Image Sets Alignment for Video-based Face Recognition

Zhen Cui[1,2], Shiguang Shan[1], Haihong Zhang[3], Shihong Lao[3], Xilin Chen[1]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[2]Scholl of Computer Science and Technology, Huaqiao University, Xiamen 361021, China
[3]Omron Social Solutions Co., LTD., Kyoto, Japan
zhen.cui@vipl.ict.ac.cn; sgshan@ict.ac.cn;{angelazhang,lao}@ssb.kusatsu.omron.co.jp; xlchen@ict.ac.cn

Abstract

Video-based Face Recognition (VFR) can be converted to the matching of two image sets containing face images captured from each video. For this purpose, we propose to bridge the two sets with a reference image set that is well-defined and pre-structured to a number of local models offline. In other words, given two image sets, as long as each of them is aligned to the reference set, they are mutually aligned and well structured. Therefore, the similarity between them can be computed by comparing only the corresponded local models rather than considering all the pairs. To align an image set with the reference set, we further formulate the problem as a quadratic programming. It integrates three constrains to guarantee robust alignment, including appearance matching cost term exploiting principal angles, geometric structure consistency using affine invariant reconstruction weights, smoothness constraint preserving local neighborhood relationship. Extensive experimental evaluations are performed on three databases: Honda, MoBo and YouTube. Compared with competing methods, our approach can consistently achieve better results.

1. Introduction

Face recognition has traditionally been posed as the problem of identifying a face from a single image, and many current methods assume face images are attained under controlled environments. However, facial appearance changes dramatically due to variations in illumination, pose, expression and other factors in unconstrained real-world applications such as video surveillance. Thus, face recognition algorithms learned with images captured under controlled conditions may not suffice for reliable recognition in many practical applications.

Recently, there has been an increasing interest on video-based Face Recognition (VFR) [1-15] because video cameras are commonly available and provide more information compared to still cameras. In the case of VFR, both gallery and query set are video sequences rather than still images. So VFR problem can be converted to measuring the similarity between two video sequences. Intuitively, one could build an appearance-based system by choosing a subset of representative frames (so-called key-frames or exemplars) from video sequence as models and then perform still image based recognition. Obviously, such an approach does not fully utilize spatiotemporal information. To make use of it, some techniques are developed, for instance, by using Hidden Markov Model (HMM) [1,2]. However, temporal model based approaches have not yet shown their full potentials as they also suffer from some drawbacks, such as only using global features while ignoring local information, the lack of discriminability between the facial dynamics.

On the contrary, without the temporal information, face images from a video sequence form an image set. So VFR can be generalized to image-set based classification, where each target person may be enrolled with one or even multiple image sets (so-called gallery set) and a query image set need to be assigned to the identity of its nearest gallery set by calculating its distance from each gallery image set.

Relevant approaches to image-set based classification underwent an explosive development in recent years [3-12]. Generally speaking, such approaches fall into two categories: parametric model methods and nonparametric sample methods. The former [3,12] exploit some parametric distribution (e.g. Gaussian) to represent each image set and then measure the between-distribution similarity. One limitation of the parametric methods is that they have to assume some distribution and handle the parameter estimation problem. If the data set does not follow the predefined statistic distribution, the estimated model will not consist with the data set.

More recently, some non-parametric methods attempt to represent an image set as a linear subspace [9,13,14] or a nonlinear manifold [7,8,15]. Such approaches do not impose any assumption on data distribution, and have shown many merits compared to parametric models. Meanwhile, some algorithms which measure the similarity

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

or distance between two subspaces are also developed. The representative methods are the principal angles which compute the angles between the principal components of two spaces [7,8], and the nearest points which use the nearest distance between two affine hulls [10,11]. Hu et al. [11] applied the affine hull model to account for unseen appearances and proposed Sparse Approximated Nearest Point (SANP) to measure between-set similarity, which enforces nearest points to be close to some facets by imposing sparsity constraints.

Based on the assumption that face images of the same subject are distributed on a nonlinear manifold rather than a linear subspace, Wang et al. extended Subspace-Subspace Distance (SSD) to Manifold-Manifold Distance (MMD) [7], where a nonlinear manifold is partitioned into a number of local linear models by Maximal Linear Patch (MLP) [16], and then MMD is converted to integrating the distances between pair-wise subspaces. An extension of MMD, called Manifold Discriminant Analysis (MDA) [8] is also proposed to solve the supervised between-manifold distance. These methods based on nonlinear manifold have achieved the state of the art results in several public face databases.

