Pattern Recognition Letters 71 (2016) 38–44

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A coupled discriminative dictionary and transformation learning approach with applications to cross domain matching ☆

Sivaram Prasad Mudunuri, Soma Biswas*

Department of Electrical Engineering, Indian Institute of Science, Bangalore, Karnataka 560012, India

ARTICLE INFO

Article history:
Received 19 July 2015
Available online 18 December 2015

Keywords:
Face recognition
Activity recognition
Dictionary learning
Metric learning

ABSTRACT

Cross domain and cross-modal matching has many applications in the field of computer vision and pattern recognition. A few examples are heterogeneous face recognition, cross view action recognition, etc. This is a very challenging task since the data in two domains can differ significantly. In this work, we propose a coupled dictionary and transformation learning approach that models the relationship between the data in both domains. The approach learns a pair of transformation matrices that map the data in the two domains in such a manner that they share common sparse representations with respect to their own dictionaries in the transformed space. The dictionaries for the two domains are learnt in a coupled manner with an additional discriminative term to ensure improved recognition performance. The dictionaries and the transformation matrices are jointly updated in an iterative manner. The applicability of the proposed approach is illustrated by evaluating its performance on different challenging tasks: face recognition across pose, illumination and resolution, heterogeneous face recognition and cross view action recognition. Extensive experiments on five datasets namely, CMU-PIE, Multi-PIE, ChokePoint, HFB and IXMAS datasets and comparisons with several state-of-the-art approaches show the effectiveness of the proposed approach.

1. Introduction

Many applications require matching data coming from different domains or modalities, for example, we may want to compare a low-resolution uncontrolled face image captured using surveillance camera with a high-resolution image that is stored in the database, or given a text query, we may want to retrieve images (Fig. 1). But cross domain or cross modality matching (everything referred as cross-domain matching from now) is very challenging, since the data coming from different domains is usually very different. This is in addition to the intra-class variability that is present in the data of both the domains. Recently, there have been significant research efforts in addressing this problem and many approaches have been proposed. Among other methods, coupled dictionary learning has emerged as a powerful method for matching images coming from different domains. Though initially proposed for reconstruction applications, recent approaches have shown excellent performance in both reconstruction and classification problems [14].

In this work, we build upon the success of the coupled dictionary learning approaches and propose a coupled discriminative dictionary and transformation learning approach specifically designed for the classification/recognition tasks. First, data samples in the two domains are transformed in such a manner that they share common sparse representations with respect to their own dictionaries in the transformed space. For improving discriminability for recognition tasks, the sparse coefficients are further mapped such that the k-nearest neighbors of the same class move closer and those of different classes are pushed far apart. This ensures that the final sparse coefficients obtained using the proposed approach are discriminative enough to distinguish between features from different classes. The dictionaries, sparse coefficients and all the transformation matrices are jointly updated in an iterative manner. During testing, the features from the two domains are first transformed using the learnt feature transformation matrices before computing the sparse coefficients. These coefficients are further transformed using the learnt discriminative mapping and then used for matching. The applicability of the proposed approach is illustrated by evaluating its performance on different challenging tasks: face recognition across pose, illumination and resolution, heterogeneous face recognition (matching visible with NIR images) and cross view action recognition. Extensive experiments on five datasets namely, CMU-PIE, Multi-PIE, ChokePoint, HFB and IXMAS

☆ This paper has been recommended for acceptance by Dr. L. Yin.

* Corresponding author. Tel.: +91 7204012320.
E-mail address: soma.biswas@ee.iisc.ernet.in (S. Biswas).

http://dx.doi.org/10.1016/j.patrec.2015.12.003
0167-8655/

Fig. 1. Applications of cross domain matching. (a) high-resolution controlled image with low-resolution uncontrolled image as in surveillance scenario; (b) cross view action recognition; (c) NIR images with visible light images; (d) Object recognition from retailer sites and consumer images.

datasets and comparisons with several state-of-the-art approaches show the effectiveness of the proposed approach. The proposed formulation shares similarities with the seminal work by Huang and Wang [14]. The differences of the proposed approach over [14] are as follows:

- The original features are transformed using the feature transformation matrices.
- Both the domains in the transformed space have the same sparse coefficients.
- Learning a classifier on the sparse coefficients is not always possible where the test classes are not the same as the training classes as in face recognition applications as done in [14].
- There is an explicit discriminative term which significantly improves the performance of coupled dictionary approaches for recognition and classification tasks as shown by the experimental results.

The rest of the paper is organized as follows. Section 2 discusses the related work. The proposed approach is presented in Section 3 and the experimental results are presented in Section 4. The paper concludes with a discussion and conclusion section.

2. Related work

In this section, we discuss the papers in the literature which are closely related to the proposed approach. Dictionary learning-based algorithms have recently been successfully applied in many applications. Wright et al. [35] propose a sparse representation based classification approach for robust face recognition. Since then, several approaches have been proposed to learn discriminative dictionaries which can help in distinguishing between different classes [15,25,39,44] and [38]. A sparse modeling and dictionary learning based method for clustering and classification of different classes is presented in [27]. Wang et al. [33] propose a semi-coupled dictionary learning method that aims at reducing the distance between the sparse coefficients of the same subject belonging to different domains by transforming one of the domain features to the new space. Huang and Wang [14] propose a coupled dictionary and feature space learning algorithm that iteratively updates two separate dictionaries and transformation matrices to transform the sparse coefficients. Matching cross-domain data using joint dimensionality reduction techniques like Canonical Correlation Analysis [12], Partial Least Squares [28], etc. have also been quite successful. In [29], the covariance between the sets is jointly optimized and the classes are also separated in their respective feature spaces. Multi-view Discriminant Analysis [17] finds a discriminative common space for all the views by jointly learning multiple view-specific linear transforms. Domain adaptation methods like [21] have also been successfully applied for cross-domain matching tasks.

