TRANSDUCTIVE VIS-NIR FACE MATCHING

Jun-Yong Zhu[1], *Wei-Shi Zheng*[2,*], *Jianhuang, Lai*[2]

1 School of Mathematics and Computational Science, Sun Yat-Sen University, China
2 School of Information Science and Technology, Sun Yat-Sen University, China
jonesjunyong@ieee.org, wszheng@ieee.org, stsljh@mail.sysu.edu.cn

ABSTRACT

The visual-near infrared (VIS-NIR) face matching, sharing the illumination-invariant property of NIR face image and remaining the use of existing VIS face images as enrollment, has been a popular issue in recent years. However, existing techniques assume that there are sufficient pairwise VIS and NIR images for each person during training, which is not realistic in VIS-NIR matching problem, as no NIR images are available for people who have already been registered in the existing face recognition system and only a handful of pairwise VIS and NIR face images captured from new people are available. To address this problem, we formulate the VIS–NIR matching as a transductive learning problem, which is a first attempt to our best knowledge. Moreover, we propose a transductive method named *Transductive Heterogeneous Face Matching (THFM)* by alleviating the domains difference and learning the discriminative model for target simultaneously, making it possible to take the query/probe NIR images into account in a transductive way. Experimental results validate the effectiveness of our approach on the heterogeneous face biometric database.

Index Terms— Heterogeneous face recognition, VIS-NIR face matching, Transductive learning

1. INTRODUCTION

The illumination problem [10], as a major challenge for face recognition, has largely restricted the use of traditional visual (VIS) image based face recognition in practical applications. Recently, a solution is given by using active near infrared (NIR) imaging which is proved to be invariant to visible light illumination changes [11].

However, many real world face recognition systems have already got lots of people enrolled using their VIS face images and it is hard to re-enroll these people using NIR images. Moreover, in many real-world applications, such as E-passport, machine readable traveling document (MRTD), ATM, etc, the ability of matching NIR probe images to the gallery VIS images is of significant importance since the query/probe face images are always acquired/captured in poor illumination conditions. Therefore, it is worth taking advantage of the NIR based approach and extending the VIS-VIS face recognition system to a heterogeneous form using NIR-VIS faces matching.

The difficulties of matching heterogeneous images are mainly due to the matching across image modalities. A few works have been proposed to handle this problem. Yi et al. introduced canonical correlation analysis (CCA) to learn the correlation between NIR and VIS faces from NIR-VIS face pairs and the learned correlation is used to evaluate similarity between an NIR face and a VIS face [2]. In order to tackle the inter-modality problem, Lin & Tang [3] considered the empirical discriminative power and the local smoothness of the feature transformation and proposed a common discriminant feature extraction (CDFE), in which both inter-modality discriminant information and intra-modality local smoothness are involved. Recently, Lei & Li [4] suggested solving this challenging problem in a more efficient manner. In their coupled spectral regression (CSR), a low dimensional representation for each face is first computed using discriminative graph embedding method and then two associated projections are learned respectively to project heterogeneous data into the discriminative common subspace for final classification.

Most existing methods assume that the people in testing stage are included in the training set, which is an inductive learning procedure. However, as mentioned above, a large amount of people have only registered their VIS images in the existing face systems (i.e. the corresponding NIR images were not registered). While we only collect a handful of available new people samples having pairwise VIS and NIR face images. Hence, all learning methods aforementioned are focusing on training set only and not directly designed to handle this realistic problem.

In this paper, we address the above problem by introducing a transductive VIS-NIR matching approach. As far as we know, it is the first time to formulate the VIS-NIR matching problem in a transductive framework. Our proposed THFM, different from previous methods, seeks to find out a feature space that alleviates the heterogeneous difference and meanwhile preserves discriminative information of testing target people by using both training and gallery sets. As a result, the heterogeneous face matching problem can be seem as a homogeneous face recognition problem in such feature space, which makes it possible to utilize newly captured unlabelled probe NIR images in a transductive form. Following this intuition, we impose a

[*] Corresponding author

penalty term on the difference between probe NIR images and gallery VIS images to constrain the bias caused by heterogeneous difference. The proposed THFM is proved effective and performs much better than existing methods in the heterogeneous face database.

Thought there is existing work on transductive face recognition [13, 15], the motivation is different and more important they cannot address the heterogeneous face matching problem.

