Pose-Robust Face Verification by Exploiting Competing Tasks

Boyu Lu, Jingxiao Zheng, Jun-Cheng Chen, Rama Chellappa
Center for Automation Research, UMIACS
University of Maryland
{bylu, jxzheng}@umiacs.umd.edu, pullpull@cs.umd.edu, rama@umiacs.umd.edu

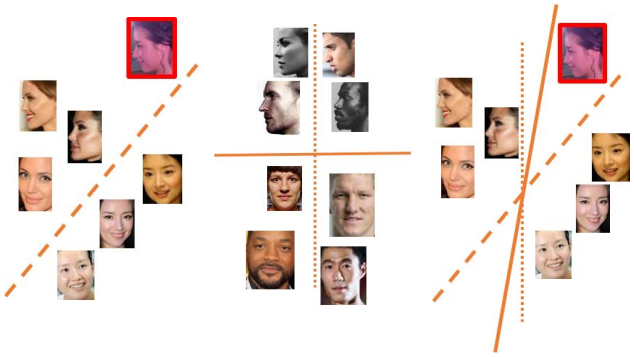
Abstract

In this paper, we propose a pose-robust metric learning framework for unconstrained face verification by jointly optimizing face and pose verification tasks. We learn a joint model for these two tasks and explicitly discourage the information sharing between pose and identity verification metrics so as to mitigate the information contained in the pose verification task leading to making the identity metrics for face verification more pose-robust. Specifically, we use the joint Bayesian metric learning framework to learn the metrics for both tasks and enforce an orthogonal regularization constraint on the learned projection matrices for the two tasks. The pose labels used for training the joint model are automatically estimated and do not require extra annotations. An efficient stochastic gradient descent (SGD) algorithm is used to solve the optimization problem. We conduct extensive experiments on three challenging unconstrained face datasets and show promising results compared to state-of-the-art methods.

1. Introduction

Due to recent advances in deep convolution neural networks, research in unconstrained face verification has achieved great progress [27, 22]. Although deep features show some robustness to face variations [26], the face verification performance when challenging conditions in pose, illumination, aging and expression are present is still far from satisfactory [15]. Among these factors, pose variation is one of the most difficult challenges to deal with as face image from various poses lie in a highly nonlinear manifold [28]. Therefore, pose-invariant face identification and verification has attracted significant attention [6]. Some previous works seek to learn pose-invariant representations [34, 11], while others focus on multi-view common subspace learning [13], or synthesize faces based on the generic 3D models [7, 1, 17].

Different from the above works, we tackle this problem by learning pose-robust metrics in which pose-sensitive information is explicitly mitigated. To achieve this goal, we introduce an auxiliary task called pose verification (*i.e.,* checking whether the two faces are in the same pose.) and exploit the competitive relationships between the auxiliary task (pose verification) and the main task of face verification.



(a) (b) (c)

Figure 1: Training a face recognition classifier by coordinating with pose information. (a) a face classifier trained with only identity information. The red boxed face is wrongly classified due to the bias in the training data. (b) a pose classifier trained using pose labels, and the classifier (solid line) is discriminative only with respect to poses. (c) using the normal direction of the pose classifier (vertical dashed line) to regularize the face classifier. The red boxed face is correctly classified by the new classifier (solid line) after regularization.

To better understand why the two tasks are competing, we give a simple example of face identification and pose classification, which are closely related to face and pose verification. In Fig. 1a, an identity classifier for face identification is trained to classify two different subjects. However, some of the training data is biased, *e.g.*, for some particular persons, the number of training samples is limited and most of the faces are frontal or near-frontal. In this situation, given a new profile face of the person, it is very likely that

978-1-5090-4822-9/17 $31.00 © 2017 IEEE
DOI 10.1109/WACV.2017.130

the face will be classified as someone else who has plenty of profile faces in the training set. To solve this problem, we can exploit the information from pose classification. As illustrated in Fig. 1b, the pose classifier indicates the most pose-sensitive orientation, while the normal vector of the pose classifier represents the least discriminative direction for poses (the dashed line). This observation suggests that the normal vector of the pose classifier can provide helpful information for the identity classifier to achieve pose robustness, which is shown in Fig. 1c. In other words, adding an orthogonal constraint between the classifiers for the two tasks would make the identity classifier more pose-robust.

In this paper, we propose a pose-robust metric learning framework for face verification by cooperating with the pose verification task. Based on the intuition that the metrics for the two tasks are competing with each other, we jointly learn the projection matrices for the two tasks and add an orthogonal regularization constraint. The orthogonal regularization enforces the metrics for the two tasks to be uncorrelated with each other and to capture different kinds of information in the features. Therefore, the learned metrics for the main task extracts pose-robust identity information and discounts the pose-sensitive information contained in the metrics for the auxiliary task. To show how the regularization affects the learned metrics, we conduct experiments by controlling the regularization parameters as discussed in Sec. 4. A stochastic gradient descent (SGD) algorithm is used to solve the optimization problem.