Cross-domain matching has several applications. In this work, we have focused mainly on face recognition across pose and resolution, NIR-vs-Visible face matching and cross-view action recognition and we will provide pointers to some related works in these areas. Recognizing faces across multiple variations like illumination, resolution, pose, etc. has received considerable attention [26,42,43] and [7]. A co-transfer learning framework, which combines transfer learning with co-training for matching faces across resolutions is proposed in [3]. Zou and Yuen [47] propose an algorithm to perform matching of low resolution facial images by learning the relationship between high-resolution gallery and the low-resolution probe images. For matching near infra-red and visible facial images, Zhu et al. [46] propose a transductive heterogeneous face matching approach that can reduce the modality gap by extracting the domain invariant and target-related discriminative features. Hou et al. [13] propose an approach to derive a common space which can relate and represent facial images of different modalities. Jin et al. [16] propose a method that can learn several image filters to simultaneously utilize discriminative information and reduce the appearance difference of facial images captured across different modalities. Lu et al. [23] propose a compact binary face descriptor (CBFD) for matching facial images of different domain. Lei et al. [19] propose a discriminant face descriptor (DFD) that can enhance the discriminative capacity of face representation. The method also formulates a coupled DFD feature that can further improve the performance of matching across different modalities.

For cross-view action recognition, Wang et al. [31] propose a method to learn action units using the graph regularized nonnegative matrix factorization from the extracted novel spatial-temporal descriptors. A multiview spatio-temporal representation based approach to handle the problem of cross view action representation is discussed in [32] and [22]. An approach to learn view-invariant sparse representations to perform cross-view action matching is described in [45]. An approach that constructs animated pose templates to detect short-term, long-term, and contextual actions from cluttered scenes in videos can be found in [40]. Wu et al. [36] propose an approach to construct a common feature space to link source view and target view for transferring knowledge between them. Yeh et al. [41] propose a method that can exploit the domain transfer ability in the correlation subspace.

3. Proposed approach

In this section, we present the proposed joint dictionary and transformation learning algorithm for matching data from two different domains.

3.1. Problem formulation

Let $\mathbf{X}_1 \in \mathbf{R}^{d_1 \times N}$ and $\mathbf{X}_2 \in \mathbf{R}^{d_2 \times N}$ be the two matrices that represent features computed from the two domains. Here, d_1 and d_2 are the length of the feature vectors and N denotes the number of training images. The goal is to learn transformation matrices, dictionaries and a mapping function such that the following two criteria are satisfied:

- The transformation matrices should transform the input features from the different domains such that they have the same sparse representation with respect to their own dictionaries in the transformed space.
- The mapping function should be capable of moving the k-nearest sparse coefficient vectors of the same class closer and simultaneously those of different classes apart.

Based on the above two criteria, we propose to minimize an objective function whose general form is given as follows:

$$\min_{\mathbf{U}_1,\mathbf{U}_2,\mathbf{D}_1,\mathbf{D}_2,\mathbf{\Lambda},\mathbf{L}} E^1_{DR} + E^2_{DR} + E_D \tag{1}$$

Here E^k_{DR} denotes the energy term based on the data reconstruction error for domain $k = 1, 2$ and E_D denotes the discriminative term in the objective function. Now, we describe each of the terms in (1) in detail.

Data Reconstruction Term: The data reconstruction term ensures that the transformed data in the two domains are sparsely reconstructed using their respective dictionaries and also, the transformed data share the same sparse coefficient vector across the two domains. This is given by the following objective function

$$\begin{aligned} E^k_{DR} &= \arg\min_{\mathbf{U}_k,\mathbf{D}_k,\mathbf{\Lambda}} \| \mathbf{U}_k\mathbf{X}_k - \mathbf{D}_k\mathbf{\Lambda}\|_2^2 + \alpha\| \Lambda\|_1^2 \\ &\text{subject to } \| \mathbf{d}^i_k \|_2 \leq 1; \forall i \qquad (k = 1, 2) \end{aligned} \tag{2}$$

Here, $\mathbf{D}_1 \in \mathbf{R}^{d_1 \times K}$ and $\mathbf{D}_2 \in \mathbf{R}^{d_2 \times K}$ are the dictionaries corresponding to the two domains with K atoms each. $\mathbf{U}_1 \in \mathbf{R}^{d_1 \times d_1}$ and $\mathbf{U}_2 \in \mathbf{R}^{d_2 \times d_2}$ are transformation matrices and $\mathbf{\Lambda} \in \mathbf{R}^{K \times N}$ is the common sparse coefficient matrix.

Discriminative Term: Though (2) ensures that there is good reconstruction of the data, it may not be sufficiently discriminative for recognition tasks. So we add a discriminative term to the objective function which ensures that the k-nearest sparse coefficient vectors for the same class come closer to one another and k-nearest neighbors from different classes move apart. Inspired by the success of metric learning approaches, we learn a mapping function $\mathbf{L}$ by using either of the two approaches Large Margin Nearest Neighbor (LMNN) [34] and Large Scale Metric Learning (LSML) [18].