However, MMD and MDA ignore the correspondence between the subspaces from the manifolds (of the two image sets) when calculating similarity. They divide each image set into several subspaces by clustering and then perform pair-wise subspaces comparison. Thus, these methods have the following disadvantage: given two sequences, face images under the same condition might be partitioned differently because of the uncertainty of the clustering. For example, face images with yaw within [10^o, 20^o] might be partitioned to one cluster for one sequence, while the face image with yaw within [15^o, 25^o] might fall into one cluster in another sequence. This implies that the distance between two manifolds in above methods might not be well defined. In addition, MDA [8] only learns one linear transformation, which is not sufficient to capture the discriminant information because the face images in a sequence are distributed on a nonlinear manifold.

To avoid the above bias originated from clustering, alignment between two image sets is a possible solution. One scheme is aligning the test image set to each gallery image set respectively, and then comparing them directly. However, such strategy is unreasonable in practical VFR for two reasons: (1) in many cases, the query image set does not wholly but only partially corresponds with the gallery image set, which implies difficult alignment; (2) it is too time-consuming to align the query set with each of the gallery sets online. To address the above issues, a reference set is introduced to bridge the query set and the gallery set. Furthermore, to obtain more discriminant features, multiple linear transformations can be learned from corresponded local models which are structured by aligning all gallery image sets with the pre-partitioned reference set.

To solve this alignment between two image sets, some previous methods assume each image set is distributed on a nonlinear manifold, and then do manifold alignment by utilizing dimensionality reduction algorithms [17]. But most of existing manifold alignment algorithms [18-25] fall into supervised or semi-supervised category, which require some known matching points obtained from manual annotation or other prior knowledge. Therefore, they are unsuitable for our question because it is intractable to get the matching point for the large scale video database by manual annotation.

Confronted with this problem, an alternative is to align two sequences without any prior information, i.e. unsupervised image sets alignment. However, to our best knowledge, few studies discussed this problem. Recently, Wang et al. [26] proposed an unsupervised alignment method without correspondence, which learns a projection transforming instances from two subspaces to a lower dimensional space, and simultaneously matches the local geometry structures by the k nearest neighbors. Nevertheless, when matching k neighbors of two points, the authors considered all $k!$ permutations to find the best match, which is too time-consuming.

To address above problem, in this paper, we propose to align all image sets to a reference image set that is well-defined and pre-structured into a number of local linear models offline. In other words, given two image sets for comparison, as long as each of them is aligned to the reference set, they are mutually aligned and well structured. Therefore the similarity between them can be computed by comparing only the corresponded local models instead of all the pair-wise ones. Furthermore, instead of a global linear transformation in MDA [8], multiple linear transformations from corresponded subspaces are learned during the training, and then applied to the query image sets.

In addition, inspired by Wang's alignment method [26], we explicitly formulate the image sets alignment problem as a quadratic programming, which can be solved faster than Wang's method. Our proposed model contains three terms. The first term checks consistency of local geometric structure. The second term measures the matching cost between two points. The last term constrains the smoothness of manifold. Different from Wang's geometric structure matching methods, we exploit the local reconstruction relationship, which is affine invariant and thus does not need to consider all possible permutations. In addition to the smoothness which is also applied in [26], we add the matching cost term into our model to increase stability.

2. Overview of the proposed method

In this section, we briefly describe our proposed method. As shown in Figure 1, before images sets alignment, we

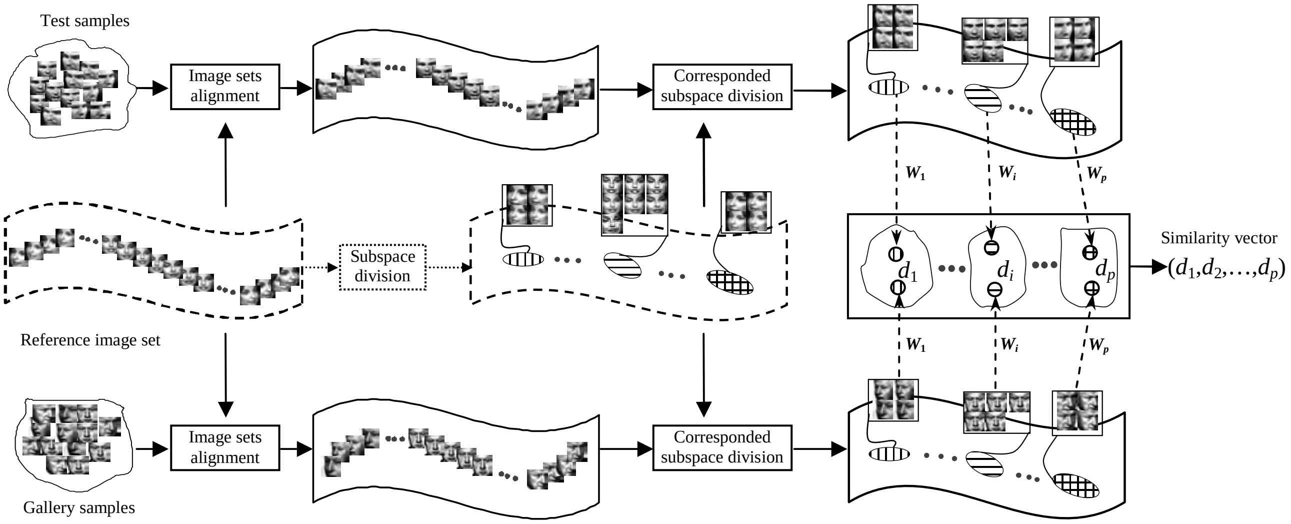

Figure 1. The framework of our proposed approach.

