Neurocomputing 151 (2015) 1500–1506

Contents lists available at ScienceDirect

Neurocomputing

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

Embedding metric learning into set-based face recognition for video surveillance

Guijin Wang [a,*], Fei Zheng [a], Chenbo Shi [a], Jing-Hao Xue [b], Chunxiao Liu [a], Li He [a]

[a] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[b] Department of Statistical Science, University College London, London WC1E 6BT, UK

ARTICLE INFO

Article history:
Received 1 April 2014
Received in revised form
27 August 2014
Accepted 11 October 2014
Communicated by J. Zhang
Available online 23 October 2014

Keywords:
Face recognition
Distance metric learning
Set-based matching
Intelligent surveillance

ABSTRACT

Face recognition in video surveillance is a challenging task, largely due to the difficulty in matching images across cameras of distinct viewpoints and illuminations. To overcome this difficulty, this paper proposes a novel method which embeds distance metric learning into set-based image matching. First we use sets of face images, rather than individual images, as the input for recognition, since in surveillance systems the former is a more natural way. We model each image set using a convex-hull space spanned by its member images and measure the dissimilarity of two sets using the distance between the closest points of their corresponding convex hull spaces. Then we propose a set-based distance metric learning scheme to learn a feature-space mapping to a discriminative subspace. Finally we project image sets into the learned subspace and achieve face recognition by comparing the projected sets. In this way, we can adapt the variation in viewpoints and illuminations across cameras in order to improve face recognition in video surveillance. Experiments on the public Honda/UCSD and ChokePoint databases demonstrate the superior performance of our method to the state-of-the-art approaches.

1. Introduction

Because of the strong demand for public security in recent years, camera networks of intelligent surveillance have been distributed all over the world. Many new issues are raised [19,20]. As one of the key technologies of intelligent surveillance, face recognition in surveillance has attracted growing interests [1–5].

In video surveillance we often need to compare face images from different cameras. For example, given a video that records a subject walking through a camera's view, we are often required to retrieve the subject from the videos captured by other cameras in a relevant camera network. However, due to the variation in the environments (e.g. viewpoints and illuminations) of the cameras, the appearances of a subject that captured by different cameras are quite different, making face recognition a challenging task.

On the other hand, many researches have shown that although single-image-based face recognition algorithms can perform well on controlled environments, their performances decrease dramatically in surveillance contexts [6]. This phenomenon has motivated the development of algorithms that make use of sets of images, rather than only individual images, that are provided by videos, to compensate for the poor viewing conditions [7]. Indeed recent development in set-based recognition has shown its excellent promise [2,5].

Although set-based methods are more promising than single-image-based methods, they also face challenges when image sets are captured by different cameras. One big challenge is the fact that, due to the large variation in viewpoints and illuminations, the similarity of image sets of the same subject becomes low and sometimes even lower than the similarity between sets of different subjects. This largely increases the misclassification errors. To overcome this challenge, a natural solution is to learn a mapping which increases the similarity of the sets of the same subject from different cameras while reducing the similarity between the sets of different subjects. However, most of such learning schemes are based on individual images.

In this paper, we propose a novel distance metric learning scheme based on image sets. Our novelty and contribution is threefold. First, our scheme aims to learn a feature-space mapping to adapt the variation in viewpoints and illuminations between cameras. The learning procedure is an extension of the large margin nearest neighbor (LMNN) [8] to image sets (LMNN was based on individual images). Secondly, although we adopt the convex hull model of [2] to represent an image set, our scheme is different from the CHISD method of [2] in that we use the learned feature-space mapping to project all face sets into a discriminative feature subspace for face recognition. Compared with the original

* Corresponding author.
E-mail addresses: wangguijin@tsinghua.edu.cn (G. Wang), zhengfei.thu@gmail.com (F. Zheng), shichenbo@gmail.com (C. Shi), jinghao.xue@ucl.ac.uk (J.-H. Xue), lcx08@mails.tsinghua.edu.cn (C. Liu), l-he10@mails.tsinghua.edu.cn (L. He).

http://dx.doi.org/10.1016/j.neucom.2014.10.032
0925-2312/

feature space, this subspace is designed to make the distances between sets of the same subject shorter and the distances between sets of different subjects longer. Thirdly, we shall use various real datasets to illustrate that, for video surveillance, face recognition using the learned feature subspace is better than that using the original feature space.

2. Related work

There are two main elements in set-based face recognition: (1) the model to represent the image sets, and (2) the distance to measure the similarity between sets. Existing set-based methods have different emphases on these two elements.

Some methods focus on one of the elements. For example, Chen et al. [3] focus on the representation model, utilizing the mean feature of a set to represent set; Stallkamp et al. [1] focus on the distance measures, manually designing three metrics that weight the distances of images between sets.

In contrast, many researches tried to combine the two elements together. Yamaguchi et al. [9] proposed a mutual subspace-based method, which uses linear subspaces to model image sets and measures the similarity between subspaces with canonical angles. Fukui et al. [10] further developed this method by projecting faces into a constraint subspace. Cevikalp and Triggs [2] claimed that models based on the affine or convex hull subspaces were more discriminative than those based on linear subspaces. In their method, the affine or convex hull is used to model image sets and a geometric closest distance is incorporated. Hu et al. [5] extended this approach by embedding sparsity into the affine hull model. Methods using nonlinear approximations of face appearances, typically with locally linear and globally nonlinear models, were also developed [4].

