Heterogeneous Face Recognition by Margin-Based Cross-Modality Metric Learning

DOI:
10.1109/TCYB.2017.2715660

Document Version
Accepted author manuscript

Heterogeneous Face Recognition by Margin-Based Cross-Modality Metric Learning

Jing Huo, Yang Gao, *Member, IEEE,* Yinghuan Shi, Wanqi Yang, and Hujun Yin, *Senior Member, IEEE*

***Abstract*—Heterogeneous face recognition deals with matching face images from different modalities or sources. The main challenge lies in cross-modal differences and variations and the goal is to make cross-modality separation among subjects. A margin-based cross-modality metric learning (MCM^{2}L) method is proposed to address the problem. A cross-modality metric is defined in a common subspace where samples of two different modalities are mapped and measured. The objective is to learn such metrics that satisfy the following two constraints. The first minimizes pairwise, intrapersonal cross-modality distances. The second forces a margin between subject specific intrapersonal and interpersonal cross-modality distances. This is achieved by defining a hinge loss on triplet-based distance constraints for efficient optimization. It allows the proposed method to focus more on optimizing distances of those subjects whose intrapersonal and interpersonal distances are hard to separate. The proposed method is further extended to a kernelized MCM^{2}L (KMCM^{2}L). Both methods have been evaluated on an ID card face dataset and two other cross-modality benchmark datasets. Various feature extraction methods have also been incorporated in the study, including recent deep learned features. In extensive experiments and comparisons with the state-of-the-art methods, the MCM^{2}L and KMCM^{2}L methods achieved marked improvements in most cases.**

***Index Terms*—Face recognition, large margin classifier, metric learning, multimodality learning.**

I. Introduction

FACE recognition under uncontrolled scenarios is challenging [1], [2]. It is also the case for heterogeneous face recognition, which deals with matching face images of different modalities or views. Previous applications include matching sketches drawn by an artist against photograph [3], [4] and matching near infrared (NIR) images to visual (VIS) images [5], [6]. There has also been an increasing need to verify low resolution ID card photograph (scanned or stored images) against images captured by high resolution cameras [7].

Due to large appearance variations of face images across different modalities, extracted features of different modalities usually lie in two separated spaces. In such case, the Euclidean distance and Mahalanobis-based distance metrics are highly influenced by modality differences, making distances of intrapersonal cross-modality pairs and interpersonal cross-modality pairs inseparable. In this paper, a cross-modality metric learning method is proposed. The goal is to learn a suitable and efficient metric function that is able to remove modality differences so that intrapersonal and interpersonal distances are separated. The problem is further cast into a framework similar to support vector machines to maximize margins between two kinds of distances. Two sets of distance constraints are adopted, pairwise intrapersonal cross-modality distance constraints and triplet-based cross-modality distance constraints. The first is to minimize intrapersonal cross-modality distances. The second is to make intrapersonal and interpersonal distances separated. Specifically, with each sample being a focal sample, a triplet is formed of this sample, a sample of the same label and a sample of different label from the other modality. With the focal sample being either of the two modalities, two sets of triplets are formed. A margin is forced between the interpersonal cross-modality distance and the intrapersonal cross-modality distance facilitated by these triplets. In methods that only use pairwise constraints, all interpersonal constraints have to be applied, while a large number of them may be already separable with intrapersonal distances. By using the hinge loss to force a margin between the two kinds of distances, only those interpersonal distances that trigger the triplet-based loss are applicable for optimization, hence making the proposed method more efficient, particularly appealing when there are large numbers of subjects or training images.

The rest of this paper is organized as follows. Section II gives a brief review of related works. The proposed framework, notations and formulation of distance constraints are provided in Section III. Problem formulations of the proposed MCM^{2}L and KMCM^{2}L are given in Section IV. The optimization method and an analysis of computational complexity are given and discussed in Sections V and VI. In Section VII, experimental results on an ID card and two benchmark datasets

Manuscript received May 25, 2016; revised February 22, 2017; accepted May 27, 2017. This work was supported in part by the National Science Foundation of China under Grant 61432008, Grant 61673203, and Grant 61603193, in part by the Young Elite Scientists Sponsorship Program by CAST under Grant YESS 20160035, and in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization. The work of J. Huo was supported by the China Scholarship Council through the University of Manchester. This paper was recommended by Associate Editor S. Zafeiriou. *(Corresponding author: Yang Gao.)*

J. Huo, Y. Gao, and Y. Shi are with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210046, China (e-mail: huojing1989@gmail.com; gaoy@nju.edu.cn; syh@nju.edu.cn).

W. Yang is with the School of Computer Science and Technology, Nanjing Normal University, Nanjing 210046, China, and also with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210046, China (e-mail: nju.yangwanqi@gmail.com).

H. Yin is with the School of Electrical and Electronic Engineering, University of Manchester, Manchester M13 9PL, U.K. (e-mail: hujun.yin@manchester.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2017.2715660

are presented, together with discussion. We conclude the study in Section VIII.

II. Related Work

In this section, two related research topics are briefly reviewed: 1) heterogeneous face recognition and 2) distance metric learning.

A. Heterogeneous Face Recognition

For heterogeneous face recognition, the main focus is to remove variations caused by modality differences. Based on the way to remove modality variations, the methods can be categorized into three groups.

1) Synthesis-Based Methods: The synthesis-based methods map the data of one modality into another by synthesizing [8]. Related work includes synthesizing sketches from photograph and then comparing synthesized images with sketches drawn by artists [3], [4], [9]–[11]. One drawback of these methods is that different synthesizing methods have to be used if the modalities of two compared images change. Besides, it is difficult to synthesize well from one modality into another. The variations introduced by different modalities are difficult to remove completely by synthesizing methods.

2) Modality Invariant Feature Extraction-Based Methods: These methods try to remove the modality variations by extracting or learning face features that are robust to modality changes. Liao *et al.* [12] proposed to use difference-of-Gaussian (DoG) filtering and multiscale block local binary pattern (LBP) to extract face features. Zhu *et al.* [5] adopted a simple modality invariant feature extraction method involving three steps: 1) log-DoG filtering; 2) local encoding; and 3) uniform feature normalization. Although hand-crafted features have achieved good performances, a number of modality-invariant face feature learning methods have been developed and they are more efficient and do not require prior domain specific knowledge. Zhang *et al.* [13] proposed a coupled information-theoretic encoding method to maximize the mutual information between two modalities in the quantized feature spaces. A coupled discriminative feature learning (CDFL) method is proposed in [14]. It learns a few image filters to maximize interclass variations and minimize intraclass variations of the learned feature in a new feature space. Yi *et al.* [15] proposed to use restricted Boltzmann machines to learn a shared representation and achieved good performance.

3) Common Subspace-Based Methods: In these methods, data of different modalities are mapped into a new, common subspace, so that they become comparable. Klare and Jain [16] proposed to represent face images of different modalities in terms of their similarities to a set of prototype face images. The prototype-based face representation was further projected to a linear discriminant subspace where the recognition was performed. In [17], a common discriminant feature extraction (CDFE) was proposed to learn a common subspace to attain both intraclass compactness and interclass dispersion. Lei and Li [18] proposed a spectral regression-based method to learn a discriminative subspace. Its objective was similar to that of linear discriminant analysis (LDA) [19]. Huang *et al.* [6] further extended this method by adding two regularization terms to force data from different classes to be separate and data of the same class to be close. The aim of the above methods is to maintain intraclass compactness and interclass separability of entire dataset. For face recognition, the goal is to make the intrapersonal and interpersonal distances separable. Maintaining intraclass compactness and interclass separability can be inconsistent with this objective to certain extent. Besides, for those subjects whose interclass separabilities are already large, further forcing them to be even larger can be inappropriate and unnecessary. Compactness (and separability) should be a relative from subject to subject. Therefore, such relative constraint-based metric learning is explored in this paper.

B. Distance Metric Learning

The goal of metric learning is to learn a distance function to satisfy a set of distance constraints defined on the training data [20]. The commonly used distance constraints include pairwise must-link/cannot-link constraints and triplet-based relative constraints. Previously, most effort has been on learning Mahalanobis distance-based metrics, which are for data of single modality.

1) Mahalanobis Metric Learning: The existing Mahalanobis metric learning can be classified into two categories [21], global-based and local-based. Global-based methods try to make the samples of same class close and the samples of different class apart by using only pairwise distance constraints. The work in [22] is an example. On the other side, local-based methods refer to those that use local neighborhood information to learn a metric. Such methods are able to deal with data that are globally nonlinear but can be seen as locally linear. Most of the previous methods are local-based [23]–[26]. For example, in [23], Fisher discriminant analysis was reformulated by assigning higher weights to neighboring pairs. Goldberger *et al.* [24] proposed to learn a distance to maximize the performance of the nearest neighbor classification by optimizing the leave-one-out classification rate on training data. In [25], a large margin nearest neighbor (LMNN) method was proposed to force a margin between a sample's nearest neighbors of same class and its nearest neighbors of different classes. The proposed method in this paper is similar to the framework of LMNN as both pairwise and triplet-based constraints are used. The difference is that the proposed method takes modality information into consideration and the constructed pairs and triplets are all cross-modality-based and the learned metric is for cross-modality distance matching.

2) Cross-Modality Metric Learning: Since the Mahalanobis distance metric is developed for data of a single modality and is hence unable to remove variations across modalities. There have been several attempts to learn cross-modality metrics in the literature. In [27], a cross modal metric learning (CMML) method was proposed to learn metrics by using pairwise constraints. Our proposed method differs from the CMML in the constraints adopted in learning the metric. Besides, the CMML is in fact a global-based cross-modality metric learning method, while the proposed method is local-based. In [28],

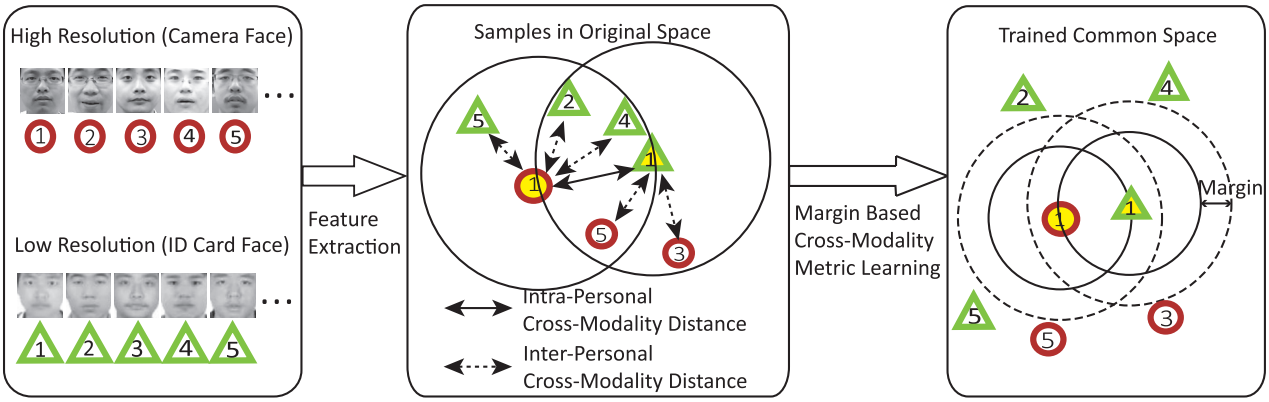


Fig. 1. Illustration of the objective in the proposed methods. 1) Intrapersonal cross-modality distances are minimized on training set (circles and triangles denote two modalities). 2) For each focal sample (the triangle and the circle filled with yellow are two examples), its interpersonal cross-modality distances are constrained to be greater than its intrapersonal cross-modality distances plus a margin.

