (i.e., maximizing inter-class variation, minimizing intra-class variation, minimizing intra-pose variation, and preserving pose continuity) to learn the ConvNet. Moreover, two-stage learning strategy is proposed to minimize the error functions in learning the deep ConvNet. The extensive experimental results (with the challenging CMU MultiPIE dataset containing pose variations) show that the proposed method outperform state-of-the-art in terms of FR accuracy. Furthermore, the proposed method shows significant improvement even for the face images whose poses are not included in training set.**

***Keywords*-multi-view face recognition; deep learning; deep convolutional neural networks; deep feature representation; multi-task learning;**

I. Introduction

For the past few decades, automatic face recognition (FR) has been achieved great progress [1-9]. Recently, FR under uncontrolled environments has attracted considerable attention [10, 11]. In the uncontrolled environments, face images have large intra-class variations caused by different facial poses, varying illumination conditions, low resolution, blur, occlusion, etc. [12]. Among them, pose variation is known as one of the crucial factors that influences FR performance [10, 11]. Since face images with different poses span quite various spaces [3], they cannot be directly matched to frontal face images in gallery set. Therefore, the pose variation problem has to be resolved for the real-world FR scenario.

In order to reduce the discrepancy between face images acquired from the different views, the multi-view common space learning approaches [1-4] have been proposed. These approaches basically split face manifold into a discrete set of poses and learn a projection for each pose to the common feature space in which the face images with different poses become highly close. The conventional unsupervised common space learning methods, such as canonical correlation analysis (CCA) [1] and partial least squares (PLS) [2], tried to find view-specific linear projections to maximize the cross correlation between views. In contrast, to obtain more discriminative power, generalized multi-view analysis (GMA) [3] and multi-view discriminant analysis (MvDA) [4] learned the common space by making use of class label as supervision. Although the face images from multi-view are distributed on highly complex nonlinear manifold [10], the linearity and limited nonlinearity by kernelization could not effectively represent the complex nonlinearity.

In order to deal with the complex nonlinearity in the manifold generated by face images with different views, deep learning-based approaches [5-9] have been spotlighted. Among them, deep CCA (DCCA) [5] and deeply coupled auto-encoder networks (DCAN) [6] utilized the deep learning to learn the multi-view common space. Thanks to the capability of a highly nonlinear representation of the deep learning, the aforementioned problems (i.e., complex nonlinearity) could be resolved. However, pose information of a testing face image should be known *a priori* due to the pose-specific projection. More recently, for the facial feature representation robust to pose variation, the authors in [7-9] proposed deep architectures to reconstruct the face images with canonical view (i.e., frontal pose and neutral illumination) from the input face images with arbitrary poses and illumination conditions. By doing so, face images with various poses and illuminations could be effectively normalized to the canonical view without labeled pose or illumination information in the testing phase. However, the discriminative power among classes is not thoroughly investigated. Moreover, since a discrete set of poses is available in the training set despite the fact that a facial pose is continuous in nature, there is limitation in addressing face images with unseen poses. Therefore, the problem for the unseen pose should be addressed as a practical issue. In this paper, in order to deal with the abovementioned limitations, we propose a multi-task learning in deep convolutional neural networks (ConvNet throughout the rest of the paper) to learn a discriminative and pose-robust feature representation for multi-view FR. The main contributions of the proposed method are summarized as two folds:

- For learning a deep ConvNet, we devise four tasks and corresponding error functions which are designed to maximize inter-class variation, minimize intra-class variation and intra-pose variation, and preserve pose continuity, respectively. Thanks to the first two tasks, the discriminative feature space can be learned. By utilizing the remaining two tasks as regularization, pose-robust feature space can be learned. Note that the continuous poses unseen in the training set can be effectively ad-

[†] Both authors contributed equally to this work.
[‡] Corresponding author (ymro@kaist.ac.kr)