choose an image set as the reference set so that all image sets can mutually align to it. The motivation of utilizing a reference set lies in three folds: (1) avoiding the bias from clustering in MMD and MDA as mentioned above; (2) addressing the difficulty due to possible partial correspondence between the query and gallery set; (3) reducing the high computational cost for online alignment of the query set and all the gallery sets.

For above purposes, the reference image set should cover as many variations as possible including pose and illumination and other factors. In our experiments, we select randomly one video sequence of one person from the gallery sets with enough frames covering diverse variations. An elaborately designed reference set can possibly further improve the recognition performance, which will be our future work.

After the reference set is selected, an offline process is applied to it, e.g. MLP [7], which partitions the reference image set into a number of local linear subspaces. Then both the test samples and the gallery samples are aligned to the reference set by our proposed image sets alignment algorithm (details in Section 3). Note that, as the test set and the gallery set might have no common intersection set due to non-overlapping poses for instance, we mirror all the samples before alignment. Subsequently, corresponded subspace division is performed on the test and gallery image sets by exploiting the pre-partition of reference set. Thus, the subspaces from the test and gallery sets are aligned naturally.

At last, the distances only between the corresponded subspaces rather than all pair-wise ones are calculated as the similarity vectors by a series of corresponded mappings (e.g. Fisherfaces) pre-learned from a training image sets. The similarity score vectors can be further fed into a classifier, or simply pooled by SUM or MIN rule to get the final score for face recognition.

3. Unsupervised image sets alignment

Video sequences may be captured in different sceneries with complicated variations in pose, illumination and expression so on. Thus, it is difficult to directly evaluate the appearance similarity between two images from two different sequences respectively. Usually, the angles between two subspaces may reflect the variant modes, which have been widely used to measure the similarity between two image sets [5,7,8,13]. Therefore, we can describe a point more robust with a subspace spanned by this point and its neighbors. Furthermore, the appearance matching cost between the two samples can be represented by the maximum principal angle.

An important assumption of manifold alignment is the consistency of the local geometry structure [26]. One may search all possible permutations to verify the consistency, which however is too time-consuming. To reduce the computational time, we formulate the local structure as a quadratic programming term by exploiting the affine invariant reconstruction weight.

In addition, the manifold should be smooth. Thus, we also add the smoothness constraint to the proposed method to decrease mismatching.

In what follows, we first develop a primary formulation to solve the image sets alignment problem, and then detailed descriptions about each term of the model are provided. At last, an efficient algorithm is summarized.

3.1. Problem formulation

Without loss of generality, in this paper, we concatenate the intensity of all pixels to form the feature vector representing any face image. Formally, two image sets, one target (gallery or test) set and the reference set are denoted respectively by $X=\{\boldsymbol{x}_i \mid i=1,2,\ldots,m\}$ and $Y=\{\boldsymbol{y}_i \mid i=1,\ldots,n\}$, where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ represent the samples, m and n are the sample numbers in X and Y respectively. Our goal is to seek a mapping function f, the so-called alignment function, which maps any target image $\boldsymbol{x}$ in X to some reference image $\boldsymbol{y}$ in Y. We formulate this task as an optimization problem,

$$\hat{f} = \arg\min \left\{E_g + \lambda_1 E_c + \lambda_2 E_s\right\}, \tag{1}$$

where

$$E_g = \sum_{\boldsymbol{x}_i \in X} g(\boldsymbol{x}_i, N_{\boldsymbol{x}_i}; f(\boldsymbol{x}_i), N_{f(\boldsymbol{x}_i)}) \ , \tag{2}$$

$$E_c = \sum_{\boldsymbol{x}_i \in X} c(\boldsymbol{x}_i, f(\boldsymbol{x}_i)), \tag{3}$$

$$E_s = \sum_{\boldsymbol{x}_i \in X} \sum_{\boldsymbol{x}_j \in N_{\boldsymbol{x}_i}} s(f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)). \tag{4}$$

In above equations, $N_{\boldsymbol{x}_i}$ denotes the neighbors of $\boldsymbol{x}_i$, λ_1 and λ_2 balance the effect of the three terms. The first term E_g represents the geometric similarity score between the two image sets, where g is the geometric consistency function measuring the geometric dissimilarity between two local models. The second term E_c reflects the appearance similarity, where c is the matching cost function between two points. The third term E_s is used to keep the smoothness, i.e. the neighborhood relationship in the target set should be preserved in the reference set.