The methods above applied the set model to the original feature space. All the dimensions of feature were equality weighted when computing the similarity between sets, which may not be suitable when sets were collected under different conditions [11]. To tackle this problem, researchers embedded distance metric learning into set matching to explore the discriminative directions in the feature space. Lu et al. [12] explored multiple order statistics as features of image sets and proposed a localized multi-kernel learning scheme, which involved various order statistics information. Mian et al. [13] first added constrains of self-regularization and non-negativity into affine hull subspaces, and then adopted LMNN [8] to learn a Mahalanobis distance by using the nearest points between a query set and all the gallery sets. It needs to learn a distance metric only for recognizing this query set, which increases the computation costs of recognition. Based on the individual feature points, the training methods in [12,13] still optimized the distance metric between two points. To optimize the distance metric between two sets, Wu et al. [11] learned a feature space projection based on discriminative ranking, to enhance the discrimination between sets of different subjects. In the same spirit with Wu et al. [11] to learn a discriminative mapping but using different techniques, we propose a novel subspace-learning scheme based on distance metric learning, which is easier to implement and also effective.

3. The proposed method

The framework of our proposed method is shown in Fig. 1. We use a convex hull model to represent an image set and use the closest distance between convex hull models to measure the similarity between sets. In the training phase, we learn a set-based discriminative subspace to adapt the variation across cameras. In the recognition phase, face images are projected into the learned subspace and the final identity of the probe set is established by a nearest neighbor classifier.

3.1. Image set model

Denote an image set by $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_n]$, where $\mathbf{x}_i \in R^d$ represents a d-dimensional feature vector of the ith face image in set $\mathbf{X}$, $i=1,\ldots,n$, and n is the number of images in the set. The feature of a face image can simply be the raw pixel values or the Local Binary Pattern (LBP) [14] of the image.

As with [2], we utilize the convex hull $H(\mathbf{X})$ to model image set $\mathbf{X}$:

$$H(\mathbf{X})=\left\{\mathbf{x}:\mathbf{x}=\sum_{i=1}^{n}a_i\mathbf{x}_i\,\middle|\,\sum_{i=1}^{n}a_i=1,\ 0\le a_i\le 1\right\}. \tag{1}$$

Let $\mathbf{a}=[a_1,\ldots,a_n]^T$, (1) can be written in a compact form:

$$H(\mathbf{X})=\left\{\mathbf{x}:\mathbf{x}=\mathbf{X}\mathbf{a}\,\middle|\,\sum_{i=1}^{n}a_i=1,\ 0\le a_i\le 1\right\}. \tag{2}$$

The convex hull synthesizes new instances by using convex combinations of the feature vectors in the set.

Given two sets $\mathbf{X}_i$ and $\mathbf{X}_j$, their distance can be defined as the distance of the closest points between their convex hulls:

$$d(\mathbf{X}_i,\mathbf{X}_j)=\min_{\mathbf{x}_1\in H(\mathbf{X}_i),\mathbf{x}_2\in H(\mathbf{X}_j)}\|\mathbf{x}_1-\mathbf{x}_2\|_2. \tag{3}$$

It follows that we can obtain this distance through a convex optimization:

$$(\hat{\mathbf{a}}_i,\hat{\mathbf{a}}_j)=\arg\min_{\mathbf{a}_i,\mathbf{a}_j}\|\mathbf{X}_i\mathbf{a}_i-\mathbf{X}_j\mathbf{a}_j\|_2^2,$$
$$\text{s.t.}\ \sum_{k=1}^{n_i}a_{ik}=1=\sum_{k'=1}^{n_j}a_{jk'},\quad 0\le a_{ik},a_{jk'}\le 1. \tag{4}$$

Then, the distance between $\mathbf{X}_i$ and $\mathbf{X}_j$ becomes

$$d(\mathbf{X}_i,\mathbf{X}_j)=\|\mathbf{X}_i\hat{\mathbf{a}}_i-\mathbf{X}_j\hat{\mathbf{a}}_j\|_2. \tag{5}$$

3.2. Learning of distance metric

In this section, we propose our learning scheme based on the convex hull model. The basic concept of the learning scheme is illustrated in Fig. 2.

Suppose the probe set $\mathbf{P}$ and an image set $\mathbf{X}_1$ are from the same subject that is different from the subject of another image set $\mathbf{X}_2$. In the original feature space, due to the large variation in viewpoints and illustration, $\mathbf{P}$ may be further away from $\mathbf{X}_1$ than from $\mathbf{X}_2$, resulting in misclassification of $\mathbf{P}$. Our aim is to learn a projection matrix $\mathbf{L}$, by which we can project the features into a learned space where $\mathbf{P}$ can be correctly classified, as shown in Fig. 2(b). More specifically, we want to ensure that the squared distance between sets of different subjects is larger than the squared distance between sets of the same subject plus one, such that the recognition performance can be improved. This idea is inspired by the work of [8], which was based on individual images. Here we extend the idea to image sets as follows.