a method, termed multiview metric learning with global consistency and local smoothness, was derived to learn cross-view metric but was designed under a semisupervised setting. It learns a projection function for each unlabeled sample. Such settings make it unsuitable for heterogeneous face recognition. Zhou *et al.* [29] extended the locally linear embedding to handle heterogeneous data. By fusing the locally linear information and pairwise distance constraints into a single framework, a heterogeneous metric was learned. In [30], a low rank bilinear cross-modality similarity learning method was proposed. The method adopts a logistic loss and pairwise distance constraints to learn a bilinear similarity function. Siena *et al.* [31] proposed a maximum-margin coupled mappings (MMCM) method to learn two projections with the objective to force margins between pairs of same class cross-modal samples and pairs of different classes samples. MMCM also uses both pairwise and triplet-based constraints. However, the main difference to our method is that it only sets samples of one modality as focal samples when constructing triplets. Thus, samples of the opposite modality do not get the same separability. Besides, we have also extended our method to a kernelized version to handle nonlinear data.

III. Framework, Notations, and Distance Constraints

A. Framework

Fig. 1 provides an overview of the proposed method. Take the ID card face recognition problem as an example, the numbers in the figure denote labels of the subjects and triangles and circles are samples of high and low resolution modalities, respectively. For heterogeneous face recognition, two kinds of cross-modality distances (intrapersonal cross-modality distance and interpersonal cross-modality distance) need to be distinguished. They are shown in Fig. 1(middle).

In the proposed method, a cross-modality metric is optimized to meet two sets of constraints.

1) The intrapersonal cross-modality distances are minimized. For example, for samples of subject 1, the intrapersonal cross-modality distance is the distance between the triangle and the circle labeled as 1 in Fig. 1.
2) Take each sample in the training set as a focal sample, its interpersonal cross-modality distances are constrained to be larger than its intrapersonal cross-modality distances plus a margin. For example, with the focal sample being subject 1 of the high resolution modality, it can be seen that samples of labels 2, 4, and 5 of the low resolution modality have smaller distances compared with the intrapersonal distance in the original feature space.

Hence, during the optimization process, the samples that violate this constraint will be pushed out of a small radius of the focal sample. After optimization, as shown in Fig. 1(right), samples of the same label (subject 1) in the common space lie closer and samples of different labels have been pushed out of the radius of the focal samples.

B. Notations

For a given training image set of n different persons, after feature extraction, one obtains two sets of training samples corresponding to two modalities, denoted as $\mathcal{X} = \{(\mathbf{x}_i, l_i^x)|i = 1, 2, \ldots, N_x\}$ and $\mathcal{Y} = \{(\mathbf{y}_i, l_i^y)|i = 1, 2, \ldots, N_y\}$. $\mathbf{x}_i \in \mathbb{R}^{d_x}$ is the ith training sample of the first modality of dimension d_x and $\mathbf{y}_i \in \mathbb{R}^{d_y}$ is the ith training sample of the second modality of dimension d_y. $l_i^x \in \{1, 2, \ldots, n\}$ is the label of the sample $\mathbf{x}_i$ and $l_i^y \in \{1, 2, \ldots, n\}$ is the label of the sample $\mathbf{y}_i$. N_x and N_y are, respectively, the total numbers of training samples of the two modalities. Suppose that each person i has n_i^x training images of the first modality and n_i^y training images of the second modality, then N_x and N_y can be calculated as $N_x = \sum_i^n n_i^x$ and $N_y = \sum_i^n n_i^y$. Denote $\mathbf{W}_x \in \mathbb{R}^{d_x \times d_c}$ and $\mathbf{W}_y \in \mathbb{R}^{d_y \times d_c}$ two projection matrices that are used to map the training samples of respective modalities into a common space, where d_c is the dimension of the common space.

C. Construction of Cross-Modality Distance Constraints

To minimize the intrapersonal cross-modality distances on the training set, pair-based constraints are formulated where each pair is formulated by two samples of the same label but different modalities. Two sets of such constraints can be constructed as there are two modalities. Define their indices as $\mathcal{S}_1 = \{(i, j)|l_i^x = l_j^y\}$ and $\mathcal{S}_2 = \{(i, j)|l_i^y = l_j^x\}$. The number of

pairs in both $\mathcal{S}_1$ and $\mathcal{S}_2$ are the same, $\sum_{i=1}^{n} n_i^x \times n_i^y$. The pairs defined by $\mathcal{S}_1$ and $\mathcal{S}_2$ fully overlap. Hence, the two sets can be combined or one can just use one set $\mathcal{S} = \mathcal{S}_1 = \{(i,j)|l_i^x = l_j^y\}$. For datasets containing multiple samples of each subject, nearest neighbors can be used to reduce the index sets. In this case, two sets of indices of the intrapersonal cross-modality sample pairs $\mathcal{S}_1 = \{(i,j)|l_i^x = l_j^y$ and $\mathbf{y}_j \in \mathcal{K}_s(\mathbf{x}_i, k)\}$ and $\mathcal{S}_2 = \{(i,j)|l_i^y = l_j^x$ and $\mathbf{x}_j \in \mathcal{K}_s(\mathbf{y}_i, k)\}$ can be formed, where $\mathcal{K}_s(\mathbf{x}_i, k)$ denotes the k cross-modal nearest neighbors of $\mathbf{x}_i$ from the same class. The two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ partly overlap, and the maximum number of pairs in the two combined sets are $\sum_{i=1}^{n}(n_i^x + n_i^y) \times k$. We use $\mathcal{S} = \{(i,j)|l_i^x = l_j^y$ and $(\mathbf{y}_j \in \mathcal{K}_s(\mathbf{x}_i, k)$ or $\mathbf{x}_i \in \mathcal{K}_s(\mathbf{y}_j, k))\}$ to denote the combined set.

To meet the second objective, triplets are constructed. By picking each sample in the training set as a focal sample, a triplet is formed of this sample, a sample of the same class and a sample of different class both from the other modality. With the focal sample being either of the two modalities, two sets of indices of triplets $\mathcal{D}_1 = \{(i,j,k)|l_i^x = l_j^y, l_i^x \neq l_k^y\}$ and $\mathcal{D}_2 = \{(i,j,k)|l_i^y = l_j^x, l_i^y \neq l_k^x\}$ are formed. The numbers of triplets in the two sets are $\sum_{i=1}^{n} n_i^x \times n_i^y \times (N_y - n_i^y)$ and $\sum_{i=1}^{n} n_i^y \times n_i^x \times (N_x - n_i^x)$, respectively. The two sets do not overlap. As $(N_y - n_i^y)$ and $(N_x - n_i^x)$, the numbers of cross-modality samples of different classes can be fairly large. So for computational efficiency, k cross-modality neighbors of different classes can be used instead. The numbers of triplets in the two sets then reduce to $\sum_{i=1}^{n} n_i^x \times n_i^y \times k$ and $\sum_{i=1}^{n} n_i^y \times n_i^x \times k$, respectively. The two sets become $\mathcal{D}_1 = \{(i,j,k)|l_i^x = l_j^y, l_i^x \neq l_k^y, \mathbf{y}_k \in \mathcal{K}_d(\mathbf{x}_i, k)\}$ and $\mathcal{D}_2 = \{(i,j,k)|l_i^y = l_j^x, l_i^y \neq l_k^x, \mathbf{x}_k \in \mathcal{K}_d(\mathbf{y}_i, k)\}$, with $\mathcal{K}_d(\mathbf{x}_i, k)$ denotes the k cross-modal nearest neighbors of $\mathbf{x}_i$ from different classes.

IV. Problem Formulation

With the above definitions, we can now formulate the proposed metric learning scheme. The linear margin-based cross-modality metric learning (MCM^{2}L) is first presented. Then extension to kernelized MCM^{2}L (KMCM^{2}L) is derived.

A. Margin-Based Cross-Modality Metric Learning

The objective of MCM^{2}L has two parts. For the first part, the intrapersonal cross-modality distance constraints are used for minimizing intrapersonal distances. For the second part, a margin is forced between the intrapersonal cross-modality and interpersonal cross-modality distances, only those interpersonal cross-modality distances that are inseparable are useful for learning the metric. The role of those interpersonal cross-modality pairs is similar to that of support vectors in the support vector machines.

For cross-modality sample pairs sharing the same label indexed by set $\mathcal{S}$, the objective is to minimize their distances as follows:

$$\mathcal{L}_p(\mathbf{W}_x, \mathbf{W}_y) = \sum_{(i,j)\in\mathcal{S}} \left\| \mathbf{W}_x^T \mathbf{x}_i - \mathbf{W}_y^T \mathbf{y}_j \right\|^2. \tag{1}$$

In (1), $\|\mathbf{W}_x^T \mathbf{x}_i - \mathbf{W}_y^T \mathbf{y}_j\|^2$ is the distance between x_i and y_j, measured by projecting them into a common space. $\mathcal{L}_p$ is a loss function defined on pairwise cross-modality constraints. It penalizes large distances between intrapersonal cross-modality samples in the optimization process.

For the second part of the objective, it is to make the interpersonal cross-modality distances indexed by triplets greater than the corresponding intrapersonal cross-modality distances plus a margin. A penalty term is defined to penalize triplets that violate the objective

$$\begin{aligned}\mathcal{L}_t(\mathbf{W}_x, \mathbf{W}_y) &= \sum_{(i,j,k)\in\mathcal{D}_1} \left[1 + \left\| \mathbf{W}_x^T \mathbf{x}_i - \mathbf{W}_y^T \mathbf{y}_j \right\|^2 - \left\| \mathbf{W}_x^T \mathbf{x}_i - \mathbf{W}_y^T \mathbf{y}_k \right\|^2 \right]_+ \\ &\quad + \sum_{(i,j,k)\in\mathcal{D}_2} \left[1 + \left\| \mathbf{W}_y^T \mathbf{y}_i - \mathbf{W}_x^T \mathbf{x}_j \right\|^2 - \left\| \mathbf{W}_y^T \mathbf{y}_i - \mathbf{W}_x^T \mathbf{x}_k \right\|^2 \right]_+ \end{aligned} \tag{2}$$

where $\mathcal{L}_t$ denotes the loss function defined on the triplet-based constraints and $[a]_+ = \max(a, 0)$ the hinge loss. The first term sets the samples of the first modality as the focal samples and the resulting triplet indices are in set $\mathcal{D}_1$; whilst the second term sets the samples of the second modality as the focal samples and the corresponding indices are in set $\mathcal{D}_2$. Different with the MMCM method proposed in [31], MMCM uses either set $\mathcal{D}_1$ or set $\mathcal{D}_2$ by setting samples of one modality as focal samples. Thus samples of the other modality do not have the same separability.