The following three subsections detail the above three terms respectively.

3.2. Local geometry consistency

Inspired by Locally Linear Embedding (LLE) [27] and the recent literature [28], we introduce a locally invariant geometric constraint for image sets alignment.

As mentioned above, we represent each image set as a manifold. To characterize the geometric properties of the neighborhood of each point in the manifold, we assume each $\boldsymbol{x}_i$ can be approximately represented by an affine combination of its neighbor points,

$$\boldsymbol{x}_i = \sum_{\boldsymbol{x}_j \in N_{\boldsymbol{x}_i}} W_{ij} \boldsymbol{x}_j \ , \tag{5}$$

where $W = \{W_{ij}\}$ is the reconstruction weight matrix for all points and the i-th row of W stores all reconstruction coefficients for the i-th point $\boldsymbol{x}_i$ with $\sum_j W_{ij}=1$. Specifically, least squares is exploited to describe the local geometric properties of each point, i.e.,

$$\| \ \boldsymbol{x}_i - \sum_{\boldsymbol{x}_j \in N_{\boldsymbol{x}_i}} W_{ij} \boldsymbol{x}_j \ \| = 0 \ . \tag{6}$$

Obviously, Eq. (6) is affine invariant approximately. Thus, we can further formulate E_g in (2) as the following object function by the weight matrix W,

$$E_g = \sum_{\boldsymbol{x}_i \in X} \| \ f(\boldsymbol{x}_i) - \sum_{\boldsymbol{x}_j \in N_{\boldsymbol{x}_i}} W_{ij}^x f(\boldsymbol{x}_j) \ \|^2 \ , \tag{7}$$

where W^x represents the reconstruction weight matrix of the image set X. If we mark the mapping relation of each point as a vector, the function f can be represented by a $\{0,1\}$ binary matrix $F_{m\times n}$. Thus, the function (7) can be rewritten as the matrix formulation,

$$E_g = \| \ (I - W^x) F Y^T \ \|^2 = \| \ L^x F Y^T \ \|^2 . \tag{8}$$

Because of the sum of each row in W equals to 1, L^x can be treated as the Laplacian matrix of a graph, where the edge may be constructed by W^x. Note I is an identity matrix.

Compared with Wang's method [26], which exploits the Euclidean distance matrix of the k nearest neighbors to describe the local geometry structure, where all $k!$ possible permutations are considered in the image matching with the cost $O(k!)$, our model is locally affine invariant and easy to solve the mapping F.

3.3. Appearance matching cost

In order to measure the similarity between two images coming from different image sets with variations in pose, illumination, expression and other factors, we exploit the maximum principal angle of their corresponded local linear subspaces as the appearance matching cost.

Formally, given two linear subspaces S_1 and S_2, the principal angles $0 \le \alpha_1 \le \ldots \le \alpha_r \le \pi/2 (r=\min(\dim(S_1),\dim(S_2)))$ are uniquely defined as following [7],

$$\cos(\alpha_k) = \max_{\boldsymbol{u}_k \in S_1 \backslash \{\boldsymbol{u}_1,\cdots,\boldsymbol{u}_{k-1}\}} \ \max_{\boldsymbol{v}_k \in S_2 \backslash \{\boldsymbol{v}_1,\cdots,\boldsymbol{v}_{k-1}\}} (\boldsymbol{u}_k)^T \boldsymbol{v}_k \ , \tag{9}$$

where $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are called the k-th pair of canonical vectors, "\" means the subtracting operation on subspaces. The cosine values of the principal angles are called canonical correlations. Obviously, the smaller the maximum principal angle is, the closer the subspaces are. Generally, we select the distance between the first pair of canonical vectors, corresponding to the most similar modes, as the distance between the two subspaces. To solve this model, a numerically stable method based on Singular Value Decomposition (SVD) is exploited from [29].

Given the above definition of subspace distance, the appearance matching cost of two images can be calculated based on the maximum principal angle between two local linear subspaces, expanded respectively by the neighbors of the two images. Formally, for two images $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ respectively from X and Y, their k-NN neighbors on X and Y expand subspace S_x and S_y. Then, the matching score between $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ is computed by the above subspace distance and denoted by C_{ij}. Then, we mark the matching scores between two image sets X and Y as the matrix $C=\{ C_{ij} \}$. Then the function E_c in (3) can be re-written as,

$$E_c = \mathrm{tr}(C^T F) \ . \qquad (10)$$

where “tr” means the trace operation.

3.4. Smoothness constraint

Intuitively, manifolds should be smooth, curved surfaces embedded in higher dimensional Euclidean space. Thus, the local neighborhood relationship should be preserved when aligning two image sets. That is, if two images in X are neighbors, their corresponding images in Y should also be neighbors.