Let $\{(\mathbf{X}_c,y_c)\}_{c=1}^{C}$ denote a training collection of C labeled image sets, where $\mathbf{X}_c$ represents an image set and y_c is the label of set $\mathbf{X}_c$. We want to learn a projection matrix $\mathbf{L}$ to compute the distance between $\mathbf{X}_i$ and $\mathbf{X}_j$ in the projected space:

$$d_{\mathbf{L}}(\mathbf{X}_i,\mathbf{X}_j)=\min_{\mathbf{x}_1\in H(\mathbf{X}_i),\mathbf{x}_2\in H(\mathbf{X}_j)}\|\mathbf{L}(\mathbf{x}_1-\mathbf{x}_2)\|_2. \tag{6}$$

For each set $\mathbf{X}_i$ in the training collection $\{(\mathbf{X}_c,y_c)\}_{c=1}^{C}$, we define "target neighbor" sets of $\mathbf{X}_i$ as $N(\mathbf{X}_i)$: if $\mathbf{X}_j$ shares the same label with $\mathbf{X}_i$ ($j\neq i$), we write as $\mathbf{X}_j\in N(\mathbf{X}_i)$. We also define "imposter" sets of $\mathbf{X}_i$ as $Im(\mathbf{X}_i)$: if a set $\mathbf{X}_k\notin N(\mathbf{X}_i)$ ($k\neq i$) and

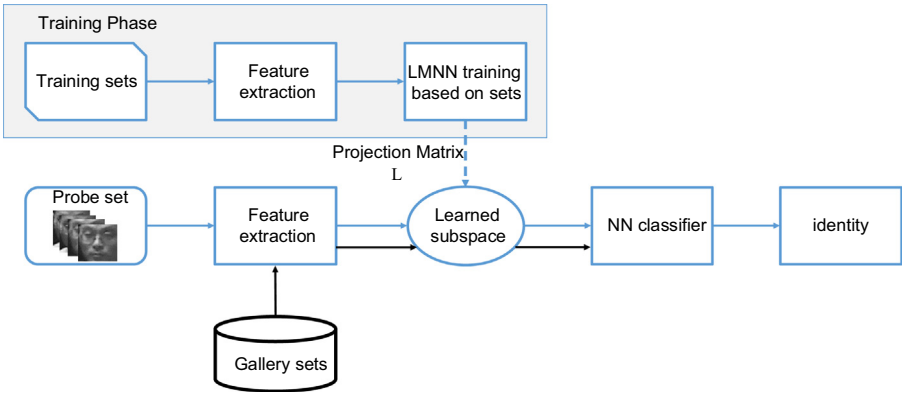


Fig. 1. Diagram of the proposed method.

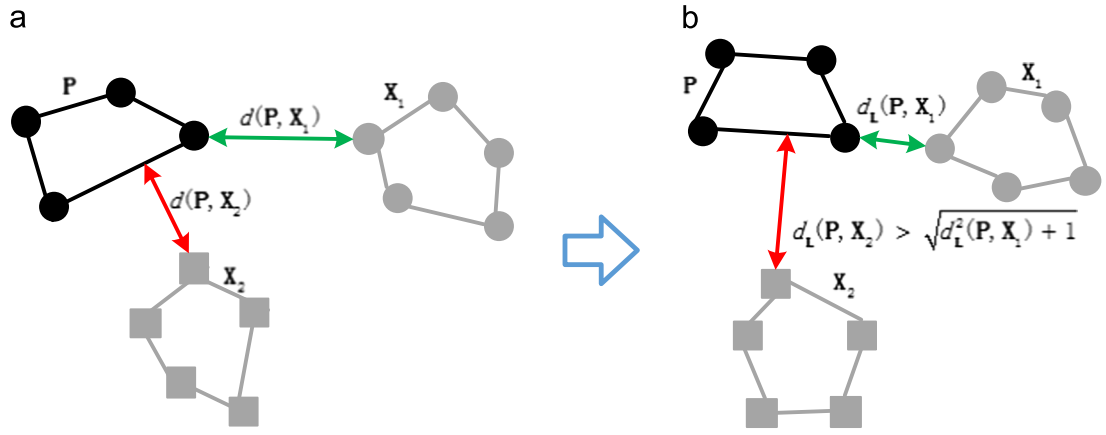


Fig. 2. Illustration of the basic concept of the proposed learning scheme. Suppose image sets $\mathbf{P}$ and $\mathbf{X}_1$ are from the same subject that is different from the subject of image set $\mathbf{X}_2$. The scheme aims to make the distance between sets of different subjects ($\mathbf{P}$ and $\mathbf{X}_2$) larger than the distance between sets of the same subject ($\mathbf{P}$ and $\mathbf{X}_1$).

$d^2_{\mathbf{L}}(\mathbf{X}_i,\mathbf{X}_k) < \max_{\mathbf{X}_j \in N(\mathbf{X}_i)} d^2_{\mathbf{L}}(\mathbf{X}_i,\mathbf{X}_j)+1$, set $\mathbf{X}_k$ is an "imposter" set of $\mathbf{X}_i$, which is written as $\mathbf{X}_k \in Im(\mathbf{X}_i)$.