Experimental results of the proposed algorithm on three challenging face datasets demonstrate promising performances as compared to other competing methods. The main contributions are summarized as follows:

- We present a novel metric learning approach for face verification, and an optimization algorithm is derived. The learned metric is robust to pose variations by mitigating the pose-sensitive information from the competing task.
- We show that the proposed method yields promising experimental results on three challenging face datasets.

The rest of the paper is organized as follows: A brief review of metric learning and pose-invariant face verification methods is presented in Section 2. In Section 3, we briefly recap the joint Bayesian metric learning approach and detail the proposed algorithm. Experimental results are given in Section 4 and finally we conclude in Section 5.

2. Related Work

Pose-Invariant Face Recognition: Pose-invariant face recognition has generated growing interest in recent years [6]. Li *et al.* [16] designed a pose-invariant representation for faces by extracting densely sampled local features and training a Gaussian mixture model (GMM) on them. The GMM captures the spatial-appearance distribution of face images by augmenting local features with their locations. Zhu *et al.* [34] proposed a two-stage deep neural network to frontalize the off-frontal face images. The first module was used for feature extraction while the second module reconstructed the faces in the canonical view. Kan *et al.* [13] learned a discriminant common space for faces from different poses by maximizing the between-class variations and minimizing the within-class variations. Ding *et al.* [7] generated a generic 3D model and transformed the profile faces to synthesized partial frontal faces. Then patch-based face representations are used for face matching. AbdAlmageed *et al.* [1] utilize generic 3D models to generate synthetic faces in different poses and use pose-specific CNNs to extract features. The similarity between two face images is computed as fusion of the pose-specific feature similarities. Recently, Masi *et al.* [17] augmented the training set by synthesizing new faces with different pose, shape and expression variations. A CNN model is trained with the augmented training data and show better performance.

Metric Learning: Metric learning approaches have been widely used in classification and verification tasks to improve the performance. Weinberger *et al.* [29] proposed the Large Margin Nearest Neighbor (LMNN) method in a large margin framework. Chen *et al.* [3] proposed a joint Bayesian approach where the joint distribution of a pair of face images is learned using an EM-based algorithm. Huang *et al.* [12] learned a distance metric across the Euclidean space and a Riemannian manifold which facilitates video-to-still face recognition. In [21], a discriminant low-rank embedding is learned using a probabilistic model for face verification and clustering. Song *et al.* [25] proposed a method to fully utilize the training data in a batch by considering the full pairwise distances between samples. Yang *et al.* [32] proposed an efficient algorithm to learn a similarity measure by only using similar pairs.

Multi-Task Learning: Jointly modeling the related tasks can exploit extra information among the tasks and the performance has been shown to be better than single task learning [2]. Parameswaran *et al.* [18] proposed a multi-task large margin metric learning (mt-LMNN) method which is an extension of the work in [29]. mt-LMNN learned a common projection for all the tasks and task-specific projections for each task. Bhattarai *et al.* [2] improved the mt-LMNN approach where low-rank embeddings are learned for the projection matrices. Kang *et al.* [14] investigated the feature sharing among the tasks and an algorithm was developed to determine which tasks should share the representations. Yang *et al.* [31] proposed to couple the multiple metric learning tasks with Neumann divergence, which preserves the geometry of the original data. Another class of works in multi-task learning ex-

ploits the competing relationships among tasks which discourages feature sharing or projection coupling among multiple tasks. Du *et al.* [8] proposed a boosting framework which excludes feature sharing among face verification and age verification tasks. In [9], Hwang *et al.* learned a tree of metrics where the children metrics are enforced to be disjoint with their ancestors. Romera-Paredes *et al.* [20] proposed to impose an orthogonal regularization on unrelated tasks which leads to a more informative representation for each task.

The proposed approach shares a similar spirit with works which exploit to jointly learn competing tasks. However, to the best of our knowledge, none of the above works focus on the pose-robust face verification task which is very challenging. In addition, we propose a pose-robust metric learning approach and introduce an orthogonal regularization on the projection matrices, which is different from all the works discussed above.

3. Proposed Approach

In this section, we describe the proposed metric learning framework. The learned metrics are robust to pose variations by coordinating with competing tasks, which exploit pose-sensitive information. After the metrics are learned, we demonstrate how to use them for pose-invariant face verification. In the following subsections, we first briefly review the joint Bayesian metric learning technique as the baseline method followed by the details of the proposed algorithm.

3.1. Joint Bayesian Metric Learning

Joint Bayesian method is widely used and shows good performance for face verification tasks [3, 24, 4]. The main idea of joint Bayesian is to model the joint distribution of a pair of feature vectors and maximize the log likelihood ratio of intra-class and inter-class distributions [3]. The final formulation of joint Bayesian can also be interpreted as a combination of Mahalanobis distance and projected cosine similarity. Instead of using statistical techniques to model the problem, Chen *et al.* [4] directly optimized the distance in a large-margin framework as follows:

$$\underset{\mathbf{W},\mathbf{V},b}{\operatorname{argmin}} \sum_{ij} \max\{0, \alpha - l_{ij}(b - d_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j) + 2s_{\mathbf{V}}(\mathbf{x}_i, \mathbf{x}_j))\} \quad (1)$$

where $d_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}^T \mathbf{W}(\mathbf{x}_i - \mathbf{x}_j)$ is the Mahalanobis distance and $s_{\mathbf{V}}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{V}^T \mathbf{V} \mathbf{x}_j$ is the projected similarity. Both $\mathbf{W} \in \mathbb{R}^{n \times d}$ and $\mathbf{V} \in \mathbb{R}^{n \times d}$ are the projection matrices. Here the projection matrices are either low rank embeddings ($n < d$) or full rank transformations ($n = d$). $l_{ij} = 1$ if $\{\mathbf{x}_i, \mathbf{x}_j\}$ is a positive pair and $l_{ij} = -1$, otherwise. b is the bias and α is the margin parameter. The optimization problem in (1) can be efficiently solved by SGD. The details can be found in [4].