2. TRANSDUCTIVE FORMULATION

Most existing systems only register the VIS face images for existing people (denoted as set A), while we only have a few newly registered people (denoted as set B) having both VIS and NIR images. The main objective of developing a matching technique between VIS and NIR face images is to match a probe NIR face image of any person in Set A to its gallery (registered) VIS image using information of both set A and set B (as shown in Fig.1). In this section, we try to formulate the VIS-NIR face matching problem using transduction.

Assume that we have a set of gallery VIS images denoted by $\{x_{p,i}^{Gallery} \mid p \in C_{Test}\}$ from subjects enrolled in the existing face recognition system, where $x_{p,i}^{Gallery}$ is the *i*th sample of class *p* in the gallery set. Also, in order to extend the face recognition system to the form of VIS-NIR matching, we collect a small set of pairwise VIS-NIR face images belonging to those who are used for learning the relationship between VIS and NIR domains, denoted as $\{x_{q,i}^{Tr\_VIS} \mid q \in C_{Train}\}$ and $\{x_{q,j}^{Tr\_NIR} \mid q \in C_{Train}\}$ respectively. The task is to match any NIR image $x_{p,j}^{probe}$ of the probe set $\{x_{p,i}^{Probe} \mid p \in C_{Test}\}$ to its corresponding gallery image $x_{p,i}^{Gallery}$ in VIS domain, where $x_{p,j}^{Probe}$ and $x_{p,i}^{Gallery}$ are sharing the same labels. It should note that, the labels of gallery samples are available but those of probe set are unknown.

In contrast to inductive inference approximating a functional dependency firstly and then using it to evaluate the values of a function at the points of interest, transductive inference estimates the values of a function at the points of interest in one step [12]. Two assumptions must be satisfied in traditional transductive methods: *a) the labels of probe samples should be included in the gallery set; b) samples in probe and gallery sets are independent and identically distributed (i.i.d.).*

In the context of VIS-NIR matching, the testing set (set A) contains two parts: gallery set containing VIS images used for registration and probe NIR images (captured by a series of cameras). We assume that, for a given probe image $x_{c_1+q,j}^{NIR}$, its corresponding gallery image $x_{c_1+q,i}^{VIS}$ has already existed though the label is assumed unknown. Thus, the only thing we concerned is the latter assumption. However, we find that distributions of gallery and probe sets are probably not consistent in the pixel space since images captured in VIS and NIR lighting condition differ from each other. That is, traditional transductive methods like KNN and TSVM could not be directly applied here. Nevertheless, given a training set of pairwise VIS and NIR images (set B), it is possible for us to alleviate the domain distinction and formulate the procedure of matching the probe images to the VIS images in the gallery set in a transductive framework.

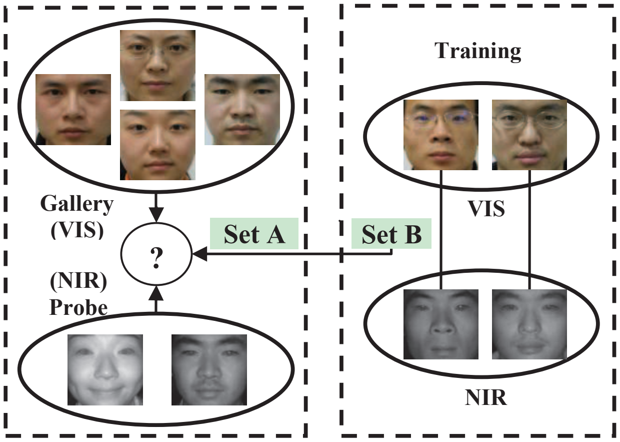

Fig. 1. Illustration of VIS-NIR face matching

Our main idea is to develop a matching technique from training data and adapt it to the matching problem between gallery and probe sets. In particular, we would like to search for a feature space in which the difference caused by domain distinction can be alleviated and meanwhile discriminant information for target classifying is persevered. Though a two step approach could be adopted here by first searching a domain-invariant feature space and then learning a discrimimant model by the projected gallery samples, discriminant information would be likely lost in the first step when the amount of training samples is limited. Hence, in order to avoid this drawback, we formulate the VIS-NIR face matching problem in a unified framework using transduction.