Our objective is that, in the $\mathbf{L}$-projected feature space with the distance defined in (6), we should preserve as many "target neighbor" sets of $\mathbf{X}_i$ as possible in the nearest neighborhood of $\mathbf{X}_i$, while at the same time we should remove as many "imposter" sets as possible from the neighborhood. A loss function for this objective can be formulated as

$$\epsilon(\mathbf{L}) = (1-\lambda)\sum_{i=1}^{C}\sum_{\mathbf{X}_j \in N(\mathbf{X}_i)} d^2_{\mathbf{L}}(\mathbf{X}_i,\mathbf{X}_j) + \lambda\sum_{i=1}^{C}\sum_{\mathbf{X}_k \in Im(\mathbf{X}_i)}\left[1+\max_{\mathbf{X}_j \in N(\mathbf{X}_i)} d^2_{\mathbf{L}}(\mathbf{X}_i,\mathbf{X}_j) - d^2_{\mathbf{L}}(\mathbf{X}_i,\mathbf{X}_k)\right], \quad (7)$$

where $\lambda \in (0,1)$ is a constant. The first term of (7) penalizes large distances between sets and their "target neighbor" sets, and the second term penalizes small distance between sets and their "imposter" sets.

To minimize the loss function (7), we rewrite the square of (6) as

$$d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_j) = \min_{\mathbf{x}_1 \in H(\mathbf{X}_i), \mathbf{x}_2 \in H(\mathbf{X}_j)} (\mathbf{x}_1-\mathbf{x}_2)^T\mathbf{M}(\mathbf{x}_1-\mathbf{x}_2), \quad (8)$$

where matrix $\mathbf{M} = \mathbf{L}^T\mathbf{L}$ is a positive semi-definite matrix, and rewrite the loss function as a function of $\mathbf{M}$:

$$\epsilon(\mathbf{M}) = (1-\lambda)\sum_{i=1}^{C}\sum_{\mathbf{X}_j \in N(\mathbf{X}_i)} d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_j) + \lambda\sum_{i=1}^{C}\sum_{\mathbf{X}_k \in Im(\mathbf{X}_i)}\left[1+\max_{\mathbf{X}_j \in N(\mathbf{X}_i)} d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_j) - d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_k)\right]. \quad (9)$$

That is, we learn $\mathbf{M}$ through

$$\arg\min_{\mathbf{M}} \epsilon(\mathbf{M}), \quad \text{s.t. } \mathbf{M} \succcurlyeq 0. \quad (10)$$

To solve (10) we design an iterative algorithm, as illustrated in Table 1 and described as follows.

When $\mathbf{M}$ is fixed, we obtain $d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_j)$ by solving the following convex optimization problem:

$$(\hat{\mathbf{a}}_i,\hat{\mathbf{a}}_j) = \arg\min_{\mathbf{a}_i,\mathbf{a}_j} (\mathbf{X}_i\mathbf{a}_i - \mathbf{X}_j\mathbf{a}_j)^T\mathbf{M}(\mathbf{X}_i\mathbf{a}_i - \mathbf{X}_j\mathbf{a}_j),$$
$$\text{s.t.} \sum_{k=1}^{n_i} a_{ik} = 1 = \sum_{k'=1}^{n_j} a_{jk'}, \quad 0 \le a_{ik}, a_{jk'} \le 1. \quad (11)$$

That is, $d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_j)$ is obtained as

$$d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_j) = (\mathbf{X}_i\hat{\mathbf{a}}_i - \mathbf{X}_j\hat{\mathbf{a}}_j)^T\mathbf{M}(\mathbf{X}_i\hat{\mathbf{a}}_i - \mathbf{X}_j\hat{\mathbf{a}}_j). \quad (12)$$

With all pairs of $d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_j)$ obtained, we can compute $\epsilon(\mathbf{M})$ in (9).

To further update $\mathbf{M}$, we compute the partial derivatives

$$\frac{\partial d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_j)}{\partial \mathbf{M}} = (\mathbf{X}_i\hat{\mathbf{a}}_i - \mathbf{X}_j\hat{\mathbf{a}}_j)(\mathbf{X}_i\hat{\mathbf{a}}_i - \mathbf{X}_j\hat{\mathbf{a}}_j)^T \quad (13)$$

and then the gradient

$$\frac{\partial \epsilon(\mathbf{M})}{\partial \mathbf{M}} = (1-\lambda)\sum_{i=1}^{C}\sum_{\mathbf{X}_j \in N(\mathbf{X}_i)} \frac{\partial d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_j)}{\partial \mathbf{M}} + \lambda\sum_{i=1}^{C}\sum_{\mathbf{X}_k \in Im(\mathbf{X}_i)}\left[\frac{\partial \max_{\mathbf{X}_j \in N(\mathbf{X}_i)} d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_j)}{\partial \mathbf{M}} - \frac{\partial d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_k)}{\partial \mathbf{M}}\right]. \quad (14)$$

When all pairs of $d^2_{\mathbf{M}}(\mathbf{X}_i,\mathbf{X}_j)$, $\epsilon(\mathbf{M})$ and $\partial\epsilon(\mathbf{M})/\partial\mathbf{M}$ are computed, $\mathbf{M}$ can be updated by $\mathbf{M} + s\partial\epsilon(\mathbf{M})/\partial\mathbf{M}$, where s is a step-size initialized with a sufficiently small value and updated by a backtracking

method as with [8]. After $\mathbf{M}$ is updated, all pairs of $d^2_{\mathbf{M}}(\mathbf{X}_i, \mathbf{X}_j)$, $\epsilon(\mathbf{M})$ and $\partial\epsilon(\mathbf{M})/\partial\mathbf{M}$ are updated further. This iterative procedure continues until no further reduction in loss can be gained.

In the iterative procedure, we first compute the nearest distance between each pair of two image sets with a fixed $\mathbf{M}$. Then, the gradient method is used to find the direction of the steepest descent. This procedure finally converges. Fig. 3 shows the loss function values $\epsilon(\mathbf{M})$ over iterations in one of our experiments on the public Honda/UCSD [15] dataset.