Formally, we denote the k-th neighborhood relationship of each image in X as a matrix R^k,

$$(R^k)_{ij} = \begin{cases} 1, & \boldsymbol{x}_j \text{ is the } k\text{-th neighbor of } \boldsymbol{x}_i \\ 0, & \text{or else} \end{cases} \ . \qquad (11)$$

Thus, the third term E_s in (4) is formulated as follows,

$$E_s = \sum_{k=1}^{K} \left\| FY^T - R^k FY^T \right\|^2 = \sum_{k=1}^{K} \left\| L^k FY^T \right\|^2 \ , \qquad (12)$$

where L^k=I-R^k, K is the local neighbor number.

3.5. Efficient solution

Followed by above analysis, the object function (1) can be formulated as a quadratic programming with an integer constraint as follows,

$$\begin{aligned} \hat{F} = \arg\min_{F} & \left\| L^x FY^T \right\|^{\boldsymbol{2}} + \lambda_1 \mathrm{tr}(C^T F) + \lambda_2 \sum_{k=1}^{K} \left\| L^k FY^T \right\|^2 \\ s.t. \ & F\mathbf{1}_{n\times1} = \mathbf{1}_{m\times1}, \\ & F \in \{0,1\}^{m\times n}, \\ & F^T \mathbf{1}_{m\times1} \le l. \end{aligned} \qquad (13)$$

The variable F is an m×n binary assignment matrix that represents the image matching function f. Each row of F contains only one 1, which means any points in X must be projected to one and only one point in Y. There are three constrains in (13). The first one guarantees all images in X should be matched to Y. The second one shows that the matching between X and Y’s points is either “Yes” or “No”. The third constraint allows that at most l images in X can be permitted to match the same image in Y.

The problem (13) is a quadratic object function with integer constraints, which is NP-complete and cannot be efficiently solved. We relax the integer constraint and meanwhile simplify the object function as follows,

$$\begin{aligned} \hat{F} = \arg\min_{F} & \left\| UFY^T \right\|^2 + \lambda_1 \mathrm{tr}(C^T F) \\ s.t. \ & F\mathbf{1}_{n\times1} = \mathbf{1}_{m\times1}, \\ & F \ge 0, \\ & F^T \mathbf{1}_{m\times1} \le l. \end{aligned} \qquad (14)$$

where U can be calculated by SVD with the following equation,

$$U^T U = (L^x)^T L^x + \lambda_2 \sum_{k=1}^{K} (L^k)^T L^k \ . \qquad (15)$$

Algorithm 1. Unsupervised image sets alignment

```
Input: $X$={$\boldsymbol{x}_1$,⋯,$\boldsymbol{x}_m$},$Y$={$\boldsymbol{y}_1$,⋯,$\boldsymbol{y}_n$},$m \leqslant n$
Output: binary matrix $F_{m\times n}$
1. Find the $K$ neighbors of each point in $X$ and $Y$ respectively.
2. Calculate the weight matrix $W^x$ by (6).
3. Calculate the appearance matching cost $C$ by principle angles (Section 3.3).
4. Initialize the trust region $T_i$={$y_{i1}$,$y_{i2}$,⋯,$y_{it}$}for each point $\boldsymbol{x}_i \in X$.
5. While trust region is enough large
6.     Solve $F$ by (14).  (Section 3.5)
7.     Shrinkage the trust region by removing these points whose value is very low in $F$.
8. End
9. Solve $F$ in (14) by the final trust region.
10.Quantify $F$ to {0,1} matrix.
```

We exploit “interior-point” method [30] to optimize the object function (14) with Matlab toolboxes. The entire algorithm of unsupervised image sets alignment is summarized in Algorithm 1.

In order to accelerate the algorithm, we utilize the trust region shrinkage method (as in [9]) to solve the convex optimization problem approximately. We initialize the trust region with the most similar t samples taken from Y by the geometric structure and the appearance similarity. Thus, the main time cost is to solve the object function (14). Fortunately, this object function is convex, which therefore can converge to the minimum fast. Typically, aligning hundreds of samples takes only several seconds on standard PC with Matlab programming.

4. Experiments

In this section, we first perform image set alignment experiments to demonstrate the effectiveness of the proposed alignment method. Then, we apply the proposed alignment method to video-based face recognition.

4.1. Image sets alignment

To validate the efficacy of the proposed alignment algorithm, we execute face image matching across poses. Multi-PIE database [36] is exploited to evaluate our method. Here, we compare the proposed method with [26] and direct matching in original feature space (DM). We randomly collect 50 subjects. For each subject, 83 face images under 7 poses with yaw within [-45^o, +45^o] (with 15^o intervals) and with different expressions and illuminations are selected. These images are cropped to 20×30 pixels to simulate the low quality video face images.