978-1-5090-0379-2/15 $31.00 © 2015 IEEE
DOI 10.1109/ISM.2015.93

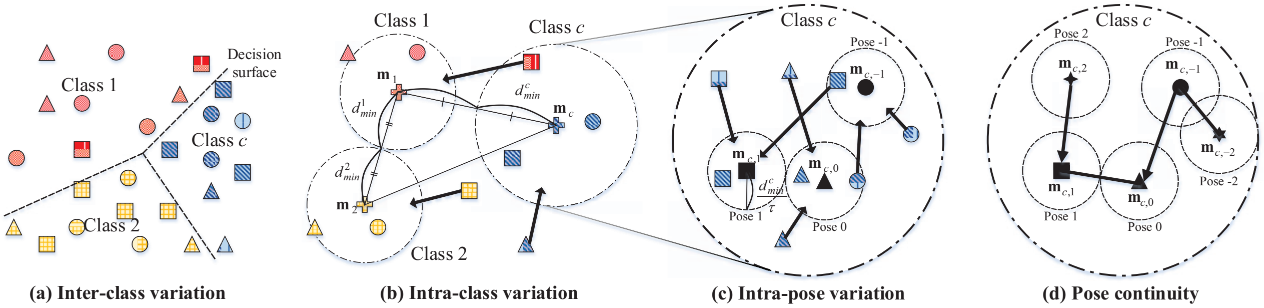


Figure 1. Visualization of error functions for our multi-task learning in deep ConvNet. The feature vector of the samples in each class are represented by the same color, and various poses are represened by different shapes.

dressed by the pose continuity term.

- To effectively learn our deep ConvNet minimizing the proposed multi-task objective functions, a two-stage learning strategy is proposed. By separately optimizing the tasks in the stages, the proposed learning strategy is quite simple and easy to implement.

The exhaustive experiments with the publicly available CMU MultiPIE dataset [13] containing variations of the pose and illumination conditions show that the proposed method outperforms state-of-the-art methods. In particular, the proposed method brings significant improvements in case that face images with unseen poses in training phase are fed as the input images in testing phase.

We organized the reminder of this paper as follows. In Section II, we introduce our proposed multi-task learning in deep ConvNet and its effective learning scheme. In Section III, we discuss the experimental results. Finally, we draw conclusions in Section IV.

II. Proposed Multi-task Learning in Deep Convolutional Neural Networks

In this section, we explain the proposed multi-task learning in deep ConvNet to learn a discriminative and pose-robust feature representation for multi-view FR. In Section A, we describe four tasks and corresponding error functions. In Section B, we introduce two-stage learning strategy to learn the deep ConvNet to minimize the error functions effectively.

A. Multi-task Learning in Deep ConvNet

Our multi-task learning in the deep ConvNet is composed of four tasks: 1) maximizing inter-class variation, 2) minimizing intra-class variation, 3) minimizing intra-pose variation, and 4) preserving pose continuity, which are illustrated in Fig. 1. Herein, we call each task as task1, task2, task3, and task4, respectively. By taking task1 and task2 into account, since the inter-class variation is maximized and the intra-class variation is minimized, the discriminative power for the learned feature representation is significantly enhanced. Additionally, since the intra-pose variation within a class is forced to be minimized, the intra-class variation is much more minimized. Finally, by regularizing the pose continuity, the unseen pose encountered in real-world scenario is able to be readily covered.

For maximizing inter-class variation (task1), a cross-entropy error function is adopted as follows:

$$E_1 = -\sum_{c,p,i}\sum_k t^k_{c,p,i} \log \hat{t}^k_{c,p,i} = -\sum_{c,p,i} \log \hat{t}^c_{c,p,i}, \quad (1)$$

where $\mathbf{t}_{c,p,i} = \left[t^1_{c,p,i}, t^2_{c,p,i}, ..., t^C_{c,p,i}\right]^T$ is the target probability distribution about $\mathbf{x}_{c,p,i}$ which is the i-th training sample of c-th class and p-th pose, $\hat{\mathbf{t}}_{c,p,i}$ and C denote the predicted probability distribution and the number of classes in training set. In order to obtain the correct label from the output of the last layer in the deep ConvNet, the samples with different classes should be linearly separated in the feature space (i.e., the output space of the last hidden fully-connected layer) as shown in Fig. 1(a) because the last layer functions as a linear classifier, thus obtaining the discriminative power.

