ENVIRONMENT COUPLED METRICS LEARNING FOR UNCONSTRAINED FACE VERIFICATION

Xinyuan.Cai Chunheng.Wang Baihua.Xiao Ji.Zhou Xue.Chen

State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Science, Beijing, 100190, China
{xinyuan.cai, chunheng.wang, baihua.xiao, ji.zhou, xue.chen}@ia.ac.cn

ABSTRACT

Making recognition more reliable under unconstrained environment is one of the most important challenges for real-world face recognition. In this paper, we propose a novel approach for unconstrained face verification. First, we use a spectral-clustering method based on Structural Similarity index to estimate the captured environments of facial images. Then for each pair of environments, we learn two coupled metrics, such that facial images captured in different environments can be transformed into a media subspace, and high recognition performance can be achieved. The coupled transformations are jointly determined by solving an optimization problem in the multi-task learning framework. Experimental results on the benchmark dataset (LFW) show the effectiveness of the proposed method in face verification across varying environments.

Index Terms— Face verification, metric learning, unconstrained environment, multi-task learning

1. INTRODUCTION

In the past several decades, face recognition has received a great deal of attention from the scientific and industrial communities, due to its wide range of applications including access control, security and surveillance. With thousands of published papers, face recognition under well controlled environment is relatively mature and provides high recognition rates. However, when images are collected under uncontrolled environment, such as uncontrolled lighting, pose variations, the performance decreases significantly.

In this paper, we address the unconstrained face verification. Face verification is a binary classification problem over pairs of face images, and we have to decide whether the same person is depicted in both images. One typical application for face verification is self-serviced immigration clearance using E-passport. Obviously, it is not a good choice to directly pass the query image to the face verification system enrolled with normal images, because the variations between images of the same face due to illuminations or viewing direction are almost always larger than image variations due to change in face identity[1].(As in Fig.1)

There has been a lot of prior works on comparing images under unconstrained environments. A commonly used way is to restore the probe image into the same condition of gallery. For this purpose, many image processing techniques for illumination normalization [1, 2] or illumination invariant features [3, 4] have been proposed. And for the problem of pose variation, some researchers proposed feasible approaches to synthesize virtual images across pose in 2D space as pose transformation [5] or in 3D space as 3D face reconstruction and projection. For the problem of expression, 2D warps or 3D morphable model [6] is used to generate some virtual expression images.

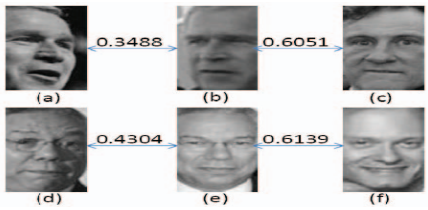

Figure 1: Illustration of the similarity between face images across pose and illumination. The similarity between images is computed using SSIM measure (Eq.1), where 1 is the score for identical images and 0 is the lowest score. (a) and (b) ((d) and (e)) are face images of the same identity, but are captured under different pose (illumination). The similarity between (a) and (b) ((d) and (e)) are smaller than (b) and (c) ((d) and (e)) of different identity.

Besides explicitly dealing with the problem of pose or illumination, many researchers apply some statistical machine learning methods for uncontrolled face recognition. Distance metric learning [7] is one of the most widely used approaches. The goal of distance metric learning is to learn a proper transformation matrix, so that in the transformed space, the distance between the images of the same identity is small, while the distance between different identities is large. Recently, some researchers apply some other statistical approaches, such as canonical correlation analysis (CCA) [8], and Partial Least Squares (PLS) [9]. Their objective is to find basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors is mutually maximized.

In this paper, inspired by previous works [7,8,9], we propose a novel way for unconstrained face verification. The main contributions of our approach lie in two aspects: Firstly, we propose a data-driven approach to estimate the captured environment. We use the Structural SIMilarity index (SSIM) [10] to measure the pixelwise similarity of two face images, and then a spectral clustering method is

applied to cluster the training images into different environments. Secondly, for comparing images captured in two different environments, we find two coupled transformation matrices(as illustrated in Fig.2), such that in a common subspace, the similarity between faces of the same identity is near one, while the similarity of different identity is near zero. We formulate the problem of simultaneously learning the environment coupled metrics for several pairs of environments as a multi-task learning problem, and an alternating method is applied to solve the problem efficiently.

2. STRUCTURAL SIMILARITY INDEX BASED ENVIRONMENT CLUSTERING

In a real face recognition system, face images may be captured under different environments, where viewpoint, illumination, expression, and occlusion can vary considerably. All these factors are confounded in the image data.