In LMNN approach [34], the matrix $\mathbf{L}$ is learned such that the following objective function is minimized [34]

$$\begin{aligned} E_D = \sum_{j \in kNN(i)} \| \mathbf{L}(\lambda_i - \lambda_j)\|_2^2 + \beta \sum_{j \in kNN(i)} \sum_l (1 - y_{il}) \dots \\ \left[1 + \| \mathbf{L}(\lambda_i - \lambda_j)\|_2^2 - \| \mathbf{L}(\lambda_i - \lambda_l)\|_2^2\right]_+ \end{aligned} \tag{3}$$

The matrix $\mathbf{L}$ is a square matrix of size $K \times K$, where K is the length of the sparse vector of both the domains. Here, λ_i is the sparse coefficient vector of the ith sample of both domains (since they have same coefficients) and the term $[p]_+ = \max(p, 0)$. The term $j \in kNN(i)$ means that j belongs to the k-nearest neighbor of i of the same class, i.e. they are the target neighbors. The indicator variable $y_{il} = 1$ if the class labels are the same i.e. $y_i = y_l$, otherwise $y_{il} = 0$.

The LSML algorithm is formulated based on equivalence of constraints in such a way that the distance metric is learned from the covariance matrices of matched pairs and non-matched pairs. Please refer to [18] for further details.

3.2. Optimization

Using these definitions for the data reconstruction and discriminative term, we would like to solve for all the unknown quantities, specifically, the two transformation matrices $\mathbf{U}_1$, $\mathbf{U}_2$, the two dictionaries $\mathbf{D}_1$, $\mathbf{D}_2$, mapping $\mathbf{L}$ and the sparse coefficient $\mathbf{\Lambda}$. We propose an iterative algorithm to solve for the different unknowns as described below.

Initialization: First the dictionaries and coefficient matrix are initialized by solving the following objective function

$$\arg\min_{\mathbf{D}_1,\mathbf{D}_2,\mathbf{\Lambda}} \left\| \begin{bmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{bmatrix} - \begin{bmatrix}\mathbf{D}_1 \\ \mathbf{D}_2\end{bmatrix}\mathbf{\Lambda} \right\|_2^2 + \alpha\| \mathbf{\Lambda}\|_1^2 \tag{4}$$

This is the standard KSVD formulation [6]. The transformation matrices and the class centroids (described later) required for updating the coefficients are initialized using (9) and (10). The matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ are initialized to identity matrices of size $d \times d$, where d is the length of the feature vectors. The matrix $\mathbf{L}$ is initialized to the LMNN/LSML metric that is learned on the sparse coefficients of the original features (without multiplying with $\mathbf{U}_1$ and $\mathbf{U}_2$). Then the iteration proceeds as follows.

Update dictionaries: When the dictionaries are updated, the transformation matrices, sparse coefficient and the mapping are fixed at the value of the previous iteration. The two terms containing the dictionary terms in (2) are combined as follows:

$$\arg\min_{\mathbf{D}_1,\mathbf{D}_2} \left\| \begin{bmatrix}\mathbf{U}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_2\end{bmatrix}\begin{bmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{bmatrix} - \begin{bmatrix}\mathbf{D}_1 \\ \mathbf{D}_2\end{bmatrix}\mathbf{\Lambda} \right\|_2^2 + \alpha\| \mathbf{\Lambda}\|_1^2 \tag{5}$$

Now combining the dictionaries, input features and the transformation matrices, the objective function takes the form

$$\arg\min_{\mathbf{D}} \| \mathbf{UX} - \mathbf{D}\mathbf{\Lambda}\|_2^2 + \alpha\| \mathbf{\Lambda}\|_1^2 \tag{6}$$

where $\mathbf{D} = [\mathbf{D}_1]$, $\mathbf{X} = [\mathbf{X}_1]$ and $\mathbf{U} = [\mathbf{U}_1 \quad \mathbf{0}]$. This is the same problem that is solved by K-SVD.

Update sparse coefficient: In this step, the transformation matrices, dictionaries and the mapping are kept fixed. For updating the sparse coefficients, we propose an approximation of the discriminative term (3) so that the resulting objective function can be solved using existing efficient algorithms like SPAMS [24]. The role of the mapping is to move k-nearest sparse coefficients closer and move k-nearest neighbor of different classes apart. We approximate it using the assumption, that because of the mapping function, most of the sparse coefficients of each class move closer to one another, i.e. closer to the class centroid. Since it may not hold accurately for all the data samples of all the classes, we down weight this approximation as shown below. But we see that, this approximation not only makes the update of the sparse coefficient very efficient, this approach also gives good performance for many applications. The sparse coefficients are updated by solving following objective function

$$\arg\min_{\mathbf{\Lambda}} \| \mathbf{UX} - \mathbf{D}\mathbf{\Lambda}\|_2^2 + \gamma\| \mathbf{C} - \mathbf{L}\mathbf{\Lambda}\|_2^2 + \alpha\| \mathbf{\Lambda}\|_1^2 \tag{7}$$

The first term in the objective function ensures that the data can be reconstructed using the dictionaries. The second term is used to provide the required discriminability with a weighting factor denoted by γ. The parameter γ controls the relative contribution of the discriminability and reconstruction terms. Each column of $\mathbf{C}$ contains the centroid of the class of the corresponding sparse coefficient. Finally, the data reconstruction and the discriminative terms of Eq. (7) can be combined to obtain the objective function

$$\arg\min_{\mathbf{\Lambda}} \left\| \begin{bmatrix}\mathbf{UX} \\ \sqrt{\gamma}\mathbf{C}\end{bmatrix} - \begin{bmatrix}\mathbf{D} \\ \sqrt{\gamma}\mathbf{L}\end{bmatrix}\mathbf{\Lambda} \right\|_2^2 + \alpha\| \mathbf{\Lambda}\|_1^2 \tag{8}$$

We solve the above equation using standard SPAMS solvers [24].

