Ensemble of Sparse Cross-Modal Metrics for Heterogeneous Face Recognition

Jing Huo[†], Yang Gao[†], Yinghuan Shi[†], Wanqi Yang[‡,†], Hujun Yin[§]
[†]State Key Laboratory for Novel Software Technology, Nanjing University, China
[‡]School of Computer Science and Technology, Nanjing Normal University, China
[§]School of Electrical and Electronic Engineering, The University of Manchester, UK
huojing1989@gmail.com, gaoy@nju.edu.cn, syh@nju.edu.cn
nju.yangwanqi@gmail.com, hujun.yin@manchester.ac.uk

ABSTRACT

Heterogeneous face recognition aims to identify or verify person identity by matching facial images of different modalities. In practice, it is known that its performance is highly influenced by modality inconsistency, appearance occlusions, illumination variations and expressions. In this paper, a new method named as ensemble of sparse cross-modal metrics is proposed for tackling these challenging issues. In particular, a weak sparse cross-modal metric learning method is firstly developed to measure distances between samples of two modalities. It learns to adjust rank-one cross-modal metrics to satisfy two sets of triplet based cross-modal distance constraints in a compact form. Meanwhile, a group based feature selection is performed to enforce that features in the same position of two modalities are selected simultaneously. By neglecting features that attribute to "noise" in the face regions (eye glasses, expressions and so on), the performance of learned weak metrics can be markedly improved. Finally, an ensemble framework is incorporated to combine the results of differently learned sparse metrics into a strong one. Extensive experiments on various face datasets demonstrate the benefit of such feature selection especially when heavy occlusions exist. The proposed ensemble metric learning has been shown superiority over several state-of-the-art methods in heterogeneous face recognition.

Keywords

Multi-modal learning; metric learning; feature selection; ensemble learning; heterogeneous face recognition

1. INTRODUCTION

Face recognition has been extensively researched over the past few decades and satisfactory performances have been achieved mostly under controlled environments. In practical situations, the problem often becomes heterogeneous where faces are of different modalities and performance degrades considerably. This is mainly caused by drastic appearance variations between face images of different modalities. There are several studies in heterogeneous face recognition in the literature such as matching sketches to photos, low resolution images to high resolution videos, visible light to near infrared images. Such situations are commonly encountered in multimedia applications. A general face recognition process often has two stages. The first is feature extraction and the second is learning a similarity measurement to measure similarities between the face features. Therefore, there are mainly two ways to alleviate modality variations for robust cross-modal face matching. One is to extract modality invariant face features and the other is to design cross-modal similarity measurements. This work focuses on the second one.

For learning a similarity measurement, metric learning is widely used [3]. However the currently widely used metric learning methods are Mahalanobis distance based. The main disadvantage when applied to cross-modal face recognition is that it is not able to remove modality variations, which influence recognition performance the most. Although there exist cross-modal metric learning methods [26], there are still serious issues to be addressed. One is that not all the face features are informative when faces are occluded or influenced by expressions or illumination variations. Thus it is beneficial to perform feature selection. The existing sparse Mahalanobis metric learning methods [35] are able to do feature selection, but it is hard to interpret which features are more important from the parameters learned. Besides, for face recognition, as features are extracted on local patches, it is beneficial to take the implicit spatial information of features into consideration with feature selection.

We therefore propose to learn a cross-modal metric that is able to remove modality variations and persist the ability of informative feature selection. Different with most sparse metric learning methods, the proposed method also has good interpretability. A simple cross-modal metric is firstly defined with two sets of parameters. They can be interpreted as two projection vectors to project face features of two different modalities into a common subspace and the distance is measured in the new subspace. Learning such a cross-modal metric is efficient as it has much fewer parameters compared with metrics parameterized by matrices. The goal of metric learning is to adjust the metric to satisfy two sets of triplet based cross-modal distance constraints. Each triplet includes a sample of one modality and a pair of samples from the other modality with one being the same label and the other of different label. From the triplet, one same labeled and one differently labeled cross-modal sample pairs can be constructed. The goal is to make the distances of the same labeled pairs smaller than that of the differently labeled pairs. Besides, another characteristic of the cross-modal metric is that it is learned by incorporating group based feature selection. Two kinds of

MM '16, October 15-19, 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2964284.2964311

groups are defined. One is single feature based group that takes two elements respectively from two projection vectors in the same position. The other is patch based group and is defined as two sets of elements from the two projection vectors corresponding to features extracted from patches of the same position. By using group based feature selection, relations of features of two modalities are incorporated. Such a learned metric is named as sparse cross-modal metric (SPAC). Once learned, a SPAC selects a set of features to measure the distances between samples.

Since one SPAC has few parameters, it may not fit well for samples of all kinds of variations. A simple idea is to learn groups of SPACs that are able to deal with different kinds of variations and are complementary. Therefore, an ensemble scheme is further applied. A set of SPACs are learned on differently weighted triplets with different variations. Once a SPAC is learned, the weights of triplets are reassigned according to whether the constraint encoded in a triplet is met by the previously learned SPACs. The learned SPACs are therefore complementary. Each SPAC is also assigned a weight when the outputs of all the SPACs are combined to make the final decision. The reweighting and combining scheme of the Adaboost framework [28] is used to learn SPACs and combine them into a strong metric. The ensemble framework of SPACs is termed as ESPAC.

2. RELATED WORK

Three related topics are reviewed, heterogeneous face recognition, single-modal metric learning and cross-modal metric learning.

2.1 Heterogeneous Face Recognition

Methods for solving heterogeneous face recognition problems can be categorized into three categories. The first category is the synthesis based methods which map data of one modality into another by synthetic methods. Related work includes synthesizing sketches from photos and then comparing synthesized images with sketches drawn by artists [31, 36]. One drawback of this kind of methods is that different synthetic methods have to be used if the modalities change.

The second category is modality invariant feature extraction based methods. In [34], Gabor features together with restricted boltzmann machines were used to learn shared representations. Jin *et al.* [12] proposed to learn a set of image filters to reduce the appearance differences of cross-modality images by maximizing inter-class variations and minimizing intra-class variations.

The last is the common subspace based methods [20, 22, 13] and the proposed method belong to this kind. Data of different modalities are mapped into a new, common subspace, so that data of different modalities become comparable. In [22], a common discriminant feature extraction (CDFE) was proposed to learn a common subspace for samples of two modalities to attain both intra-class compactness and the inter-class dispersion. In [20], Lei *et al.* proposed a spectral regression based method (CSR) to learn a discriminative subspace. The objective is similar to linear discriminant analysis (LDA) [2]. These methods all learn common subspaces to remove modality variations. However, none of them have considered the cases where there are also occlusions, expressions and illumination changes, which widely exist in real-world heterogeneous face data.

2.2 Single-modal Metric Learning

Metric learning for single modal data has been extensively studied over the past decade. The existing approaches can be mainly grouped into two categories; Mahalanobis metric and bilinear metric based. A comprehensive review can be found in [3]. We herein mainly review related work regarding ensemble based metric learning and sparse metric learning. For ensemble based methods, Shen *et al.* proposed a PSDBoost method [30] to learn a positive semidefinite (PSD) matrix by column generation for Mahalanobis metrics. The method was later extended into a more efficient version in [29]. The main idea of the method is that a PSD matrix, the parameter of Mahalanobis based metric, can be decomposed into a linear combination of trace-one rank-one matrices. Therefore, rank-one PSD matrices parameterized weak metrics are learned sequentially. In [4], Bi *et al.* proposed to learn Mahalanobis metrics in an Adaboost manner. The same decomposition property of PSD matrix [30] was used and weak metrics were also parameterized by rank-one PSD matrices. Another boosting framework was proposed in [33] for combining weak metrics with binary outputs. These existing ensemble based metric learning methods are designed for single modal data.