We carry out between-set alignment and report the quantitative results in Figure 2. The nearest poses are also viewed as a correct matching when *r* equals to 1, while *r*=0 means only the corresponded pose is counted as correct matching. As we can see, our approach can achieve an accuracy large than 98% when *r*=1, which means that nearly all the matching results lie in $\pm 15^{o}$ pose error at most. It is worth pointing out that the dataset might favor DM more because all face images come from the same scenes. Wang's method [26] only used the geometric similarity without appearance matching cost. Our method achieves a higher accuracy than Wang's method and DM, probably because we utilize appearance matching cost and geometric structure similarity simultaneously. In addition, our approach only needs about 3 seconds to finish the matching between two image sets, which is much faster than Wang's method. Some alignment results are shown in Figure 3, where the third images are incorrectly aligned: a front face image is aligned to a face image with yaw +15^{o}.

4.2. Face recognition

Datasets: We used three public datasets: Honda/UCSD [31], CMU MoBo [32] and YouTube Celebrities [33].

Honda/UCSD was collected by Lee et al. for video-based face recognition. In our experiment, we exploit their first subset containing 59 videos of 20 subjects (at least 2 videos for each subject). Each video sequence has different pose and expression vibrations. A cascaded face detector [34] is applied to detect faces in each video sequence. Then all faces are resized to a 20×20 pixels gray image as [7]. The length of the videos varies from 12 to 645 frames. To eliminate the lighting effects, histogram equalization is employed to pre-process the images.

CMU MoBo database was originally created for human pose identification. This database includes 96 sequences of 24 different subjects, i.e. 4 videos per subject. Each video was captured from walking on treadmill and has 300 frames. We exploit the same strategy as we did for the Honda dataset to obtain the face images with 30×30 pixels.

YouTube Celebrities was collected from YouTube for face tracking and recognition in real world applications. The dataset contains 1910 video sequences of 47 celebrities (actors, actresses and politicians). Each sequence has hundreds of frames with low resolution and high compression rates. Compared with Honda and MoBo databases, this database is much more challenging because of noises and complicated variations in pose, illumination and expression. Face images are also cropped as above and resized to 30×30 pixels.

On all of three datasets, we conduct five-fold cross validation experiments, i.e. five randomly selected training and testing combinations, to evaluate our proposed method. For Honda and MoBo database, one video sequence per subject is used for training and the rest sequences for testing. For YouTube, each person has 41 clips on average which cover 3 sections. We randomly choose 9 clips per person, 3 clips per section, as the experimental data. One clip per section is used for training and 6 for testing for each subject.

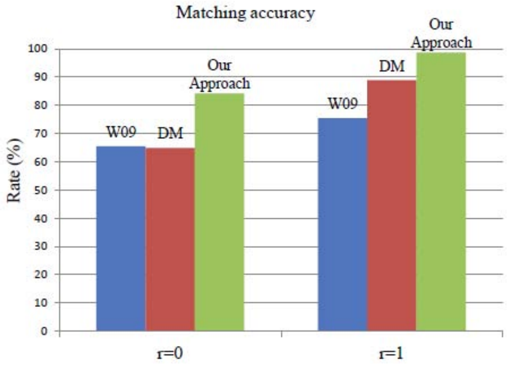

Figure 2. The alignment accuracy across poses.

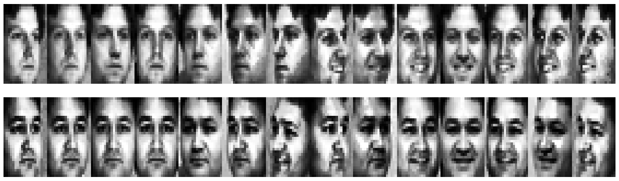

Figure 3. Examples of alignment under different poses.

Comparison with existing methods: We compare our proposed approach with several image-set based methods proposed in recent years, including Mutual Subspace Method (MSM) [35], Discriminant Canonical Correlation Analysis (DCC) [5], Manifold-to-Manifold Distance (MMD) [7], Manifold Discriminant Analysis (MDA) [8], and Sparse Approximated Nearest Points (SANP) [11]. Here we do not provide more experimental results of exemplar-based methods except LDA because recent literatures [5,7,8,11] have shown that image-set based methods are generally superior to exemplar based methods.

Our implementation: For LDA and MSM, we adopt the techniques in accordance with [7] and [5] respectively. The source codes of MMD, MDA and SANP are downloaded from the original authors, and referred parameters are followed by the referred literatures [7,8,11].

The important parameters in our proposed method include: (i) the nearest neighbors size is set to 10; (ii) the control parameters: λ_1=2, λ_2=0.1; (iii) the dimension of PCA is set to 70, 60, 80 for three database respectively when projecting the gray features of local linear model; (iv) the dimension of LDA is set to the number of classes minus 1. In our experiments, we utilize the Euclidean distance to calculate the similarity between two corresponded linear models after mapping, and then the minimum value is exploited as the final between-set distance.