Update transformation matrices and mapping: In this step, all the other parameters are kept fixed and the transformation matrices are updated as follows:

$$\mathbf{U}_k = \mathbf{D}_k\mathbf{\Lambda}\mathbf{X}_k^T(\mathbf{X}_k\mathbf{X}_k^T)^{-1} \qquad k = 1, 2 \tag{9}$$

During updating $\mathbf{L}$, since all the other parameters are kept fixed, the objective function that needs to be minimized is (3). For this, we use two different approaches: (1) algorithm in [34] and (2) a recent robust metric learning approach Large Scale Metric Learning (LSML) [18]. The main objective of both the approaches is to learn a transformation matrix such that transformed feature sets

Table 1
Training stage of the proposed algorithm.

- ***Input: Initialization:*** *The different parameters of the objective function* $\mathbf{D}_k^0$, $\mathbf{U}_k^0$, $\mathbf{\Lambda}^0$, $\mathbf{L}^0$ *and* $\mathbf{C}^0$ *are initialized. Here* $k = 1, 2$.
- *For iteration* $i = 1, 2, \ldots,$ *do until convergence*
  1. ***Update the dictionaries:*** *The dictionaries* $\mathbf{D}_1^{i+1}$ *and* $\mathbf{D}_2^{i+1}$ *are updated according to (5) using* $\mathbf{U}_1^i$, $\mathbf{U}_2^i$ *and* $\mathbf{\Lambda}^i$.
  2. ***Update the sparse coefficient matrix:*** *The sparse coefficient matrix* $\mathbf{\Lambda}^{i+1}$ *is updated according to (7) using* $\mathbf{D}_1^{i+1}$, $\mathbf{D}_2^{i+1}$, $\mathbf{U}_1^i$, $\mathbf{U}_2^i$, $\mathbf{C}^i$ *and* $\mathbf{L}^i$.
  3. ***Update the transformation matrices:*** *The two transformation matrices* $\mathbf{U}_1^i$ *and* $\mathbf{U}_2^i$ *are updated according to (9) using* $\mathbf{D}_1^{i+1}$, $\mathbf{D}_2^{i+1}$ *and* $\mathbf{\Lambda}^{i+1}$.
  4. ***Update the mapping:*** *The mapping* $\mathbf{L}^{i+1}$ *is updated either by using the algorithm in [34] or [18] using* $\mathbf{\Lambda}^{i+1}$.
- *Check for convergence: Check if the difference in the estimates of consecutive iterations is less than some threshold. Otherwise, go to Step 1.*
- ***Output:*** $\mathbf{D}_1$, $\mathbf{D}_2$, $\mathbf{U}_1$, $\mathbf{U}_2$, $\mathbf{\Lambda}$ *and* $\mathbf{L}$.

of the same class come closer to one another and those of different classes move further away. The class centroids (for class i) are then computed as given below

$$\mathbf{c}_i = \sum_{class_i} \mathbf{L}\lambda_i \tag{10}$$

These steps are repeated until the change in the parameters in the consecutive iterations is less than a pre-specified threshold. The various steps of the iterative optimization described here are summarized in Table 1.

3.3. Testing

The dictionaries, transformation matrices and the mapping that are learnt in the training stage are used for matching data from two different domains during testing. The data from the two domains $\mathbf{x}_i$ and $\mathbf{x}_j$ are first mapped to the transformed space using $\mathbf{U}_1$ and $\mathbf{U}_2$ respectively. The corresponding sparse coefficient vectors λ_i and λ_j are then computed using the learnt dictionaries $\mathbf{D}_1$ and $\mathbf{D}_2$ respectively. The distance between $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_j$ is computed as follows

$$d_{i,j} = \|\mathbf{L}(\lambda_i - \lambda_j)\|_2^2 \tag{11}$$

We used 1-NN classifier that finds the closest neighbor of the probe data among the gallery data. During testing, the Euclidean distance between the transformed sparse coefficient vectors of the probe data and the gallery data are computed. The gallery data with the smallest distance is chosen.

4. Experimental results

Extensive experiments are conducted to evaluate the usefulness of the proposed approach for different applications. Specifically, we evaluate the proposed approach for the task of (1) Face recognition across pose variations; (2) Face recognition across pose, illumination and resolution as captured in surveillance scenario; (3) Heterogeneous face recognition: near infra-red vs visible facial images and (4) cross-view action recognition.

4.1. Face recognition across pose variations

To study the effectiveness of our proposed approach in recognizing facial images of different poses, we perform an experiment on the CMU-PIE dataset [30]. We follow the same protocol as in [25] and use all the 68 subjects under 4 different poses and frontal illumination for this experiment. The frontal images are used as the gallery and the non-frontal images under the different poses are used as the probe images. The results of the proposed approach for this experiment are reported in Table 2.

