HETEROGENEOUS FACE RECOGNITION VIA GRASSMANNIAN BASED NEAREST SUBSPACE SEARCH

Yuan Tian[1], *Cheng Yan*[1], *Xiao Bai*[1], *Jun Zhou*[2]

[1]School of Computer Science and Engineering, Beihang University, Beijing, China
[2]School of Information and Communication Technology, Griffith University, Nathan, Australia

ABSTRACT

Heterogeneous face recognition involves matching faces in different image modalities, such as near infrared images to visible images or sketch images to photos. This challenging task has attracted increasing attention in recent years. This paper presents, for the first time, a subspace based method to tackle the problem of face recognition between visible images (VIS) and near infrared (NIR) images. Subspace is used to extract essential attributes from VIS and NIR images. We adopt Grassmannian radial basis function (RBF) kernel to keep the relationship between subspaces, and use kernel canonical correlation analysis (KCCA) to handle correlation mapping between VIS and NIR domains. After mapping both VIS and NIR images to the common space, the heterogeneous face recognition problem can be easily completed by the nearest search. We evaluate the proposed method on the CASIA NIR-VIS 2.0 dataset. The experimental results demonstrate that our method is very effective for NIR-VIS face recognition.

Index Terms— Subspace, Heterogeneous, Grassmannian, Face Recognition, Common Space

1. INTRODUCTION

Heterogeneous face recognition has attracted increasing attention in the past several years [1, 2]. Its purpose is to recognize face images obtained from different modalities, such as those captured in the visible (VIS) light spectrum and near infrared (NIR) spectrum. This setting is useful in many real world applications, for example surveillance, which have to handle NIR images but most accessible datasets only contain VIS images. This is a more challenging task compared with face recognition in the same data modality.

Many methods have been proposed for the heterogeneous face recognition task [3, 4, 5, 6]. However, most of these methods directly use single image of a face as a data point for analysis, which may lose the global or structure information of a face. Linear or affine subspace representation [7, 8, 9] can capture the structure and the global feature of a face from several images of the face. Since the face images in different modalities are very different, using linear or affine subspaces will better represent the structural information of face than using only one image. This forms the motivation that we use subspace to extract the basic attributes of a face.

For heterogeneous face recognition, mapping two different modalities to a common space is a major solution for the recognition problem. To minimize the intra-class difference of two modals, canonical correlation analysis (CCA) [10, 11] is a classic and widely used solution. It aims at maximizing the correlation of the projection of two modalities and has shown effectiveness in heterogeneous face recognition [3, 12]. Moreover, to ensure high accuracy of face recognition and to consider that the distribution of subspaces conforms to manifold, kernel CCA is often adopted with nonlinear kernels [13]. For the subspaces, the most important steps are mapping subspaces to a common space and then calculating the distances between subspaces in the common space. An effective solution for these steps would allow more accurate face recognition.

In this paper, we propose a new approach for heterogeneous face recognition. We use subspace to capture the essential information of each face. Then we apply kernel CCA as the correlation mapping method to learn the subspace correlation between NIR and VIS images, in which the Grassmannian [14] radial basis function has been used for nonlinear modelling. We define the Grassmannian distance as the distance metric between subspaces in order to effectively find the nearest subspace of query. Our method shows better performance in the experiments than several alternative methods. The contribution of this paper are three fold. First, according to our knowledge, it is the first time that a subspace based solution is proposed to handle the heterogeneous face recognition problem. Second, we take kernel into consideration and combine kernel canonical correlation analysis with subspace to minimize the intra-class difference between NIR and VIS faces for the mapping problem. Third, we define a distance metric to measure the distance between subspaces for the nearest subspace search.

The rest of the paper is organized as follows. The proposed method is described in Section 2. The details of datasets and experimental results are presented in Section 3. The conclusions are drawn in Section 4.