In addition to the inter-class variation, the intra-class variation within the same identity is also critical factor determining the discriminative power in the feature space. Therefore, the error function for minimizing the intra-class variation is newly defined in Eq. (2). Note that this error function is defined in terms of feature vectors (i.e., the output of the last hidden layer) of the samples in each class, which means the learning in terms of representation.

$$E_2 = \frac{1}{2}\sum_{c,p,i} g\left(\|\mathbf{m}_c - \mathbf{y}_{c,p,i}\|^2_2 - (d^c_{min})^2\right), \quad (2)$$

where $\mathbf{m}_c$ denotes the mean vector of all feature vectors of the training samples in c-th class, $\mathbf{y}_{c,p,i}$ is the feature vector of the training samples $\mathbf{x}_{c,p,i}$, d^c_{min} is the half of the minimum distance between $\mathbf{m}_c$ and $\mathbf{m}_j$ for all j except for c-th class, $g(z) = \frac{1}{\beta}\log\left(1 + \exp(\beta z)\right)$ the smoothed approximation of $[z]_+ = \max(z, 0)$ [14], and β denotes a sharpness parameter. By minimizing Eq. (2), the feature vectors included in each identity are forced to be gathered in each multi-dimensional sphere, where the radius and the center of the sphere are denoted by d^c_{min} and $\mathbf{m}_c$, respectively.

To effectively address pose variation within a class, two tasks for minimizing intra-pose variation and preserving pose continuity are newly designed. First of them, to minimize intra-pose variation within a class, the error function for task3 is defined as follows:

$$E_3 = \frac{1}{2}\sum_{c,p,i} g\left(\|\mathbf{m}_{c,p} - \mathbf{y}_{c,p,i}\|_2^2 - \left(\frac{d_{min}^c}{\tau}\right)^2\right), \qquad (3)$$

where $\mathbf{m}_{c,p}$ is the mean vector of all feature vectors of training samples in the p-th pose of the c-th class, τ denotes the constant determining the range of pose distribution. Figure 1(c) shows the effect of the error function E_3. The largest circle denotes the multi-dimensional sphere for class c, and the smaller circles are the multi-dimensional spheres for poses within the c-th class in the feature space. Originally, the feature vectors of the samples with the same pose in each class without considering the intra-pose variation are distributed unmethodically. However, as described by the dotted arrows, task3 can rearrange the feature vectors in a pose-wise manner. In other words, the intra-class variation problem caused by the pose variation within class can be much more mitigated.

Even if feature vectors of samples with the same pose within a class form a cluster in the feature space, the clusters configuration can be disordered by the training of deep ConvNet. However, since face images with the neighboring poses (e.g., +45° and +30°) share common facial information to a certain extent (i.e., correlated), the correlation between adjacent poses in the feature space should be considered for the deep ConvNet learning. Herein, we define task4 for preserving the pose continuity, where the error function for the task4 is formulated as follows:

$$E_4 = \frac{1}{2}\sum_{c}\sum_{p,p\neq 0} g\left(\|\mathbf{m}_{c,q} - \mathbf{m}_{c,p}\|_2^2 - \left(\frac{2d_{min}^c}{\tau}\right)^2\right),$$

$$q = \begin{cases} p+1, & \text{if } p < 0 \\ p-1, & \text{if } p > 0 \end{cases}, \qquad (4)$$

where the p denotes the pose index included in the training phase, and the index is labeled with $2P-1$ discrete poses considering poses angular changes in yaw, i.e., $p \in \{0, \pm 1, ..., \pm P\}$. When $p = 0$, the face image is frontal face (the angle of facial pose is zero in yaw). If p increases or decreases, the angle of facial pose in yaw increases and decreases from 0, respectively. By the neighborhood index q, $\mathbf{m}_{c,p}$ is able to head for the adjacent pose which is nearer to the frontal pose. This directivity is necessary for good convergence. Concerning the relation between the directivity and the convergence, Fig. 2 illustrates the necessity of the directivity. If $\mathbf{m}_{c,2}$ is intended to approach to both adjacent poses (i.e., $\mathbf{m}_{c,1}$ and $\mathbf{m}_{c,3}$), the directions of the gradients are represented as two blue arrows. Then, the final direction becomes the red arrow by the summation of the vectors. Likewise, $\mathbf{m}_{c,3}$ moves to the direction described by the yellow arrow. Since $\mathbf{m}_{c,2}$ and $\mathbf{m}_{c,3}$ are not able to get close, it results in poor convergence of the learning. In order to prevent this problem, the directivity should be considered. By utilizing the error function E_4 the feature vectors of the samples with different poses within a class can be rearranged as meaningful order (i.e., pose continuity) as shown in Fig. 1(d). In other words, by preserving the pose continuity, we can effectively address the input images with unseen poses in training set. When a face image with unseen pose is fed into the deep ConvNet, the learned feature becomes located around the region between the two closest poses.