In this section, we use a clustering method to estimate the environments implicitly. The key step for clustering is to define the similarity. We apply the Structural SIMialrity index (SSIM) [10] to measure the similarity between images. According to [11], SSIM yielded the most robust performance across multiple pose and illumination. The SSIM measure between pixels in the same location x in image I_1 and I_2 is given by

$$S_{I_1,I_2}(x) = \frac{u_1(x)u_2(x)\sigma_{12}(x)}{[u_1^2(x)+u_2^2(x)][\sigma_1^2(x)+\sigma_2^2(x)]} \quad (1)$$

where μ_i, σ_i are the local mean and standard deviation of a fixed window around x.The value σ_{12} is the correlation of pixels in this window in both images. The values of S_{I_1,I_2} are averaged across all pixel location to produce the final similarity measure. This measure ranges between [0, 1], where 1 is achieved for identical images.

After the definition of similarity, we can measure the similarities between all the training facial images. It results in a similarity matrix $W \in R^{n\times n}$, where n is the number of training images. If n is very large, the similarity matrix W will be very dense and large. So we define a sparse similarity matrix, where just the similarity between one image and its K nearest neighbors are considered (the k nearest neighbors are selected based on SSIM). It can be formulated as:

$$W_{i,j} = \begin{cases} SSIM(I_i, I_j) & if\ j \in knn(i) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

And then a spectral clustering method is applied to get the implicit environments. Some example images of clustering results are shown in Fig 3.

3. ENVIRONMENT COUPLED METRIC LEARNING

The output of last section is the environment label that each image belongs to. So if there are n facial images in the training set, it can be represented as $\{x_i, I_i, e_i\}_{\text{i=1...n}}$, where $x_i \in R^m$ is the feature vector of one image; $I_i \in R^1$ ($I_i = 1, \dots, C$)is the identity label, C is the number of individuals; and $e_i = 1, \dots, E$ is the environment label, E is the number of environments.

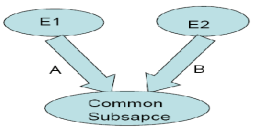

Figure 2: An illustration of coupled transformation matrices {A, B} for one pair of environment $\{E_1, E_2\}$

3.1. Coupled Metric Learning for One Pair of Environments

For simplicity, we first consider one pair of ments$\{e_1, e_2\}$.$\{x_i^1, I_i^1, e_1\}_{\text{i=1,...,n}_1}$ represents the training images in environment e_1, and$\{x_i^2, I_i^2, e_2\}_{\text{i=1,...,n}_2}$ in environment e_2. The basic idea of our method is to learn two coupled metrics, which map x_i^1 and x_j^2 respectively into a joint new subspace, where the new distance measure is more ideal for face recognition.

We represent two coupled metrics as A and B, each for one environment. And we define the linear transformation in matrix form as follows:

$$y_i^1 = Ax_i^1, \qquad y_j^2 = Bx_j^2 \quad (3)$$

where $A \in R^{d\times m}, B \in R^{d\times m}$, and y represent the new feature in the d dimensional transformed subspace. The objective is that in the new subspace, the similarity of two face images representing the same individual should be near one, and the similarity of two face images representing different individuals should be near zero. In our experiment, we use Euclidean distance measure between two images in the transformed subspace, and a sigmoid function is applied to obtain a probabilistic estimation of whether the two images depict the same individual. The probability can be modeled as:

$$p(x_i^1, x_j^2) = p(I_i^1 = I_j^2 \mid x_i^1, x_j^2, A, B) = \sigma(b - d_{A,B}(x_i^1, x_j^2)) \quad (4)$$

Where $$\sigma(z) = (1+\exp(-z))^{-1} \quad (5)$$

$$d_{A,B}(x_i^1, x_j^2) = (Ax_i^1 - Bx_j^2)^T (Ax_i^1 - Bx_j^2) \quad (6)$$

and b is a bias term.

We formulate the coupled metric learning as an optimization problem. The objective function to be maximized is defined as:

$$\max_{A,B,b} L(A,B,b) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}[\delta(I_i^1, I_j^2)\ln p(x_i^1, x_j^2) + (1-\delta(I_i^1, I_j^2))\ln(1-p(x_i^1, x_j^2))] \quad (7)$$

where$\delta(I_i^1, I_j^2) = 1$,if $I_i^1 = I_j^2$, otherwise $\delta(I_i^1, I_j^2) = 0$. This objective function can be seen as the log-likelihood of all the images pairs between two environments, and it is very similar to Logistic Discriminate Metric Learning[7], but we learn two coupled metrics instead of just one metric.