By using the hinge loss, if the interpersonal cross-modality sample pairs of a focal sample have smaller distances than its intrapersonal cross-modality distances, the interpersonal cross-modality sample pairs will trigger a loss while other pairs make no contribution to the loss. During the optimization, the samples that trigger a loss will generate a push force to repel these samples away from the focal sample. Without loss of generality, a unit margin is used in the method as indicated in (2). In fact, using any margin of size $m > 0$ will result in the same results, since the margin size only affects the scale of the squared distance. This has been discussed in [25].

Combining these two loss functions, we have the final objective for the proposed method

$$\min_{\mathbf{W}_x, \mathbf{W}_y} \mathcal{L}(\mathbf{W}_x, \mathbf{W}_y) = \mu \mathcal{L}_p(\mathbf{W}_x, \mathbf{W}_y) + (1-\mu)\mathcal{L}_t(\mathbf{W}_x, \mathbf{W}_y) \tag{3}$$

where μ is a tradeoff parameter. The two terms are complementary. The first term pulls cross-modality samples of same labels closer, while the second pushes the nearest cross-modal samples of different labels apart.

B. Kernelized Margin-Based Cross-Modality Metric Learning

High-dimensional face features often lie on nonlinear manifolds. Using linear projection functions may cause performance degradation. The proposed method is further integrated with the kernel tricks to project face features into an implicit high-dimensional feature space to make face features of different person more separable.

Suppose that training samples of two modalities are mapped into a high-dimensional space by mapping function ϕ. Mapped samples are represented as $\Phi_x = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_{N_x})]$ and $\Phi_y = [\phi(\mathbf{y}_1), \phi(\mathbf{y}_2), \ldots, \phi(\mathbf{y}_{N_y})]$. Suppose that the projection matrices of the high-dimensional space can be represent as a linear combination of high-dimensional samples, $\mathbf{W}_x = \Phi_x \mathbf{A}_x$, with $\mathbf{A}_x \in \mathbb{R}^{N_x \times d_c}$ and $\mathbf{W}_y = \Phi_y \mathbf{A}_y$, with $\mathbf{A}_y \in \mathbb{R}^{N_y \times d_c}$.

Then, the loss function defined on the pairwise constraints is changed to

$$\mathcal{L}_p(\mathbf{A}_x, \mathbf{A}_y) = \sum_{(i,j)\in\mathcal{S}} \left\| \mathbf{A}_x^T \mathbf{k}_i^x - \mathbf{A}_y^T \mathbf{k}_j^y \right\|^2 \tag{4}$$

where $\mathbf{k}_i^x = [\phi(\mathbf{x}_1)^T\phi(\mathbf{x}_i), \phi(\mathbf{x}_2)^T\phi(\mathbf{x}_i), \ldots, \phi(\mathbf{x}_{N_x})^T\phi(\mathbf{x}_i)]^T \in \mathbb{R}^{N_x}$, $\mathbf{k}_j^y = [\phi(\mathbf{y}_1)^T\phi(\mathbf{y}_j), \phi(\mathbf{y}_2)^T\phi(\mathbf{y}_j), \ldots, \phi(\mathbf{y}_{N_y})^T\phi(\mathbf{y}_j)]^T \in \mathbb{R}^{N_y}$. For the training samples of the first modality, denote $\mathbf{K}^x \in \mathbb{R}^{N_x \times N_x}$ the kernel matrix with $\mathbf{K}^x_{i,j} = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$. Denote $\mathbf{K}^y \in \mathbb{R}^{N_y \times N_y}$ the kernel matrix of the second modality with $\mathbf{K}^y_{i,j} = \phi(\mathbf{y}_i)^T\phi(\mathbf{y}_j)$, then $\mathbf{k}_i^x$ and $\mathbf{k}_j^y$ are the ith and the jth columns of $\mathbf{K}^x$ and $\mathbf{K}^y$, respectively.

Similarly, the loss function defined on the triplet-based constraints becomes

$$\begin{aligned}
&\mathcal{L}_t(\mathbf{A}_x, \mathbf{A}_y) \\
&= \sum_{(i,j,k)\in\mathcal{D}_1} \left[1 + \left\| \mathbf{A}_x^T \mathbf{k}_i^x - \mathbf{A}_y^T \mathbf{k}_j^y \right\|^2 - \left\| \mathbf{A}_x^T \mathbf{k}_i^x - \mathbf{A}_y^T \mathbf{k}_k^y \right\|^2 \right]_+ \\
&\quad + \sum_{(i,j,k)\in\mathcal{D}_2} \left[1 + \left\| \mathbf{A}_y^T \mathbf{k}_i^y - \mathbf{A}_x^T \mathbf{k}_j^x \right\|^2 - \left\| \mathbf{A}_y^T \mathbf{k}_i^y - \mathbf{A}_x^T \mathbf{k}_k^x \right\|^2 \right]_+ .
\end{aligned} \tag{5}$$

The final objective of the KMCM^{2}L method is therefore defined as

$$\min_{\mathbf{A}_x, \mathbf{A}_y} \mathcal{L}(\mathbf{A}_x, \mathbf{A}_y) = \mu \mathcal{L}_p(\mathbf{A}_x, \mathbf{A}_y) + (1-\mu)\mathcal{L}_t(\mathbf{A}_x, \mathbf{A}_y). \tag{6}$$

After learning, in the testing stage, with $\mathbf{A}_x$ and $\mathbf{A}_y$ obtained, the distance between two samples $\mathbf{x}$ and $\mathbf{y}$ is calculated by $\|\mathbf{A}_x^T\mathbf{k}^x - \mathbf{A}_y^T\mathbf{k}^y\|^2$, where $\mathbf{k}^x = [\phi(\mathbf{x}_1)^T\phi(\mathbf{x}), \phi(\mathbf{x}_2)^T\phi(\mathbf{x}), \ldots, \phi(\mathbf{x}_{N_x})^T\phi(\mathbf{x})]^T \in \mathbb{R}^{N_x}$ and $\mathbf{k}^y = [\phi(\mathbf{y}_1)^T\phi(\mathbf{y}), \phi(\mathbf{y}_2)^T\phi(\mathbf{y}), \ldots, \phi(\mathbf{y}_{N_y})^T\phi(\mathbf{y})]^T \in \mathbb{R}^{N_y}$.

V. Optimization

Comparing (4) and (5) with (1) and (2), the only difference is that the original samples $\mathbf{x}_i$, $\mathbf{y}_i$ are now changed to $\mathbf{k}_i^x$ and $\mathbf{k}_i^y$. Therefore, the two optimization problems defined in (3) and (6) can be solved using the same algorithm. As the hinge loss adopted in (2) and (5) is not smooth, a subgradient descent is used. The optimization procedure is to compute the subgradients of two projection matrices separately and perform subgradient descent. Notice that the objective functions are nonconvex with respect to both $\mathbf{W}_x$ and $\mathbf{W}_y$ or $\mathbf{A}_x$ and $\mathbf{A}_y$. However, such an optimization procedure works well in practise. The detailed optimization procedure is as follows (taking the optimization of KMCM^{2}L as an example).

Differentiating (4) with respect to $\mathbf{A}_x^{(t)}$ and $\mathbf{A}_y^{(t)}$ results in the following gradient terms:

$$\begin{aligned}
\frac{\partial \mathcal{L}_p}{\partial \mathbf{A}_x^{(t)}} &= 2\left(\sum_{(i,j)\in\mathcal{S}} \mathbf{k}_i^x \mathbf{k}_i^{xT} \right) \mathbf{A}_x^{(t)} - 2\left(\sum_{(i,j)\in\mathcal{S}} \mathbf{k}_i^x \mathbf{k}_j^{yT} \right) \mathbf{A}_y^{(t)} \\
&= \mathbf{P}_x^1 \mathbf{A}_x^{(t)} - \mathbf{P}_x^2 \mathbf{A}_y^{(t)}
\end{aligned} \tag{7}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}_p}{\partial \mathbf{A}_y^{(t)}} &= 2\left(\sum_{(i,j)\in\mathcal{S}} \mathbf{k}_j^y \mathbf{k}_j^{yT} \right) \mathbf{A}_y^{(t)} - 2\left(\sum_{(i,j)\in\mathcal{S}} \mathbf{k}_j^y \mathbf{k}_i^{xT} \right) \mathbf{A}_x^{(t)} \\
&= \mathbf{P}_y^1 \mathbf{A}_y^{(t)} - \mathbf{P}_y^2 \mathbf{A}_x^{(t)}
\end{aligned} \tag{8}$$

where t denotes the iteration. The optimization process results in smaller intrapersonal cross-modality distances. By precalculating $\mathcal{S}$ and keeping it fixed, $\mathbf{P}_x^1$ and $\mathbf{P}_x^2$ in (7) do not change during the optimization and they can be calculated before the iteration begins. At each iteration, they are, respectively, multiplied with $\mathbf{A}_x^{(t)}$ and $\mathbf{A}_y^{(t)}$ to obtain the gradient. The gradient in (8) is calculated in the same manner. Similarly, to optimize MCM^{2}L, the gradient of (1) with respect to $\mathbf{W}_x^{(t)}$ and $\mathbf{W}_y^{(t)}$ can be obtained in the same way.