Once $\mathbf{M}$ is learned, we do eigendecomposition of $\mathbf{M}$ to obtain L: first

$$\mathbf{M} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T, \tag{15}$$

Table 1
The iterative algorithm for learning $\mathbf{M}$.

Input: $\{(\mathbf{X}_c, y_c)\}_{c=1}^{C}$, λ, s
Output: $\mathbf{M}$

1	Initialize $\mathbf{M}=\mathbf{I}$, $t=0$
2	While $t <$ maximum number of iterations
3	Compute all pairs of $d^2_{\mathbf{M}}(\mathbf{X}_i, \mathbf{X}_j)$ by (11) and (12)
4	Compute $\epsilon(\mathbf{M})$ by (9) and $\frac{\partial\epsilon(\mathbf{M})}{\partial\mathbf{M}}$ by (14)
5	$O_t = \epsilon(\mathbf{M})$
6	If $t>1$ and $\frac{\lvert O_t - O_{t-1}\rvert}{O_t} <$ tolerance
7	Break
8	End if
9	Update $\mathbf{M}$ with $\mathbf{M} + s\frac{\partial\epsilon(\mathbf{M})}{\partial\mathbf{M}}$
10	Update s with a backtracking method
11	$t = t+1$
12	End while

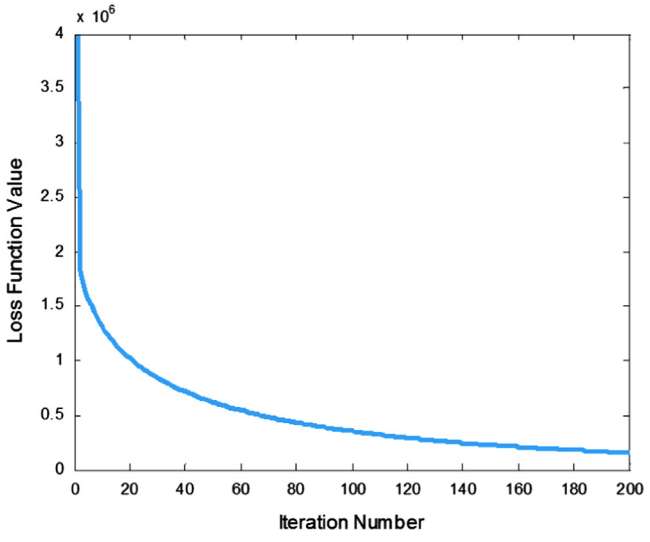


Fig. 3. Illustration of convergence of our learning method.

where the columns of $\mathbf{A}$ are eigenvectors and $\mathbf{\Lambda}$ is a diagonal matrix whose diagonals are the corresponding eigenvalues, then $\mathbf{L}$ is obtained as

$$\mathbf{L} = \mathbf{\Lambda}^{1/2}\mathbf{A}^T. \tag{16}$$

In the recognition phase, we carry out the nearest neighbor classification to identify the label of probe image sets. That is, given a probe set $\mathbf{P}$ and gallery sets $\{(\mathbf{G}_c, y_c)\}_{c=1}^{C}$, $\mathbf{P}$ will be labeled by $y_{\hat{c}}$ of the $\mathbf{G}_{\hat{c}}$ closest to $\mathbf{P}$ in the projected subspace: $y_{\hat{c}}$ with $\hat{c} = \arg\min_c d_{\mathbf{L}}(\mathbf{P}, \mathbf{G}_c)$.

4. Experiments

We evaluate our proposed method on two public databases of faces, Honda/UCSD [15] and ChokePoint [16], and compare it with some state-of-the-art set-based methods including CHISD [2] as the baseline.

System parameters: There are two parameters we should consider: the penalty factor λ in (9) and the initial step-size s to update $\mathbf{M}$. We set the penalty factor λ to 0.5 to balance the cost of "target neighbor" sets and the cost of "imposter" sets. The initial step-size s needs to be a sufficiently small value for convergence; we set it to 10^{-9}. Parameters are set without further tuning. In addition, we set the maximum iteration number of each learning process to 100.

4.1. Experiments on the Honda/UCSD dataset

The Honda/UCSD database is a widely used database on video-based face recognition. It consists of 59 video sequences, involving 20 individuals. The length of each video sequence varies from 13 to 782. Each sequence contains large pose variations and moderate expression variations. We used the version from [17], which had the faces detected by [18] and resized to gray-scale images of size 20×20 with histograms equalized. As the images are of low resolution and large pose variations, they are very similar to those captured in video surveillance.

The methods termed AHISD [2], CHISD [2], SNAP [5], $\text{SBDR}_{\text{CHISD}}$ [11] and $\text{SBDR}_{\text{SNAP}}$ [11] have achieved the state-of-the-art results on the Honda/UCSD database. We compared our method with them. In addition, we compared the original LMNN training scheme (termed **LMNN+CHISD**) with ours. In LMNN+CHISD, a projection matrix was obtained by following the original LMNN method and setting in [8]. Then, the projection matrix was utilized to compute the distance between sets as we did in (6).