3.2. Learning by Exploiting Competing Tasks

In order to fully exploit pose-sensitive information and coordinate with face verification task, we construct an auxiliary competing task called pose verification. Different from the main task of face verification, pose verification aims to learn the pose-sensitive information in features. More specifically, given a pairs of features $\{\mathbf{y}_i, \mathbf{y}_j\}$, the algorithm generates large (small) similarity scores when $\mathbf{y}_i$ and $\mathbf{y_j}$ have similar (different) poses. One key property is that the similarity scores should only depend on the similarity of poses, regardless of whether the features come from the same person or not.

For the main task of face verification, we rewrite the hinge-loss objective function in (1) as $\mathcal{L}_f(\mathbf{W_f}, \mathbf{V_f}, b_f)$. Similarly, we can denote the objective function for pose verification as $\mathcal{L}_p(\mathbf{W_p}, \mathbf{V_p}, b_p)$. Intuitively, the competing relationships between the main task and the auxiliary task suggest that the projections for face verification and that for pose verification should be uncorrelated. In addition, the features used for both tasks should be extracted from the same feature pool, which makes the projection matrices for different tasks comparable. The joint multi-task model is formulated by:

$$\underset{\substack{\mathbf{W}_f, \mathbf{V}_f, b_f, \\ \mathbf{W}_p, \mathbf{V}_p, b_p}}{\operatorname{argmin}} \mathcal{L}_f(\mathbf{W}_f, \mathbf{V}_f, b_f) + \mathcal{L}_p(\mathbf{W}_p, \mathbf{V}_p, b_p) + \lambda_1 \|\mathbf{W}_f^T \mathbf{W}_p\|_F^2 + \lambda_2 \|\mathbf{V}_f^T \mathbf{V}_p\|_F^2 \quad (2)$$

where λ_1, λ_2 are regularization parameters. The projection matrices are chosen to be low-rank embeddings and can be initialized by PCA of the training data. The low rank embeddings not only efficiently simplify the computation complexity, but also eliminate the underlying noise and provide better performance [21]. Although the optimization function for the projection matrices is non-convex, the algorithm still achieves good results [2].

The above objective function has two parts. The first two terms jointly minimize the verification errors for both tasks, while the last two terms enforce the orthogonal regularizations on the projection matrices for face and pose verification. Compared to the baseline method, the projection matrices for face verification learned by the proposed framework are more robust to pose variations because they not only encode the identity-sensitive information, but also mitigate the pose-sensitive information by coordinating with the pose verification task.

We use the stochastic gradient descent (SGD) method to optimize the objective function in (2). In each iteration, we randomly pick up a positive or negative pair of training samples $\{\mathbf{x}_i, \mathbf{x}_j\}$ for face verification and $\{\mathbf{y}_i, \mathbf{y}_j\}$ for

pose verification. If the similarity condition is violated, we update $\mathbf{W}_f, \mathbf{W}_p, \mathbf{V}_f, \mathbf{V}_p, b_f, b_p$ as follows:

$$\begin{aligned}
\mathbf{W}_f^{t+1} &= \begin{cases} \mathbf{W}_f^t, & \text{if } l_{ij}\rho_{ij} \geq \alpha_f \\ \mathbf{W}_f^t - \tau_f(l_{ij}\mathbf{W}_f^t\mathbf{\Psi}_{ij} + \lambda_1\mathbf{W}_p^t{\mathbf{W}_p^t}^T\mathbf{W}_f^t), & \text{otherwise,} \end{cases} \\
\mathbf{V}_f^{t+1} &= \begin{cases} \mathbf{V}_f^t, & \text{if } l_{ij}\rho_{ij} \geq \alpha_f \\ \mathbf{V}_f^t + \tau_f(l_{ij}\mathbf{V}_f^t\mathbf{\Gamma}_{ij} + \lambda_2\mathbf{V}_p^t{\mathbf{V}_p^t}^T\mathbf{V}_f^t), & \text{otherwise,} \end{cases} \\
b_f^{t+1} &= \begin{cases} b_f^t, & \text{if } l_{ij}\rho_{ij} \geq \alpha_f \\ b_f^t + \tau_f l_{ij}, & \text{otherwise,} \end{cases}
\end{aligned} \tag{3}$$