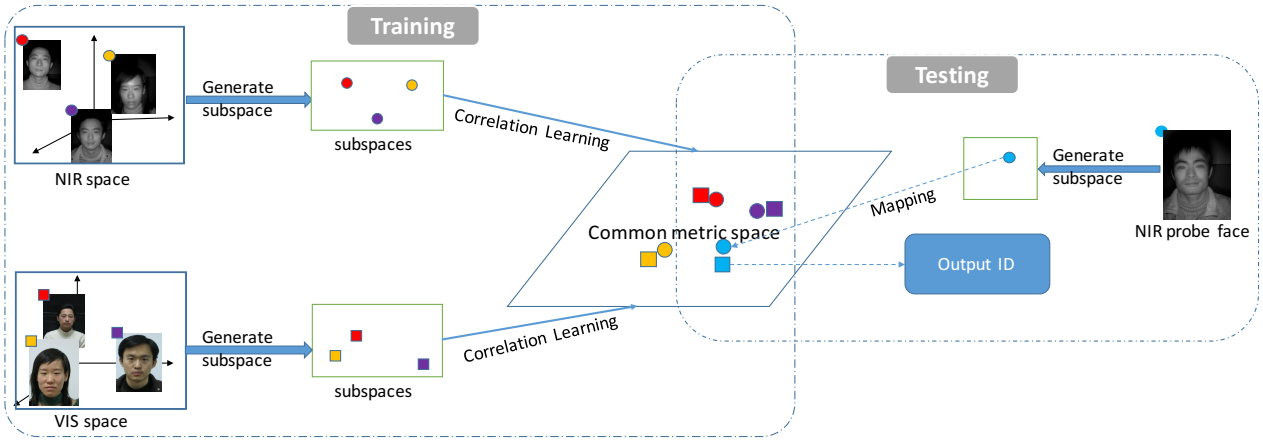


Fig. 1. This figure shows the basic framework of our method. During the training phase, we extract subspaces of faces in two modalities, and then learn the correlation between NIR and VIS spaces by mapping two different modalities into a common space. For testing, we utilize the correlation matrix to map the probe subspace into the common space and find the nearest gallery subspace to get the corresponding face ID.

2. PROPOSED METHOD

As is shown in Fig. 1, we first extract the subspace information of a face from its images in the same modality (in Section 2.1). Then two projection matrices, which map the subspaces of NIR and VIS images to a common space respectively, are learned by using kernel canonical correlation analysis (in Section 2.2). Finally, the distance between subspaces is defined, and a subspace searching strategy is employed for face recognition (in Section 2.3).

2.1. Subspace Generation

As illustrated in [15, 16], subspace representation has many advantages in capturing the global information of a face dataset. Several vectors of the same face in one modal constitute a subspace. We utilize principal component analysis (PCA) [17] to generate the subspace from face images.

Each face I consists of several images of the same person. $I = (\xi_1, \xi_2, \cdots, \xi_{n_i})$ is an $m \times n_i$ matrix, where m is the dimensionality of each vectorized face image, n_i is the number of images, and $\xi_1, \xi_2, \cdots, \xi_{n_i}$ are the vectorized images of face i. We first subtract the mean face by

$$\begin{aligned} \hat{I} &= (\xi_1 - \overline{\xi}, \xi_2 - \overline{\xi}, \cdots, \xi_{n_i} - \overline{\xi}), \\ \overline{\xi} &= \frac{1}{n_i} \sum_{k=1}^{n_i} \xi_k \end{aligned} \tag{1}$$

where $\overline{\xi}$ is the mean value of the vectorized face images. Then we generate the subspace x_i of each face based on the eigenvalue decomposition of $\hat{I}\hat{I}^T$. We use x_i to present the samples in the subspace of face i. x_i is an $m \times d$ matrix, which is composed of the top d principal components of the eigenvectors of covariance matrix $\hat{I}\hat{I}^T$.

2.2. Subspace Mapping

When we have generated the subspaces of each face in both NIR and VIS modalities, we map these subspaces into a common space. To deal with the nonlinear problem caused by the distribution of subspaces of face, we use Grassmannian radial basis function (RBF) kernel [18] for subspace embedding, which is defined as

$$\kappa(x_i, x_j) = \varphi(x_i)\varphi(x_j)^T = \exp(\|x_i^T x_j\|_F^2) \tag{2}$$

where x_i, x_j are the generated subspaces, $\|\cdot\|_F^2$ is the Frobenius norm.

Since we take kernel into consideration, we use kernel canonical correlation analysis (KCCA) [13] for correlation learning. The subspaces can be used to maximize the correlation between two modalities $X = (x_1, x_2, \cdots, x_n)$ and $Y = (y_1, y_2, \cdots, y_n)$ in KCCA, we want to find matrices W_x and W_y that project the embedded item $\varphi(\cdot)$ from each modal into a low dimensional common space such that the distance in the resulting space between each pair of modals for the same face is minimized. The similarity between the items in the same modal is defined by a kernel function $\kappa(x_i, x_j)$.