To calculate the subgradients of $\mathcal{L}_t$ in (5) with respect to $\mathbf{A}_x^{(t)}$ and $\mathbf{A}_y^{(t)}$, denote $\tilde{\mathcal{D}}_1^{(t)}$ and $\tilde{\mathcal{D}}_2^{(t)}$ two subsets of $\mathcal{D}_1^{(t)}$ and $\mathcal{D}_2^{(t)}$ that contain the indices of triplets those trigger the hinge loss defined in (5). The subgradients are as follows:

$$\begin{aligned}
\frac{\partial \mathcal{L}_t}{\partial \mathbf{A}_x^{(t)}} &= 2\left[\sum_{(i,j,k)\in\tilde{\mathcal{D}}_2^{(t)}} \left(\mathbf{k}_j^x \mathbf{k}_j^{xT} - \mathbf{k}_k^x \mathbf{k}_k^{xT} \right) \right] \mathbf{A}_x^{(t)} \\
&\quad + 2\left[\sum_{(i,j,k)\in\tilde{\mathcal{D}}_1^{(t)}} \left(\mathbf{k}_i^x \mathbf{k}_k^{yT} - \mathbf{k}_i^x \mathbf{k}_j^{yT} \right) \right] \mathbf{A}_y^{(t)} \\
&\quad + 2\left[\sum_{(i,j,k)\in\tilde{\mathcal{D}}_2^{(t)}} \left(\mathbf{k}_k^x \mathbf{k}_i^{yT} - \mathbf{k}_j^x \mathbf{k}_i^{yT} \right) \right] \mathbf{A}_y^{(t)} \\
&= \mathbf{Q}_x^{1(t)} \mathbf{A}_x^{(t)} + \mathbf{Q}_x^{2(t)} \mathbf{A}_y^{(t)}
\end{aligned} \tag{9}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}_t}{\partial \mathbf{A}_y^{(t)}} &= 2\left[\sum_{(i,j,k)\in\tilde{\mathcal{D}}_1^{(t)}} \left(\mathbf{k}_j^y \mathbf{k}_j^{yT} - \mathbf{k}_k^y \mathbf{k}_k^{yT} \right) \right] \mathbf{A}_y^{(t)} \\
&\quad + 2\left[\sum_{(i,j,k)\in\tilde{\mathcal{D}}_2^{(t)}} \left(\mathbf{k}_i^y \mathbf{k}_k^{xT} - \mathbf{k}_i^y \mathbf{k}_j^{xT} \right) \right] \mathbf{A}_x^{(t)} \\
&\quad + 2\left[\sum_{(i,j,k)\in\tilde{\mathcal{D}}_1^{(t)}} \left(\mathbf{k}_k^y \mathbf{k}_i^{xT} + \mathbf{k}_j^y \mathbf{k}_i^{xT} \right) \right] \mathbf{A}_x^{(t)} \\
&= \mathbf{Q}_y^{1(t)} \mathbf{A}_y^{(t)} + \mathbf{Q}_y^{2(t)} \mathbf{A}_x^{(t)}.
\end{aligned} \tag{10}$$

Similarly, calculation of the subgradient in (9) can be decomposed into computing $\mathbf{Q}_x^{1(t)}$ and $\mathbf{Q}_x^{2(t)}$. $\mathbf{Q}_x^{1(t)}$ is then multiplied with $\mathbf{A}_x^{(t)}$ and $\mathbf{Q}_x^{2(t)}$ multiplied with $\mathbf{A}_y^{(t)}$. However, $\mathbf{Q}_x^{1(t)}$ and $\mathbf{Q}_x^{2(t)}$ do change during the optimization, as the sets of triplets $\mathcal{D}_1^{(t)}$ and $\mathcal{D}_2^{(t)}$ that trigger the loss terms vary at each iteration. Directly recomputing $\mathbf{Q}_x^{1(t)}$ and $\mathbf{Q}_x^{2(t)}$ would be very

Algorithm 1 KMCM^{2}L

1: Initialize $\mathbf{A}_x^{(0)}$ and $\mathbf{A}_y^{(0)}$ using KPCA and coupled spectral regression and compute the kernel matrix $\mathbf{K}^x$ and $\mathbf{K}^y$
2: Compute $\mathcal{S}$, $\mathcal{D}_1^{(0)}$ and $\mathcal{D}_2^{(0)}$
3: Compute $\mathbf{P}_x^1$, $\mathbf{P}_x^2$, $\mathbf{P}_y^1$ and $\mathbf{P}_y^2$ in Eqs. (7)(8)
4: Compute $\mathbf{Q}_x^{1(0)}$, $\mathbf{Q}_x^{2(0)}$, $\mathbf{Q}_y^{1(0)}$ and $\mathbf{Q}_y^{2(0)}$ in Eqs. (9)(10)
5: Initialize $t = 0$
6: **while** not converged **do**
7: Compute sub-gradients $\mathbf{G}_x^{(t)}$ and $\mathbf{G}_y^{(t)}$ in Eqs. (11)(12)
8: $\mathbf{A}_x^{(t+1)} = \mathbf{A}_x^{(t)} - \alpha \mathbf{G}_x^{(t)}$
9: $\mathbf{A}_y^{(t+1)} = \mathbf{A}_y^{(t)} - \alpha \mathbf{G}_y^{(t)}$
10: Update sets $\tilde{\mathcal{D}}_1^{(t+1)}$ and $\tilde{\mathcal{D}}_2^{(t+1)}$
11: Update $\mathbf{Q}_x^{1(t+1)}$, $\mathbf{Q}_x^{2(t+1)}$, $\mathbf{Q}_y^{1(t+1)}$ and $\mathbf{Q}_y^{2(t+1)}$ in Eqs. (9)(10)
12: $t = t + 1$
13: **end while**
14: Output $\mathbf{A}_x$ and $\mathbf{A}_y$

costly since the two sets can be very large. A few techniques given in [32] can be used to efficiently update $\mathbf{Q}_x^{1(t)}$ and $\mathbf{Q}_x^{2(t)}$. Similar to the subgradients of KMCM^{2}L, the subgradients of $\mathcal{L}_t$ in (2) with respect to $\mathbf{W}_x^{(t)}$ and $\mathbf{W}_y^{(t)}$ can be calculated in the same way.

By putting the two terms together, we have the subgradients of (6) with respect to $\mathbf{A}_x^{(t)}$ and $\mathbf{A}_y^{(t)}$

$$\mathbf{G}_x^{(t)} = \frac{\partial \mathcal{L}}{\partial \mathbf{A}_x^{(t)}} = \mu \frac{\partial \mathcal{L}_p}{\partial \mathbf{A}_x^{(t)}} + (1-\mu)\frac{\partial \mathcal{L}_t}{\partial \mathbf{A}_x^{(t)}} \tag{11}$$

$$\mathbf{G}_y^{(t)} = \frac{\partial \mathcal{L}}{\partial \mathbf{A}_y^{(t)}} = \mu \frac{\partial \mathcal{L}_p}{\partial \mathbf{A}_y^{(t)}} + (1-\mu)\frac{\partial \mathcal{L}_t}{\partial \mathbf{A}_y^{(t)}}. \tag{12}$$

The detailed procedure of the proposed method is described in Algorithm 1. KMCM^{2}L can be initialized using coupled spectral regression (CSR) [18] to find the projection matrices of kernel principal component analysis (KPCA) [33] for two modalities. For MCM^{2}L, it is directly initialized using principal component analysis (PCA) by regarding all the samples as from one modality. The optimization procedure of MCM^{2}L is the same as that of KMCM^{2}L.

VI. Complexity Analysis

For the training process, the main computational cost lies in the while loop in Algorithm 1. The complexity of the initialization steps compared with that of the while loop is low and thus is omitted in discussion. Table I provides the complexity of the steps in the while loop of both MCM^{2}L and KMCM^{2}L together with the complexity of the testing procedure. In the table, k is the number of nearest neighbors and is usually much smaller than the number of dimensions d or the number of samples N.

From Table I, if $d \gg N$, the complexity of both MCM^{2}L and KMCM^{2}L is dominated by step 7 which is of $O(d^3)$ and $O(dN^2)$. However, the complexity can be reduced by performing PCA to reduce the dimension of features before applying the proposed methods. On the other side, if $N \gg d$, the complexity of MCM^{2}L and KMCM^{2}L is, respectively, dominated by steps 10 and 11 which is of $O(d^2N + dN^2)$ and $O(k^2N^3)$. Steps 10 and 11 are related to recalculating the different class nearest neighbors to find triplets that trigger losses and update $\mathbf{Q}_x^{1(t+1)}$, $\mathbf{Q}_x^{2(t+1)}$, $\mathbf{Q}_y^{1(t+1)}$, and $\mathbf{Q}_y^{2(t+1)}$. The complexity of the two steps can be largely reduced by using the active set method and tree-based search suggested in [32]. In this paper, we adopted the active set method to boost the speed.

TABLE I
Complexity Analysis of MCM^{2}L and KMCM^{2}L

	Step 7	Step 8,9	Step 10	Step 11	Testing
MCM^{2}L	$O(d^3)$	$O(d^2)$	$O(d^2N + dN^2)$	$O(d^2k^2N)$	$O(d^2)$
KMCM^{2}L	$O(dN^2)$	$O(dN)$	$O(dN^2)$	$O(k^2N^3)$	$O(dN)$

For the testing procedure, the complexity involves projecting two samples into the common space and compute their distance. The complexity is $O(d^2)$ or $O(dN)$ for MCM^{2}L or KMCM^{2}L, which is fairly low.

VII. Experimental Results

In this section, experimental results on three datasets of different heterogeneous face recognition scenarios are presented. First, the proposed method was evaluated on an ID card face dataset collected in Nanjing University (NJU-ID dataset[1]). Detailed information of the dataset is given below. To demonstrate the applicability of the proposed methods to other heterogeneous face recognition scenarios, the widely adopted CUHK Face Sketch FERET (CUFSF) dataset [9], [13] and CASIA NIR-VIS 2.0 dataset [34] were also used. The proposed methods were compared with several common subspace methods and state-of-the-art heterogeneous face recognition methods. The effectiveness of different feature extraction methods was also evaluated along with the use of the proposed methods. Detailed experiments and results are given below.

A. Datasets and Evaluation Protocols

1) NJU-ID Dataset: The ID cards used were the second generation of resident ID cards of China. A noncontact IC chip is embedded in the card. On the chip, a low resolution photograph of the card owner is stored and can be obtained by IC card reader. NJU-ID dataset contains images of 256 persons. For each person, there are one card image and one image collected from a high resolution digital camera. The ID card image is of resolution 102×126, while the camera image is of resolution 640×480. Exemplar pairs from the dataset are shown in Fig. 2. To evaluate on this dataset, we randomly divided the dataset into tenfolds according to identity information. The tenfolds were fully independent and nonoverlapping. On the testing fold, each person had one intrapersonal cross-modality image pair. Then interpersonal cross-modality image pairs were randomly selected to make the two kinds of cross-modality image pairs of same number.

2) CUHK Face Sketch FERET Dataset [9], [13]: The CUFSF dataset was used for photograph to sketch face matching. It includes 1194 persons from the FERET dataset [35]. For

[1]The dataset is available from http://cs.nju.edu.cn/rl/Data.html.

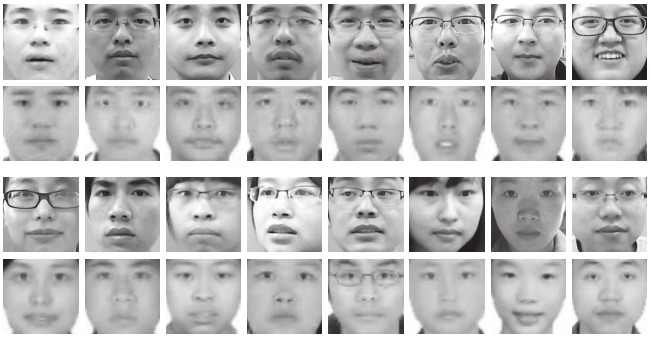


Fig. 2. Exemplar face pairs of NJU-ID dataset. The first and third rows are aligned face images of high resolution modality (captured by a digital camera) and the second and fourth rows are the corresponding low resolution face images (stored on the ID cards).