Table 2
Rank-1 recognition accuracies (%) for face recognition across pose variations on the PIE dataset [30].

Method	c11	c29	c05	c37	Average
K-SVD [6]	48.5	76.5	80.9	57.4	65.8
Eigen light-field [10]	78.0	91.0	93.0	89.0	87.8
SGF [9]	58.8	89.7	89.7	72.1	77.6
GFK [8]	63.2	92.7	92.7	76.5	81.3
Subspace interp. via DL [25]	76.5	98.5	98.5	88.2	90.4
Proposed approach	**95.6**	**98.5**	**98.5**	**97.1**	**97.4**

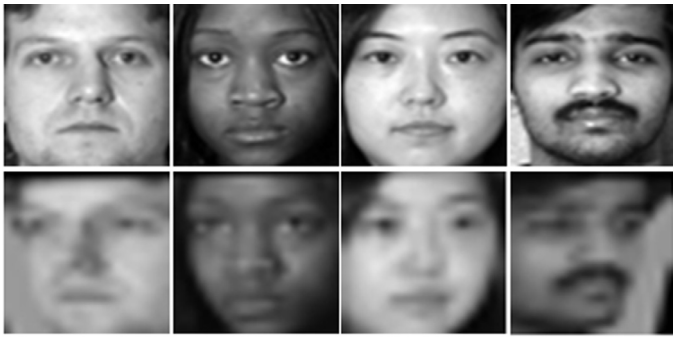

Fig. 2. Example images from the Multi-PIE data [11]. (Top row) Frontal HR images used as gallery; (Bottom row) LR images under non-frontal pose (pose 13_0, 14_0, 05_0 and 04_1 as given in the dataset) and different illumination conditions used as probe images.

Comparison with several other approaches, namely K-SVD [6], SGF [9], GFK [8] and Eigen-field approach [10] are also shown. The recognition accuracies of all the other approaches are taken directly from [25].

For this experiment, we have used 100 subjects from the Multi-PIE data whose images have been captured under very similar conditions as the PIE data for training. We see that the proposed approach performs significantly better than all the other approaches for the task of recognizing faces across pose.

4.2. Face recognition across resolution, pose, illumination

In this experiment, the gallery consists of high resolution (HR) images under frontal pose and frontal illumination, while the probe consists of low-resolution (LR) images captured under non-frontal pose and different illumination conditions.

Results on the Multi-PIE dataset: The experiments are conducted on the Multi-PIE dataset [11] which contains images of 337 subjects under different poses, illumination conditions and expressions. For our experiments, we use images under frontal pose and frontal illumination as gallery. Images taken under pose 04_1, 05_0, 13_0 and 14_0 (as indicated in the dataset) under all the 20 illuminations are used as the probe. Fig. 2 shows some sample gallery and probe images under the four different probe poses. We use a similar experimental setup as in [7] with 100 randomly chosen subjects for training and the remaining subjects for testing. HR images of size 60×55 and LR images of size 20×18 (i.e. scale factor of 3) are used for all the experiments.

SIFT descriptors computed from 15 fiducial locations (corners of eyes, nose and lip) are concatenated to form the feature vector for each image. The recognition accuracy is computed as the average rank-1 recognition over all the 20 different probe illuminations. Table 3 compares the recognition accuracy of the proposed approach with several recent approaches for four different probe poses. We observe that the proposed approach performs better than the state-of-the-art approaches for the task of matching facial images across pose, illumination and resolution.

Table 3
Rank-1 recognition accuracy of the proposed approach with HR frontal gallery images and LR non-frontal probe images for different probe poses on the Multi-PIE dataset [11]. Comparison with the other approaches are also shown.

Method	Pose 13_0	Pose 14_0	Pose 05_0	Pose 04_1	Avg.
Baseline approach [7]	57	68	65	57	62
MDS learning [7]	76	87	83	75	80
SCDL [33]	69	75	76	70	73
LSML (M_y=1) [18]	57	72	74	75	70
LSML [18]	61	75	76	77	72
GMA-LPP [29]	62	76	80	67	71
GMA-MFA [29]	73	81	84	74	78
CDL [14]	77	81	82	78	80
Proposed approach with LMNN	**82**	**89**	**87**	**81**	**85**
Proposed approach with LSML	**83**	**90**	**91**	**89**	**88**

Table 4
Rank-1 Recognition (%) of the proposed approach and comparison with existing algorithms on ChokePoint database [37].

Resolution		Algorithm															
Gallery	Probe	LPQ	SIFT	E1	E2	Fusion	MDS	CTL	HR/ LR TL (LPQ)	HR/ LR TL (SIFT)	HR/LR TL LPQ+ (SIFT)	HR/ LR CT	COTS	CTL+ MDS	CTL+ COTS	CDL	Proposed with LMNN
48 x 48	24x24	23.2	20.4	27.4	24.8	29.5	30.2	33.1	31.6	29.1	32.8	27.1	11.8	32.6	37.2	40.4	**42.0**
	16x16	17.6	14.5	21.8	19.6	24.1	26.3	28.3	25.8	23.6	26.5	22.7	4.7	27.5	31.6	33.6	**37.1**
32 x 32	24x24	20.4	14.8	23.4	18.7	24.3	28.6	31.6	26.2	25.6	29.4	21.3	16.4	30.8	35.4	36.7	**37.4**
	16x16	14.6	9.6	17.3	13.4	19.6	21.9	23.1	21.1	19.2	21.8	15.6	3.5	22.5	26.0	35.2	**39.1**
24 x 24	16x16	19.4	15.6	22.7	18.6	25.8	28.7	30.5	27.4	24.2	29.1	20.8	13.5	31.4	35.8	33.2	**38.1**