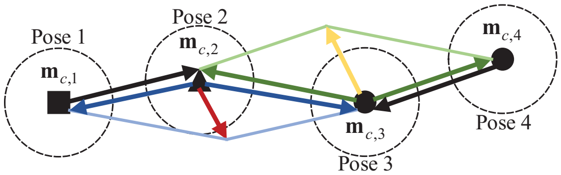


Figure 2. Illustration of the necessity of the directivity for pose continuity (best viewed in color).

B. *Two-stage Learning Strategy*

In order to learn the proposed deep ConvNet with multi-task by mini-batch gradient descent with backpropagation algorithm [15, 16], gradients in respect of the parameters (i.e., weights and biases) are necessary. In this section, we first present the gradients of the four error function. Next, the proposed two-stage learning strategy is described in detail.

For the first task, the gradients of the error function with respect to the parameters of the last layer (i.e., the L-th layer) are described as follows [17]:

$$\frac{\partial E_1}{\partial \mathbf{b}^{(L)}} = \sum_{c,p,i} \hat{\mathbf{t}}_{c,p,i} - \mathbf{t}_{c,p,i}, \qquad (5)$$

$$\frac{\partial E_1}{\partial \mathbf{W}^{(L)}} = \mathbf{x}^{(L-1)}\left(\frac{\partial E_1}{\partial \mathbf{b}^{(L)}}\right)^T, \qquad (6)$$

where $\mathbf{b}^{(L)}$ and $\mathbf{W}^{(L)}$ denote biases and weights of the last layer, respectively, and $\mathbf{x}^{(L-1)}$ denotes the output of the last hidden layer (i.e., the $(L-1)$th layer). As can be seen in Eq. (6), since the gradient with respect to the weight is dependent on the gradient for the bias in Eq. (5), we present the gradients with respect to the bias for the following equations. The gradients with regard to the parameters of the other layers can be computed by the backpropagation algorithm [15, 16].

For the second task, we assume the target vector $\mathbf{m}_c$ in (Eq. (2)) as a constant vector to calculate the gradients simply, and then the gradients of E_2 with respect to the bias of the last hidden layer are described as follows:

$$\frac{\partial E_2}{\partial \mathbf{b}^{(L-1)}} = \sum_{c,p,i} g'(e_2)\left(\mathbf{y}_{c,p,i} - \mathbf{m}_c\right) \circ \sigma'\left(\mathbf{u}^{(L-1)}\right), \qquad (7)$$

$$\text{where } e_2 = \|\mathbf{m}_c - \mathbf{y}_{c,p,i}\|_2^2 - (d_{min}^c)^2, \qquad (8)$$

'$\circ$' denotes element-wise multiplication, $\sigma(\cdot)$ is an activation fuction, and $\mathbf{u}^{(L-1)} = \mathbf{W}^{(L-1)}\mathbf{x}^{(L-2)} + \mathbf{b}^{(L-1)}$. Likewise, the gradients for task3 and task4 with respect to the bias can be arranged as follows:

$$\frac{\partial E_3}{\partial \mathbf{b}^{(L-1)}} = \sum_{c,p,i} g'(e_3)\left(\mathbf{y}_{c,p,i} - \mathbf{m}_{c,p}\right) \circ \sigma'\left(\mathbf{u}^{(L-1)}\right), \qquad (9)$$

$$\frac{\partial E_4}{\partial \mathbf{b}^{(L-1)}} = \sum_{c,p,p\neq 0} \frac{g'(e_4)}{N_{c,p}}\left(\mathbf{m}_{c,p} - \mathbf{m}_{c,q}\right) \circ \sigma'\left(\mathbf{u}^{(L-1)}\right), \qquad (10)$$

where $\mathbf{m}_{c,p}$ in Eq. (9) and $\mathbf{m}_{c,1}$ in Eq. (10) are also regarded as constant vector. Herein, e_3 in Eq. (9) and e_4 in Eq. (10) are presented like:

$$e_3 = \|\mathbf{m}_{c,p} - \mathbf{y}_{c,p,i}\|_2^2 - \left(\frac{d_{min}^c}{\tau}\right)^2, \qquad (11)$$

$$e_4 = \|\mathbf{m}_{c,q} - \mathbf{m}_{c,p}\|_2^2 - \left(\frac{2d_{min}^c}{\tau}\right)^2. \qquad (12)$$