For a fair comparison, our experiments on the Honda/UCSD database strictly followed the settings in [11]. The standard training/test configuration was used: 20 sequences as gallery and the remaining 39 sequences as probe. The raw pixel value vector was used as the feature for each image. The strategy of [11] on constructing the training collection $\{(X_c, y_c)\}_{c=1}^{C}$ and test sets were also followed: 50/100 frames randomly sampled from each sequence to form a training image set $\mathbf{X}_i$, and 50/100 frames from the remaining frames of the sequence for testing. For those short sequences which do not have enough frames, we sampled them as many as possible while making

Fig. 4. The training and test sets of a sequence. Each set contains 50 frames. Large pose variations in sets can be observed.

the training and test sets balanced. An example of training and test sets of a sequence is depicted in Fig. 4.

The results in Table 2 are the average recognition rates over 10 trials, such that the influence of random sampling can be largely reduced.

Compared with the state-of-the-art set-based methods (AHISD, CHISD and SNAP) that do not learn projection matrices, the performance of our proposed method was better. In particular, compared with the baseline method CHISD, our performance was much superior, indicating the discriminative power of the feature subspace learned by our method.

Compared with the original learning scheme of LMNN, our method was more effective and achieved better performance. Compared with the two SBDR methods which embed a ranking-based learning scheme into set-based matching, we achieved the best result when using 50 frames to construct a set and the second best result when using 100 frames to construct a set. It should be noticed that the performance of computing SBDR in CHISD was lower than ours. Only the performance of the method that computed SBDR in SNAP and used 100 frames to constructed a set was slightly better than ours. This is because SNAP not only uses hulls to match but also uses the samples obtained to formulate a sparse representation to help recognition. However, it needs a long observation to make the sparse representation complete, which is not always possible in real video surveillance due to the track loss in tracking.

Table 2
Average recognition rate (%) on the Honda/UCSD dataset. The star indicates that the results are slightly different from those in [11] because the random sampling results may be different.

Methods	50 samples	100 samples
AHISD [2]	91.54	92.31
CHISD [2]	91.28	92.31
SANP [5]	91.03	92.31
$SBDR_{CHISD}$ [11]	96.41	95.73
$SBDR_{SANP}$ [11]	95.64	**97.95**
LMNN+CHISD	93.59	94.62
Our proposal	**98.72**	96.67

4.2. Experiments on the chokepoint dataset

The ChokePoint [16] database is designed for evaluating face recognition algorithms under real-world surveillance conditions. A subset of the video sequences (S1) from three cameras (C1, C2, C3) in one portal (P1) in two directions (E and L) was used: P1E_S1_C1, P1L_S1_C2 and P1L_S1_C3. Each sequences recorded 25 subjects walking through a portal in turn, within about 30–60 frames for each subject. If the directions (E and L) are different, the views of the cameras are non-overlapping. Thus, the view of P1E_S1_C1 is different from the views of P1L_S1_C2 and P1L_S1_C3. The face images are detected and aligned by a commercial system in [16]. Samples of them are shown in Fig. 5, from which we can observe apparent variation in viewpoints and illuminations of different cameras.

For this database, we evaluated the performances of our proposed method in dealing with the variation in viewpoints and illuminations of different cameras. We used two sequences from different cameras as gallery and probe, respectively. All images of a subject in a sequence were used to construct an image set of the subject. The face images were resized to 40×40. We partitioned the resized image into non-overlapping blocks of size 8×8 and extracted the LBP histograms from each block to concatenate a long feature vector as the representation of the image.

Because SBDR and SANP have not reported results for this database, we just compared our method with the methods, AHISD and CHISD, which are also the state-of-the-art set-based methods. We also compared our method with LMNN+CHISD: train a projection matrix by LMNN and used it to compute the distance between sets.

To construct the training collection $\{(X_c, y_c)\}_{c=1}^{C}$, we randomly selected 5 subjects' sets from two cameras for training, and the sets of the remaining 20 subjects were for testing. Two experiments were performed: one used sequence P1E_S1_C1 as gallery and P1L_S1_C2 as probe, and the other used P1E_S1_C1 as gallery and P1L_S1_C3 as probe. This setting is to ensure that the views of gallery sequences and probe sequences are different.

Fig. 6 shows the Cumulative-Matching-Curve (CMC) results of the two experiments, each of them was the averaged results over 10 trails. The CMC represents the possibilities of finding the correct match in top rank n matches.