For sparse metric learning, there have been several attempts to learn sparse low-rank metric matrices [35]. Such settings are related to finding low dimensional subspaces instead of finding informative features. For finding informative features and feature relationships, Liu *et al.* [23] proposed to learn a sparse bilinear similarity function by decomposing the PSD matrix as a combination of rank-one sparse PSD matrices. The rank-one sparse PSD matrices were defined with certain structures with only four non-zero elements related to a pair of features. Such rank-one matrices were greedily added to the final similarity function during learning. In [1], a PSD matrix was decomposed into a diagonal matrix and a set of all-zeros diagonal matrices with only one row and one column have non-zero elements. The learning procedure lead to many of the all-zeros diagonal matrices to zeros and the final combined matrix is thus sparse. Our proposed feature selection scheme differs from these methods in that sparsity is forced not on the metric matrix but on projection vectors. As the projection vectors directly relate to the weights of features, the proposed feature selection scheme thus has better interpretability on which features being more important. Besides, the proposed feature selection scheme differs from these methods in that feature selection is done jointly with the modality relations of features and the structures of face features considered.

2.3 Cross-modal Metric Learning

With the development of multi-modal data, there has been an increase in the study of cross-modal metric learning. Mignon *et al.* [26] proposed a common subspace based cross-modal metric learning method to satisfy both the pairwise similar/dissimilar cross-modal distance constraints. In [38], Zhou *et al.* proposed to incorporate both homogeneous local information and heterogeneous constraints into a whole framework to learn a cross-modal metric by using an extension of locally linear embedding (LLE). A low-rank bilinear cross-modal similarity method is proposed in [14]. Our method differs with these methods in several aspects. One is the distance constraints used. These methods use pairwise cross-modal distance constraints, while triplet based distance constraints are used in our proposed methods. The second is that the proposed metric is learned in a boosting manner with each weak metric focusing on different groups of features thus eliminating noise features for different groups of inseparable triplets.

3. FRAMEWORK OF ESPAC

Fig. 1 shows the framework of the proposed method. In the training stage, face alignment and face feature extraction are firstly performed on the training face images. Local feature extraction

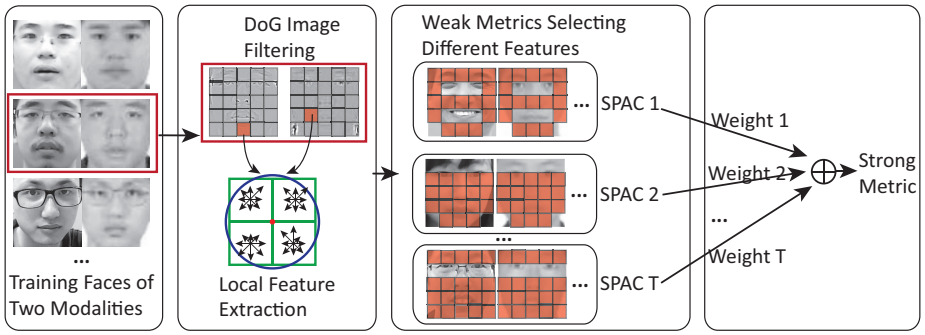


Figure 1: Framework of ESPAC. Local based feature extraction is firstly performed on given training data. Then a set of SPACs together with their weights are learned using an ensemble learning framework with each focusing on different features to measure cross-modal distances. The final output of ESPAC is the sum of weighted outputs of all the learned SPACs.

method is adopted. For faces of different modalities, same feature extraction process is applied. The extracted face features can be highly influenced by modality differences.

Then a set of weak sparse cross-modal metrics (SPACs) are learned. There are several properties of the learned SPACs. 1) The cross-modal metric is able to remove modality variations in the extracted features. 2) Each SPAC is learned with the weights of triplet based constraints differently so that the cross-modal metrics are complementary. The reweighting scheme of the Adaboost framework is used to assign different weights to different triplets. Bisection method is used to find the weight of a SPAC (or the contribution of a SPAC) to the final strong metric by minimizing the training error of the final strong metric. 3) For different appearance variations existing in different triplets, a group based feature selection scheme is used while learning SPACs. Two kinds of group based feature selection methods are proposed. In the first setting, two features of two modalities at the same position are simultaneously selected. In the second setting, features extracted in the patches of same position of two modalities are forced to be selected together. We prove that the two feature selection schemes achieve better results compared with no feature selection based one. Specifically, the second one can highly improve recognition performances in high occlusion cases.

In the testing stage, the same feature extraction process is applied. The outputs of all the learned sparse cross-modal metrics are weighted and combined to make the final prediction.

4. LOCAL FACE FEATURE EXTRACTION

For face normalization, faces are firstly rotated so that two eyes are located on a horizontal line, and then resized to make the distances between two pupils of 75 pixels. A face region of 160×160 is 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 adopted is image filtering techniques (difference of gaussian, DoG) to help compensate illumination variations and also to reduce the variations caused by modality difference [16].

The last step is to extract features of these filtered face images. Each face image is firstly partitioned into patches. As is shown in Fig. 2, the face image is partitioned into 5×5 patches, with each patch of size 32×32. In the experiments, different partition configurations are tested. The 160×160 face images are also partitioned into 3×3, 4×4, 5×5, 6×6 and 7×7. For some partition settings, patches may be of different size, but the difference is only one pixel in row size or column size. In the final experiments, we used the setting of partition images into 6×6 which produces relative good performance. After partition, local feature extraction methods are then used to extract features on each patch, in our case, scale-invariant feature transform (SIFT) [9] is adopted following an empirical comparison with other features. By concatenating all the local features into a whole vector, the final face feature is obtained. In the final feature vector, the features therefore have implicit spatial positions and patch-based structures.

5. PROBLEM FORMULATION OF SPAC

5.1 Notations

Suppose the training face image sets of two modalities are denoted as $\mathcal{X} = \{(\mathbf{x}_i, l_i^x)|i = 1, 2, ..., N_x\}$ and $\mathcal{Y} = \{(\mathbf{y}_i, l_i^y)|i = 1, 2, ..., N_y\}$, where $\mathbf{x}_i \in \mathbb{R}^{d_x}$ 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. In our case, as same feature extraction process is applied to both modalities, $d_x = d_y$. Suppose there are n different persons within the training set. $l_i^x \in \{1, 2, ..., n\}$ is the label of the sample $\mathbf{x}_i$ and $l_i^y \in \{1, 2, ..., n\}$ is the label of the sample $\mathbf{y}_i$. N_x and N_y are sample sizes of two modalities respectively.

To learn a metric on the given training datasets, a set of distance constraints should be constructed. Relative distance constraints encoded in triplets are used in the proposed method. Different with metric learning for single modality dataset, there are two kinds of triples for datasets of two modalities. One is of the form $(\mathbf{x}_i, \mathbf{y}_j, \mathbf{y}_k)$, with $l_i^x = l_j^y$ and $l_i^x \neq l_k^y$. The constraint is to force $d(\mathbf{x}_i, \mathbf{y}_j) < d(\mathbf{x}_i, \mathbf{y}_k)$. It can be constructed by setting every sample of the first modality as a focal sample. Then the second sample and the third sample are selected from samples of the second modality with the second one of the same label and the third one of different label. Another type of triplet is of the form $(\mathbf{y}_i, \mathbf{x}_j, \mathbf{x}_k)$ with $l_i^y = l_j^x$ and $l_i^y \neq l_k^x$. And it can be constructed similarly by setting the focal sample of the second modality. Suppose two sets of triplets $\mathcal{T}_1 = \{(\mathbf{x}_i, \mathbf{y}_j, \mathbf{y}_k)|l_i^x = l_j^y, l_i^x \neq l_k^y\}$ and $\mathcal{T}_2 = \{(\mathbf{y}_i, \mathbf{x}_j, \mathbf{x}_k)|l_i^y = l_j^x, l_i^y \neq l_k^x\}$ are constructed. The goal is to learn a cross-modal metric which meets the constraints defined in the two sets of triplets.