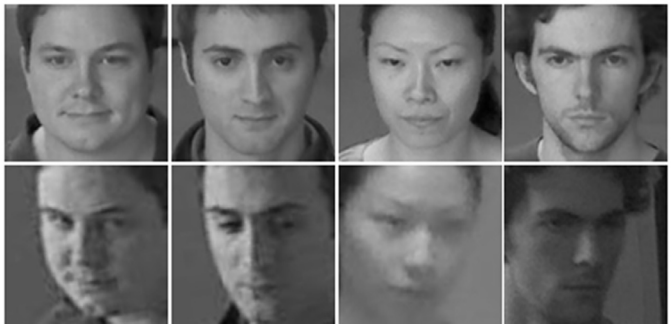

Fig. 3. Example facial images of Choke Point database [37]. Top row: frontal gallery images, second row: corresponding probe images.

Performance on surveillance quality images: In this experiment, we test the effectiveness of the proposed approach on real-surveillance quality facial images of the ChokePoint Dataset [37] which contains facial images of 29 persons captured with three cameras. The captured low resolution images have wide variations in pose, illumination and expressions (Fig. 3). We followed the same protocol as in [3] for our experiments in which we have used images of Multi-PIE dataset for training and all the images of ChokePoint database for testing. During testing, a single image per each subject is taken as gallery and randomly selected five images per each subject are used as probe images. The experiment is repeated five times with different random sampling of the subjects and the mean accuracy of the proposed approach is reported in Table 4. The results of all the other approaches (except CDL) are directly taken from [3]. We observe that the proposed approach significantly outperforms all the existing approaches for surveillance quality data. We also observed that the results of proposed approach increases by 1% if LSML is incorporated in place of LMNN.

4.3. Heterogeneous face recognition

In this section, we test the proposed approach for the task of heterogeneous face recognition in which the gallery and probe images are from different modalities, namely near infra-red (NIR) and visible (VIS) which makes the matching task very challenging.

Fig. 4. Sample images from Heterogeneous Face Biometrics database [5]. Top row: VIS face images and Bottom row: Corresponding NIR images.

We use the Heterogeneous Face Biometrics (HFB) database [5] which contains images of 100 subjects. Each subject has 4 NIR and corresponding 4 VIS light facial images (Fig. 4). We follow the same protocol as in [17] in which the images from the first 70 subjects are used for training, and the images from the remaining subjects are used for testing. The results of all the other approaches (except CDL where we used the authors code) are directly taken from [17]. In Table 5, We see that the proposed approach performs significantly better than the state-of-the-art algorithms for the task of heterogeneous face recognition.

Experiments on HFB Version 2 Database: The proposed approach is evaluated on HFB version 2 dataset [20] that contains NIR and VIS facial images of 202 persons. We followed the same protocol as that of [23] in which the VIS images are used as gallery and NIR images are used as probe. The first 100 subjects are used for training and the remaining 102 subjects are used for testing, thus there is no overlap between the training and testing subjects. The performance comparison of the proposed approach (both with LMNN and LSML) with many recent algorithms is reported in Table 6 (the results of all the other approaches are taken from the respective papers). From Table 6, we observe that, the proposed approach with LSML gives superior performance over C-CBFD +

Table 5
Performance comparison with state-of-the-art-algorithms on HFB dataset [5].

	CCA	CCA+LDA	CDFE	CSR	PLS	U-LDA	GMA	MvDA	MvDA-VC	CDL	Ours-LMNN	Ours-LSML
NIR-VIS	36.7	40.0	40.8	26.7	38.3	39.1	47.5	53.3	59.2	60.5	**70.83**	**80.9**
VIS-NIR	30.0	40.0	36.7	32.5	40.8	40.0	45.0	50.0	59.2	66.0	**70.16**	**78.0**

Table 6
Performance comparison with state-of-the-art-algorithms on HFB version 2 dataset [20]. The results of C-CBFD and DFD approaches are directly taken from [23] and [19] respectively.

Method	C-CBFD	C-CBFD + LDA	DFD (S=3)	DFD (S=5)	DFD (S=7)	C-DFD (S=3)	C-DFD (S=5)	C-DFD (S=7)	Ours with LMNN	Ours with LSML
Rank-1 Accuracy	56.6	81.8	91.5	69.3	58.8	92.2	85.2	74.5	**80.9**	**91.2**

Table 7
Recognition results on IXMAS dataset for cross-view action recognition. The proposed approach is compared against A:[1], B: [4], C: [14], D: [41]. Each row corresponds to a source view camera and each column a target view camera.