3. PROPOSED ALGORITHM

In this section, we elaborate our proposed THFM in three folds: domain invariant feature extraction, target related discriminant model learning and cross domain penalization.

3.1. Domain Invariant Feature Extraction

Searching for the domain invariant feature or common features in VIS and NIR domains is a key ingredient for the VIS-NIR face matching problem [2-4]. Naturally, since there are a small amount of available training samples (set B) for us to investigate the relationship between two domains, we hope that the expected feature mapping $f: X \rightarrow Z$ could draw points of the same class together despite which domain them belong. We attempt to realize it by minimizing the intra-class variation of training samples in feature space.

Suppose that we are given the training VIS-NIR images $\{x_{q,k}^{Tr} \mid q \in C_{Train}\} = \{x_{q,i}^{Tr\_VIS} \mid q \in C_{Train}\} \cup \{x_{q,j}^{Tr\_NIR} \mid q \in C_{Train}\}$, where $x_{q,k}^{Tr}$ is the *k*th sample of class *q*. The intra-class variation can be quantified by the trace of average intra-class scatter matrix

$$S_{intra} = \frac{1}{N_{Tr}} \sum_{q \in C_{Train}} \sum_{k} (x_{q,k}^{Tr} - \bar{m}_q^{Tr})(x_{q,k}^{Tr} - \bar{m}_q^{Tr})^T \qquad \mathbf{(1)}$$

where $\bar{m}_q^{Tr}$ represent the mean of class q in training set, N_{Tr} denotes the number of classes in training set and N_q is the total amount of samples in class q.

3.2. Target Related Discriminant Model Learning

Matching probe images to the gallery images can be seen as applying a classifier trained on the gallery set to those probe images. However, it works only when the samples in the gallery and probe set are coming from the same domain. Fortunately, as we consider domain invariant feature extraction and target classification simultaneously, we are able to cope with the heterogeneous matching problem in a homogeneous way. That is, the classifier trained by gallery set would be adaptively fit for the probe images in the feature space.

Thus, we hope that the discriminant information of gallery set could be preserved in the desiring feature space. Lots of work has been done on the discriminant analysis, such as FDA [13], MMC [15], SVM [14] etc. Without loss of generalization, we adopt the between-class variation to measure the separability of subjects in gallery set, which is proved effective in FDA described by the inter-class scatter matrix.

Using denotation in section 2, the average between-class variation in gallery set is calculated by the trace of

$$S_{inter} = \frac{1}{N_G} \sum_{i \in C_{Test}} N_i (\bar{m}_i^G - \bar{m}^G)(\bar{m}_i^G - \bar{m}^G)^T \qquad \mathbf{(2)}$$

where N_i represents the amount of samples in class i, N_G is the sum of N_i, m_i^G denotes the mean of class i and m^G is the mean of all gallery samples.

3.3. Cross Domain Penalization

In section 3.1, we have considered the relation between VIS and NIR domains using training set. However, since our goal is to perform matching on the testing data (set A), we have to guarantee that the VIS and NIR images of the same person in testing set have the same or similar representation in the feature space. It should note that, label information of probe images is missing, which leads to the failure of aligning VIS and NIR face images according to their labels.

In this part, we attempt to incorporate the Maximum Mean Discrepancy (MMD) [8] to model and penalize the cross domain difference in testing set for its briefness and effectiveness. Other works can be found for modeling distribution difference, including kernel mean matching (KMM) [7], Bregman Divergence based regularization [9]. However, these methods assume that all samples are come from the same domain, which is different from that in heterogeneous face matching.

Consequently, given samples $X^G = \{x_{p,i}^{Gallery} \mid p \in C_{Test}\}$ and $X^P = \{x_{p,i}^{Probe} \mid p \in C_{Test}\}$ drawn from two different domains and the feature map f, the empirical estimate of MMD between the two distributions in the feature space is defined as

$$MMD(X^G, X^P) = \left\| \frac{1}{N_G} \sum_{p,i} f(x_{p,i}^{Gallery}) - \frac{1}{N_P} \sum_{p,j} f(x_{p,j}^{Probe}) \right\|^2 \qquad \mathbf{(3)}$$

It can be verified that $MMD(X^G, X^P)$ can be written as:

$$\Omega(f) = MMD(X^G, X^P) = tr(ZLZ^T) \qquad \mathbf{(4)}$$

where $Z = \{f(x_k) \mid x_k \in X^G \cup X^P\}$, $N_G = |X^G|$, $N_P = |X^P|$, $L = [L_{ij}]$ with $L_{ij} = 1/N_G^2$ if x_i, x_j are gallery samples, $L_{ij} = 1/N_P^2$ if x_i, x_j belong to the probe set, otherwise $L_{ij} = -1/N_G N_P$.