5.2 Sparse Cross-Modal Metric Learning

Suppose $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{y} \in \mathbb{R}^{d_y}$ are the face features of two modalities. The cross-modal metric is defined as:

$$d(\mathbf{x}, \mathbf{y}) = [f(\mathbf{x}) - g(\mathbf{y})]^2. \quad (1)$$

In case where both $f(\mathbf{x})$ and $g(\mathbf{y})$ are linear functions, they can be defined as $f(\mathbf{x}) = \mathbf{w}_x^\top \mathbf{x}$ and $g(\mathbf{y}) = \mathbf{w}_y^\top \mathbf{y}$. $\mathbf{w}_x \in \mathbb{R}^{d_x}$ and $\mathbf{w}_y \in \mathbb{R}^{d_y}$ are parameters. Such a metric can be seen as projecting samples into a one dimensional common subspace and then measuring their Euclidean distance. As the number of parameters is small, such metric is weak. The ultimate goal of the proposed method is to combine a set of these weak metrics into a strong one. The ensemble scheme will be illustrated later. We firstly present how to learn a weak sparse metric.

As being mentioned previously, the goal is to learn week metrics that satisfy the two sets of triplet based constraints. For the first set of triplets, the goal is to maximize the value of $d(\mathbf{x}_i, \mathbf{y}_k) - d(\mathbf{x}_i, \mathbf{y}_j)$, which can be represented as:

$$\rho_r = d(\mathbf{x}_i, \mathbf{y}_k) - d(\mathbf{x}_i, \mathbf{y}_j) = (\mathbf{w}^\top \mathbf{z}_r^d)^2 - (\mathbf{w}^\top \mathbf{z}_r^s)^2, \quad (2)$$

where $\mathbf{w} = [\mathbf{w}_x; \mathbf{w}_y] \in \mathbb{R}^{d_x+d_y}$ is the concatenation of the two parameters and r is the index of the triplet in $\mathcal{T}_1$. $\mathbf{z}_r^s = [\mathbf{x}_i; -\mathbf{y}_j] \in \mathbb{R}^{d_x+d_y}$ is formed by concatenating the same labeled sample pair of the rth triplet into one. $\mathbf{z}_r^d = [\mathbf{x}_i; -\mathbf{y}_k] \in \mathbb{R}^{d_x+d_y}$ is formed using the differently labeled sample pair. Similarly, for triplets in the second set, $d(\mathbf{x}_k, \mathbf{y}_i) - d(\mathbf{x}_j, \mathbf{y}_i) = (\mathbf{w}^\top \mathbf{z}_r^d)^2 - (\mathbf{w}^\top \mathbf{z}_r^s)^2$ can be represented in the same form, with $\mathbf{z}_r^s = [\mathbf{x}_j; -\mathbf{y}_i]$ and $\mathbf{z}_r^d = [\mathbf{x}_k; -\mathbf{y}_i]$. By using the above formulation and combining the two sets into one, the loss of a weak metric on the defined triplets can be represented as:

$$J = \sum_{r=1}^{|\mathcal{T}_1|+|\mathcal{T}_2|} \ell(\rho_r), \quad (3)$$

where $\ell(a)$ is a convex loss function which is selected as exponential loss $\exp(-a)$.

In weak cross-modal metric learning, merely minimizing the loss defined in Eq. (3) will suffer from server overfitting problem and regularization terms need to be added. In practice, the final objective of the no feature selection based weak cross-modal metric learning is defined as minimizing the following function:

$$\min_{\mathbf{w}} \sum_{r=1}^{|\mathcal{T}_1|+|\mathcal{T}_2|} \ell(\rho_r) + \tau \sum_{k=1}^{|\mathcal{S}|} \delta_k + \frac{\eta}{2} \mathbf{w}^\top L \mathbf{w} + \frac{\lambda}{2} \|\mathbf{w}\|^2, \quad (4)$$

where τ, η and λ are regularization parameters. $\mathbf{w}$ is constrained to be unit vector.

The first term in Eq. (4) is to minimize the loss defined in Eq. (3). The second term is a regularization term to control the distances of same labeled cross-modal pairs to be small, with $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_j) | l_i^x = l_j^y\}$ the set of same labeled cross-modal sample pairs. $\delta_k = (\mathbf{w}^\top \mathbf{z}_k)^2$ is the distance of the kth same labeled cross-modal sample pair, $\mathbf{z}_k = [\mathbf{x}_i; -\mathbf{y}_j] \in \mathbb{R}^{d_x+d_y}$ is the concatenation of the kth same labeled cross-modal pair in set $\mathcal{S}$. Thus $\sum_{k=1}^{|\mathcal{S}|} \delta_k$ is the sum of all the same labeled cross-modal sample distances. This regularization term is used; because during the learning to force the distances of differently labeled pairs to be larger than the distances of the same labeled pairs, the distances of both the same labeled pairs and the differently labeled pairs may increase together, leading to poor generalization ability. The third term in Eq. (4) is also a regularization term. $L = D - S$ is a Laplacian matrix, with $S = \begin{bmatrix} \mathbf{0} & I \\ I & \mathbf{0} \end{bmatrix}$, $\mathbf{0} \in \mathbb{R}^{d \times d}$ is a zero matrix and $I \in \mathbb{R}^{d \times d}$ is an identity matrix, where d equals to d_x or d_y. D is a diagonal matrix whose elements are column sums of S. This term is equivalent to $\|\mathbf{w}_x - \mathbf{w}_y\|_2^2$ and the objective is to penalize $\mathbf{w}_x$ and $\mathbf{w}_y$ from differing too much. This is based on the prior knowledge that face features at the same position of two modalities are correlated, so the parameters of $\mathbf{w}_x$ and $\mathbf{w}_y$ at the same position should also be correlated. The last regularization term is related to control the l_2 norm of $\mathbf{w}$ to avoid overfitting.

Group based Feature Selections. By using non-sparse projection vectors $\mathbf{w}_x$ and $\mathbf{w}_y$, all the extracted face features of two modalities are used for measuring distances. However, due to expression differences, high occlusions and face images captured under different illumination environments, not all patches of face images are useful for identifying the identities of face images. For cross modal face matching, once a feature in the first modality is selected, the corresponding feature of the same position of the second modality should also be selected. Therefore, group based feature selection is incorporated to select face features that are useful for distinguishing persons' identities. As the values in $\mathbf{w}_x$ and $\mathbf{w}_y$ indicate the importance of face features, group sparsity is forced on $\mathbf{w}_x$ and $\mathbf{w}_y$ to perform feature selection. Two schemes are proposed. An illustration is shown in Fig. 2. The first one is single feature based groups. In this case, a group contains two elements of $\mathbf{w}$ with one element corresponds to $\mathbf{w}_x$ and the other corresponds to $\mathbf{w}_y$. Use $G_i = \{G_i^x, G_i^y\}, i = 1, 2, ..., d$ to denote the set of indices within $\mathbf{w}$ corresponding to the ith features of two modalities. Therefore, there is a total number of d single feature based groups. In the second case, as for faces of heavy occlusions or with expression changes, some patches of the face images are completely unusable. We therefore propose patch based groups. The ith group is defined as the elements of both $\mathbf{w}_x$ and $\mathbf{w}_y$ corresponding to the features of the ith patch. $G_i = \{G_i^x, G_i^y\}, i = 1, 2, ..., p$ is used interchangeably to denote the set of indices within $\mathbf{w}$ which correspond to features of ith patches of two modalities. p is the total number of patches. With such a definition, the learned cross-modal metric uses the features selected from either the same positions or the same face regions of two cross-modal face images, leading to the following final objective function to learn a SPAC:

$$\begin{aligned} \min_{\mathbf{w}} & \sum_{r=1}^{|\mathcal{T}_1|+|\mathcal{T}_2|} \ell(\rho_r) + \gamma \sum_{i=1}^{N_G} \|\mathbf{w}_{G_i}\|^2 \\ & + \tau \sum_{k=1}^{|\mathcal{S}|} \delta_k + \frac{\eta}{2} \mathbf{w}^\top L \mathbf{w} + \frac{\lambda}{2} \|\mathbf{w}\|^2, \end{aligned} \quad (5)$$

where γ is a balance parameter and N_G is the number of groups which equals to either d or p. The objective function defined in Eq. (5) balances satisfying distance constraints and group based feature selection. Once a SPAC is learned, it selects a few face features or face regions for measuring cross-modal distances.