Since the error function of task1 is defined about the label information and the remaining error functions are defined about the feature vectors of the samples in training set, task1 and the other tasks are separately learned. That is, in the first stage of the learning, the deep ConvNet is learned by the conventional mini-batch gradient descent with backpropagation algorithm [15, 16] using the gradients of task1 in Eq. (5) and Eq. (6) until convergence. In the second stage of the learning, the remaining three tasks are optimized simultaneously. Note that since three tasks are not related to the class label, the last fully-connected layer is removed. Additionally, since $\mathbf{m}_c$, $\mathbf{m}_{c,p}$, and $\mathbf{m}_{c,q}$ should be calculated as constant vectors with samples of the same class in advance in the second stage learning, the random batch sequence cannot be straightforwardly applied. However, if the planned batch sequence is adopted for the learning, it results in the poor convergence [18]. In order to resolve this problem, the deep ConvNet is learned by minimizing E_2 with a random batch sequence first (with fixed $mathbf{m}_c$ computed in the first stage). Then, the deep ConvNet is learned by minimizing Eq. (13) with the batch sequence containing all samples from each class (i.e., class-wise batch sequence).

$$E = E_2 + \lambda E_3 + \gamma E_4, \qquad (13)$$

where λ and γ are regularization parameters. We repeat the steps until convergence. By inducing the learning procedure to randomness, we are able to achieve good convergence.

III. Experiments

A. Experimental Setups

In order to evaluate the effectiveness of the proposed method in multi-view FR scenarios, publicly available CMU MultiPIE dataset [13] was adopted. In our experiments, we used a subset of CMU MultiPIE which consisted of the images with 13 poses in yaw angle ranging from $-90°$ to $90°$ with $15°$ interval, 20 illumination conditions, neutral expression, and 337 subjects across the 4 sessions. All face images used in our experiments were cropped based on the facial landmarks in [7]. Each cropped facial image was resized to 32×32 pixels. We conducted the experiments with two different settings: Setting-I and Setting-II. In Setting-I, the face images with 13 poses from $-90°$ to $90°$ and neutral illumination were used. The first 200 subjects were used for training and remaining 137 subjects were chosen for testing. In the testing set, one frontal image of each subject was selected for the gallery and the remaining images were selected for probes. To evaluate the FR accuracy in the scenario such that unseen poses in training set present in testing set, three settings were defined according to the form of training set: Setting-I-1, Setting-I-2, and Setting-I-3. The training sets with the different pose sets were generated as $\Theta_{train}^1 = \{0°, \pm15°, \pm30°, \pm45°, \pm60°, \pm75°, \pm90°\}$, $\Theta_{train}^2 = \{0°, \pm30°, \pm60°, \pm90°\}$, and $\Theta_{train}^3 = \{0°, \pm45°, \pm90°\}$.

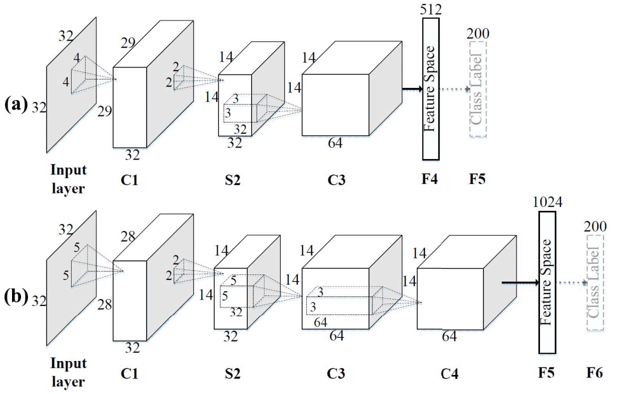


Figure 3. Architectures of ConvNets in our experiments. (a) ConvNet for Setting-I. (b) ConvNet for Setting-II. (C: convolutional layer, S: subsampling layer, F: fully-connected layer)