$$\begin{aligned}
\mathbf{W}_p^{t+1} &= \begin{cases} \mathbf{W}_p^t, & \text{if } a_{ij}\theta_{ij} \geq \alpha_p \\ \mathbf{W}_p^t - \tau_p(a_{ij}\mathbf{W}_p^t\mathbf{\Phi}_{ij} + \lambda_1\mathbf{W}_f^{t+1}{\mathbf{W}_f^{t+1}}^T\mathbf{W}_p^t), & \text{otherwise,} \end{cases} \\
\mathbf{V}_p^{t+1} &= \begin{cases} \mathbf{V}_p^t, & \text{if } a_{ij}\theta_{ij} \geq \alpha_p \\ \mathbf{V}_p^t + \tau_p(a_{ij}\mathbf{V}_p^t\mathbf{\Delta}_{ij} + \lambda_2\mathbf{V}_f^{t+1}{\mathbf{V}_f^{t+1}}^T\mathbf{V}_p^t), & \text{otherwise,} \end{cases} \\
b_p^{t+1} &= \begin{cases} b_p^t, & \text{if } a_{ij}\theta_{ij} \geq \alpha_p \\ b_p^t + \tau_p a_{ij}, & \text{otherwise,} \end{cases}
\end{aligned} \tag{4}$$

where τ_f, τ_p are the learning rates, l_{ij}, a_{ij} are training labels, $\mathbf{\Psi}_{ij} = (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$, $\mathbf{\Phi}_{ij} = (\mathbf{y}_i - \mathbf{y}_j)(\mathbf{y}_i - \mathbf{y}_j)^T$, $\mathbf{\Gamma}_{ij} = \mathbf{x}_i\mathbf{x}_j^T + \mathbf{x}_j\mathbf{x}_i^T$, $\mathbf{\Delta}_{ij} = \mathbf{y_i}\mathbf{y}_j^T + \mathbf{y}_j\mathbf{y}_i^T$, $\rho_{ij} = b_f - d_{\mathbf{W}_f}(\mathbf{x}_i, \mathbf{x}_j) + 2s_{\mathbf{V}_f}(\mathbf{x}_i, \mathbf{x}_j)$, $\theta_{ij} = b_p - d_{\mathbf{W}_p}(\mathbf{y}_i, \mathbf{y}_j) + 2s_{\mathbf{V}_p}(\mathbf{y}_i, \mathbf{y}_j)$. Instead of updating at every iteration, the regularization terms are updated only when the similarity condition is violated. In practice, this strategy significantly reduces the computation complexity but yields similar results.

Although the same deep features are used for both tasks, the difficulties for the main task (face verification) and the auxiliary task (pose verification) are very different since the deep neural networks are trained solely by the identity labels. Therefore, the features are more specific to identity information. To solve this problem, we pre-train the pose verification model using the pose labels. The pre-trained model can thus encode more pose information from the features and provide a good initialization of the pose metrics for multi-task learning. The whole procedure is summarized in Algorithm 1.

3.3. Pose-Robust Face Verification

Although the joint model learns two metrics, one for the main task and the other for the auxiliary task, we only utilize the face verification model to achieve better performance on the main task. Once the projection matrices $\mathbf{W}_f, \mathbf{V}_f$ are learned, we calculate the similarity scores of the testing pairs $\{\mathbf{x}_i, \mathbf{x}_j\}$ as

$$sim(\mathbf{x}_i, \mathbf{x}_j) = 2s_{\mathbf{V}_f}(\mathbf{x}_i, \mathbf{x}_j) - d_{\mathbf{W}_f}(\mathbf{x}_i, \mathbf{x}_j) \tag{5}$$

The learned bias b_f is not included in the final formulation of the similarity scores because the bias is only an uniform offset and will not change the final performance.

Algorithm 1 Stochastic Gradient Descent (SGD) for the Proposed Multi-Task Metric Learning

Input: Training pairs X with the associated labels L for the face verification and pairs Y with the labels A for the pose verification, margin α, parameter λ_1, λ_2, maximum iteration number N

1: **Pre-train Pose Model**: Pre-train the pose model $\mathbf{W}_{p0}, \mathbf{V}_{p0}, b_{p0}$ using (1)
2: **Initialization**: Initialize $\mathbf{W}_{f0}, \mathbf{W}_{f0}$ using PCA, $b_{f0} = 0$, $\mathbf{W}_{p0}, \mathbf{V}_{p0}, b_{p0}$ from the pre-trained model
3: **for** t = 1:N **do**
4: Randomly pick up a pair $\{\mathbf{x}_i, \mathbf{x}_j\}$, update the face verification model $\mathbf{W}_f^t, \mathbf{V}_f^t, b_f^t$ using (3)
5: Randomly pick up a pair $\{\mathbf{y}_i, \mathbf{y}_j\}$, update the pose verification model $\mathbf{W}_p^t, \mathbf{V}_p^t, b_p^t$ using (4)
6: **end for**

Output: Projection matrices $\mathbf{W}_f, \mathbf{W}_p, \mathbf{V}_f, \mathbf{V}_p$ and biases b_f, b_p

4. Experiments

In this section, we evaluate the proposed approach on three challenging datasets: IARPA Janus Benchmark A (IJB-A), Janus Challenging Set 3 Covariates (CS3_cov), and Celebrities in Frontal-Profile (CFP). We begin with introducing the details of the datasets and the experimental settings. Then discussions on the experimental results are presented.