$L(A,B,b)$is differentiable with regard to matrix A, B and b, so we can optimize it using a gradient based optimizer. The gradient of $L(A,B,b)$ can be computed as:

$$\frac{\partial L(A,B,b)}{\partial A} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\frac{\delta(I_i^1, I_j^1) - p(x_i^1, x_j^2)}{p(x_i^1, x_j^2)(1-p(x_i^1, x_j^2))} \times \frac{\partial p(x_i^1, x_j^2)}{\partial A} \quad (8)$$

$$\frac{\partial L(A,B,b)}{\partial b} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\delta(I_i^1, I_j^2) - p(x_i^1, x_j^2) \quad (9)$$

where $\frac{\partial p(x_i^1, x_j^2)}{\partial A} = 2p(x_i^1, x_j^2)(p(x_i^1, x_j^2)-1)(Ax_i^1 - Bx_j^2)(x_i^1)^T \quad (10)$

The gradient of $L(A,B,b)$ with regard to B can be calculated as similar as A. After obtaining the gradient, A, B and b can be updated by gradient ascend.

3.2. Multi-task Metric Learning for Several Pairs of Environments

In the last part, we just consider the problem of coupled metric learning for one pair of environments. But in practice, facial images may be captured under several environments, and we should do face verification across different environments.

Let $\{A_{p,q}, B_{p,q}\}$ denote the coupled metrics for one pair of environments $\{e_p, e_q\}$. Inspired by the methodology of multi-task learning [14], we consider the coupled metric learning for one pair of environments as one task, and model the commonalities between various tasks through shared metrics $\{A_0, B_0\}$ and the task-specific idiosyncrasies with additional matrices $\{\hat{A}_{p,q}, \hat{B}_{p,q}\}_{p=1,\ldots,E,q=1,\ldots,E}$. The final optimization problem is defined as:

$$\max_{\substack{\{A_{p,q},B_{p,q}\}\\ \{A_0,B_0\}}} \sum_{p=1}^{E}\sum_{q=p}^{E}[L(A_{p,q},B_{p,q},b_{p,q}) - \gamma_{p,q}(\|\hat{A}_{p,q}\|_F^2 + \|\hat{B}_{p,q}\|_F^2)] \quad (11)$$

$$\text{s.t. } A_{p,q} = \hat{A}_{p,q} + A_0 \qquad B_{p,q} = \hat{B}_{p,q} + B_0 \quad (12)$$

$$A_{p,q} = B_{p,q} \text{ (if p=q)} \qquad \gamma_{p,q} \geq 0 \quad (13)$$

The first term is the log-likelihood of all training samples. The second term encourages that the coupled matrixes for each pair of environments are not far away from the shared metrics, which solve the problem between small sample size and large number of parameters in some way. $\gamma_{p,q}$ is the weighting parameter, and it is set by cross validation.

To solve this problem, we use the alternating method to iteratively optimize over two set of variables: the common metric $\{A_0, B_0\}$ and the task specific metrics $\{A_{p,q}, B_{p,q}\}_{q=1\ldots E}^{p=1\ldots E}$. Specifically, given initial $\{A_0^{(0)}, B_0^{(0)}\}$, we fix $\{A_0, B_0\}$ and optimize over $\{A_{p,q}, B_{p,q}\}_{q=1\ldots E}^{p=1\ldots E}$, and then fix $\{A_{p,q}, B_{p,q}\}_{q=1\ldots E}^{p=1\ldots E}$ and optimize over $\{A_0, B_0\}$. This process is repeated until convergence or a pre-defined number of steps.

With $\{A_0, B_0\}$ fixed, the resulting problem can be decomposed into several optimization problems, each of which optimizes over $\{A_{p,q}, B_{p,q}\}$. It can be formulated as:

$$\max_{A_{p,q},B_{p,q},b_{p,q}} f(A_{p,q},B_{p,q},b_{p,q}) = L(A_{p,q},B_{p,q},b_{p,q}) - \gamma_{p,q}(\|A_{p,q}-A_0\|_F^2 + \|B_{p,q}-B_0\|_F^2) \quad (14)$$

It is an extension of problem (7) with some regularization terms, and it can be solved by gradient ascend.