5.3 Optimization

As there is a group sparsity term in Eq. (5), alternating direction method of multipliers (ADMM) [5] is used to efficiently minimize the objective function. We firstly transform the minimization problem of Eq. (5) into an equivalent form:

$$\min_{\mathbf{w},\mathbf{v}} F(\mathbf{w}) + H(\mathbf{v}) \quad \text{s.t. } \mathbf{w} = \mathbf{v}, \quad (6)$$

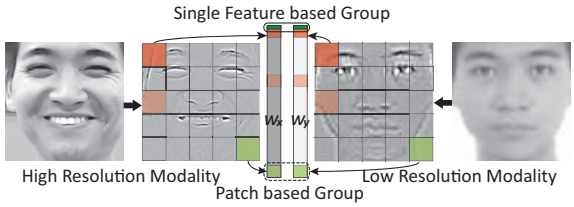


Figure 2: Illustration of two kinds of groups. A single feature based group contains two elements with one from $\mathbf{w}_x$ and one from $\mathbf{w}_y$ which are of the same position. Patch based group is defined as two groups of elements of $\mathbf{w}_x$ and $\mathbf{w}_y$ corresponding to features of two modalities extracted in patches of same position.

where

$$F(\mathbf{w}) = \sum_{r=1}^{|\mathcal{T}_1|+|\mathcal{T}_2|} \ell(\rho_r) + \tau \sum_{k=1}^{|\mathcal{S}|} \delta_k + \frac{\eta}{2}\mathbf{w}^\top L\mathbf{w} + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

$$H(\mathbf{v}) = \gamma \sum_{i=1}^{p} \|\mathbf{v}_{G_i}\|^2.$$

The augmented Lagrangian for Eq. (6) is:

$$\mathcal{L}_\varrho = F(\mathbf{w}) + H(\mathbf{v}) + \nu^\top(\mathbf{w}-\mathbf{v}) + \frac{\varrho}{2}\|\mathbf{w}-\mathbf{v}\|^2, \quad (7)$$

where ϱ is a scale parameter and ν is the lagrange multiplier. The ADMM optimization procedure for the above problem is therefore as follows:

$$\begin{aligned} \mathbf{w}^{k+1} &:= \arg\min_{\mathbf{w}} F(\mathbf{w}) + \frac{\varrho}{2}\|\mathbf{w}-\mathbf{v}^k+\mathbf{u}^k\|^2 \\ \mathbf{v}_{G_i}^{k+1} &:= S_{\frac{\gamma}{\varrho}}(\mathbf{w}_{G_i}^{k+1}+\mathbf{u}_{G_i}^k) \\ \mathbf{u}^{k+1} &:= \mathbf{u}^k + \mathbf{w}^{k+1} - \mathbf{v}^{k+1}, \end{aligned} \quad (8)$$

where $\mathbf{u} = \frac{1}{\varrho}\nu$. S_κ is a vector soft thresholding operator which is defined in [5]. For the first minimization problem to find optimal $\mathbf{w}$, a solution is to use the gradient descent. However, as the gradient calculation of the term $\sum_{r=1}^{|\mathcal{T}_1|+|\mathcal{T}_2|}\ell(\rho_r)$ with respect to $\mathbf{w}$ is very time consuming. A modified scheme proposed in [25] is adopted. Then, the gradient term is of the form $\begin{bmatrix} \mathbf{X}\mathbf{A}\mathbf{X}^T & \mathbf{X}\mathbf{B}\mathbf{Y}^T \\ \mathbf{Y}\mathbf{B}^T\mathbf{X}^T & \mathbf{Y}\mathbf{C}\mathbf{Y}^T \end{bmatrix}\mathbf{w}$, where $\mathbf{X} \in \mathbb{R}^{d_x \times N_x}$ and $\mathbf{Y} \in \mathbb{R}^{d_y \times N_y}$ are matrices of samples of two modalities. By first calculating $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, the gradient can be efficiently updated. A summary of the algorithm is given in Algorithm 1.

Algorithm 1 ADMM for Learning Sparse Cross-Modal Metric (SPAC)

1: Two sets of constructed triplets $\mathcal{T}_1$ and $\mathcal{T}_2$.
2: Initialize $\mathbf{w}^1$, $\mathbf{v}^1$ and $\mathbf{u}^1$.
3: Initialize $k = 1$.
4: **while** not converged **do**
5: $\quad \mathbf{w}^{k+1} := \arg\min_{\mathbf{w}} F(\mathbf{w}) + \frac{\varrho}{2}\|\mathbf{w}-\mathbf{v}^k+\mathbf{u}^k\|^2$.
6: $\quad \mathbf{v}_{G_i}^{k+1} := S_{\frac{\gamma}{\varrho}}(\mathbf{w}_{G_i}^{k+1}+\mathbf{u}_{G_i}^k)$.
7: $\quad \mathbf{u}^{k+1} := \mathbf{u}^k + \mathbf{w}^{k+1} - \mathbf{v}^{k+1}$.
8: $\quad k = k+1$.
9: **end while**
10: Output $\mathbf{w}$.

6. ENSEMBLE OF SPACS

An ensemble framework is used to further combine a set of weak SPACs into a strong one. Suppose the final strong metric is represented as $D(\mathbf{x},\mathbf{y}) = \sum_{t=1}^{T}\alpha_t d_t(\mathbf{x},\mathbf{y})$, where α_t is the coefficient of the tth SPAC. Then the subtraction of the distances of a same labeled pair and a differently labeled pair measured by the strong metric is $D(\mathbf{x}_i,\mathbf{y}_k) - D(\mathbf{x}_i,\mathbf{y}_j) = \sum_{t=1}^{T}\alpha_t\rho_r^t$, where $\rho_r^t = d_t(\mathbf{x}_i,\mathbf{y}_k) - d_t(\mathbf{x}_i,\mathbf{y}_j)$. Therefore the objective of the final strong metric is to minimize the following loss:

$$J = \sum_{r=1}^{|\mathcal{T}_1|+|\mathcal{T}_2|}\left[s_r\ell\left(\sum_{t=1}^{T}\alpha_t\rho_r^t\right)\right], \quad (9)$$

where s is a distribution over the two sets of triplets. s_r is the weight of the rth triplet and is initialized as $1/(|\mathcal{T}_1|+|\mathcal{T}_2|)$. The ensemble scheme is summarized in Algorithm 2.

Theorem 1. *With Z_t defined in step 7 of Algorithm 2, the following equation holds for the initial distribution s:*

$$J = \sum_{r=1}^{|\mathcal{T}_1|+|\mathcal{T}_2|}\left[s_r\exp\left(-\sum_{t=1}^{T}\alpha_t\rho_r^t\right)\right] = \prod_{t=1}^{T}Z_t.$$

Proof: According to the definition of distribution:

$$s_r^{T+1} = \frac{s_r\exp(-\sum_{t=1}^{T}\alpha_t\rho_r^t)}{\prod_{t=1}^{T}Z_t}.$$

Therefore

$$\begin{aligned} J &= \sum_{r=1}^{|\mathcal{T}_1|+|\mathcal{T}_2|}\left[s_r\exp\left(-\sum_{t=1}^{T}\alpha_t\rho_r^t\right)\right] \\ &= \sum_{r=1}^{|\mathcal{T}_1|+|\mathcal{T}_2|} s_r^{T+1}\prod_{t=1}^{T}Z_t \\ &= \prod_{t=1}^{T}Z_t. \end{aligned}$$

Therefore, minimizing J is equivalent to minimizing Z_t at each iteration. The step 5 in Algorithm 1 for learning a SPAC is therefore changed slightly in the ESPAC framework. The first term in SPAC's objective function Eq. (5) is changed to Z_t, a weighted loss, so as to meet the objective of ESPAC. This leads to $F(\mathbf{w})$ in Eq. (6) becoming the following form:

$$F(\mathbf{w}) = \sum_{r=1}^{|\mathcal{T}_1|+|\mathcal{T}_2|} s_r\ell(\rho_r) + \tau\sum_{k=1}^{|\mathcal{S}|}\delta_k + \frac{\eta}{2}\mathbf{w}^\top L\mathbf{w} + \frac{\lambda}{2}\|\mathbf{w}\|^2. \quad (10)$$

However the optimization procedure of Algorithm 1 remains unchanged. The only change is $F(\mathbf{w})$ in step 5 of Algorithm 1 is changed to Eq. (10) and it can be optimized in the same way.