IARPA Janus Benchmark A [15]: This dataset contains 500 subjects with a total of 5,397 images and 2,042 videos. For the evaluation purpose, ten splits are generated based on different training / test set division. Each training set and test set contains 333 subjects and 167 subjects, respectively. The dataset contains many extreme pose and illumination variations and some sample images are shown in Fig. 2. The IJB-A verification protocol has around 11,748 pairs of gallery-to-probe templates (1,756 positive and 9,992 negative pairs), with each templates containing a mixture of images and video frames.

Janus Challenging Set 3 Covariates: The Janus Challenging Set 3 (CS3) dataset contains 1871 subjects and 68716 images and video frames. The covariates protocol aims to focus on the effect of eight different covariates (age, eyes visible, facial hair, forehead visible, gender, indoor, nose and mouth visible and skin tone) on the verification performance. The protocol evaluates 20,866,895 pair of templates (5,961,839 positive and 14,905,056 negative pairs) where each template contains one image or frame. Some sample images are shown in Fig. 3.

Celebrities in Frontal-Profile [23]: This dataset investigates the influence of extreme pose variations on the face verification performance. The dataset contains 500 sub-

Figure 2: Sample images in IJB-A dataset.

Figure 3: Sample images in CS3 dataset.

Figure 4: Sample images in CFP dataset.

jects, with ten frontal and four profile images for each subject. Most of the profile images are in extreme poses and some sample images are shown in Fig. 4. For the evaluation protocol, there are two settings: frontal-to-frontal and frontal-to-profile face verification. For each setting, it consists of ten disjoint splits and each split has 350 positive and 350 negative pairs. The final performance is averaged over ten splits. In this paper, we focus on the pose-varied face verification and thus only run the experiments for the frontal-to-profile protocol.

4.1. Experimental Setup

Features: The deep CNN features used in all the experiments of this work are extracted using the architecture proposed in [4]. The model consists of ten convolutional layers, five pooling layers and one fully connected layer and is trained using the CASIA-WebFace dataset [33]. The output of the pool5 layer is used as the final features and the dimensionality of the features is 320. All the features are l_2 normalized before computing the similarity score. In the IJB-A dataset, there are more than one samples in each templates. We perform media averaging similar to the one reported in [21].

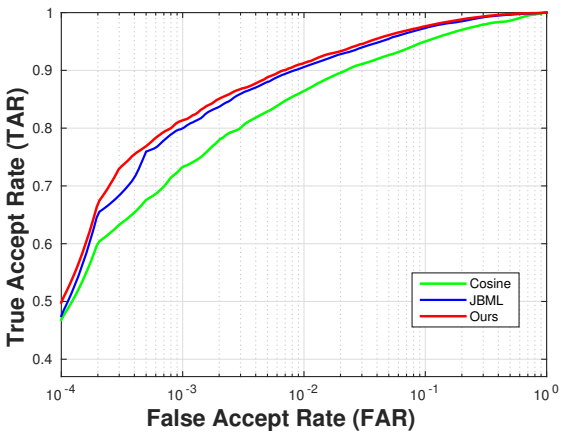
Auxiliary Task Design: The face poses used in this paper are estimated by the approach discussed in [19]. Since the estimated poses may not be perfectly accurate, we cluster the poses into groups and treat the poses equally within each group. For CS3 and IJB-A datasets, we divide the poses into four groups and for the CFP dataset three groups are generated. In order to avoid the identity bias in the pose groups (some subjects may have more large poses than others), we randomly choose samples from different subjects for each pose group. The positive pairs are selected by randomly picking up two samples in the same group and the negative pairs consist of samples picked from different groups.

Accuracy Metrics: To evaluate the CS3 Covariates and IJB-A verification performance, we follow the evaluation protocol defined in [15]. The original protocol reports ROC curves as well as the True Acceptance Rate (TAR) when False Alarm Rate (FAR) equals $10^{-3}, 10^{-2}, 10^{-1}$. For the CS3 Covariates protocol, the total number of pairs is extremely large (about 20 million pairs), thus we also report the TAR at FAR$= 10^{-5}, 10^{-4}$. In addition, we also analyze the performance under different covariates that are related with poses (eyes visibility, forehead visibility). The accuracy metrics used for the CFP dataset follow the protocol in [23]. Area under the curve (AUC) and Equal Error Rate (EER) are computed for each split, as well as the classification accuracy. The performance is reported by averaging over ten splits. For classification accuracy, we select the threshold that provides highest accuracy on the training set.