	cam0					cam1					cam2					cam3					cam4				
	A	B	C	D	Ours	A	B	C	D	Ours	A	B	C	D	Ours	A	B	C	D	Ours	A	B	C	D	Ours
cam0	-	-	-	-	-	72	75.5	75.8	75.6	**79.4**	61	64.4	74.0	78.2	**76.7**	62	67.7	63.9	78.3	**78.1**	30	66.0	72.5	70.1	**72.2**
cam1	69	75.7	76.8	76.0	**78.7**	–	–	–	–	–	64	64.2	68.2	76.0	**78.0**	68	68.1	65.4	76.6	**77.3**	41	56.0	61.1	71.3	**71.8**
cam2	62	70.3	79.0	75.8	**77.4**	67	66.3	74.2	75.0	**75.8**	–	–	–	–	–	67	71.3	81.8	79.6	**81.0**	43	62.4	66.9	74.2	**74.4**
cam3	63	73.7	71.9	77.8	**77.5**	72	65.6	64.9	75.5	**75.6**	68	71.3	77.8	77.9	**80.6**	–	–	–	–	–	44	58.0	59.9	74.1	**73.5**
cam4	51	71.3	69.4	75.4	**78.0**	55	66.3	68.9	74.1	**75.0**	51	70.9	69.7	77.6	**79.3**	53	63.5	65.9	74.2	**77.3**	-	-	-	-	-
Avg.	61.3	72.8	74.3	76.2	**77.9**	66.5	68.4	70.9	75.0	**76.5**	61	67.7	72.4	77.5	**78.7**	62.5	67.6	69.2	77.2	**78.4**	39.5	60.6	65.1	72.4	**73.0**

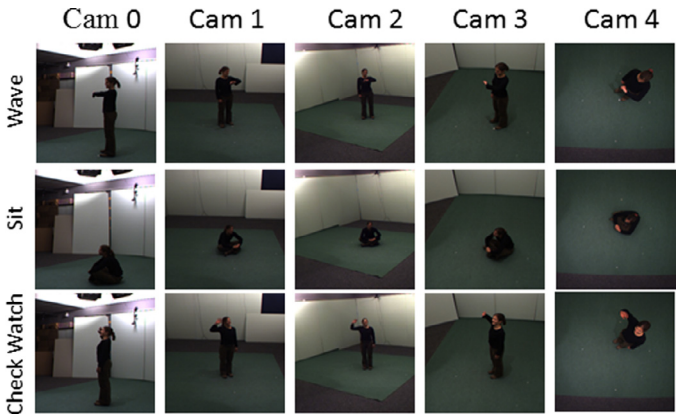

Fig. 5. Sample frames from IXMAS multi-view action dataset [2]. Each row shows one action captured across different views.

LDA but slightly worse than C-DFD (S=3). However, both the approaches [23] and [19] are designed specifically for face recognition application, whereas the proposed approach is useful for other general applications.

4.4. Cross-view activity recognition

Now, we test the effectiveness of the proposed approach for cross-view action recognition. The experiments are conducted on IXMAS multi-view action database [2] that contains action videos of 11 classes. Each action is performed by twelve actors for three times, and is captured from four side views and one top view (Fig. 5).

We use the same bag-of-features model as in [14] and [41], where the features in both the source and target domains are of 1000 dimensions. Following the same protocol with partially labeled data, first the unlabeled pair data is used to find all parameters (except the mapping **L**) of the algorithm using (2). The partially labeled data of the source domain is then used to learn the mapping function which is then used for transforming the sparse coefficients computed from the target data. The performance of the proposed approach along with comparisons with other approaches is reported in Table 7. All the other numbers are taken from their respective papers. We see that the proposed approach performs favorably with respect to the state-of-the-art approaches.

5. Computational cost

The details of the computational complexity of the proposed approach is discussed in this section. In our experiment on face recognition across resolution, pose and illumination (Section 4.2 - MultiPIE), the proposed approach takes 0.07 s to compare a probe image with one gallery image. Since there are 237 gallery subjects, total time taken to identify the probe is around 17 s. For the same experiment, the training stage which includes extracting sift features, learning dictionaries (D_1 and D_2), transformation matrices (U_1 and U_2) and mapping function takes around 27 min. Note that the training can be done offline. All the experiments are conducted using an unoptimized Matlab code on a i7 PC.

6. Summary and discussion

In this work, we have proposed a joint discriminative dictionary and transformation learning approach for the task of matching cross domain data. We propose an iterative solution for updating the coupled dictionaries, transformation matrices and the mapping function. Extensive experiments on different datasets and comparisons with the state-of-the-art approaches justify the effectiveness of the proposed approach.

References

[1] A. Farhadi, M. Tabrizi, Learning to recognize activities from the wrong view point, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 154–166.
[2] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Comput. Vis. Image Underst. 104 (2) (2006) 249–257.
[3] H. Bhatt, R. Singh, M. Vatsa, N. Ratha, Improving cross-resolution face matching using ensemble based co-transfer learning, IEEE Trans. Image Process. 23 (12) (2014) 5654–5669.