The weight of a weak metric α_t (step 6 of Algorithm 2) is calculated by taking the derivative of Z_t with respect to α_t and using bisection search for the value that meets $\frac{\partial Z_t}{\partial \alpha_t} = 0$.

7. EXPERIMENTAL RESULTS

Experimental results on four face recognition datasets are presented. The AR dataset [24] was used for an illustration of the effect of patch-based feature selection. As AR is a single modal dataset, comparisons of recognition performance with two single-modal boosting based metric learning methods are also given. The other three datasets are cross-modal based. The first is

Algorithm 2 Ensemble of Sparse Cross-Modal Metrics (ESPAC)

1: Construct two sets of triplets $\mathcal{T}_1$ and $\mathcal{T}_2$.
2: Initialize $s_r^1 = 1/(|\mathcal{T}_1| + |\mathcal{T}_2|), r = 1, 2, ..., |\mathcal{T}_1| + |\mathcal{T}_2|$.
3: Initialize $t = 1$.
4: **while** not converged **do**
5: Train a weak metric $d_t(\mathbf{x}, \mathbf{y})$ using distribution s^t by Algorithm 1.
6: Choose the weight of the weak metric α_t using bisection.
7: Update the distribution using:

$$s_r^{t+1} = \frac{s_r^t \exp\left(-\alpha_t \rho_r^t\right)}{Z_t}$$

where $Z_t = \sum_{r=1}^{|\mathcal{T}_1|+|\mathcal{T}_2|} s_r^t \exp\left(-\alpha_t \rho_r^t\right)$ is a normalization factor such that s^{t+1} is a distribution.
8: $t = t + 1$.
9: **end while**
10: Output $D(\mathbf{x}, \mathbf{y}) = \sum_t \alpha_t d_t(\mathbf{x}, \mathbf{y})$.

a newly collected Chinese resident identity card face dataset (NJU-ID) which was used for evaluating the performance of matching high resolution faces to low resolution faces. CUHK face sketch FERET (CUFSF) dataset [31, 37] was also used to evaluate the performance of recognizing sketches from photos. Whilst the CASIA NIR-VIS 2.0 dataset [21] was evaluated for matching visible light images (VIS) to near infrared images (NIR).

7.1 Datasets and Evaluation Protocols

AR Dataset: The AR dataset consists of over 4000 facial images with different variations, including facial expressions, illumination and occlusion by sunglasses or scarf. Two subsets of the dataset with occlusion variations of sunglasses and scarf were used in experiments, denoted as AR-Sunglasses and AR-Scarf in the following experiments. In total, the two subsets are from 111 subjects with each having 4 images, two are frontal images with no other variations and the two images are from two sessions. The other two are face images with occlusions of sunglasses or scarf. Example images are given in Fig. 3. For result reporting, both AR-Scarf and AR-Sunglasses were randomly divided into four parts for ten times. Each time, three parts were used for training and the remaining part for testing.

NJU-ID Dataset: The NJU-ID dataset contains images of 256 persons. For each person, there are one low resolution NJU-ID card image and fifty images collected from a high resolution digital camera. Only randomly chosen one of the fifty high resolution images was used in the experiment (Available from: http://cs.nju.edu.cn/rl/Data.html). The resident identity card image is of resolution 102×126 and the high resolution image is of resolution 640×480. Exemplar pairs from the dataset are shown in Fig. 7. To report results on this dataset, we randomly divided the dataset into 10 folds according to identity information and 10-fold cross-validation was used.

CUHK Face Sketch FERET Dataset (CUFSF): The CUFSF dataset was used for photo to sketch face recognition. It includes 1194 persons from the FERET dataset [27]. Samples are shown in Fig. 8. For each person, there are one photo and one sketch drawn by an artist after viewing the photo. 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.

The CASIA NIR-VIS 2.0 Dataset: The CASIA NIR-VIS 2.0 dataset was used for evaluating the VIS-NIR face recognition. Example images in Fig. 9. It contains 725 subjects. We followed the same evaluation protocol on this dataset by [21]. The dataset was divided into two views. View 1 was used for parameter tuning and view 2 for reporting results.

Figure 3: Examples of aligned face images of AR-Sunglasses and AR-Scarf. The first row are face images of AR-Sunglasses of two persons with each having four images. The second row are from AR-Scarf.

On all the four datasets, the feature extraction method described in Section 4 was applied. To maintain a processable number of features for our method and the compared methods, all the extracted local features were applied with principal component analysis (PCA) [2] locally with the features of a patch of dimension 15. When applying PCA locally, the first three components of PCA were removed.

Verification rates with equal error rates (EER) denoted as (VR-EER) and verification rates with false positive rates (FPR) equal to 0.1% denoted as (VR-FPR0.1%) are reported and we also provide the value of area under the receiver operating characteristic (AUC). Rank-1 recognition rates on NJU-ID, CUFSF and CASIA NIR-VIS 2.0 are also reported. To report rank-1 on NJU-ID dataset, high resolution face images in the test folds are registered as gallery images and low resolution images as probe images. On CUFSF dataset, photos are gallery images and sketches are probe images. On CASIA NIR-VIS 2.0, visible light images are registered as gallery images and near-infrared images are used as probe images.

7.2 Parameter Settings

There are a few parameters to tune in the proposed methods. The parameters of τ and η were tuned by setting γ and ϱ to zero without the sparse feature selection term. They were tuned separately; when tuning τ, η was set to zero and vise versa. After tuning τ and η, γ and ϱ were tuned together, with τ and η setting to the best parameters on each dataset. For tuning parameters, on AR and CUFSF, a separate split of data was used different from the ten partitions for results reporting. On NJU-ID dataset, a separate cross validation was used. On CASIA, the parameters were tuned on View 1 as required by the evaluation protocol.

Other parameters include the numbers of triplets used for training. In all the experiments, triplets were constructed by finding focal samples' intra-personal and inter-personal k-nearest neighbors. AR is a single modal dataset. On AR, intra-personal nearest neighbors denoted as $K1$ was set to 3 as there are four images per person. Inter-personal nearest neighbors denoted as $K2$, set to 100. On both NJU-ID and CUFSF, $K1$ was set to 1. $K2$ was set to 100 for NJU-ID dataset and 150 for CUFSF dataset. The setting of $K2$ on these three datasets were tuned. On CASIA NIR-VIS 2.0 dataset, both $K1$ and $K2$ were set to 10. Generally, larger $K1$ and $K2$ can lead to better results but it is not always the case as the results may become saturated. The settings on CASIA NIR-VIR 2.0 dataset need to balance between the performance and training time.

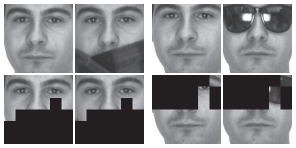


Figure 4: The first row shows example training images. In the second row, black patches on images correlate to zeros in the parameters of $\mathbf{w}_x$ and $\mathbf{w}_y$ of the weak learners. The two weak metrics successfully identify the occluded structures of faces.