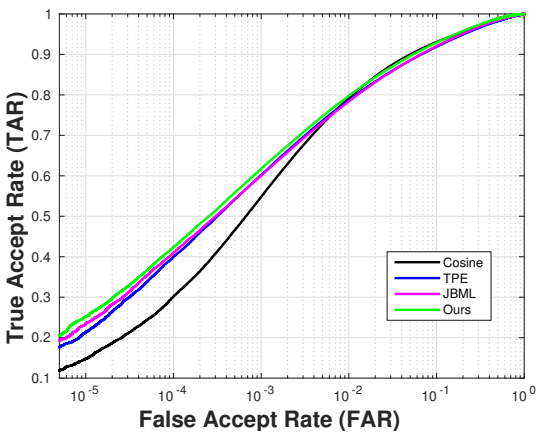
Parameters: We set the margin $\alpha_f = \alpha_p = 0.001$. Intuitively, a small margin encourages the projection matrices to be updated only by the hard negative/positive pairs since small margins result in less strict condition than large margins. The hard negative mining yields a similar idea and has been widely used for SGD updating. Based on the above observation, we choose the margin to be a small value. The initialization of the projection matrices for CS3 dataset is the whitening PCA of the training data while for IJB-A and CFP datasets, we find that initialization using WCPA makes the projection matrices have very large values and thus they become unstable. Therefore, we use PCA to initialize the projection matrices. The learning rate is set to be $3 \times 10^{-4}, 5 \times 10^{-3}, 3 \times 10^{-3}$ for CS3, IJB-A and CFP respectively.

4.2. Evaluation Results on IJB-A dataset

Compared Methods: The experimental results of the proposed approach are compared with two baseline methods, the cosine similarity and the joint Bayesian metric learning (JBML). The cosine similarity measure is computed directly from the raw features without and training while JBML is learned by using the identity labels of the training data. We also compare with the triplet probabilistic embedding (TPE) method [21], which is the state-of-the-art metric learning approach. We use the same features to



(a) ROC curves for IJB-A



(b) ROC curves for CS3 Covariates

Figure 5: ROC curves for CS3 Covariates and IJB-A dataset. The results are averaged over 10 splits for IJB-A dataset.

Method	TAR@FAR = 10^{-3}	TAR@FAR = 10^{-2}	TAR@FAR = 10^{-1}
Cosine	0.734±0.042	0.864±0.014	0.950±0.006
Pooling Face[10]	0.631	0.819	-
Augmented Face[17]	0.725	0.886	-
Template Adaptation [5]	-	**0.939±0.013**	-
TPE [21]	0.813±0.020	0.900±0.010	0.964±0.005
JBML	0.799±0.022	0.906±0.010	0.973±0.004
Proposed Method	**0.814±0.027**	0.913±0.010	**0.977±0.003**

Table 1: Verification results for IJB-A dataset. Results are averaged over ten splits.

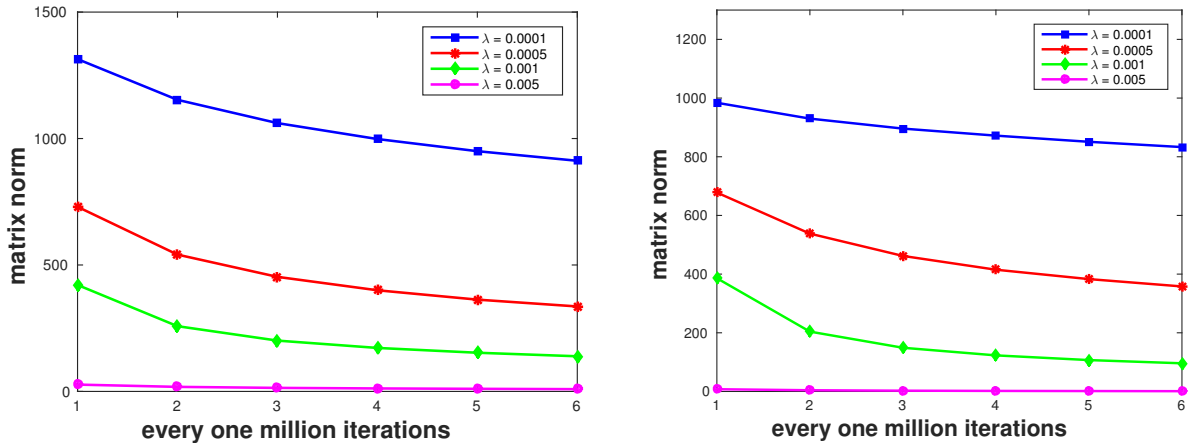
compute the similarity scores for cosine similarity, JBML and our method. In contrast, the results for TPE are directly cited from [21] and it is noticed that the raw features used in [21] have a better baseline performance than our features. For better comparison, we also include the experimental results of some recently proposed methods on IJB-A datasets.

Table 1 summarizes the results for the IJB-A dataset. To better visualize the performance, ROC curves are shown in Fig. 5a. It can be seen that the proposed approach performs better than methods based on cosine similarity, JBML and slightly outperforms TPE [21], though the DCNN features used in [21] obtain better cosine baseline than the one used in this paper. Although [17] mitigates the problem due to pose variations by augmenting test samples in different poses, the proposed approach still outperforms it. Although [5] achieves better results than the proposed method, it adopts on-the-fly training during testing time with the help of a pre-selected large negative set as one-shot similarity kernel [30] and thus requires much more testing time than ours. These highlight the effectiveness of the proposed approach by mitigating pose-sensitive information.