Fig. 3. Examples of heterogeneous face image pairs. The first and second rows are aligned face photograph and sketches, respectively, from CUFSF dataset. The third and fourth rows are aligned face images of visible light and near-infrared, respectively, from CASIA NIR-VIS 2.0 dataset.

each person, there are one photograph and one sketch drawn by an artist after viewing the photograph. Exemplar pairs are shown in the first two rows of Fig. 3. To evaluate on this dataset, we randomly split the dataset into two parts for ten times and each time the first part was used for training and the other part was used for testing.

3) CASIA NIR-VIS 2.0 Dataset [34]: The CASIA NIR-VIS 2.0 dataset was used for evaluating the visible to NIR face image recognition. It contains 725 subjects. Examples of aligned face images are given in the last two rows of Fig. 3. We followed the same evaluation protocol on this dataset by [34]. The dataset was divided into two views. View 1 was used for parameter tuning and view 2 for testing.

On NJU-ID and CUFSF datasets, face verification rates (VR) are reported and we also provide the receiver operating characteristic (ROC) curves for completeness and the value of area under the ROC curve (AUC) [36]. ROC curves and AUC values can incorporate results of various thresholds and thus serve as a complementary evaluation to the verification performance. On CASIA NIR-VIS 2.0 dataset, following the evaluation protocol used for this dataset, rank-1 recognition rates are reported and the cumulative match characteristic (CMC) curves are also given.

B. Face Feature Extraction

In the experiments, the following face normalization and feature extraction methods were adopted for all the three datasets. For normalization, the faces were first rotated so that the two eyes were located on a horizontal line, and then resized to make the distances between two pupils of 75 pixels. A face region of 160×160 was cropped out, with the eye central to the region's upper edge by 35 pixels and to the region's left edge by 80 pixels.

The second step applied an image filtering technique (self-quotient image) [37] to help compensate illumination variations and also to reduce the variations caused by modality difference. Similar filtering scheme has also been used in [16].

The last step was to extract features of these face images. Three kinds of local feature descriptors were tested, including LBP [38], Gabor [39], and scale invariant feature transform (SIFT) [40]. The raw filtered gray image was also directly used as the gray feature. For gray, LBP and SIFT, the filtered 160×160 face image was also resized to 96×96 and 32×32. So the filtered face images of three scales were used. The gray feature was the images of three scales reshaped into vectors and concatenated into one. For LBP features, uniformed LBP was adopted, extracted by dividing the image into patches of 32×32 with a spacing of 8 pixels. All the local features of an image of three scales were then concatenated into one. The SIFT features were also extracted in patches of 32×32 and all the features were concatenated. For Gabor features, 40 Gabor filers were used (8 orientations and 5 scales) to filter the original 160×160 face image. Then all the filtered images were down sampled to two scales of 16×16 and 8×8 and all the images were reshaped to vectors and concatenated to form the Garbor features. After feature extraction, PCA was applied to all the features so as to retain a processable number of features.

Deep learned features were also used in the study and comparison (see Section VII-G for details).

C. Parameter Analysis

1) Step Size for Gradient Update: The gradient update step size was set to $\alpha_t = \min(\alpha_{t-1} \times 1.01, \alpha_{\max})$. On NJU-ID and CUFSF datasets, for both MCM^{2}L and KMCM^{2}L, $\alpha_1 = 10^{-9}$ and $\alpha_{\max}$ was set 10^{-6}. On CASIA dataset, $\alpha_1 = 10^{-10}$ and $\alpha_{\max}$ was set 10^{-8}.

2) Kernel Parameters: For KMCM^{2}L, on all the three datasets, the kernel type was selected as radial basis function kernel which is defined as $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/(2\sigma^2))$, where σ is the kernel parameter. On both NJU-ID and CUFSF datasets, this parameter was set to 0.5. On CASIA NIR-VIS 2.0 dataset, this was set to 1.5. This parameter was selected form a set of $\{0.1, 0.5, 1, 1.5, 2, 5\}$. On NJU-ID, a separate cross-validation was used for tuning the parameter. On CUFSF, it was tuned on a separate split of data. View 1 was used for tuning the parameter on CASIA NIR-VIS 2.0.

3) Number of Nearest Neighbors: For both NJU-ID and CUFSF datasets, each person has only one image per modality. Therefore the number of the same class cross-modality nearest neighbors must be one. Besides, as the two datasets are relatively small, all the different class cross-modality samples were used to form triplets. For CASIA dataset, as each

TABLE II
COMPARISON OF PERFORMANCE ON NJU-ID DATASET

	Methods	Gray		LBP		Gabor		SIFT		All Features	
		VR(%)	AUC	VR(%)	AUC	VR(%)	AUC	VR(%)	AUC	VR(%)	AUC
Single-Modal	PCA	60.9 ± 6.3	0.578	59.0 ± 2.8	0.540	64.1 ± 4.2	0.616	71.5 ± 6.4	0.716	62.9 ± 5.5	0.609
	LDA	57.8 ± 2.9	0.530	59.8 ± 2.9	0.553	64.6 ± 2.0	0.624	66.6 ± 4.5	0.645	67.2 ± 4.9	0.635
	KPCA	61.1 ± 6.2	0.575	59.9 ± 4.1	0.543	63.1 ± 5.1	0.606	69.1 ± 6.8	0.690	61.5 ± 5.5	0.593
	KDA	62.7 ± 5.8	0.600	71.1 ± 3.5	0.718	70.7 ± 4.1	0.726	76.2 ± 5.7	0.784	74.7 ± 7.2	0.766
	NCA	62.1 ± 5.6	0.580	70.5 ± 3.4	0.708	73.8 ± 6.1	0.769	73.8 ± 5.7	0.768	73.8 ± 6.7	0.764
	LMNN	60.9 ± 5.1	0.575	70.1 ± 3.6	0.697	72.1 ± 4.6	0.735	75.2 ± 4.5	0.777	74.0 ± 5.5	0.759
Multi-Modal	$MMCM_h$	64.5 ± 6.2	0.604	65.2 ± 3.3	0.607	67.2 ± 5.6	0.666	74.2 ± 6.4	0.741	66.6 ± 7.3	0.661
	$MMCM_l$	63.7 ± 6.4	0.597	63.1 ± 2.6	0.602	67.0 ± 5.3	0.662	73.0 ± 6.6	0.739	66.4 ± 7.1	0.655
	CSR	64.3 ± 4.0	0.626	71.5 ± 4.0	0.719	73.6 ± 4.8	0.743	76.0 ± 5.9	0.795	77.0 ± 8.1	0.801
	KCSR	64.9 ± 5.2	0.642	68.0 ± 5.4	0.694	69.1 ± 5.1	0.695	74.8 ± 7.0	0.770	72.4 ± 8.2	0.743
	CCA	64.3 ± 6.5	0.627	67.0 ± 3.7	0.672	65.4 ± 4.0	0.622	72.1 ± 6.3	0.730	71.2 ± 7.9	0.729
	KCCA	65.1 ± 5.2	0.630	57.8 ± 2.0	0.537	65.0 ± 4.7	0.628	67.6 ± 3.8	0.682	65.0 ± 2.8	0.619
	CDFE	66.0 ± 7.8	0.638	68.8 ± 3.0	0.674	68.2 ± 5.2	0.680	75.6 ± 5.1	0.790	70.9 ± 6.4	0.734
	MvDA	66.1 ± 6.1	0.628	66.6 ± 3.8	0.653	64.9 ± 3.1	0.635	69.8 ± 4.7	0.709	68.6 ± 6.8	0.709
	MCM^2L	$\mathbf{67.8 \pm 6.3}$	**0.657**	71.1 ± 4.1	0.691	73.4 ± 6.3	0.749	79.3 ± 4.5	0.816	$\mathbf{77.8 \pm 7.0}$	**0.802**
	$KMCM^2L$	66.0 ± 6.2	0.631	$\mathbf{73.5 \pm 3.6}$	**0.729**	$\mathbf{75.2 \pm 5.4}$	**0.771**	$\mathbf{79.9 \pm 3.4}$	**0.824**	76.8 ± 5.2	0.770

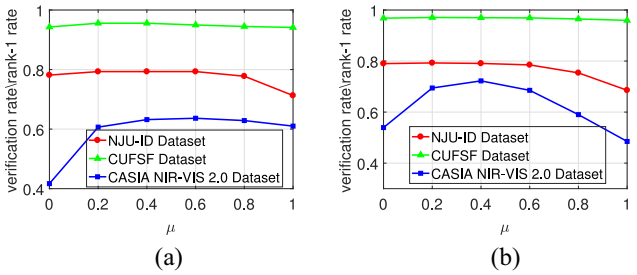


Fig. 4. Influence of parameter μ on VRs on NJU-ID dataset and CUFSF dataset. Together with the influence of parameter μ on rank-1 rates on CASIA NIR-VIS 2.0 dataset. Results of (a) MCM^2L and (b) $KMCM^2L$.

person has more than one image per modality, all the same class cross-modality samples were used to construct pairs. Each different class cross-modality sample together with all the same class cross-modality samples were used to form triplets. The constructed triplets were of a large number, subsampling was performed at each iteration to reduce the number of constructed triplets to a processable number.

4) Influence of Parameter μ: In Fig. 4, the effect of μ from a set of {0, 0.2, 0.4, 0.6, 0.8, 1} is illustrated. On all three datasets, when setting μ to 0 or 1, the performance of both MCM^2L and $KMCM^2L$ suffers a loss; this means both terms in (3) and (6) are needed. On CUFSF dataset, the influence of μ is quite small, as the performances vary little. On NJU-ID dataset, setting μ to 1 means removing the second parts of (3) and (6) and leads to considerable performance reduction. But on the contrary, on CASIA NIR-VIS 2.0 dataset, setting μ to 0 causes performance degradation. This is mainly due to that the first terms in (3) and (6) relate to removing intrapersonal variations. On NJU-ID dataset, each person has only a pair of images; this means the intrapersonal variation is of a small scale. On the other side, on CASIA NIR-VIS 2.0 dataset, each person has 1–22 VIS and 5–50 NIR face images, so the intrapersonal variations are much larger. For the experiments, μ was set to 0.4 on both NJU-ID and CUFSF datasets, and to 0.5 on CASIA NIR-VIS 2.0 dataset.