Figure 5: Results of patch based feature selection of the first two weak metrics on CUFSF dataset.

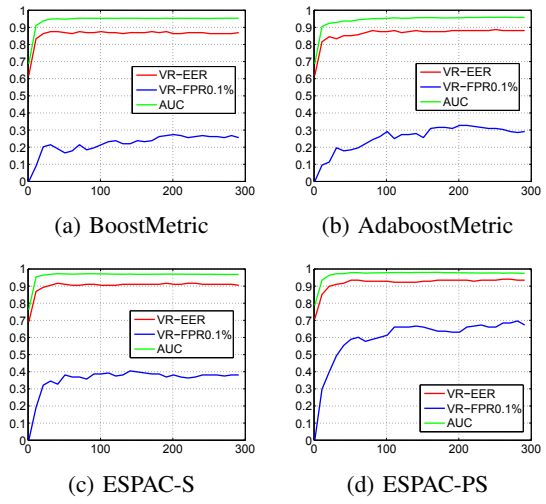
7.3 Visualization of Sparse Structures of Learned Weak Metrics

Fig. 4 shows the visualization results of the sparse structures of the patch-based weak metrics on AR-Sunglasses and AR-Scarf. The first row shows example training images. In the second row, the black patches correspond to zero values of the first learned weak metrics on two datasets. From the left four images in Fig. 4, the first weak metric on AR-Scarf does not use the lower parts of the face images for measuring distances as the lower parts are occluded. Same results are observed on AR-Sunglasses.

In Fig. 5, sparse structures of the first two weak metrics on CUFSF dataset are given. In the first row, the first three face pairs are those with the largest distances by using the first weak metric. In the second row, corresponding sparse structure of the first weak metric is given. The first weak metric does not use the features corresponding to eyes and jaws. By carefully observing the faces of this dataset, eye regions are less informative in distinguish person identities possibly because humans' eyes are hard to capture in sketches (Note that we have manually remove all the samples with eyeglasses in this experiment). For the jaws, as can be seen in the first three face pairs, since our alignment procedure is done by setting eyes in the same position, due to the error estimation of the ratio of faces' heights and widths, the faces in the sketches are shorter than their corresponding faces in photos, leading to misalignment in the jaw part. The third row also shows three face pairs with the largest distances by using the second weak metric. The forth row shows the structure of the second weak metric. The second weak metric does not use most of the jaw part. This is because the first weak learner is not sufficient to handle those images that have misalignment problems. The first and second weak learners are complementary as they use different face regions for measuring distances and the resulting same labeled pairs with large distances are different.

7.4 Influences of Different Number of Weak Metrics

Fig. 6 shows the influences of different number of weak metrics



(a) BoostMetric (b) AdaboostMetric

(c) ESPAC-S (d) ESPAC-PS

Figure 6: Influence of different number of weak metrics on AR-Scarf.

Table 1: Comparison with other boosting based metric learning methods on AR-Sunglasses dataset.

Methods	VR-EER(%)	VR-FPR0.1%(%)	AUC
BoostMetric	79.9 ± 1.7	18.0 ± 3.5	0.885
AdaboostMetric	80.2 ± 1.7	18.2 ± 3.4	0.884
ESPAC-NS	81.7 ± 2.0	18.0 ± 6.5	0.896
ESPAC-S	81.5 ± 2.2	18.1 ± 5.4	0.898
ESPAC-PS	$\mathbf{85.2 \pm 1.6}$	$\mathbf{29.2 \pm 5.9}$	**0.926**

on AR-Scarf and also comparative convergence results with BoostMetric [29] and AdaboostMetric [4]. In the following, ESPAC-NS denotes the proposed method with no feature selection. ESPAC-S is the proposed method with single-feature based group selection and ESPAC-PS uses the patch-based feature selection. As can be seen, the proposed methods, ESPAC-S and ESPAC-PS can obtain similar convergence results compared with BoostMetric and AdaboostMetric while achieving better results. Both ESPAC-S and ESPAC-PS converged with about 100 weak metrics. In the following, for reporting results, on all the four datasets, we used 150 weak metrics.

Table 2: Comparison with other boosting based metric learning methods on AR-Scarf dataset.

Methods	VR-EER(%)	VR-FPR0.1%(%)	AUC
MetricBoost	89.1 ± 1.3	34.8 ± 3.8	0.961
AdaboostMetric	89.2 ± 1.1	32.4 ± 3.6	0.960
ESPAC-NS	90.7 ± 1.2	40.6 ± 4.3	0.969
ESPAC-S	92.0 ± 0.9	45.3 ± 5.0	0.974
ESPAC-PS	$\mathbf{93.8 \pm 1.6}$	$\mathbf{64.1 \pm 6.0}$	**0.982**

7.5 Results on AR Dataset

Table 1 and Table 2 present results on AR-Sunglasses and AR-Scarf. The proposed methods have been compared with two boosting based metric learning methods. The code of BoostMetric [29] was provided by the authors. AdaboostMetric [4] was implemented by ourselves. The parameters of the two methods were tuned to their optimal. The best results were achieved by ESPAC-PS, followed by ESPAC-S. All three versions of the proposed method performed better than BoostMetric and

AdaboostMetric, except for VR-FPR0.1% on AR-Sunglasses, with the results of ESPAC-NS and ESPAC-S similar to those of BoostMetric and AdaboostMetric. ESPAC-PS improves the results to a large extent. This is mainly because that the face data of this dataset is under heavy occlusion and ESPAC-PS can successfully identify such occluded sparse structures.

7.6 Results on NJU-ID Dataset

Table 3 presents results on NJU-ID dataset. The proposed method has been compared with a set of state-of-the-art methods for face recognition. The parameters of the compared methods were adjusted to their optimal. In the table, the first six compared methods are single-modal based, including PCA [2], LDA [2], kernel principal component analysis (KPCA) [15], kernel discriminant analysis (KDA) [6], neighbourhood components analysis (NCA) [10] and large margin nearest neighbor (LMNN) [32]. Among them, PCA, LDA and KPCA are relatively worse than the other three. This is mainly that because PCA and KPCA are unsupervised. The performance of LDA is also unsatisfactory, as there are only two images per person. This leads to the within-class scatter matrix of LDA becomes singular and the performance degrades. The following six methods are multi-modal based, including CSR [20], kernel coupled spectral regression (KCSR) [20], canonical correlation analysis (CCA) [11], kernel canonical correlation analysis (KCCA) [18], CDFE [22] and multi-view discriminant analysis(MvDA) [13]. CSR, KCSR, CDFE and MvDA are widely used for multi-modal face recognition. Among all the compared methods, ESPAC-PS is the best with respect to all the evaluation protocols. Among the proposed three versions of our methods, ESPCA-NS is the worst. On NJU-ID dataset, as the face data has many other variations besides modality variation, this proves that by using feature selection together with the proposed boosting based cross-modal metric learning, the performances can be improved. The proposed methods are able to capture the variations in the extracted face features and neglect these noisy features.

Figure 7: Examples of aligned face images. The first row are samples of high resolution and the second row are corresponding samples of low resolution.

7.7 Results on CUFSF Dataset

Table 4 presents the recognition results on CUFSF dataset. The six single-modal methods and six multi-modal methods are also compared. Compared with these state-of-the-art methods, the proposed ESPAC-PS is the best among three of all the evaluation protocols (though for rank-1, CSR achieved similar performance). Among the six single-modal methods, the best results were achieved by KDA. The results of KDA are comparable with our methods. As KDA is a non-linear method, this may illustrate that the data distribution has non-linear properties, while the proposed method is linear-based. It is likely that extending our method to non-linear based by design non-linear weak metrics can further improve the recognition results.

Figure 8: Examples of aligned face images. The first row are samples of photos and the second row are corresponding sketches.

Table 3: Comparison with other methods on NJU-ID dataset.