The objective function for this optimization problem is given by

$$\max_{W_x, W_y} \frac{W_x^T K_X K_Y W_y}{\sqrt{W_x^T K_X K_X W_x}\sqrt{W_y^T K_Y K_Y W_y}} \tag{3}$$

where the K_X and K_Y are kernel matrices of modal X and Y respectively, which can be calculated by Equation (2). $K_X K_X$ and $K_Y K_Y$ represent the empirical covariance matrices for the two modalities of data respectively, while $K_X K_Y$ represents the cross-covariance matrix between them.

KCCA maximizes the correlation between the projected matrices $W_x^T\varphi(X)$ and $W_y^T\varphi(Y)$ to keep the similarity of intra-class in the embedded common space. The standard minimization process for KCCA can be found in [10]. Finally, we can get the mapping matrices W_x and W_y for these two modalities.

2.3. Nearest Subspace Search

After mapping, the problem of face recognition is to find the nearest subspace P_i in one modal of the given probe subspace Q in another modal. We can formulate this problem as

$$\widehat{i} = \arg\min_i \ d_G(P_i, Q) \tag{4}$$

where $\widehat{i}$ is the ID of the nearest gallery subspace to the probe subspace, d_G is the geodesic distance.

Geodesic distance is a formal measure on the distance between two subspaces. It is the length of the shortest path connecting two points on the Grassmannian manifold [19]. Let $G(D,d)$ denote the Grassmannian manifold which is a set of d-dimensional linear subspaces of the $\mathbb{R}^D$. Let $0 < d_1 \leq d_2 \leq D$, $x_1 \in G(D,d_1)$ and $x_2 \in G(D,d_2)$, the principal angles $\theta_1 \geq \cdots \geq \theta_{d_1}$ are defined as follows: for $i = 1, \cdots, d_1$ let $\delta(x_1^T x_2)$ denote the i-th largest singular value of the matrix $x_1^T x_2$. The principal angles $0 \leq \theta_d \leq \cdots \leq \theta_2 \leq \theta_1 \leq \frac{\pi}{2}$, are

$$\theta_i = \arccos(\delta_{d-i}(x_1^T x_2)), i = 1, \cdots, d_1 \tag{5}$$

and the geodesic distance between x_1 and x_2 is

$$d_G(x_1, x_2) = (\sum_{i=1}^{d_1} \theta_i^2)^{\frac{1}{2}} \tag{6}$$

If $d_1 = d_2 = d$, it is a metric. If $d_1 \neq d_2$, it is still a good method to measure the distance between subspaces when the dimensions of subspaces are different. If $d_1 = 1$, $d_G(x_1, x_2)$ is the elevation angle between the line (single image) and the subspace [14]. For all of these searches, the query time is $O(Ddn^\rho)$, where d is the largest dimension of subspaces among both query elements and the database elements, D is the ambient dimension and $\rho < 1$.

With respect to our framework, subspaces from different modals are mapped into a common space, so the distance between the probe subspace and the gallery dataset turns into the distance between embedded subspaces. Let X represent the VIS gallery subspaces and Y represent the NIR probe subspaces, as defined previously, we can obtain the mapping matrices W_x and W_y using Equation (3) to embed X and Y into the learnt common space. The mapped subspace of VIS and NIR space is

$$\begin{aligned} u_j &= \sum_i W_{xi} K(x_i, x_j) \\ v_j &= \sum_i W_{y_i} K(y_i, y_j) \end{aligned} \tag{7}$$

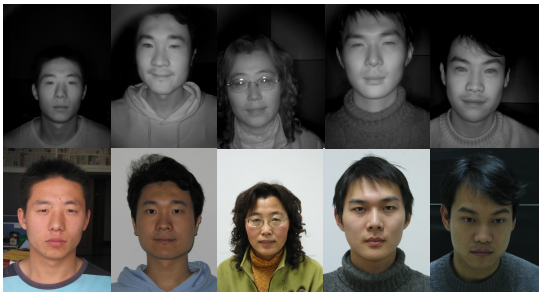


Fig. 2. Face image samples in the CASIA NIR-VIS 2.0 dataset. Each column represents one face, the top row contains near-infrared images and the bottom row are the visible images.

where x_i, x_j are the subspaces of the i-th face and the j-th face respectively from the VIS dataset, y_i, y_j are the subspaces of the i-th face and the j-th face in the NIR dataset, u_j is the embedded subspace of the j-th face in VIS dataset, and v_j is the embedded subspace of the j-th face in the NIR dataset.