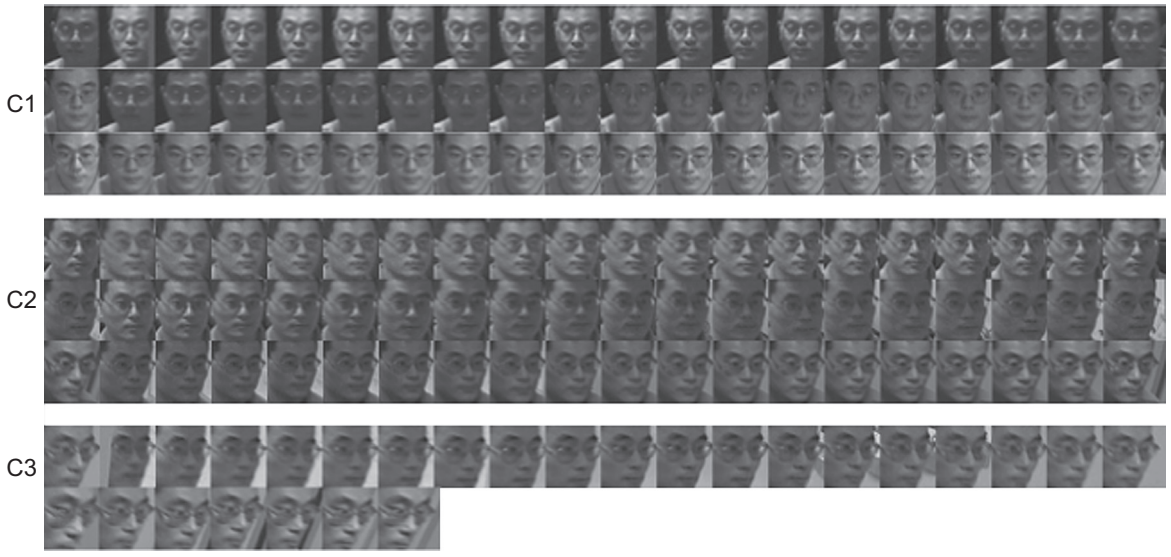


Fig. 5. Sets from three sequences of one subject: C1 from P1E_S1_C1, C2 from P1L_S1_C2, and C3 from P1L_S1_C3.

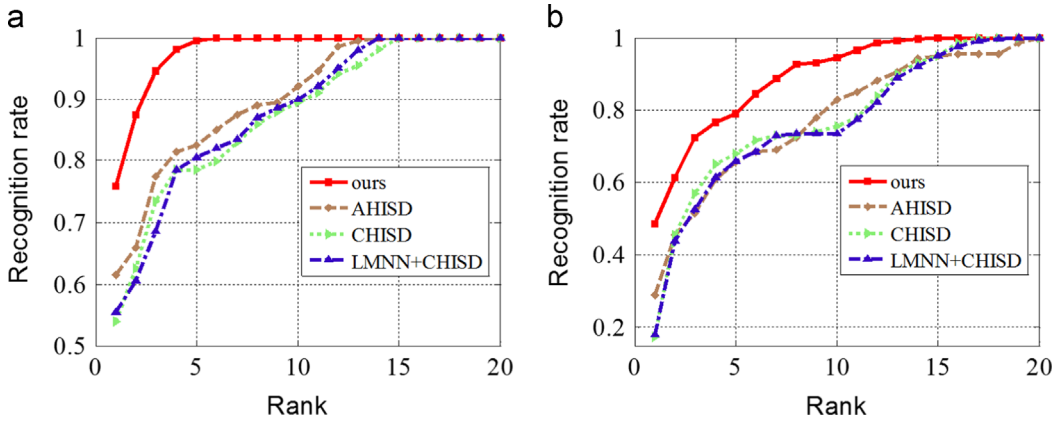


Fig. 6. Cumulative match curves (CMC) of face recognition rate for: (a) P1E_S1_C1 as gallery, P1L_S1_C2 as probe; and (b) P1E_S1_C1 as gallery, P1L_S1_C3 as probe.

Table 3
Computation time (seconds) on Chokepoint dataset.

Methods	AHISD	CHISD	LMNN+CHISD	Our approach
Training time	NA	NA	82.8	227.0
Testing time	0.0078	0.0109	0.0308	0.0310

If the variation in viewpoints and illuminations across cameras was too large (see the result of P1E_S1_C1 as gallery and P1L_S1_C3 as probe), the performances of the set-based methods (AHISD and CHISD), that do not learn a feature-mapping between cameras, were very low. In contrast to them, our proposed method increased the recognition rate by about 20 percent for top ranked test subjects, not only when the variation was little (P1E_S1_C1 as gallery and P1L_S1_C2 as probe) but also when the variation was large (P1E_S1_C1 as gallery and P1L_S1_C3). This demonstrated the ability of the proposal method to adapt to the variation in viewpoints and illuminations. Compared with LMNN+CHISD, our method also achieved better performance. This indicates that the original LMNN is not very effective when directly embedded into set-based matching for the sets collected from different conditions.

Computation time: In Table 3 we report the computation time of different set-based methods on the Chokepoint dataset. For testing, we report the time for matching one probe image set with one gallery image set. The results were obtained by using Matlab in a personal computer with a 3.4-GHz CPU and 8 GB RAM. It is noted that the training time is only for our approach and the original LMNN learning scheme. In our training, we need to solve $N(N-1)/2$ (N is the number of image sets used for training) convex optimization problems in each iteration, leading to a high computational complexity. When matching a probe set with a gallery image set, our method takes more time than CHISD due to the computation of transforming the features to the learned subspace.

5. Conclusions

In this paper we have proposed a method embedding distance metric learning into set-based face recognition for video surveillance. The idea was motivated by the recognition difficulty due to viewpoint and illumination variations across multiple cameras in surveillance networks. Experiments on public databases showed that the proposed method was superior to the state-of-the-art methods in overcoming this difficulty. The method can be applied to surveillance networks with fixed cameras; future work includes making it more scalable across more complicated networks.

Acknowledgments

The work was partially sponsored by National Natural Science Foundation of China (Nos. 61132007 and 61271390).