5) Influence of Number of Reduced Dimensions: We also investigated the effect of dimension reduction. Fig. 5 shows the number of dimensions against the performance on three datasets while using SIFT features. In Fig. 5(a), using a larger number of dimensions achieved relatively better verification results on NJU-ID Dataset. In the rest of the experiments, on NJU-ID dataset, the dimension number was chosen as 450 for both MCM^2L and $KMCM^2L$. Fig. 5(b) shows the influence of dimensions on CUFSF dataset. Similarly, larger dimensions tended to achieve better VRs. However, little improvement was achieved after the dimension exceeding 450. Therefore, on this dataset, the dimension number of 450 was also chosen for MCM^2L and $KMCM^2L$. Fig. 5(c) is the number of dimensions against rank-1 rate on the CASIA NIR-VIS 2.0 dataset. On this dataset, the dimension was set to 2000 for MCM^2L and 1000 for $KMCM^2L$.

D. Results on NJU-ID Dataset

1) Comparison With Single-Modality Methods: MCM^2L and $KMCM^2L$ have been compared with some state-of-the-art single-modality methods, including PCA [41], LDA [19], KPCA [33], kernel discriminant analysis (KDA) [42], neighborhood components analysis (NCA) [24], and LMNN [25]. When applying these methods, the parameters of these methods were adjusted to their optimal. Among the six compared methods, the first four are global methods and the last two are local-based. PCA and KPCA are unsupervised and the other four are supervised. While testing these single modal methods, the data of two modalities are treated as from single modality.

The first part of Table II provides the mean VRs and standard deviations of tenfold cross validation of these methods. AUC values are also provided. The following observations can be made.

1) When using gray and all features, MCM^2L achieved the best results. All features stand for all the four kinds of features combined. $KMCM^2L$ was the best when LBP, Gabor, and SIFT were used. Among the six compared single-modality methods, the best results were achieved by KDA with SIFT feature used. However, MCM^2L and $KMCM^2L$ achieved 3.1% and 3.7% higher VRs than KDA.

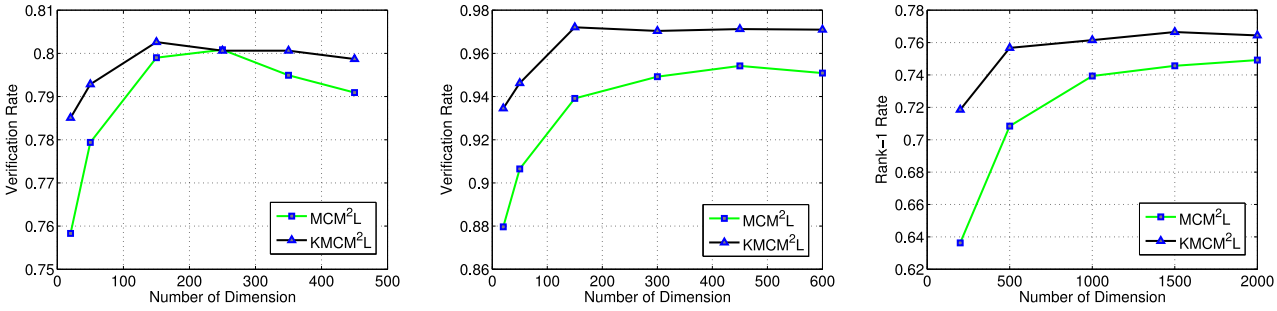


Fig. 5. Influence of dimensions on VRs on (a) NJU-ID dataset, (b) CUSFS dataset, and (c) CASIA NIR-VIS 2.0 dataset with rank-1 rates.

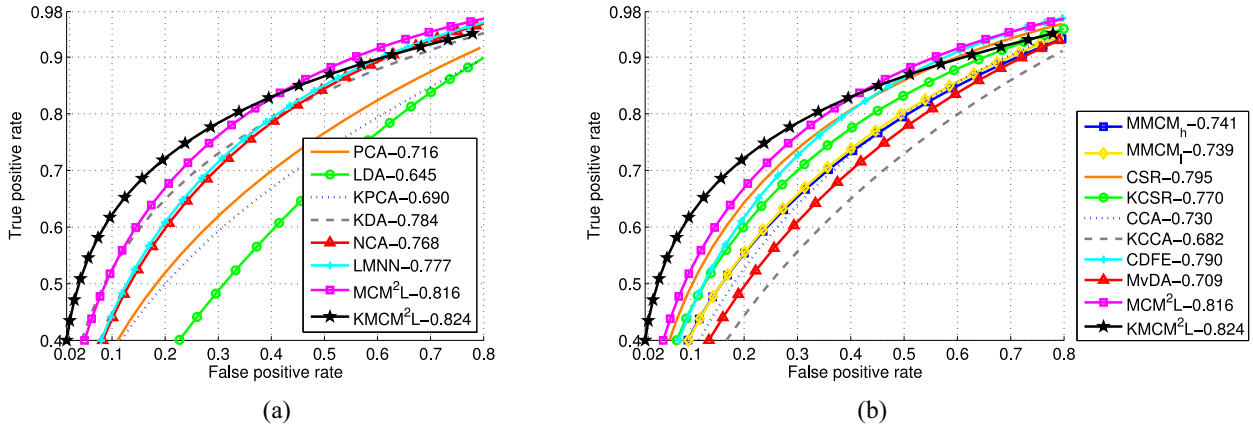


Fig. 6. ROC curves on NJU-ID dataset using SIFT features. Comparison with (a) single-modality methods and (b) multimodality methods.

2) PCA, LDA, KPCA, and KDA are four most commonly used dimension reduction methods. Unsupervised PCA and KPCA did not perform as well as the supervised methods. But for this experiment, the within-class scatter matrix of LDA becomes singular and its performance degrades. KDA is the kernelized version of LDA. By using the kernel trick, the projection matrix is learned in a new feature space with nonlinear mapping. The results of KDA were the best among the four methods. NCA and LMNN are local-based metric learning methods which seem to attribute to their rather good performance.
3) Among all these features, gray features performed the worst. SIFT was the best and Gabor was relatively worse compared to SIFT. A conclusion is that SIFT features are effective for low to high resolution face verification. All four types of features were also combined for verification test. The results of MCM^{2}L and KMCM^{2}L were slightly worse than using SIFT features. Similar results were observed with PCA, KPCA, NCA, and LMNN. This is partly because that using all the features introduces a great deal of redundancy and noise. As the results of using gray, LBP, or Gabor features are poor on this dataset, it seems to indicate that there is a great deal of noise and redundancy in these features. When combined with SIFT, the resulting features are still noisy and indiscriminative, making the performance worse than that of using SIFT alone.

Fig. 6(a) shows the ROC curves of the compared single-modality methods with SIFT features. The superiority of the proposed methods can be clearly seen.

2) Comparison With Multimodality Methods: Since MCM^{2}L and KMCM^{2}L also belong to the category of multimodality methods, they have been compared with seven state-of-the-art multimodality methods in the ID card face verification task. The methods compared include MMCM [31], CSR [18], kernel CSR (KCSR) [18], canonical correlation analysis (CCA) [43], kernel CCA (KCCA) [44], CDFE [17], and multiview discriminant analysis (MvDA) [45]. The source codes of CCA, KCCA, and MvDA were available at their authors websites.[2,3] The MMCM, CSR, KCSR, and CDFE were implemented by ourselves. The parameters of these methods were adjusted to their best for a fair comparison.

The second part of Table II presents the VRs of all the compared methods together with the AUC values. Key results or observations are summarized as follows.

1) First, MCM^{2}L achieved the best results on two out of the five types of features used (i.e., gray and all features). When using LBP, Gabor, and SIFT, the best

[2]http://www.public.asu.edu/~jye02/Software/CCA/index.html
[3]http://vipl.ict.ac.cn/resources/codes

TABLE III
COMPARISON OF PERFORMANCE ON CUFSF DATASET

	Methods	Gray		LBP		Gabor		SIFT		All Features	
		VR(%)	AUC	VR(%)	AUC	VR(%)	AUC	VR(%)	AUC	VR(%)	AUC
Single-Modal	PCA	58.5 ± 1.6	0.597	78.6 ± 0.7	0.863	84.8 ± 0.8	0.923	76.8 ± 0.7	0.826	80.2 ± 0.8	0.882
	LDA	56.8 ± 0.6	0.561	70.8 ± 1.0	0.764	73.8 ± 0.8	0.806	72.4 ± 0.8	0.784	79.3 ± 0.9	0.864
	KPCA	57.5 ± 1.4	0.591	76.6 ± 0.8	0.844	82.2 ± 1.0	0.902	74.4 ± 0.8	0.810	77.0 ± 0.8	0.853
	KDA	72.7 ± 1.1	0.789	94.1 ± 0.7	0.983	96.3 ± 0.6	0.992	94.9 ± 0.7	0.988	97.2 ± 0.5	0.994
	NCA	75.4 ± 1.3	0.831	93.0 ± 0.6	0.979	94.9 ± 0.8	0.987	94.9 ± 0.6	0.987	96.1 ± 0.4	0.993
	LMNN	76.5 ± 1.3	0.840	93.0 ± 0.7	0.978	94.8 ± 0.6	0.987	94.5 ± 0.8	0.986	95.9 ± 0.4	0.991
Multi-Modal	$MMCM_p$	76.1 ± 0.7	0.834	93.7 ± 0.6	0.982	95.7 ± 0.7	0.989	95.3 ± 0.6	0.988	97.0 ± 0.3	0.993
	$MMCM_s$	76.0 ± 0.7	0.834	93.6 ± 0.5	0.981	95.7 ± 0.7	0.989	95.3 ± 0.6	0.988	97.0 ± 0.3	0.993
	CSR	71.5 ± 0.9	0.774	93.9 ± 0.4	0.983	96.2 ± 0.6	0.992	95.5 ± 0.5	0.990	97.5 ± 0.4	0.995
	KCSR	77.1 ± 1.1	0.847	93.7 ± 0.7	0.981	95.8 ± 0.4	0.989	94.3 ± 0.6	0.985	96.2 ± 0.3	0.990
	CCA	66.6 ± 1.4	0.720	85.5 ± 0.7	0.926	89.9 ± 0.8	0.962	91.3 ± 0.7	0.969	94.7 ± 0.5	0.985
	KCCA	69.0 ± 1.0	0.748	73.3 ± 0.7	0.794	91.2 ± 0.7	0.970	91.4 ± 0.9	0.969	90.4 ± 0.8	0.962
	CDFE	77.6 ± 0.6	0.854	86.4 ± 0.7	0.931	90.5 ± 0.7	0.964	91.3 ± 1.1	0.970	93.8 ± 0.6	0.982
	MvDA	70.0 ± 0.9	0.760	90.6 ± 0.7	0.965	93.6 ± 0.7	0.980	93.9 ± 0.8	0.982	95.4 ± 0.6	0.989
	MCM^2L	77.7 ± 0.8	0.855	94.2 ± 0.6	0.985	$\mathbf{96.4 \pm 0.5}$	**0.992**	96.5 ± 0.5	0.993	$\mathbf{97.8 \pm 0.2}$	**0.996**
	$KMCM^2L$	$\mathbf{79.2 \pm 0.7}$	**0.868**	$\mathbf{94.7 \pm 0.3}$	**0.988**	96.2 ± 0.4	0.992	$\mathbf{97.1 \pm 0.5}$	**0.995**	97.8 ± 0.4	0.996

results were obtained by $KMCM^2L$. The reason that MCM^2L and $KMCM^2L$ performed well attributes to that they take into account local information of focal samples in both same and different classes, while all the other six compared methods (excluding MMCM) are holistical-based without considering such local information. MMCM is the most related to MCM^2L. As it only sets samples of one modality as focal sample, in this experiment, we tested two settings. One sets high resolution modality as focal modality and the other sets low resolution modality as focal modality, respectively, denoted as $MMCM_h$ and $MMCM_l$ in Table II. As can be seen, the results of $MMCM_h$ and $MMCM_l$ are similar, with $MMCM_h$ slightly better. Both our MCM^2L and $KMCM^2L$ markedly improved the results of MMCM.