3.4. The Proposed Criterion

Based on the analysis above, we attempt to search a feature space in which S_{intra} and $\Omega(f)$ are minimized and meanwhile S_{inter} is maximized. If we consider the feature mapping f as the linear function, then $f(x) = W^T x$, we form our objective function as follow:

$$\max_{W} \frac{tr(W^T S_{inter} W)}{tr(W^T (S_{intra} + M + \eta I) W)} \qquad \mathbf{(5)}$$

where $M = XLX^T, X = [X^G, X^P]$, η is a constant for Tikhonov regularization which is used to avoid degeneration in the generalized Eigen-decomposition problem. Note that, nonlinear situation could be extended straightforwardly.

Finally, the solution to (5) is equal to the leading eigenvectors computed by the following general eigenvalue problem

$$S_{inter} w = \lambda (S_{intra} + M + \eta I) w \qquad \mathbf{(6)}$$

Note that, only eigenvectors with non-zero eigenvalue are kept in our final result.

4. EXPERIMENTS

In this section, we apply the proposed algorithm to the VIS-NIR face images matching. We used the heterogeneous face biometric (HFB) dataset published in the 2009 IEEE CVPR [1], which contains 100 persons, each with 4 VIS face images and 4 NIR face images. All faces were manually aligned according to the eye coordinates and cropped in 128x128 pixels. In order to reduce the computational cost, we simply resized the face images into 32x32 and transformed each image to form a 1024 dimension input feature. Cosine distance was used to measure the similarity between samples in feature space and nearest neighbour (NN) classifier was adopted for classification.

We evaluate the proposed method by comparing several related VIS-NIR methods, including FDA[13], LSCR[4], CDFE[3], PCA+CCA[2], LDA+CCA[2]. Note that all these methods are not transductive. In our method, we simply set the parameter μ *as* a small value, e.g. 0.1 (μ=0.1 here), since it is only applied to avoid degeneration. Besides, we discarded those eigenvectors with small eigenvalue (say 10e-3). For the PCA related approach, the radio of preserving principal components is set to be 99%; In CDFE, α, β and k are set to 1, 0.5 and 2 respectively; In addition,

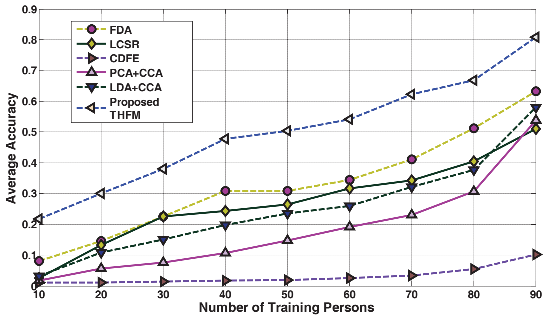

Fig. 2. Average recognition accuracy with different training set

the regularized coefficients $\{\lambda, \eta\}$ we utilized in LSCR is $\{0.001, 0.01\}$, which is suggested in their paper.

In the following experiment, we randomly selected K persons as the training set, and the rest persons for testing, where K is varying from 10 to 90 with step 10. Note that, there is no overlap in the training set and the testing set. For each K, we repeated experiments with different training sets 10 times and finally obtained the average recognition rate.

As illustrated in Fig. 2, the CCA based methods did not perform well except that there were sufficient training data. The result of LCSR was a little worse than that of FDA, the reason might be that the two steps approach may suffer from overfitting for its variances are more than those in FDA . For our proposed method, by considering domain invariant feature extraction and target-related discriminant model learning in a transductive framework, it outperforms other algorithms although they were reported to perform well when subjects of testing are included in the training set [4]. That is to say, the results verify our analysis in the context of transductive VIS-NIR face matching.