Setting-II introduced in [8] was designed to evaluate the effectiveness of the proposed method in challenging conditions under various poses and illumination. The face images under 9 poses from $-60°$ to $60°$ with 20 illuminations were used for the training and testing sets. The other setups were the same as Setting-I. Since severely large or small ConvNet architecture compared to the size of training set could result in overfitting or underfitting [19], we made use of two deep ConvNet architectures as shown in Fig. 3 by considering the size difference of training set. For Setting-I, approximately 3,000∼8,000 face images were utilized for training whereas 36,000 images were used in Setting-II. Therefore, the relatively shallower ConvNet architecture as shown in Fig. 3(a) was adopted for Setting-I, and deeper ConvNet architecture was utilized for Setting-II as shown in Fig. 3(b). The ReLU [20] was adopted for the activation function of all convolutional layers except for the last hidden fully-connected layer (i.e., sigmoid function for the layer). The softmax function was adopted for representing the probability distribution over the classes [21].

In the first stage of our learning strategy, the initial learning rate η was set to 0.1. On the other hand, in the second stage, initial learning rate η was set adaptively according to the structure of ConvNet. η was set to 0.05 for shallower one, and 0.002 for deeper one, empirically. The learning rate η decayed exponentially as $\eta^t = \eta^{t-1} \times 0.99$ [22], where η^{t-1} denotes the learning rate of previous epoch, and η^t the learning rate of current epoch. The parameter τ was set to 4 and the regularization parameters λ and γ were set to 0.1.

B. Results of Setting-I

Table I shows the results of Setting-I, where FR performances for each pose and each task are represented. Among

Table I

FR ACCURACY (%) OF SETTING-I. THE HIGHEST RECOGNITION ACCURACY IS HIGHLIGHTED IN BOLD.

	$-90°$	$-75°$	$-60°$	$-45°$	$-30°$	$-15°$	$15°$	$30°$	$45°$	$60°$	$75°$	$90°$	Average
						Setting-I-1							
Task1	21.2	43.8	63.5	81.0	96.4	98.5	98.5	92.0	83.2	61.3	47.4	29.2	68.0
Task1 and 2	55.5	67.9	86.1	**94.2**	97.8	99.3	**100.0**	**99.3**	95.6	84.7	**75.2**	60.6	84.7
Task1, 2, and 3	51.1	**69.3**	86.1	**94.2**	**98.5**	**100.0**	**100.0**	**99.3**	**97.1**	**86.9**	74.5	59.9	84.7
All tasks	**56.2**	66.4	**86.9**	**94.2**	97.8	**100.0**	**100.0**	**99.3**	96.4	83.9	**75.2**	**65.0**	**85.1**
						Setting-I-2							
Task1	19.7	34.3	57.7	79.6	94.2	**99.3**	98.5	94.2	81.0	57.7	35.8	24.1	64.7
Task1 and 2	48.9	59.1	74.5	86.1	96.4	97.1	99.3	95.6	86.1	75.2	56.2	48.9	76.9
Task1, 2, and 3	49.6	58.4	75.9	**87.6**	95.6	96.4	98.5	95.6	88.3	73.7	51.8	51.8	76.9
All tasks	**54.0**	**63.5**	**81.0**	**87.6**	**97.8**	**99.3**	**99.3**	**96.4**	**90.5**	**81.0**	**59.9**	**54.0**	**80.4**
						Setting-I-3							
Task1	12.4	25.5	44.5	67.9	89.1	97.1	96.4	82.5	63.5	40.9	27.7	18.2	55.5
Task1 and 2	40.9	54.7	65.0	88.3	**97.1**	**99.3**	**100.0**	**96.4**	89.8	64.2	59.1	49.6	75.4
Task1, 2, and 3	40.9	55.5	**67.2**	88.3	**97.1**	**99.3**	99.3	95.6	89.1	67.9	**61.3**	**52.6**	76.2
All tasks	**47.4**	**56.2**	65.7	**89.8**	**97.1**	98.5	**100.0**	95.6	**92.7**	**70.8**	58.4	51.1	**76.9**

Table II

FR ACCURACY (%) OF SETTING-II. THE FIRST AND SECOND HIGHEST RECOGNITION ACCURACY OF EACH COLUMN ARE HIGHLIGHTED IN BOLD.