With fixed $\{A_{p,q}, B_{p,q}\}_{q=1\ldots E}^{p=1\ldots E}$, we need to solve the following problem:

$$\max_{\{A_0,B_0\}} \sum_{p=1}^{E}\sum_{q=p}^{E} -\gamma_{p,q}(\|A_{p,q}-A_0\|_F^2 + \|B_{p,q}-B_0\|_F^2) \quad (15)$$

It can be solved efficiently with closed-form solution.

$$A_0 = \Big(\sum_{p=1}^{E}\sum_{q=p}^{E}\gamma_{p,q}A_{p,q}\Big) \Big/ \Big(\sum_{p=1}^{E}\sum_{q=p}^{E}\gamma_{p,q}\Big) \qquad B_0 = \sum_{p=1}^{E}\sum_{q=p}^{E}\gamma_{p,q}B_{p,q} \Big/ \sum_{p=1}^{E}\sum_{q=p}^{E}\gamma_{p,q} \quad (15)$$

4. EXPERIMENTS

We conduct experiments on the LFW dataset [12]. The LFW benchmark defines two evaluation protocols: the restricted and unrestricted setting. In our experiment, we follow the unrestricted setting. Under this setting, arbitrary number of training pairs can be generated based on the given face labels. The performance of our method is measured by 10-fold cross validation procedure. Note that, the people for testing are different from the training, so the experiment can verify the generalization ability of our model.

4.1. Experimental setting

Following [7], Viola Jones's face detector [13] is run on the original images, which gives out an approximate location and scale of the face. And then we cut the face image and normalize to 80×150 pixels. These cropped images are used for environment clustering. Nine facial points are located and then SIFT descriptors are computed at three scales, centered on these nine facial points, which are available from[7]. The dimension of the result descriptor for each point is $128 \times 3 = 384$. For testing, we first determine the captured environment for each image, and then transform the original descriptor by the learned environment coupled metrics. Finally, we compute the probability that the two faces belong to the same individual. We implement our method in MATLAB, and the source code is available upon request.

4.2. Experimental results

In order to get good initialization for the alternating optimization procedure, we use nine of the ten subsets for estimating the $\{A_0, B_0\}$. The subsets are provided by the author under restricted mode, and each subset consists of 300 same-person pairs and 300 different-person pairs. We discard the environment of each image, and get $\{A_0, B_0\}$ by solving the optimization problem (7) with the constraint $A_0 = B_0$. And we denote $M_{init} = A_0$.

In order to estimate the captured environment of each image, we cluster all the face images belonging to the nine training folders into E centers. In our experiment, according to [11], we set k=100 for constructing the similarity matrix W. And we set $E = 6$ by considering the trade-off between the number of images belonging to each cluster and the interpretation of each cluster (the sample images of clustering result and interpretation are shown in Fig(3)). When a new image comes, we use the majority voting of k nearest neighbors to determine its environment class.

After estimating the environment, we randomly generate N intra-personal pairs and N extra-personal pairs for each pair of environments, which are used for training the environment coupled metrics. (In experiment, we set N=1000). If the overall number of intra-personal pairs that can be generated from one pair of environments is smaller than 200,

we won't learn the coupled metric for this pair of environments, and we set the metric with M_{init}. The performance of M_{init} is $(79.27\pm0.60)\%$, which is equal to the result of LDML[7] under restricted setting. Table 1 shows the performance of our environment coupled metrics for the corresponding pairs of environments. It can be seen that the accuracy of comparing pairs of images under the same environment is higher than under different environments, and the performance of our learned coupled metric is much better than M_{init}. The overall accuracy of our method's verification rate at FPR=0.1 is **$(88.9\pm0.37)\%$**. In Table 2, we compare our method with other state-of-art methods (we copy the results from[15,17].). With only SIFT descriptor, our method significantly outperforms [7,16,17]. When combining four descriptors, the results of [16] and [17] are improved to$(89.50\pm0.51)\%$ and $(90.07\pm0.51)\%$ respectively, while our method with only one descriptor is comparable to them. The results verify the effectiveness of our method. The ROC curves of our method and others are depicted in Fig 4. Complete benchmark results can be found on the LFW website [15].

5. CONCLUSION

In this paper, we present a new approach named environment coupled metric learning for unconstrained face verification. First, we use SSIM based spectral clustering to estimate the image captured environments. Second, for comparing images captured in different environments, different from some previous work, our approach tends to find two coupled transformation matrixes, so that in the transformed subspace, the similarity of intra-person is near one, while the similarity of inter-person is near zero. We further formulate the problem of learning multi pairs of environment coupled metrics in the multi-task learning framework. Experiment results show the effectiveness of our proposed method. In the future work, we will investigate the performance of our method with different descriptors.