For complement, we compared the performance of each part in section 3 (denoted as S_intra, S_inter, MMD respectively) to our whole algorithm. In particular, the number of leading/last eigenvectors used in each part was set to the same as that in THFM. Mean accuracy and standard variation on 20 randomly split set (K=50) were reported (Table 1). We could find that unsatisfying results were gained when considering them individually. However, our proposed THFM, taking them all into account using transduction, had achieved great improvement.

5. CONCLUSION

This work has formulated the VIS-NIR face matching as a transductive learning problem. To our best knowledge, it is the first attempt in this field. In particular, we have developed a novel transductive subspace learning method for heterogeneous face matching by considering the domain invariant feature extraction, target related discriminant model learning and cross domain difference on testing set at the same time. Experiment results show that our proposed method has outperformed the related subspace learning based VIS-NIR matching approaches. In future, we will go on investigating a more suitable face representation for cross domain matching, such as some image descriptors in [5, 6].

Table **1.** Comparison among parts of THFM

	S_intra	S_inter	MMD	THFM
Mean	0.1820	0.0918	0.0383	0.4913
Std	0.0305	0.0313	0.0205	0.0522

6. ACKNOWLEDGEMENTS

This research was supported by the National Natural Science of Foundation of China (Nos.61102111, 61173084), Specialized Research Fund for the Doctoral Program of Higher Education (No.20110171120051), and the 985 Project at Sun Yat-sen University under Grant No.35000-3181305.

7. REFERENCES

[1] S.Z. Li, Z. Lei, M. Ao, “The HFB Face Database for Heterogeneous Face Biometrics Research,” IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2009.

[2] D. Yi, R. Liu, R. Chu, Z. Lei, S. Z. Li, “Face Matching between Near Infrared and Visible Light Images,” IAPR/IEEE International Conference on Biometrics , 2007.

[3] D. Lin, X. Tang, “Inter-modality Face Recognition,” European Conference on Computer Vision, pp. 13–26 , 2006.

[4] Z. Lei, S. Liao, S.Z. Li, “Coupled spectral regression for matching heterogeneous faces,” IEEE Conference on Computer Vision and Pattern Recognition ,2009.

[5] S. Liao, D. Yi, Z. Lei, R. Qin, S. Z. Li, “Heterogeneous Face Recognition from Local Structures of Normalized Appearance,” IAPR/IEEE International Conference on Biometrics , 2009.

[6] B. Klare, A. K. Jain, “Heterogeneous Face Recognition: Matching NIR to Visible Light Images,” International Conference on Pattern Recognition, pp.1513-1516, 2010.

[7] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, B. Scholkopf, “Covariate Shift By Kernel Mean Matching,” J. Dataset Shift in Machine Learning, Citeseer, pp. 131-160, 2009.

[8] K.M. Borgwardt, A. Gretton, M.J. Rasch, H.P. Kriegel, B. Sch{\\"o}lkopf, A.J. Smola, “Integrating Structured Biological Data by Kernel Maximum Mean Discrepancy,” J. Bioinformatics, Oxford Univ Press, 22(14):49–57, 2006.

[9] S. Si, and D. Tao, B. Geng, “Bregman Divergence-based Regularization for Transfer Subspace Learning,” IEEE Trans on KDE, pp. 929-942, 2009.

[10] Y. Adini, Y. Moses and S. Ullman. “Face Recognition: The Problem of Compensating for Changes in Illumination Direction”. IEEE Trans on PAMI, 19(7):721-732, 1997.

[11] S.Z. Li, R. Chu, S. Liao, L. Zhang, “Illumination Invariant Face Recognition using Near-Infrared Images”. IEEE Trans on PAMI, 29(4):627-639, 2007.

[12] F. Li and H. Wechsler, “Open Set Face Recognition using Transduction” IEEE Transa on PAMI, 27(11):1686-1697, 2005.

[13] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, “Eigenfaces versus Fisherfaces: Recognition Using Class Specific Linear Projection,”, IEEE Transa on PAMI, 19(7): 711-720, July, 1997.

[14] C.Cortes and V. Vapnik, “Support-vector networks”, Machine Learning, 20(3): 273-297, November 1995.

[15] H. Li, T. Jiang, and K. Zhang, “Efficient and robust feature extraction by maximum margin criterion”, IEEE Trans. on Neural Networks 17, 157-165 (2006).