2) The seven methods compared performed similarly with CSR being the best. From both parts of Table II, the multimodality methods outperformed most of the single modality methods such as PCA, KPCA and LDA. But the results of NCA and LMNN are comparable with that of the multimodality methods. The main factor is that although NCA and LMNN are single-modality methods, they are local-based. This further illustrates the benefit of local and cross-modality-based metric learning.
3) Similarly, among all four kinds of features, SIFT is the best. Besides, using combined features will not always yield better verification results compared to using single features. Further research into why these features perform differently can provide a guidance on how to design specific feature extraction methods for heterogeneous face matching problem. Since feature extraction is imperative, it remains a focus of our future work.

Fig. 6(b) shows that the result of the proposed methods were markedly better than others. It is also shown that the results of KCCA and MvDA were among the worst and CSR the closest to the proposed methods.

E. Results on CUFSF Dataset

On the CUFSF dataset, the proposed methods were also compared with both single-modality and multimodality methods. The dataset was randomly divided into two parts of equal size. One part was used for training and the other for testing. This process was repeated for ten times and the average VRs and AUC values are presented in Table III.

1) As is shown, on all kinds of features, the best results were achieved by either MCM^2L or $KMCM^2L$. Besides, the standard deviations of the proposed methods were also much smaller. $KMCM^2L$ slightly improved the results of MCM^2L. However, even without $KMCM^2L$, MCM^2L achieved the best results compared with all the other compared methods. This shows that adopting cross-modality local information helps to gain separability among subjects and improves recognition performance.
2) Among the compared single-modality methods, KDA, NCA and LMNN achieved relatively comparable results to our methods. LMNN can be seen as the closest single modality method to MCM^2L. However, MCM^2L consistently outperformed LMNN on VR by 1.2%–2%. Among all the compared multimodal methods, CSR was relatively the best. However, the proposed methods outperformed CSR by 0.2%–6.2%. $MMCM_p$ and $MMCM_s$ are two versions of MMCM by setting photograph as focal modality and setting sketch as focal modality, respectively. The results of $MMCM_p$ and $MMCM_s$ were almost the same. MCM^2L and $KMCM^2L$ were better than both $MMCM_p$ and $MMCM_s$ among all the features.
3) On this dataset, the best verification results were achieved with all features. The results achieved by Gabor and SIFT were comparable. In fact, except for the gray features, all the other three single features performed well. The combined features thus contain a large number of useful features and this is perhaps the reason why on CUFSF dataset combining features improved the performances.

F. Results on CASIA NIR-VIS 2.0 Dataset

On the CASIA NIR-VIS 2.0 dataset, we followed strictly the evaluation protocol in [34]. On view 1, parameters were tuned. Methods were then tested on view 2 with results reported.

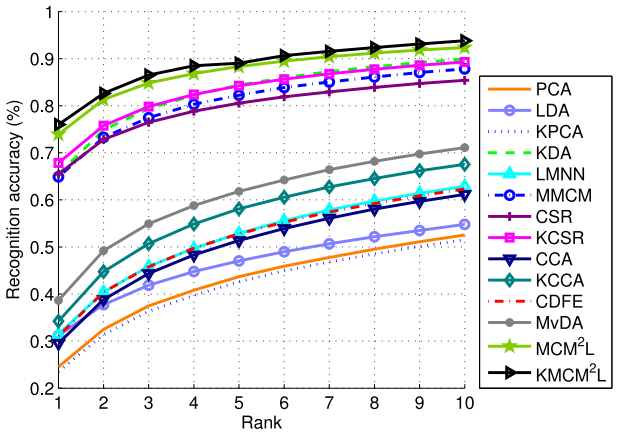


Fig. 7. CMC curves on CASIA NIR-VIS 2.0 dataset using SIFT features.

TABLE IV
COMPARISON OF RECOGNITION RATES OF MCM^{2}L AND KMCM^{2}L WITH STATE-OF-THE-ART METHODS ON CASIA NIR-VIS 2.0 DATASET

	Methods	Rank-1	Rank-10
Single-Modal	PCA	$24.5 \pm 0.8\%$	$52.5 \pm 1.1\%$
	LDA	$31.2 \pm 1.8\%$	$54.8 \pm 2.0\%$
	KPCA	$23.5 \pm 0.8\%$	$51.5 \pm 1.2\%$
	KDA	$65.5 \pm 1.4\%$	$90.0 \pm 0.9\%$
	LMNN	$31.4 \pm 1.8\%$	$62.9 \pm 2.7\%$
Multi-Modal	MMCM	$64.9 \pm 0.9\%$	$87.8 \pm 0.7\%$
	CSR	$65.5 \pm 1.2\%$	$85.4 \pm 0.8\%$
	KCSR	$67.8 \pm 0.9\%$	$89.3 \pm 0.6\%$
	CCA	$29.5 \pm 2.1\%$	$61.2 \pm 1.6\%$
	KCCA	$34.2 \pm 1.0\%$	$67.5 \pm 1.3\%$
	CDFE	$31.1 \pm 1.3\%$	$62.3 \pm 1.8\%$
	MvDA	$38.6 \pm 1.0\%$	$71.1 \pm 0.9\%$
	MCM^{2}L	$73.9 \pm 0.9\%$	$92.4 \pm 0.7\%$
	KMCM^{2}L	$\mathbf{76.0 \pm 0.7\%}$	$\mathbf{93.8 \pm 1.1\%}$
Other Methods	CDFL [14]	$71.5 \pm 1.4\%$	-
	Results of [46]	$78.46 \pm 1.67\%$	-
	Results of [15]	$\mathbf{86.16 \pm 0.98\%}$	-

Only SIFT features were used for face representation as SIFT were the best among all four kinds of features.

The CMC curves are depicted in Fig. 7. Table IV provides the rank-1 and rank-10 results on this dataset. As is shown, the proposed methods are the best among the state-of-the-art methods. On this dataset, as NCA was extremely computationally intensive, thus results of NCA were not obtained and included. The results of MMCM were obtained by setting near-infrared images as focal modality. Among the compared methods, KDA and KCSR were relatively better. The rank-1 rate of KDA was $65.5 \pm 1.4\%$ and KCSR $67.8 \pm 0.9\%$. The proposed MCM^{2}L method outperformed them by 8.4% and 6.1%, while KMCM^{2}L outperformed them by 10.5% and 8.2%.

Besides, the proposed methods are also compared with three other methods [14], [15], [46] that are not based on subspace or metric learning. The method proposed in [46] is image synthesis-based. The best result of this method is $78.46 \pm 1.67\%$. The results of KMCM^{2}L is slightly worse than this method. The other two methods are feature learning-based methods. The method in [14] achieved the rank-1 recognition rate of $71.5 \pm 1.4\%$ and in [15], the rank-1 recognition rate of $86.2 \pm 1.0\%$ was reported. While our results, $73.9 \pm 0.9\%$ and $76.0 \pm 0.7\%$ were better than that of [14] but worse than that of [15]. Despite of this, as the proposed methods are general metric learning methods, they can be combined with image synthesis-based methods and feature learning methods to take advantages of both. Due to that three methods [14], [15], [46] are not open sources, in the next subsection, we tested the proposed methods when combined with two publicly available deep features. On the CASIA NIR-VIS 2.0 dataset, KMCM^{2}L achieved a significant improvement on the rank-1 result of $96.5 \pm 0.4\%$ (see Section VII-G and Table VI for details).

TABLE V
RESULTS OF MCM^{2}L AND KMCM^{2}L COMBINED WITH DEEP FEATURES ON NJU-ID AND CUFSF DATASETS

	Feature	Methods	VR(%)	AUC
NJU-ID	VGG-Face	Euclidean	$87.3 \pm 2.7\%$	0.914
		PCA	$90.4 \pm 4.0\%$	0.950
		MCM^{2}L	$94.3 \pm 2.2\%$	0.982
		KMCM^{2}L	$96.3 \pm 2.3\%$	0.987
	WenECCV16	Euclidean	$95.5 \pm 2.7\%$	0.981
		PCA	$97.7 \pm 2.4\%$	0.993
		MCM^{2}L	$98.3 \pm 1.7\%$	**0.995**
		KMCM^{2}L	$\mathbf{98.5 \pm 2.0\%}$	**0.995**
CUFSF	VGG-Face	Euclidean	$86.4 \pm 0.8\%$	0.934
		PCA	$88.8 \pm 0.6\%$	0.955
		MCM^{2}L	$\mathbf{98.5 \pm 0.3\%}$	**0.998**
		KMCM^{2}L	$98.1 \pm 0.4\%$	**0.998**
	WenECCV16	Euclidean	$84.1 \pm 0.8\%$	0.918
		PCA	$87.4 \pm 0.8\%$	0.942
		MCM^{2}L	$94.7 \pm 0.6\%$	0.985
		KMCM^{2}L	$95.6 \pm 0.5\%$	0.991

TABLE VI
RESULTS OF MCM^{2}L AND KMCM^{2}L COMBINED WITH DEEP FEATURES ON CASIA NIR-VIS 2.0 DATASET

Feature	Methods	Rank-1	Rank-10
VGG-Face	Euclidean	$68.5 \pm 1.5\%$	$90.0 \pm 0.9\%$
	PCA	$74.1 \pm 1.4\%$	$94.9 \pm 0.6\%$
	MCM^{2}L	$92.2 \pm 1.2\%$	$97.3 \pm 0.8\%$
	KMCM^{2}L	$92.7 \pm 0.9\%$	$99.2 \pm 0.2\%$
WenECCV16	Euclidean	$84.9 \pm 1.6\%$	$97.6 \pm 0.5\%$
	PCA	$88.3 \pm 0.9\%$	$98.2 \pm 0.4\%$
	MCM^{2}L	$96.3 \pm 1.1\%$	$99.2 \pm 0.5\%$
	KMCM^{2}L	$\mathbf{96.5 \pm 0.4\%}$	$\mathbf{99.4 \pm 0.1\%}$

G. Results of the Proposed Methods Combined With Deep Features

The experiments in this section were to verify that the proposed methods are able to improve the performance of deep features. Two off-the-shelf deep models [47], [48] were used. The first is the VGG-Face [47] and the second is denoted as WenECCV16 [48] in Tables V and VI. To use these two deep features, the face alignment method was modified to be consistent with those used in [47] and [48]. Other experimental settings were the same to those in the previous sections.