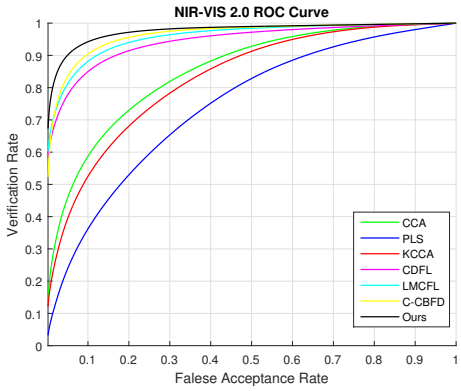
Combined with Equation (6), the distance between the subspaces in the VIS modality and the NIR modality is $d_G(u, v)$. Given a query image, the identification of the most similar subspace is the heterogeneous face recognition result.

3. EXPERIMENTS AND RESULTS

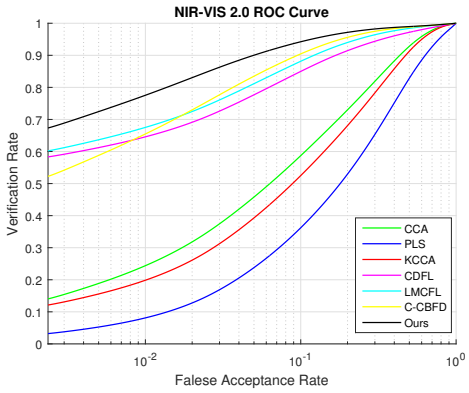
3.1. Dataset and Protocol

In this work, we used the CASIA NIR-VIS 2.0 dataset [20] for experimental validation. This is the largest publicly available heterogeneous face recognition dataset across the NIR and the VIS spectrum. It contains 17580 images of 725 faces. The dataset provides two views: View 1 is meant for algorithm development and parameters can be tuned on this subset, View 2 is to be used for performance evaluation which is divided into 10 folds. For each fold, there are 357 faces for training (learning mapping matrix in our experiment) and 358 faces for testing. Performance evaluation is obtained from the average performance of 10 folds in View 2. Figure 2 shows some sample images in the CASIA NIR-VIS 2.0 dataset. The images in the dataset have been aligned and cropped into 128x128 from their original face images. Following the protocol of many methods [3, 21], we downsampled these well-aligned images by restricting the set of pixels and then used 32×32 resized images for our experiment.

For each face in different modals, we vectorized each 32×32 training image to a 1024 dimensional vector and created 5-dimensional subspaces (one for each face) in each modality as mentioned in Section 2.1. When a face has only single image, the problem is to find its closest face (subspace) in the other modal, for which the solution is given in Section 2.3.



(a)



(b)

Fig. 3. ROC Curves on the CASIA NIR-VIS 2.0 dataset. (a) shows the comparison with other algorithms. (b) shows the comparison in semi-log scale to emphasize the performance difference at very low FAR.

3.2. Results

We have compared our method with several state-of-art methods, including CCA [10], KCCA [10], Partial Least Squares (PLS) [22], Coupled Discriminative Feature Learning (CDFL) [6], Large Margin Coupled Feature Learning (LMCFL) [5], NIR-VIS Heterogeneous face recognition [3] and Couple Compact Binary Face Descriptor (C-CBFD) [21]. Among these, as mentioned in their methods, CCA, PLS and KCCA based methods directly vectorize the 32×32 face images to a 1024 dimensional vector, and adopt a Gaussian kernel [18] for KCCA. The parameters of the compared methods are tuned on View 1 of the dataset.

The experimental results are shown in Table 1, which summarizes the rank-1 identification rates and their standard deviations crossing all ten folds, and the verification rates at 0.1% false accept rate (FAR). It is obvious that the proposed method is better than the other methods being compared with. An important reason is that the subspace representation is very suitable for face recognition. The experimental results

NIR-VIS 2.0	Rank 1	Std. Dev.	FAR=.001
CCA [10]	28.5	3.4	10.8
PLS [22]	17.7	1.9	2.3
KCCA [10]	30.3	1.0	9.4
CDFL [6]	71.5	1.4	55.1
LMCFL [5]	75.7	2.5	55.9
NIR-VIS Rec.[3]	78.5	1.7	85.8
C-CBFD [21]	81.8	2.3	47.3
Ours	82.6	2.0	67.2

Table 1. Experimental results on View 2 of the CASIA NIR-VIS 2.0 Face Dataset.

prove that the Grassmannian distance is a good metric to measure the similarity between subspaces. The results of our method are better than NIR-VIS Heterogeneous face recognition [3] and C-CBFD [21] which use normalized cosine distance for measuring the distance between face samples. Furthermore, compared with the results of KCCA [10], our method has better performance. It proves that Grassmannian RBF kernel makes correlation analysis more suitable for nonlinear problem, especially for heterogeneous face recognition.