	$-60°$	$-45°$	$-30°$	$-15°$	$0°$	$15°$	$30°$	$45°$	$60°$	Average
FIP + LDA [7]	49.3	66.1	78.9	91.4	94.3	90.0	82.5	62.0	42.5	72.9
RL + LDA [7]	44.6	63.6	77.5	90.5	94.3	89.8	80.0	59.5	38.9	70.8
MTL + RL + LDA [8]	51.5	70.4	80.1	**91.7**	93.8	89.6	83.3	63.8	50.2	74.8
MVP + LDA [8]	60.2	75.2	83.4	**93.3**	**95.7**	**92.2**	83.9	70.6	60.0	79.3
Task1	43.3	66.3	78.2	80.3	89.2	81.1	75.0	65.2	48.8	69.6
Task1 and 2	59.8	80.4	**88.7**	88.0	94.2	**90.3**	**87.2**	76.6	59.5	80.4
Task1, 2, and 3	**62.9**	**80.7**	88.4	88.5	93.0	90.1	**87.8**	**78.7**	**60.5**	**81.1**
All task	**62.8**	**80.8**	**88.6**	88.9	93.9	89.5	86.8	**78.2**	**61.9**	**81.2**

all tasks, task2 leaded to the most significant improvements on FR accuracy. For example, when task2 was additionally considered with the conventional ConvNet (i.e., Task1), the average accuracies increased with 17.1%, 15.7%, and 21.4%, in Setting-I-1, Setting-I-2, and Setting-I-3, respectively. In particular, there was larger improvement for large pose variation. It is mainly because that the images with large pose variation could be located close to the frontal face by minimizing intra-class variation

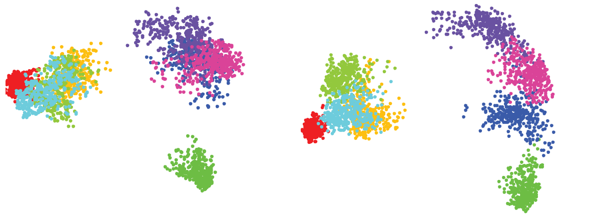
In Setting-I-2, the pose-related tasks (i.e., Task3 and Task4) achieved significant improvement on FR accuracy because they could address the face images with unseen poses in training set effectively. Specifically, the average accuracy of All tasks at $-75°$ was 4.4% higher than the case that only inter-class and intra-class variations were considered (the result of Task1 and 2). For the Setting-I-1, the pose-related tasks could obtain the slight improvement. Since all pose variations were available for the training and the testing, the Task 1 and 2 was enough to learn effectively the deep ConvNet. On the other hand, in the Setting-I-3, the FR accuracy was slightly decreased in some part. This is mainly because the objective function for the pose continuity might not take advantage of the various pose information due to the lack of pose variations in the Θ^3_{train} and large gap between adjacent poses (i.e., 0° and $45°$).

C. Results of Setting-II

Table II shows the FR accuracy of the proposed method for Setting-II. For the comparison, FIP+LDA [7], RL+LDA [7], MTL+RL+LDA [8], and MVP+LDA [8] method were used, where these approaches learned a deep model to reconstruct the canonical view face image. Among them, MVP+LDA showed the best average accuracy (79.3%). When we considered only inter-class variation, the average accuracy was 69.6% which was lower than other approaches. However, when we used all tasks to learn the deep ConvNet, we achieved the best result which was 81.2%. In case of the nearly frontal face between $-15°$ and $+15°$, the FR accuracy of our method was slightly lower than the state-of-the-art methods. It is mainly because the proposed method does not thoroughly consider the various illumination conditions by minimizing intra-class variation. On the other hand, for the face images with large pose variation, the proposed method outperformed the state-of-the-art methods. In particular, the accuracy of $45°$ in All tasks was 78.2% which is 7.6% higher than the state-of-the-art method (MVP + LDA).

D. Feature Space Analysis

Figure 4 visualizes the distributions of testing samples under various poses and illuminations in the learned space, where each dot represents a feature. And, the features from 8 different classes were plotted in different colors. Two deep ConvNets learned by Setting-II with task1 (Fig. 4(a)) and all tasks (Fig. 4(b)) were used. The feature space learned by the proposed method with all tasks showed higher level of separability compared to the conventional deep ConvNet only considering the Task1. More specifically, the samples plotted in blue were well separated from the samples of other classes in case of the proposed method with all tasks.