Methods	VR-EER(%)	VR-FPR0.1%(%)	Rank-1(%)	AUC
PCA	63.6 ± 5.2	8.2 ± 7.3	13.3 ± 6.5	0.688
LDA	59.7 ± 5.3	8.1 ± 8.7	10.9 ± 8.1	0.637
KPCA	64.0 ± 4.8	8.2 ± 6.8	14.0 ± 5.8	0.687
KDA	67.6 ± 4.7	11.7 ± 8.8	18.3 ± 8.4	0.723
LMNN	69.9 ± 3.9	13.4 ± 11.2	20.3 ± 6.4	0.737
NCA	65.6 ± 3.3	9.4 ± 6.8	15.6 ± 6.2	0.715
CSR	67.2 ± 3.9	11.2 ± 11.1	18.6 ± 8.2	0.730
KCSR	68.3 ± 4.2	16.5 ± 11.1	19.4 ± 9.4	0.727
CCA	62.5 ± 4.5	5.1 ± 5.8	10.5 ± 5.8	0.638
KCCA	62.1 ± 3.6	7.0 ± 8.6	10.6 ± 5.0	0.632
CDFE	58.2 ± 5.1	2.3 ± 2.6	6.2 ± 3.6	0.588
MvDA	62.1 ± 4.3	4.3 ± 5.6	10.2 ± 4.3	0.628
ESPAC-NS	65.0 ± 7.5	14.8 ± 6.0	19.7 ± 5.1	0.738
ESPAC-S	69.2 ± 6.9	16.6 ± 8.6	20.4 ± 5.5	0.747
ESPAC-PS	$\mathbf{70.1 \pm 7.3}$	$\mathbf{16.8 \pm 7.8}$	$\mathbf{20.8 \pm 6.2}$	**0.748**

Table 4: Comparison with other methods on CUFSF dataset.

Methods	VR-EER(%)	VR-FPR0.1%(%)	Rank-1(%)	AUC
PCA	85.6 ± 5.2	23.9 ± 7.3	37.9 ± 6.5	0.937
LDA	79.0 ± 5.3	20.2 ± 8.7	21.4 ± 8.1	0.874
KPCA	85.6 ± 4.8	23.8 ± 6.8	37.9 ± 5.8	0.937
KDA	93.9 ± 4.7	41.1 ± 8.8	53.2 ± 8.4	0.985
LMNN	93.2 ± 3.9	35.4 ± 11.2	49.2 ± 6.4	0.982
NCA	92.8 ± 3.3	31.8 ± 6.8	42.5 ± 6.2	0.979
CSR	92.8 ± 3.9	44.9 ± 11.1	$\mathbf{53.8 \pm 8.2}$	0.980
KCSR	93.7 ± 4.2	16.4 ± 11.1	45.1 ± 9.4	0.983
CCA	91.6 ± 4.5	17.5 ± 5.8	35.5 ± 5.8	0.973
KCCA	93.1 ± 3.6	14.0 ± 8.6	35.6 ± 5.0	0.980
CDFE	89.7 ± 5.1	17.6 ± 2.6	24.4 ± 3.6	0.960
MvDA	80.2 ± 4.3	7.3 ± 5.6	9.7 ± 4.3	0.884
ESPAC-NS	94.0 ± 3.8	43.4 ± 2.8	52.2 ± 5.5	0.987
ESPAC-S	94.2 ± 3.4	44.6 ± 3.8	53.3 ± 2.0	0.987
ESPAC-PS	$\mathbf{94.5 \pm 4.0}$	$\mathbf{45.1 \pm 3.1}$	53.7 ± 2.2	**0.989**

7.8 Results on CASIA NIR-VIS 2.0 Dataset

On CASIA NIR-VIS 2.0 dataset, results of the proposed methods and state-of-the-art methods are given in Table 5. Note that results of NCA is not given as the computational cost of NCA is too heavy on this dataset. The compared methods were all adjusted to their optimal. On this dataset, ESPAC-S and ESPAC-PS achieved the best AUC result. ESPAC-PS is best for the VR-EER performance measure. Among the three versions of the proposed methods, the improvement of ESPAC-S and ESPAC-PS over ESPAC-NS is relatively small. As this dataset is collected under near controlled environment, the only variations are eyeglasses and a few expression changes which are of a small portion. Another observation is that our methods are worse than LDA and CSR with respect to VR-FPR0.1% and Rank-1. This is due to that our methods use part of all the triplets constructed on this dataset. For low rank-1 rate, it means for some samples, their intra-personal distances are still larger than one of their inter-personal distances. As in ESPAC, triplets are all pre-constructed and kept fixed during optimization. Although

Figure 9: Examples of aligned faces in CASIA NIR-VIS 2.0 dataset.

Table 5: Comparison with other methods on CASIA NIR-VIS 2.0 dataset.

Methods	VR-EER(%)	VR-FPR0.1%(%)	Rank-1(%)	AUC
PCA	83.1 ± 0.5	20.8 ± 1.7	38.6 ± 1.0	0.909
LDA	88.6 ± 0.6	$\mathbf{46.8 \pm 1.5}$	$\mathbf{56.4 \pm 1.1}$	0.954
KPCA	82.2 ± 0.5	17.6 ± 1.5	34.0 ± 1.1	0.902
KDA	90.2 ± 0.4	31.6 ± 3.1	50.7 ± 1.6	0.963
LMNN	87.3 ± 0.6	27.1 ± 1.8	41.7 ± 0.8	0.946
CSR	89.3 ± 0.6	44.2 ± 1.4	55.6 ± 0.7	0.957
KCSR	90.1 ± 0.5	27.2 ± 3.3	49.6 ± 0.9	0.962
CCA	84.0 ± 0.5	23.7 ± 1.1	30.2 ± 1.0	0.917
KCCA	82.5 ± 0.4	20.6 ± 1.3	25.8 ± 1.1	0.905
CDFE	76.7 ± 0.3	1.9 ± 0.2	19.4 ± 1.2	0.832
MvDA	84.1 ± 0.4	16.2 ± 1.7	29.5 ± 1.4	0.918
ESPAC-NS	90.3 ± 0.5	33.6 ± 1.3	45.1 ± 1.6	0.972
ESPAC-S	90.8 ± 0.5	35.0 ± 1.4	46.6 ± 1.7	**0.974**
ESPAC-PS	$\mathbf{90.9 \pm 0.5}$	34.8 ± 1.2	46.5 ± 1.5	**0.974**

ESPAC can learn well on the given sets of triplets, there are still unseen triplets that the learned ESPAC not perform well. A solution is to perform ESPAC multiple passes to further enhance the performances, as is also suggested by [32, 29].

7.9 Comparison of Training Time

A comparison of training time on AR-Scarf is given in Table 6. The two compared methods, BoostMetric and AdaboostMetric, also use triplets for training. Implementation of BoostMetric was provided by its authors. AdaboostMetric was implemented by strictly following its paper. The experiments were conducted on the same computer under exactly the same setting. The training time of each of the methods is related to the number of triplets used. For 5 different numbers of triplets constructed, the training times (in seconds) are given in Table 6. The complexity of BoostMetric per iteration mainly relies on the number of triplets ($O(N \times S \times D)$, N the learned number of samples, S the number of same class neighbors, D the number of different classes neighbors). AdaboostMetric decomposes triplets into pairs ($O(N \times (S + D))$) and is more efficient. The implementation of the proposed methods uses a modification of the scheme in the work of McFee and Lanckriet [25], which is neither triplet nor pair based. ESPAC-NS with this scheme is very efficient. While ESPAC-S and ESPAC-PS use ADMM to learn SPAC which takes hundreds of iterations to converge, the training processes are longer. However, with this scheme, the training time increases little with the triplets' number increases.

Table 6: Comparison of training time (in seconds).