We show the ROC curve on View 2 of the CASIA NIR-VIS 2.0 dataset in Figure 3. In order to emphasize the performance at very low FAR, the ROC curves are shown in the semi-log scale. It shows that at very low FAR, the performance of the proposed method is still very promising.

4. CONCLUSIONS

We have introduced a new approach for NIR-VIS face recognition. It is the first time subspace is used to handle the heterogeneous face recognition problem. We extract subspace information of a face and measure the distance between subspaces on the Grassmannian manifold. The mapping matrix is learned by maximizing the intra-class similarity of subspaces from NIR and VIS face modalities. Then heterogeneous face recognition task can be implemented by the nearest neighbor search. The experiments show that the proposed method has achieved higher accuracy than several alternatives.

5. ACKNOWLEDGEMENT

This work was supported by NSFC project No.61370123, BNSF project No.4162037 and support funding from State Key Lab. of Software Development Environment.

6. REFERENCES

[1] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D

+ 2D face recognition," *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[2] B. Klare and A. K. Jain, "Heterogeneous face recognition: Matching NIR to visible light images," in *International Conference on Pattern Recognition*, 2010, pp. 1513–1516.

[3] F. Juefei-Xu, D. K. Pal, and M. Savvides, "NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *Computer Vision and Pattern Recognition Workshops*. IEEE, 2015, pp. 141–150.

[4] C. Reale, N. M. Nasrabadi, H. Kwon, and R. Chellappa, "Seeing the forest from the trees: A holistic approach to near-infrared heterogeneous face recognition," in *Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2016, pp. 320–328.

[5] Y. Jin, J. Lu, and Q. Ruan, "Large margin coupled feature learning for cross-modal face recognition," in *International Conference on Biometrics*, 2015, pp. 286–292.

[6] Y. Jin, J. Lu, and Q. Ruan, "Coupled discriminative feature learning for heterogeneous face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 3, pp. 640–652, 2015.

[7] R. Basri, T. Hassner, and L. Zelnik-Manor, "Approximate nearest subspace search with applications to pattern recognition," in *Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[8] R. Basri, T. Hassner, and L. Zelnik-Manor, "A general framework for approximate nearest subspace search," in *International Conference on Computer Vision Workshops*. IEEE, 2009, pp. 109–116.

[9] R. Basri, T. Hassner, and L. Zelnik-Manor, "Approximate nearest subspace search," *IEEE Transactions on Software Engineering*, vol. 33, no. 2, pp. 266–278, 2011.

[10] D.R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639, 2004.

[11] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *International Conference on Multimedea*, 2010, pp. 251–260.

[12] W. Yang, D. Yi, Z. Lei, and J. Sang, "2D-3D face matching using CCA," in *International Conference on Automatic Face and Gesture Recognition*. IEEE, 2008, pp. 1–6.

[13] S. Akaho, "A kernel method for canonical correlation analysis," *In Proceedings of the International Meeting of the Psychometric Society (IMPS2001*, vol. 40, no. 2, pp. 263–269, 2006.

[14] X. Wang, S. Atev, J. Wright, and G. Lerman, "Fast subspace search via Grassmannian based hashing," in *International Conference on Computer Vision*. IEEE, 2013, pp. 2776–2783.

[15] L. Wang, X. Wang, and J. Feng, "Subspace distance analysis with application to adaptive Bayesian algorithm for face recognition," *Pattern Recognition*, vol. 39, no. 3, pp. 456–464, 2006.

[16] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, "Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 2705–2712.

[17] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[18] M.T. Harandi, M. Salzmann, S. Jayasumana, R. Hartley, and H. Li, "Expanding the family of Grassmannian kernels: an embedding perspective," *Computer Science*, vol. 8695, pp. 408–423, 2014.

[19] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, "Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2273–2286, 2011.

[20] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Computer Vision and Pattern Recognition Workshops*. IEEE, 2013, pp. 348–353.

[21] J. Lu, V. E. Liong, X. Zhou, and J. Zhou, "Learning compact binary face descriptor for face recognition.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 10, pp. 2041, 2015.

[22] A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 593–600.