[4] J. Liu, M. Shah, B. Kuipers, S. Savarese, Cross-view action recognition via view knowledge transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 3209–3216.
[5] S.Z. Li, Z. Lei, M. Ao, The hfb face database for heterogeneous face biometrics research, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2009, pp. 1–8.
[6] M. Aharon, M. Elad, A. Bruckstein, K-svd: An algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311–4322.
[7] S. Biswas, G. Aggarwal, P.J. Flynn, K.W. Bowyer, Pose-robust recognition of low-resolution face images, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 3037–3049.
[8] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2066–2073.
[9] R. Gopalan, R. Li, R. Chellappa, Domain adaptation for object recognition: an unsupervised approach, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 999–1006.
[10] R. Gross, I. Matthews, S. Baker, Appearance-based face recognition and light-fields, IEEE Trans. Pattern Anal. Mach. Intell. 26 (4) (2004) 449–465.
[11] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Guide to the cmu multi-pie database, Technical report, Carnegie Mellon University, 2007.
[12] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3-4) (1936) 321–377.
[13] C.A. Hou, M.C. Yang, Y.C. Wang, Domain adaptive self taught learning for heterogeneous face recognition, in: Proceedings of the IEEE International Conference on Pattern Recognition, 2014, pp. 3068–3073.
[14] D.A. Huang, Y.C.F. Wang, Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2496–2503.
[15] Z. Jiang, Z. Lin, L.S. Davis, Label consistent k-svd: learning a discriminative dictionary for recognition., IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2651–2664.
[16] Y. Jin, J. Lu, Q. Ruan, Coupled discriminative feature learning for heterogeneous face recognition, IEEE Trans. Inf. Forensics Secur. 10 (3) (2015) 640–652.
[17] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 38 (1) (2016) 188–194.
[18] M. Kostinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.
[19] Z. Lei, M. Pietikainen, S. Li, Learning discriminant face descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 289–302.
[20] S.Z. Li, D. Yi, Z. Lei, S. Liao, The casia nir vis 2.0 face database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 348–353.
[21] W. Li, L. Duan, D. Xu, I.W. Tsang, Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation, IEEE Trans. Pattern Anal. Mach. Intell. 36 (6) (2014) 1134–1148.
[22] L. Liu, L. Shao, X. Li, K. Lu, Learning spatio-temporal representations for action recognition: a genetic programming approach, IEEE Trans. Cybern. 46 (1) (2016) 158–170.
[23] J. Lu, V. Liong, X. Zhou, J. Zhou, Learning compact binary face descriptor for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2041–2056.
[24] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online dictionary learning for sparse coding, in: Proceedings of the International Conference on Machine Learning, 2009.
[25] J. Ni, Q. Qiu, R. Chellappa, Subspace interpolation via dictionary learning for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 692–699.
[26] M. Nishiyama, H. Takeshima, J. Shotton, T. Kozakaya, O. Yamaguchi, Facial deblur inference to improve recognition of blurred faces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1115–1122.
[27] I. Ramirez, P. Sprechmann, G. Sapiro, Classification and clustering via dictionary learning with structured incoherence and shared features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3501–3508.
[28] R. Rosipal, N. Kramer, Overview and recent advances in partial least squares, Subspace Latent Struct. Feature Sel. (2006) 34–51.
[29] A. Sharma, A. Kumar, H. Daume, D.W. Jacobs, Generalized multiview analysis: a discriminative latent space, in: Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, 2012, pp. 2160–2167.
[30] T. Sim, S. Baker, M. Bsat, The cmu pose, illumination and expression database, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1615–1618.
[31] H. Wang, C. Yuan, W. Hu, H. Ling, W. Yang, C. Sun, Action recognition using nonnegative action component representation and sparse basis selection, IEEE Trans. Image Process. 23 (2) (2014) 570–581.
[32] J. Wang, X. Nie, Y. Xia, Y. Wu, S.C. Zhu, Cross-view action modeling, learning, and recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014b, pp. 2649–2656.
[33] S. Wang, D. Zhang, Y. Liang, Q. Pan, Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis, in: Proceedings of the IEEE International Conference on Computer Vision, 2012, pp. 2216–2223.
[34] K.Q. Weinberger, L.K. Saul, Fast solvers and efficient implementations for distance metric learning, in: Proceedings of the IEEE International Conference on Machine Learning, 2008, pp. 1160–1167.
[35] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[36] X. Wu, H. Wang, C. Liu, Y. Jia, Cross-view action recognition over heterogeneous feature spaces, IEEE Trans. Image Process. 24 (11) (2015) 4096–4108.
[37] Y. Wong, S. Chen, S. Mau, C. Sanderson, B.C. Lovell, Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition, in: Proceedings of the Computer Vision and Pattern Recognition Workshops, 2011, pp. 74–81.
[38] M. Yang, D. Dai, L. Shen, L.V. Gool, Latent dictionary learning for sparse representation based classification., in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 4138–4145.
[39] M. Yang, L. Van, L. Zhang, Sparse variation dictionary learning for face recognition with a single training sample per person., in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 689–696.
[40] B.Z. Yao, B.X. Nie, Z. Liu, S.C. Zhu, Animated pose templates for modeling and detecting human actions, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 436–452.
[41] Y.R. Yeh, C.H. Huang, Y.C.F. Wang, Heterogeneous domain adaptation and classification by exploiting the correlation subspace, IEEE Trans. Image Process. 23 (5) (2014) 2009–2018.
[42] P.H.H. Yeomans, S. Baker, B.V.K.V. Kumar, Simultaneous super-resolution and feature extraction for recognition of low-resolution faces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[43] H. Zhang, J. Yang, Y. Zhang, N.M. Nasrabadi, T. Huang, Close the loop: joint blind image restoration and recognition with sparse representation prior, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 770–777.
[44] Q. Zhang, B. Li, Discriminative k-svd for dictionary learning in face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2691–2698.
[45] J. Zheng, Z. Jiang, Learning view-invariant sparse representations for cross-view action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3176–3183.
[46] J.Y. Zhu, W.S. Zheng, J.H. Lai, S.Z. Li, Matching nir face to vis face using transduction, IEEE Trans. Inf. Forensics Secur. 9 (3) (2014) 501–514.
[47] W.W.W. Zou, P.C. Yuen, Very low resolution face recognition problem, IEEE Trans. Image Process. 21 (1) (2012) 327–340.