6. REFERENCES

[1] Y. Moses, Y. Adini, and S. Ullman. "Face recognition: The problem of compensating for changes in illumination direction". In ECCV(1994).

[2] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Lecture Notes in Computer Science*, 4778, 2007.

[3] H. Wang, S. Li, and Y. Wang, Face recognition under varying lighting conditions using self quotient image," In ICFG, 2004, pp. 819–824.

[4] T. Chen, W. Yin, X. Zhou, D. Comaniciu, and T. Huang, "Total variation models for variable lighting face recognition," IEEE TPAMI, vol. 28, no. 9, pp. 1519–1524, 2006

[5] D. Beymer, T. Poggio, "Face recognition from one exampl view". In ICCV(1995), pp. 500–507.

[6] S. Romdhani, T. Vetter, D. J. Kriegman, "Face recognition using 3-D models: pose and illumination", In Proc of the IEEE, vol.94, no.11, 2006.

[7] M. Guillaumin, J. Verbeek, C. Schmid. Is that you? Metric learning approaches for face identification. In ICCV(2009).

[8] A.Li, S.Shan, X.Chen and W.Gao. "Maximizing Intraindividual Cor relations for Face Recognition Across Pose Differences". In CVPR(2009).

[9] W.Schwartz, H.Guo, and L.Davis. "A Robust and Scalable aproach to face identification". In Proc.ECCV (2010).

[10] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. IEEE TIP, 13(4):600–612, 2004.

[11] F.Schroff, T.Treibitz, D.Kriegman, S.Belongie. "Pose, Illumination and Expression invariant pairwise face-similarity measure via doppelgan ger list comparison". In CVPR (2011).

[12] B. Gary, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[13] P. Viola, M.J. Jones. Robust Real-Time Face Detection. In IJCV (2004).

[14] S.Parameswaran and K.Weinberger. Large Margin Multi-task Metric Learning. In NIPS (2010).

[15] http://vis-www.cs.umass.edu/lfw/results.html:

[16] Y.Taigman, L.Wolf, and T.Hassner. Multiple One-Shots for Utilizing Class Label Information. In BMVC(2009).

[17] P.Li, Y.Fu, U.Mohammed, J.H.Elder and S.J.D.Price. Probabilistic Models for Inference About Identity. IEEE TPAMI vol.3, no.1, pp.144-157, Jan.2012.

Figure 3: Some example images of environment clustering results.(a) left profile; (b) frontal; (c) right profile: (d) most are male with smiling; (e) most are female with smiling; (f) black people or dark illumination.

	a	b	c	d	e	f
a	89.29±0.37	88.57±0.71	85.30±0.39	86.07±0.52	88.32±0.21	82.50±0.58
b		90.80±0.47	87.73±0.44	85.89±0.35	86.70±0.56	83.10±0.21
c			90.91±0.48	89.08±0.45	86.79±0.41	84.20±0.32
d				91.43±0.50	----------	86.78±0.45
e					89.74±0.34	85.36±0.60
f						88.89±0.47
Over all accuracy **88.9±0.37**						

Table 1. Performance of each pair of environment coupled metrics. The symbols (a to f) represent different environment.

Method	Accuracy
SIFT LDML, funneled [7]	83.20±0.40
SIFT ITML-MultiShot, aligned[16]	83.97±0.70
SIFT PLDA, funneled[17]	86.20±1.20
Combined LDML-MKNN,funneled[7]	87.50±0.40
Combined multishot, aligned[16]	89.50±0.51
Combined PLDA, aligned & funneled[17]	90.07±0.51
Our method, funneled	**88.90±0.37**

Table 2. Result of our method and other state-of-art methods for LFW database under unrestricted setting.

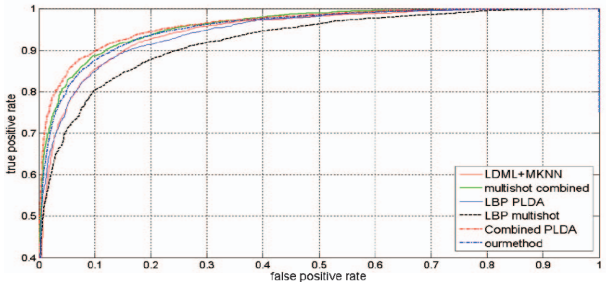

Figure 4. ROC curves averaged over 10 folds of View 2.