(a) Conventional deep ConvNet (task1) (b) Proposed Multi-task deep ConvNet (all tasks)

Figure 4. Visualization of sample distribution in learned feature space. Each dot represents a sample, and 8 classes are plotted with different colors. (a) Results of conventional deep ConvNets. (b) Results of the proposed multi-task deep ConvNets (best viewed in color).

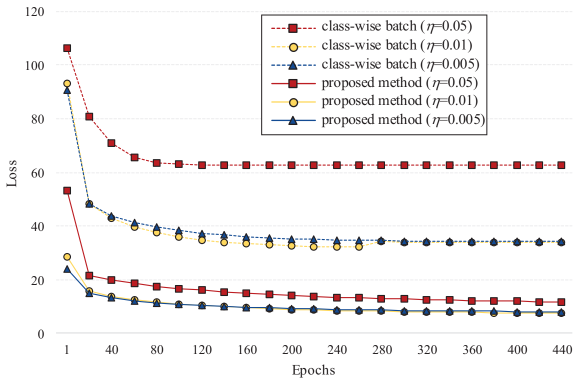


Figure 5. Convergence analysis of the proposed two step learning. Dotted line and solid line represent the results of learning with class-wise batches and proposed learning method, respectively. Lines with the same indicator represent the same initial learning rate.

E. Convergence of Learning

Figure 5 represents the value of error function E (i.e., loss) in training at each epoch in the second learning stage with the different learning rates η. The dotted line and solid line represents the results of learning with class-wise batches and proposed learning strategy which make use of two types of batches (i.e., random batch and class-wise batch), respectively. We could see that the loss obtained from learning with class-wise batch converged to higher value than the proposed learning strategy. This result showed that giving randomness on the batch is one of the important factors for convergence in mini-batch gradient descent with backpropagation algorithm.

IV. Conclusion

In this paper, we proposed a multi-task learning in deep ConvNet for a discriminative and pose-robust feature representation. For that purpose, we introduced four multi-task objective functions for maximizing inter-class variation, minimizing intra-class variation, minimizing intra-pose variation, and preserving pose continuity. By considering multi-task objectives simultaneously, the learned feature achieved highly discriminative power with the minimized intra-class variation. In particular, by considering the pose-related tasks, we could deal with the face images with unseen poses in training set. Through the comparative experiments, our proposed method outperformed the state-of-the-arts. Furthermore, our proposed deep ConvNet was able to achieve good convergence by the proposed two-stage learning strategy.

Acknowledgement

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

References

[1] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, 16(12):2639-2664, 2004.

[2] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *IEEE CVPR*, 2011.

[3] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *IEEE CVPR*, 2012.

[4] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *ECCV*, 2012.

[5] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013.

[6] W. Wang, Z. Cui, H. Chang, S. Shan, and X. Chen, "Deeply coupled auto-encoder networks for cross-view classification," *arXiv:1402.2031*, 2014.

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

[8] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: a deep model for learning face identity and view representations," in *NIPS*, 2014.

[9] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (SPAE) for face recognition across poses," in *IEEE CVPR*, 2014.

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

[11] J. Wright and G. Hua, "Implicit elastic matching with random projections for pose-variant face recognition," in *IEEE CVPR*, 2009.

[12] G. Hua, M.-H. Yang, E. Learned-Miller, Y. Ma, M. Turk, D. J. Kriegman, and T. S. Huang, "Introduction to the special section on real-world face recognition," *IEEE TPAMI*, 33(10):1921-1924, 2011.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, 28(5):807-813, 2010.

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

[15] C. M. Bishop, *Neural networks for pattern recognition*: Oxford university press, 1995.

[16] J. Bouvrie, "Notes on convolutional neural networks," MIT CBCL Tech Report, 2006.

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

[18] UFLDL Tutorial [Online] Available: http://ufldl.stanford.edu/tutorial/supervised/OptimizationStochasticGradientDescent/

[19] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, and T. Hospedales, "When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition," *arXiv:1504.02351*, 2015.

[20] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *ICML*, 2010.

[21] C. M. Bishop, *Pattern recognition and machine learning*: Springer, 2006.

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