4.3. Evaluation Results on CS3 Covariates

General Performance: For a fair comparison, the same features are used for all the methods. We plot the ROC curves for the CS3 Covariates protocol in Fig. 5b and Table 2 shows the True Acceptance Rate (TAR) when False Alarm Rate (FAR) at different values. We notice that the proposed approach performs better than JBML baseline and TPE. Interestingly, we observe that JBML and TPE perform slightly worse than the cosine baseline at FAR= $10^{-2}, 10^{-1}$. Possibly this is because the training set may not contain sufficient face images with large poses and the learned metrics is biased to frontal or near-frontal faces. When the projection matrices are applied to the test data, where many faces are in extreme poses, the performance goes down. In contrast, the proposed method explicitly avoids the pose informations in the metrics for the main task, and thus it is more pose-robust than the baseline metric.

Covariates Analysis: In order to better understand how the covariates affect the verification performance, we evaluate two pose-relevant covariates, eye visibility and forehead visibility, and present the results. Tables 3 and 4 show the experimental results for cosine, JBML and the proposed



(a) (b)

Figure 6: Frobenius norm of the regularization terms for W and V matrices over iterations.

Method	TAR@FAR = 10^{-5}	TAR@FAR = 10^{-4}	TAR@FAR = 10^{-3}	TAR@FAR = 10^{-2}	TAR@FAR = 10^{-1}
Cosine	0.148	0.302	0.548	0.792	**0.931**
JBML	0.236	0.410	0.601	0.784	0.921
TPE	0.213	0.400	0.602	0.788	0.919
Proposed Method	**0.252**	**0.424**	**0.618**	**0.798**	0.930

Table 2: Verification results for CS3 covariates protocol.

method over eye and forehead visibility. *same* represents the same visibility (both visible or non-visible) and *different* means different visibility for the compared faces. Generally, the performance for same visibility is better than that for different visibility. Since eye and forehead visibility partially reflect the pose variations, it demonstrates that pose variations indeed degrade the performance. We notice that the proposed approach consistently outperforms the joint Bayesian baseline and cosine similarity for most cases except slightly worse than cosine similarity at FAR= 10^{-1}. Moreover, the improvement of our method over the baseline shows similar trends for *same* and *different* visibility cases. This reveals that the pose variations still exist, though smaller than that in different visibility case, in the same visibility situation.

Regularization Parameter Analysis: The regularization parameter λ controls the orthogonality of the projection matrices for the two tasks. We investigate the function of the regularization terms by varying the values of λ. The Frobenius norm of the inner product of the projection matrices for the two tasks are shown in Fig. 6. We see that the Frobenius norm monotonically goes down as the iterations increase and a larger λ results in a more strict regularization on the projection matrices. When λ is large enough (typically larger than 0.005), the Frobenius norm becomes small and does not change much. We also run experiments to see how the performance changes with different λ's and does not notice much difference when λ changes from 10^{-4} to 5×10^{-3}.

4.4. Evaluation Results on CFP dataset

For the CFP dataset, the experimental results are given in Table 5. Surprisingly, we see that the proposed approach only slightly outperforms the JBML baseline on accuracy by 0.5% and performs a bit worse on EER. Intuitively, the learned metrics should alleviate the pose mismatch in the test pairs and improve the JBML performance. We further conduct experiments to see the underlying reasons for this issue. We find that the pose metric converges much faster than the identity metric. The accuracy for the pose verification is almost 100%. Considering the fact that the dataset only consists frontal and profile faces, the learned pose metric is not discriminative enough to small pose differences. At the same time, we draw the plot in Fig. 7 that the regularization term does not change much during joint training. This further demonstrates that the regularization term does not affect the face metric much.

Eye Visible	Method	TAR@FAR = 10^{-5}	TAR@FAR = 10^{-4}	TAR@FAR = 10^{-3}	TAR@FAR = 10^{-2}	TAR@FAR = 10^{-1}
Same	Cosine	0.146	0.297	0.546	0.798	**0.935**
	JBML	0.219	0.404	0.605	0.792	0.926
	Proposed Method	**0.243**	**0.418**	**0.622**	**0.805**	0.933
Different	Cosine	0.118	0.254	0.468	0.705	**0.888**
	JBML	0.220	0.344	0.503	0.694	0.874
	Proposed Method	**0.221**	**0.350**	**0.515**	**0.709**	0.886

Table 3: Covariates analysis on eye visibility. *Same* represents that the two face images in a pair are both eye visible or non-visible, and *Different* means that one of the faces is eye visible while the other is non-visible.

Forehead Visible	Method	TAR@FAR = 10^{-5}	TAR@FAR = 10^{-4}	TAR@FAR = 10^{-3}	TAR@FAR = 10^{-2}	TAR@FAR = 10^{-1}
Same	Cosine	0.145	0.305	0.559	0.796	**0.930**
	JBML	0.219	0.413	0.609	0.788	0.919
	Proposed Method	**0.245**	**0.430**	**0.624**	**0.798**	0.926
Different	Cosine	0.161	0.294	0.530	0.785	0.933
	JBML	0.260	0.404	0.586	0.777	0.923
	Proposed Method	**0.267**	**0.415**	**0.608**	**0.797**	**0.935**

Table 4: Covariates analysis on forehead visibility. *Same* represents that the two face images in a pair are both forehead visible or non-visible, and *Different* means that one of the faces is forehead visible while the other is non-visible.