Table 1. Identification rates on three databases

Dataset	The mean and standard deviation of recognition rates of different methods						
	LDA	MSM	MMD	DCC	MDA	SANP	**Proposed method**
Honda/UCSD	0.789±0.01	0.923±0.04	0.969±0.02	0.980±0.01	**0.989±0.01**	0.959±0.03	**0.989±0.01**
CMU MoBo	0.885±0.01	0.886±0.03	0.897±0.01	0.903±0.05	0.947±0.01	0.900±0.02	**0.950±0.01**
YouTube	0.604±0.03	0.616±0.04	0.634±0.02	0.673±0.03	0.676±0.02	0.634±0.03	**0.746±0.03**

Results and analysis: The recognition results of all comparative methods on the three different face databases are summarized in Table 1. From this table, we can find our proposed method achieves superior performances in most testing. Now we conclude as follows.

(1)The comparison methods based on image sets have shown distinct performance. Among them, MSM, MMD and SANP, deal with image data generatively, which makes them less appealing than DCC, MDA and the proposed method, which exploit disciminant label information. SANP is superior to MSM because the sparsity constraint enforces the nearest point to close on facets of affine hull. MMD is comparable to SANP and also exhibits more excellent recognition results than MSM because it represents the complex image set as several local linear models. This also explains the superiority of MDA and our proposed method over DCC. MDA exploits nonlinear models, but it ignores between-set correspondences, only learns a global linear transformation to extract the discriminant features and performs the comparisons of pair-wise subspaces without the alignment process. Compared with MDA, our proposed method is more superior because of eliminating the biases from clustering.

(2)Among three databases, all methods have the worst performance on YouTube database, because the video sequences are captured from real world with low quality and broad appearance variations. From Table 1, we can find our proposed method outperforms other competing methods, about 7 percents on YouTube. Note that the results of SANP are slightly lower than those reported in [11], because the testing protocol in [11] is different from ours on YouTube, and [11] exploited LBP as the face feature on MoBo.

4.3. Discussion

In our experiments on three video databases (5 random experiments per database), we only randomly selected the image set of one subject with enough images covering various variations for each experiment as the reference set. The consistently good results imply desirable insensitivity of the reference set selection. Intuitively, the reference set should be a “complete” set, which means it should be as large as possible to cover sufficient variations. Moreover, such a “complete” set could help alignment and further promote face recognition performance. To create such a reference set, a direct method is to capture all the various face images with a well-configured complicated environment by a camera, as in Multi-PIE database [36] with configurations of different illuminations and poses. However, a “complete” set is not easy to be obtained from a single video sequence in real world, even specially designing. A practical method is to build a statistical reference model from large scale video sequences, which is also our future work.

5. Conclusion

Manifold alignment (or more generally image sets alignment) facilitates video-based face recognition. In this paper, a reference set that is well pre-defined and pre-structured to a number of local models is used as the bridge aligning two image sets. Thus, similarity scores can be attained by only comparing the corresponded local models instead of all the pair-wise ones. To the best of our knowledge, the proposed unsupervised image set alignment algorithm is the first attempt to solve this problem as an optimization problem from the view of manifold. Experimental results on three public databases demonstrate that the proposed method is convincingly applicable to video-based face recognition, and achieves consistently better performance.

Acknowledgements

This work was partially supported by the National Basic Research Program of China (973 Program) under contract 2009CB320902, and the Natural Science Foundation of China under contract Nos. 60832004, 61025010 and 61173065, and the Beijing Natural Science Foundation under contract No. 4111003. Haihong Zhang and Shihong Lao are partially supported by “R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society”, Special Coordination Fund for Promoting Science and Technology of MEXT, the Japanese Government.

References

[1] X. Liu, T. Chen. Video-based Face Recognition Using Adaptive Hidden Markov Models. In CVPR, 2003.

[2] M. Kim, S. Kumar, V. Pavlovic, H. Rowley. Face Tracking and Recognition with Visual Constraints in Real-World Videos. In CVPR, 2008.

[3] O. Arandjelović, G. Shakhnarovich, G. Fisher, J.R. Cipolla, and R. A. Zisserman. Face Recognition with Image Sets Using Manifold Density Divergence. In CVPR, 2005.

[4] O. Arandjelović, R. Cipolla. A Manifold Approach to Face Recognition from Low Quality Video Across Illumination and Pose Using Implicit Super-Resolution. In ICCV, 2007.

[5] T.K. Kim, O. Arandjelović, and R. Cipolla. Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations. TPAMI, 29(6), 2007.