As can be seen, the performances of deep learning-based features in Tables V and VI are much better than those of handcrafted features in Tables II–IV. In Tables V and VI, Euclidean denotes using Euclidean distance to directly measure the similarity of deep features. As can be seen, the results of Euclidean is the worst compared with PCA and our methods. The results of PCA are better than Euclidean but worse

than MCM^{2}L and KMCM^{2}L. This illustrates that the proposed methods are able to improve the performances of deep learning features and the improvement is clear. This is mainly because that these deep models are learned on data of single modality. There may be still some modality variations existing in extracted features. With the proposed methods, modality variations can be further removed, hence increasing the performances. Another observation is that the performance of the proposed methods while combined with deep features almost saturated on all the three datasets. For examples, the best VR results of the proposed methods on NJU-ID and CUFSF were $98.5 \pm 2.0\%$ and $98.5 \pm 0.3\%$; and the best rank-1 result on CASIA NIR-VIS 2.0 was $96.5 \pm 0.4\%$.

VIII. Conclusion

In this paper, an MCM^{2}L method and a KMCM^{2}L method are proposed for heterogeneous face recognition. The proposed cross-modality metric learning aims to minimize intrapersonal cross-modality distances and force a margin between person specific intrapersonal and interpersonal cross-modality distances. Compared with existing methods that use pairwise only constraints, the proposed methods add triplet-based constraints to allow focusing efficiently on optimizing the distances of those subjects whose intrapersonal and interpersonal cross-modality distances are hard to separate. Experimental results on three datasets demonstrate the effectiveness and superiority of the proposed methods.

Future work will include developing more specifically designed and deep learned features for specific heterogeneous face recognition, as it is evident that such features are beneficial [14], [15], [47], [48]. Besides, multimetric-based methods [17], [25] have achieved better results over single metric-based methods. Extending the proposed methods to multimetric-based will also be worth pursuing.

Acknowledgment

The authors would like to thank the Editors and the anonymous reviewers for their helpful and useful comments which helped improve this paper.

References

[1] J. Wang *et al.*, "Robust face recognition via adaptive sparse representation," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2368–2378, Dec. 2014.

[2] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Robust face recognition for uncontrolled pose and illumination changes," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 1, pp. 149–163, Jan. 2013.

[3] X. Tang and X. Wang, "Face sketch recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 50–57, Jan. 2004.

[4] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, 2005, pp. 1005–1010.

[5] J.-Y. Zhu, W.-S. Zheng, J.-H. Lai, and S. Z. Li, "Matching NIR face to VIS face using transduction," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 3, pp. 501–514, Mar. 2014.

[6] X. Huang, Z. Lei, M. Fan, X. Wang, and S. Z. Li, "Regularized discriminative spectral regression method for heterogeneous face matching," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 353–362, Jan. 2013.

[7] T. Bourlai, A. Ross, and A. K. Jain, "Restoring degraded face images: A case study in matching faxed, printed, and scanned photos," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 2, pp. 371–384, Jun. 2011.

[8] Y. Xu *et al.*, "Data uncertainty in face recognition," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1950–1961, Oct. 2014.

[9] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009.

[10] W. Zhang, X. Wang, and X. Tang, "Lighting and pose robust face sketch synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 420–433.

[11] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 2216–2223.

[12] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Z. Li, "Heterogeneous face recognition from local structures of normalized appearance," in *Proc. Int. Conf. Biometr.*, Alghero, Italy, 2009, pp. 209–218.

[13] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 513–520.

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

[15] D. Yi, Z. Lei, and S. Z. Li, "Shared representation learning for heterogenous face recognition," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, vol. 1. Ljubljana, Slovenia, May 2015, pp. 1–7.

[16] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1410–1422, Jun. 2013.

[17] D. Lin and X. Tang, "Inter-modality face recognition," in *Proc. Eur. Conf. Comput. Vis.*, Graz, Austria, 2006, pp. 13–26.

[18] Z. Lei and S. Z. Li, "Coupled spectral regression for matching heterogeneous faces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 1123–1128.

[19] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[20] A. Bellet, A. Habrard, and M. Sebban. (2013). *A Survey on Metric Learning for Feature Vectors and Structured Data*. [Online]. Available: https://arxiv.org/pdf/1306.6709.pdf

[21] Y. Mu, W. Ding, and D. Tao, "Local discriminative distance metrics ensemble learning," *Pattern Recognit.*, vol. 46, no. 8, pp. 2337–2349, 2013.

[22] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2002, pp. 505–512.

[23] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, May 2007.

[24] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2004, pp. 513–520.

[25] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Dec. 2009.

[26] S. Ying *et al.*, "Manifold preserving: An intrinsic approach for semisupervised distance metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2691005.

[27] A. Mignon and F. Jurie, "CMML: A new metric learning approach for cross modal matching," in *Proc. Asian Conf. Comput. Vis.*, 2012, p. 14.

[28] D. Zhai, H. Chang, S. Shan, X. Chen, and W. Gao, "Multiview metric learning with global consistency and local smoothness," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 1–22, 2012.

[29] P. Zhou, L. Du, M. Fan, and Y.-D. Shen, "An LLE based heterogeneous metric learning for cross-media retrieval," in *Proc. SIAM Int. Conf. Data Min.*, Vancouver, BC, Canada, 2015, pp. 64–72.

[30] C. Kang *et al.*, "Cross-modal similarity learning: A low rank bilinear formulation," in *Proc. Int. Conf. Inf. Knowl. Manag.*, Melbourne, VIC, Australia, 2015, pp. 1251–1260.

[31] S. Siena, V. N. Boddeti, and B. V. K. V. Kumar, "Maximum-margin coupled mappings for cross-domain matching," in *Proc. IEEE Int. Conf. Biometr. Theory Appl. Syst.*, Arlington, TX, USA, 2013, pp. 1–8.

[32] K. Q. Weinberger and L. K. Saul, "Fast solvers and efficient implementations for distance metric learning," in *Proc. Int. Conf. Mach. Learn.*, Helsinki, Finland, 2008, pp. 1160–1167.

[33] K. I. Kim, K. Jung, and H. J. Kim, "Face recognition using kernel principal component analysis," *IEEE Signal Process. Lett.*, vol. 9, no. 2, pp. 40–42, Feb. 2002.

[34] S. Z. Li, D. Yi, Z. Lei, and S. Liao, "The CASIA NIR-VIS 2.0 face database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Portland, OR, USA, 2013, pp. 348–353.
[35] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[36] H. J. Seo and P. Milanfar, "Face verification using the LARK representation," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 4, pp. 1275–1286, Dec. 2011.
[37] H. Wang, S. Z. Li, Y. Wang, and J. Zhang, "Self quotient image for face recognition," in *Proc. Int. Conf. Image Process.*, vol. 2. Singapore, 2004, pp. 1397–1400.
[38] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
[39] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.
[40] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[41] M. A. Turk and A. P. Pentland, "Face recognition using Eigenfaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Maui, HI, USA, 1991, pp. 586–591.
[42] D. Cai, X. He, and J. Han, "Speed up kernel discriminant analysis," *VLDB J.*, vol. 20, no. 1, pp. 21–33, 2011.
[43] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3–4, pp. 321–377, 1936.
[44] P. L. Lai and C. Fyfe, "Kernel and nonlinear canonical correlation analysis," *Int. J. Neural Syst.*, vol. 10, no. 5, pp. 365–377, 2000.
[45] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, 2012, pp. 808–821.
[46] F. Juefei-Xu, D. K. Pal, and M. Savvides, "NIR-VIS heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Boston, MA, USA, 2015, pp. 141–150.
[47] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–12.
[48] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 499–515.

Jing Huo received the B.Eng. degree in computer science and technology from Nanjing Normal University, Nanjing, China, in 2011. She is currently pursuing the Ph.D. degree with the Department of Computer Science, Nanjing University.

Her current research interests include machine learning, computer vision, metric learning, subspace learning, and their applications to heterogeneous face recognition.

Yang Gao (M'05) received the Ph.D. degree in computer software and theory from Nanjing University, Nanjing, China, in 2000.

He is a Professor with the Department of Computer Science and Technology, Nanjing University. He has published over 100 papers in top conferences and journals in and outside of China. His current research interests include artificial intelligence and machine learning.

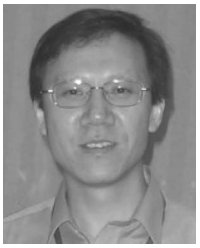
Yinghuan Shi received the B.Sc. and Ph.D. degrees from the Department of Computer Science, Nanjing University, Nanjing, China, in 2007 and 2013, respectively.

He is currently an Assistant Researcher with the Department of Computer Science and Technology, Nanjing University. He was a Visiting Scholar with the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, and the University of Technology Sydney, Ultimo, NSW, Australia. He has published over 40 research papers in related journals and conferences. His current research interests include computer vision and medical image analysis.

Dr. Shi serves as a Program Committee Member for several conferences, and a Referee for several journals.

Wanqi Yang received the Ph.D. degree from the Department of Computer Science and Technology, Nanjing University, Nanjing, China.

She is currently a Lecturer with the School of Computer Science and Technology, Nanjing Normal University, Nanjing. She has published several papers in top conferences and journals, such as the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, CVIU, and ACMMM. Her current research interests include multiview learning, feature selection, multimodal fusion, abnormal event detection, activity recognition, multiview feature selection, cross-view correlation analysis, and their applications to real-world problems in image/video analysis.



Hujun Yin (SM'03) received the Ph.D. degree in neural networks from the University of York, York, U.K., and the B.Eng. degree in electronic engineering and the M.Sc. degree in signal processing and from Southeast University, Nanjing, China.

He is a Senior Lecturer (Associate Professor) with the School of Electrical and Electronic Engineering, University of Manchester, Manchester, U.K. He has published over 150 peer-reviewed articles in a range of topics from density modeling, image processing, face recognition, text mining and knowledge management, gene expression analysis and peptide sequencing, novelty detection, to financial time series modeling and decoding neuronal responses. His current research interests include neural networks, self-organizing learning, image processing, face recognition, time series, and bio-/neuro-informatics.