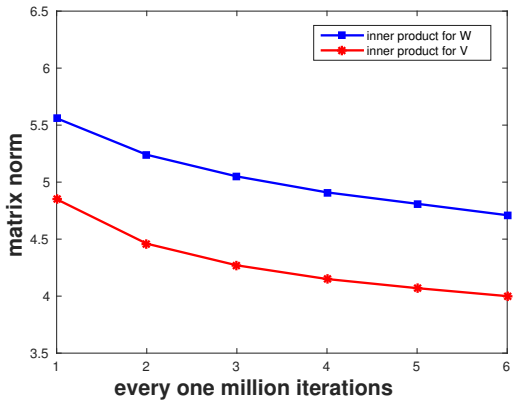


Figure 7: Frobenius norm of the regularization terms for W and V matrices over iterations.

Method	Accuracy	EER	AUC
Cosine	0.904	0.094	0.967
Sengupta *et al.* [23]	0.849	0.150	0.930
JBML	0.924	**0.068**	**0.981**
Proposed Method	**0.929**	0.071	**0.981**

Table 5: Verification results for the frontal-to-profile protocol of the CFP dataset. Results are averaged over ten splits.

5. Conclusion

In this paper, we showed the benefit of cooperating with the pose verification task for pose-robust face verification. We proposed a joint model to learn the metrics for the two tasks together and enforce an orthogonal regularization on the learned projection matrices for the two tasks. By mitigating the information contained in the auxiliary task, the learned metric for face verification is more pose-robust. We conducted extensive experiments on three challenging datasets and the experimental results show that the proposed approach improves the baseline methods and are competitive with the state-of-the-art.

Acknowledgments

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan, et al. Face recognition using deep multi-pose representations. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[2] B. Bhattarai, G. Sharma, and F. Jurie. Cp-mtml: Coupled projection multi-task metric learning for large scale face re-

trieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *European Conference on Computer Vision*, pages 566–579. Springer, 2012.

[4] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[5] N. Crosswhite, J. Byrne, O. M. Parkhi, C. Stauffer, Q. Cao, and A. Zisserman. Template adaptation for face verification and identification. *arXiv preprint arXiv:1603.03958*, 2016.

[6] C. Ding and D. Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(3):37, 2016.

[7] C. Ding, C. Xu, and D. Tao. Multi-task pose-invariant face recognition. *IEEE Transactions on Image Processing*, 24(3):980–993, 2015.

[8] L. Du and H. Ling. Cross-age face verification by coordinating with cross-face age verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2329–2338, 2015.

[9] K. Grauman, F. Sha, and S. J. Hwang. Learning a tree of metrics with disjoint visual features. In *Advances in neural information processing systems*, pages 621–629, 2011.

[10] T. Hassner, I. Masi, J. Kim, J. Choi, S. Harel, P. Natarajan, and G. Medioni. Pooling faces: template based face recognition with pooled face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 59–67, 2016.

[11] H. T. Ho and R. Chellappa. Pose-invariant face recognition using markov random fields. *IEEE transactions on image processing*, 22(4):1573–1584, 2013.

[12] Z. Huang, R. Wang, S. Shan, L. Van Gool, and X. Chen. Cross euclidean-to-riemannian metric learning with application to face recognition from video. *arXiv preprint arXiv:1608.04200*, 2016.

[13] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. In *European Conference on Computer Vision*, pages 808–821. Springer, 2012.

[14] Z. Kang, K. Grauman, and F. Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 521–528, 2011.

[15] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939. IEEE, 2015.

[16] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3499–3506, 2013.

[17] I. Masi, A. T. Tran, J. T. Leksut, T. Hassner, and G. Medioni. Do we really need to collect millions of faces for effective face recognition? *arXiv preprint arXiv:1603.07057*, 2016.

[18] S. Parameswaran and K. Q. Weinberger. Large margin multi-task metric learning. In *Advances in neural information processing systems*, pages 1867–1875, 2010.

[19] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *arXiv preprint arXiv:1603.01249*, 2016.

[20] B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil. Exploiting unrelated tasks in multi-task learning. In *AISTATS*, volume 22, pages 951–959, 2012.

[21] S. Sankaranarayanan, A. Alavi, C. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. *arXiv preprint arXiv:1604.05417*, 2016.

[22] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[23] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[24] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *BMVC*, volume 2, page 4, 2013.

[25] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. *arXiv preprint arXiv:1511.06452*, 2015.

[26] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2892–2900, 2015.

[27] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[28] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[29] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

[30] L. Wolf, T. Hassner, and Y. Taigman. The one-shot similarity kernel. In *2009 IEEE 12th International Conference on Computer Vision*, pages 897–902. IEEE, 2009.

[31] P. Yang, K. Huang, and C.-L. Liu. Geometry preserving multi-task metric learning. *Machine learning*, 92(1):133–175, 2013.

[32] Y. Yang, S. Liao, Z. Lei, and S. Z. Li. Large scale similarity learning using similar pairs for person verification. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[33] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[34] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 113–120, 2013.