Methods	9960	19920	29880	39840	49800
MetricBoost	924	1863	2724	3926	5040
AdaboostMetric	573	987	1412	1839	2203
ESPAC-NS	61	69	75	82	91
ESPAC-S	2237	2364	2513	2647	2773
ESPAC-PS	2434	2626	2768	2776	2898

7.10 Interpretations and Some Possible Extensions

Relationship with face attribute learning. In [17], the authors proposed to learn semantic face features to describe face attributes. The proposed SPAC also can be seen as learning a classifier but with the objective of separating pairs of samples encoded in triplets. SPAC automatically determines the best regions for separating the current inseparable groups of triplets. From Fig. 4 and Fig. 5, the learned SPAC also have some implicit semantic meanings such as misaligned jaw on CUFSF dataset or occluded eye areas on AR dataset. In this way, the proposed method is related to automatic face semantic feature learning, without the need of labeling attributes and manually determine attribute related face regions for training. As the method in [17] used non-linear SVM for training, this further provides us insight to make SPAC non-linear in our future work.

Relationship with Mahalanobis metric learning. As the proposed method is also applicable for single-modal metric learning. Compared with Mahalanobis metric learning, it is easy to see that Mahalanobis metric tries to separate samples in the space of $\mathbf{x}_i - \mathbf{x}_j$, while the cross-modal metric is using the space of $[\mathbf{x}_i; -\mathbf{x}_j]$. As has been discussed in [7], the former space may reduce data separability and our formulation is a more generalized one. We have partly proved this on AR dataset that our method even the non-sparse based version is either better or comparable to BoostMetric and AdaboostMetric. So it is worth trying to test it on more single-modal based datasets.

Extend to other feature extraction frameworks. In Section 4, we have introduced the local face feature extraction method used in this paper. It is clear that combinations with other facial landmark features and features at multiple scales [8] are also possible. In this way, the proposed method will be able to find more informative face landmarks for recognition. Combining with the convolutional neural networks (CNN) [19] is also an interesting new line of research as the features of CNN have invariant spatial properties.

8. CONCLUSION

In this paper, an ensemble of sparse cross-modal metric learning method is proposed for heterogeneous face recognition. A set of sparse cross-modal weak metrics are learned to remove modality variations. In addition, feature selection is performed to remove noisy features relating to the variations such as, occlusions, expressions and illumination changes. Each learned weak metric selects a few face features or a few face regions for measuring distances between samples. Although one weak cross-modal metric is not able to handle all the variations, by using an Adaboost framework to reassign weights based on triplets, the finally learned weak metrics become complementary. All the weak metrics are combined into a strong one which is able to deal with various variations. State-of-the-art performances on three heterogeneous face recognition datasets are achieved. On the AR dataset, we have shown that the proposed methods can improve the recognition results by a large extent for faces with heavy occlusions.

There are a few future directions of this work. One is to extend the linear weak metrics to non-linear ones which may further improve face recognition performances. Another is that multi-pass ensembles can be further tested. Besides, as the framework is fairly flexible, combination with other face feature extraction framework such as face landmark indexed features, can also be tested to find which face landmarks convey more information for face matching.

Acknowledgement

The work is supported by the National Science Foundation of China (Nos. 61432008, 61305068, 61321491), the Graduate Research Innovation Program of Jiangsu, China (KYLX_0047) and the Collaborative Innovation Center of Novel Software Technology and Industrialization. J. Huo is also supported by a scholarship from the China Scholarship Council as a one-year visiting student at the University of Manchester.

9. REFERENCES

[1] Y. Atzmon, U. Shalit, and G. Chechik. Learning sparse metrics, one feature at a time. In *Proc. 1st Workshop on Feature Extraction Modern Questions and Challenges*, pages 30–48, 2015.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *TPAMI*, 19(7):711–720, 1997.

[3] A. Bellet, A. Habrard, and M. Sebban. Metric learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 9(1):1–151, 2015.

[4] J. Bi, D. Wu, L. Lu, M. Liu, Y. Tao, and M. Wolf. Adaboost on low-rank psd matrices for metric learning. In *CVPR*, pages 2617–2624, 2011.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[6] D. Cai, X. He, and J. Han. Speed up kernel discriminant analysis. *VLDB J.*, 20(1):21–33, 2011.

[7] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *ECCV*, pages 566–579. 2012.

[8] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *CVPR*, pages 3025–3032, 2013.

[9] C. Geng and X. Jiang. Face recognition using sift features. In *ICIP*, pages 3313–3316, 2009.

[10] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS*, pages 513–520, 2004.

[11] H. Hotelling. Relations between two sets of variates. *Biometrika*, pages 321–377, 1936.

[12] Y. Jin, J. Lu, and Q. Ruan. Coupled discriminative feature learning for heterogeneous face recognition. *TIFS*, 10(3):640–652, 2015.

[13] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. In *ECCV*, pages 808–821. 2012.

[14] C. Kang, S. Liao, Y. He, J. Wang, W. Niu, S. Xiang, and C. Pan. Cross-modal similarity learning: A low rank bilinear formulation. In *CIKM*, pages 1251–1260, 2015.

[15] K. I. Kim, K. Jung, and H. J. Kim. Face recognition using kernel principal component analysis. *IEEE Signal Process. Lett.*, 9(2):40–42, 2002.

[16] B. F. Klare and A. K. Jain. Heterogeneous face recognition using kernel prototype similarities. *TPAMI*, 35(6):1410–1422, 2013.

[17] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, pages 365–372, 2009.

[18] P. L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *IJNS*, 10(05):365–377, 2000.

[19] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, pages 255–258, 1998.

[20] Z. Lei and S. Z. Li. Coupled spectral regression for matching heterogeneous faces. In *CVPR*, pages 1123–1128, 200.

[21] S. Li, D. Yi, Z. Lei, and S. Liao. The casia nir-vis 2.0 face database. In *CVPR Workshops*, pages 348–353, 2013.

[22] D. Lin and X. Tang. Inter-modality face recognition. In *ECCV*, pages 13–26. 2006.

[23] K. Liu, A. Bellet, and F. Sha. Similarity learning for high-dimensional sparse data. In *AISTATS*, pages 653–662, 2015.

[24] A. M. Martinez. The ar face database. *CVC Technical Report*, 24, 1998.

[25] B. McFee and G. R. Lanckriet. Metric learning to rank. In *ICML*, pages 775–782, 2010.

[26] A. Mignon and F. Jurie. Cmml: a new metric learning approach for cross modal matching. In *ACCV*, pages 1–14, 2012.

[27] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *TPAMI*, 22(10):1090–1104, 2000.

[28] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.*, 37(3):297–336, 1999.

[29] C. Shen, J. Kim, L. Wang, and A. Van Den Hengel. Positive semidefinite metric learning using boosting-like algorithms. *JMLR*, 13(1):1007–1036, 2012.

[30] C. Shen, A. Welsh, and L. Wang. Psdboost: Matrix-generation linear programming for positive semidefinite matrices learning. In *NIPS*, pages 1473–1480, 2009.

[31] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *TPAMI*, 31(11):1955–1967, 2009.

[32] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009.

[33] L. Yang, R. Jin, L. Mummert, R. Sukthankar, A. Goode, B. Zheng, S. C. Hoi, and M. Satyanarayanan. A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval. *TPAMI*, 32(1):30–44, 2010.

[34] D. Yi, Z. Lei, and S. Z. Li. Shared representation learning for heterogenous face recognition. In *FG*, volume 1, pages 1–7, 2015.

[35] Y. Ying, K. Huang, and C. Campbell. Sparse metric learning via smooth optimization. In *NIPS*, pages 2214–2222, 2009.

[36] W. Zhang, X. Wang, and X. Tang. Lighting and pose robust face sketch synthesis. In *ECCV*, pages 420–433. 2010.

[37] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, pages 513–520, 2011.

[38] P. Zhou, L. Du, M. Fan, and Y.-D. Shen. An lle based heterogeneous metric learning for cross-media retrieval. In *SDM*, pages 64–72, 2015.