[6] T.K. Kim, J. Kittler, and R. Cipolla. Incremental Learning of Locally Orthogonal Subspaces for Set-based Object Recognition. In BMVC, 2006.

[7] R. Wang, S. Shan, X. Chen and W. Gao. Manifold-Manifold Distance with Application to Face Recognition based on Image Set. In CVPR, 2008.

[8] R. Wang and X. Chen. Manifold Discriminant Analysis. In CVPR, 2009.

[9] M. Nishiyama, M. Yuasa, T. Shibata, T. Wakasugi, T. Kawahara, and O. Yamaguchi. Recognizing Faces of Moving People by Hierarchical Image-Set Matching. In CVPR, 2007.

[10] H. Cevikalp and B. Triggs. Face Recognition Based on Image Sets. In CVPR, 2010.

[11] Y. Hu, A.S. Mian, and R. Owens. Sparse Approximated Nearest Points for Image set Classification. In CVPR, 2011.

[12] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face Recognition from Long-term Observations. In ECCV, 2002.

[13] M. Nishiyama, O. Yamaguchi, and K. Fukui. Face Recognition with the Multiple Constrained Mutual Subspace Method. In AVBPA, 2005.

[14] K. Fukui and O. Yamaguchi. The Kernel Orthogonal Mutual Subspace Method and Its Application to 3D Object Recognition. In ACCV, 2007.

[15] M.T. Harandi, C. Sanderson, S. Shirazi, and B.C. Lovell. Graph Embedding Discriminant Analysis on Grassmannian Manifolds for Improved Image Set Matching. In CVPR, 2011.

[16] R. Wang, S. Shan, X. Chen, J. Chen, and W. Gao. Maximal Linear Embedding for Dimensionality Reduction. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1776-1792, 2011.

[17] T. Zhang, D. Tao, X. Li, and J. Yang. Patch Alignment for Dimensionality Reduction. IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1299–1313, Sep. 2009.

[18] C. Wang and S. Mahadevan. Manifold Alignment Using Procrustes Analysis. In ICML, 2008.

[19] J. Ham, I. Ahn, and D. Lee. Learning a Manifold -Constrained Map between Image Sets: Applications to Matching and Pose Estimation. In CVPR, 2006.

[20] H. Gong, C. Pan, Q. Yang, H. Lu, and S. Ma. A Semi-Supervised Framework for Mapping Data to the Intrinsic Manifold. In ICCV, 2005.

[21] S. Lafon, Y. Keller, and R. Coifman. Data Fusion and Multicue Data Matching by Diffusion Maps. TPAMI, 28(11), 2006.

[22] J. Ham, D. Lee, and L. Saul. Semisupervised Alignment of Manifolds. In Proc. of the Tenth Int’l Workshop on Artificial Intelligence and Statistics, 2005, pp. 120–127.

[23] L. Xiong, F. Wang, and C. Zhang. Semi-Definite Manifold Alignment. In ECML, 2007.

[24] D. Zhai, B. Li, H. Chang, S. Shan, X. Chen, and W. Gao. Manifold Alignment via Corresponding Projections. In BMVC, 2010.

[25] J.J. Verbeek. Learning Nonlinear Image Manifolds by Global Alignment of Local Linear Models. TPAMI 28(8), 2006.

[26] C. Wang and S. Mahadevan. Manifold Alignment without Correspondence. In Proc. Int’l Joint Conf. Artificial Intelligence (IJCAI), 2009, pp. 1273–1278.

[27] S. Roweis and L. Saul. Nonlinear Dimensionally Reduction by Locally Linear Embedding. Science, 290(5500): 2323-2326, 2000.

[28] H. Li, E. Kim, X. Huang, and L. He. Object Matching with a Locally Affine-Invariant Constraint. In CVPR, 2010.

[29] Å. Björck and G.H. Golub. Numerical Methods for Computing Angles between Linear Subspaces. Mathematics, of Computation, 27(123): 579-594, 1973.

[30] P.E. Gill, W. Murray, and M.H. Wright. Practical Optimization, Academic press, London, UK, 1981.

[31] K.C. Lee, J. Ho, M. H. Yang, and D. Kriegman. Video-Based Face Recognition Using Probabilistic Appearance Manifolds. In CVPR, 2003.

[32] R. Gross and J. Shi. The CMU Motion of Body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, June 2001.

[33] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face Tracking and Recognition with Visual Constraints in Real-World Videos. In CVPR, 2008.

[34] P. Viola and M. Jones. Robust Real-Time Face Detection. Int’l J. Computer Vision, vol. 57, no. 2, pp. 137–154, 2004.

[35] O. Yamaguchi, K. Fukui, K.Maeda. Face Recognition Using Temporal Image Sequence. In International Conference on Automatic Face and Gesture Recognition, 1998.

[36] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. In International Conference on Automatic Face and Gesture Recognition, 2008.