References

[1] J. Stallkamp, H.K. Ekenel, R. Stiefelhagen, Video-based face recognition on real-world data, in: ICCV, 2007, pp. 1–8.
[2] H. Cevikalp, B. Triggs, Face recognition based on image sets, in: CVPR, 2010, pp. 2567–2573.
[3] S. Chen, S. Mau, M.T. Harandi, C. Sanderson, A. Bigdeli, B.C. Lovell, Face recognition from still images to video sequences: a local-feature-based framework, EURASIP J. Image Video Process. (2011) 790598.
[4] R. Wang, S. Shan, X. Chen, W. Gao, Manifold-manifold distance with application to face recognition based on image set, in: CVPR, 2008, pp. 1–8.
[5] Y. Hu, A.S. Mian, R. Owens, Sparse approximated nearest points for image set classification, in: CVPR, 2011, pp. 121–128.
[6] S. Zhou, R. Chellappa, Beyond a single still image: face recognition from multiple still images and videos, Face Process. Adv. Model. Methods (2005) 547.
[7] J.R. Barr, K.W. Bowyer, P.J. Flynn, S. Biswas, Face recognition from video: a review, Int. J. Pattern Recognit. Artif. Intell. 26 (05) (2012) 1266002.
[8] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.
[9] O. Yamaguchi, K. Fukui, K.-i. Maeda, Face recognition using temporal image sequence, in: Automatic Face and Gesture Recognition, 1998, pp. 318–323.
[10] K. Fukui, O. Yamaguchi, Face recognition using multi-viewpoint patterns for robot vision, in: Robotics Research, Springer, 2005, pp. 192–201.
[11] Y. Wu, M. Minoh, M. Mukunoki, S. Lao, Set based discriminative ranking for recognition, in: ECCV, 2012, pp. 497–510.
[12] J. Lu, G. Wang, P. Moulin, Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning, in: ICCV, 2013, pp. 329–336.
[13] A. Mian, Y. Hu, R. Hartley, R. Owens, Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning, IEEE Trans. Image Process. 22 (12) (2013) 5252–5262.
[14] T. Ahonen, A. Hadid, M. Pietikäinen, Face recognition with local binary patterns, in: ECCV, 2004, pp. 469–481.
[15] K.C. Lee, J. Ho, M.H. Yang, D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, in: CVPR, 2003, pp. 313–320.
[16] 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: CVPRW, 2011, pp. 74–81.
[17] R. Wang, X. Chen, Manifold discriminant analysis, in: CVPR, 2009, pp. 429–436.
[18] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[19] Chunxiao Liu, Guijin Wang, Xinggang Lin., Multiple-shot person re-identification by pairwise multiple instance learning, IEICE Transactions on Information and Systems 96 (12) (2013) 2900–2903.

[20] Liu Chunxiao, Wang Guijin, Lin Xinggang, Li Liang, Person re-identification by spatial pyramid color representation and local region matching, IEICE Transactions On Information And Systems 95 (8) (2012) 2154–2157.

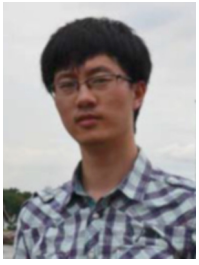
Guijin Wang received the B.S. and Ph.D. degrees in signal and information processing (with honors) from the Department of Electronics Engineering, Tsinghua University, China, in 1998 and 2003, respectively. From 2003 to 2006, he was with Sony Information Technologies Laboratories as a researcher. Since 2006, he has been with the Department of Electronics Engineering at, Tsinghua University, China, as an associate professor. He has published over 50 International journal and conference papers and holds several patents. He is the session chair of IEEE CCNC'06. His research interests are focused on wireless multimedia, image and video processing, depth imaging, pose recognition, intelligent surveillance, industry inspection, object detection and tracking and online learning.

Fei Zheng received the B.S. degree in Information and Electronics Engineering from the Department of Electronic Engineering, Tsinghua University, China in 2011. He is currently a master candidate in the Department of Electronic Engineering, Tsinghua University. His research interests are in the area of machine learning and intelligent surveillance.

Chenbo Shi received the B.S. and Ph.D. degrees from the Department of Electronics Engineering, Tsinghua University, China in 2005 and 2012 respectively. From 2008 to 2012, He has published over 10 International journal and conference papers. He is the reviewers for several international journals and conferences. Now he is a postdoctoral researcher in Tsinghua University. His research interests are focused on image stitching, stereo matching, matting, object detection and tracking, etc.

Jing-Hao Xue received the B.Eng. degree in telecommunication and information systems in 1993 and the Ph.D. degree in signal and information processing in 1998, both from Tsinghua University, the M.Sc. degree in medical imaging and the M.Sc. degree in statistics, both from Katholieke Universiteit Leuven in 2004, and the degree of Ph.D. in statistics from the University of Glasgow in 2008. He has worked in the Department of Statistical Science at University College London as a Lecturer since 2008. His research interests include statistical and machine-learning techniques for pattern recognition, data mining and image processing, in particular supervised, unsupervised and incompletely supervised learning for complex and high-dimensional data.

Chunxiao Liu is a Ph.D. candidate in the Department of Electronics Engineering, Tsinghua University, China. He received his B.S. degree in the Department of Electronics and Information Engineering, Huazhong University of Science & Technology, Wuhan, in 2008. His research interests include human re-identification, tracking, camera network activity analysis, machine learning.



Li He was born in Jilin, China, in 1986. He received the B.S. degree from the Department of Electronics Engineering, Tsinghua University, Beijing, China, in 2010. He is currently working toward the Ph.D. degree in the Department of Electronics Engineering, Tsinghua University, Beijing, China. His research interests include the applications of machine learning and pattern recognition in human pose/action recognition and tracking.