Human Action Recognition using Local Spatio-Temporal Discriminant Embedding *

Kui Jia†
Shenzhen Institute of Advanced Technology
Shenzhen, China
kui.jia@sub.siat.ac.cn

Dit-Yan Yeung
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
dyyeung@cse.ust.hk

Abstract

Human action video sequences can be considered as nonlinear dynamic shape manifolds in the space of image frames. In this paper, we address learning and classifying human actions on embedded low-dimensional manifolds. We propose a novel manifold embedding method, called Local Spatio-Temporal Discriminant Embedding (LSTDE). The discriminating capabilities of the proposed method are two-fold: (1) for local spatial discrimination, LSTDE projects data points (silhouette-based image frames of human action sequences) in a local neighborhood into the embedding space where data points of the same action class are close while those of different classes are far apart; (2) in such a local neighborhood, each data point has an associated short video segment, which forms a local temporal subspace on the embedded manifold. LSTDE finds an optimal embedding which maximizes the principal angles between those temporal subspaces associated with data points of different classes. Benefiting from the joint spatio-temporal discriminant embedding, our method is potentially more powerful for classifying human actions with similar space-time shapes, and is able to perform recognition on a frame-by-frame or short video segment basis. Experimental results demonstrate that our method can accurately recognize human actions, and can improve the recognition performance over some representative manifold embedding methods, especially on highly confusing human action types.

1. Introduction

Recognizing human activities in videos has many important computer vision applications, such as video surveillance, human-computer interaction, video browsing, analysis of sports events, etc. It still remains a very challenging computer vision problem, especially under situations when there exist non-stationary backgrounds in uncontrolled imaging conditions, intra-class variations in appearance and size of different human subjects, and action types with similar human body shapes.

Various approaches for human activity recognition have been proposed in the literature [5, 14, 15, 16, 1, 4, 3, 7, 8]. Among these approaches, a key consideration is what feature representations are extracted and used from the spatio-temporal volumes of video sequences. In particular, holistic representations extract key frames [1, 2], compute optical flow [3] and space-time gradients [4], perform feature tracking [5, 6], and part-based representations detect sparse interest points of the whole spatio-temporal volumes as feature descriptors [9, 7, 8]. Recognition using key frames ignores the motion information, and the computation of space-time gradients or other intensity-based features can be unreliable when videos are captured at lower quality or with discontinuous motions. On the other hand, the sparse representations of interest points discard global structural information which is often useful for recognition.

Recently, some researchers have used human silhouettes as features for human activity understanding [2, 10, 11, 12, 13]. A human silhouette contains detailed body shape information. A sequence of human silhouettes generates space-time shapes which contain both instant spatial information about the body pose and dynamic temporal motion information of the global body and local body parts. Human silhouettes are also easy to obtain in many scenarios, especially in the case of stationary cameras.

Human action video sequences characterized by temporally continuously deforming human silhouettes can be considered as data points on nonlinear dynamic shape manifolds. Thus manifold embedding methods can be used for the analysis of human actions. In [12], for example, locality preserving projections (LPP) [17] were used for learning and matching of human action shape manifolds. However, LPP, and also the earlier manifold learning methods such

*This research has been supported by Competitive Earmarked Research Grant (CERG) 621305 from the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China.

†This work was performed when the first author was working as a research associate at the Hong Kong University of Science and Technology.

as isometric feature map (Isomap) [18], locally linear embedding (LLE) [19], and Laplacian eigenmap [20], discover the intrinsic geometrical structure of a data manifold in an unsupervised manner. As a consequence, the derived low-dimensional features in the embedding space are not necessarily optimal for discriminant analysis of different human action classes. On the other hand, supervised manifold learning methods such as local discriminant embedding (LDE) [21], locality sensitive discriminant analysis (LSDA) [22], local Fisher discriminant analysis (LFDA) [23], marginal Fisher analysis (MFA) [24], etc, take the class label information into account. They can discover both the local geometrical and discriminant structures of the data manifold.

However, the supervised manifold learning methods above only consider discovering the local *spatial* discriminant structure, which can be used for discriminating individual silhouette frames of different action classes. On the other hand, the temporal dynamic shape variation of human silhouette sequences provides important discriminative information for action classification. Our aim in this paper is to find a manifold embedding method which can optimally make use of the discriminative temporal shape variation information between different types of actions. Our objective is that after manifold embedding, both the *local spatial* and *local temporal* discriminant structures of the data can be discovered effectively.

Motivated by the above considerations, we propose in this paper a novel manifold embedding method, called *Local Spatio-Temporal Discriminant Embedding* (LSTDE), for human action recognition. Specifically, LSTDE projects data points in a local neighborhood into the embedding space where data points of the same action class are close while those of different classes are far apart. Furthermore, in such a local neighborhood, each data point has an associated short video segment which forms a local temporal subspace on the embedded manifold. LSTDE finds an optimal embedding which maximizes the principal angles between those temporal subspaces associated with data points of different classes. Benefiting from the joint spatio-temporal discriminant embedding, our method is potentially more powerful for classifying human actions with similar space-time shapes, and is able to perform recognition on a frame-by-frame or short video segment basis.

The rest of this paper is organized as follows. We briefly introduce some related work in Section 2. In Section 3, we present our LSTDE method, followed by experimental results and comparison in Section 4. Section 5 draws the conclusion.

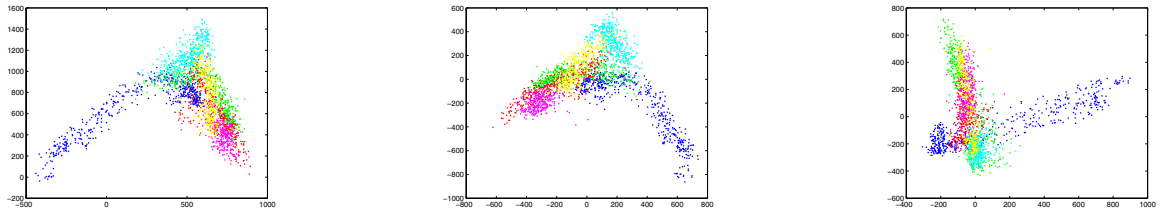
2. Related work

A few manifold learning methods have been proposed in the literature for human activity analysis. In [26], LLE was used to learn an activity manifold so that 3D body pose can be inferred by projecting a silhouette-based visual input into the learned embedding space. Sminchisescu and Jepson [27] used Laplacian eigenmap to derive a low-dimensional representation for tracking and reconstruction of 3D human motion in monocular video. The application of manifold learning methods to human action recognition was first reported in [12]. The authors used LPP to learn and match the dynamic shape manifolds of silhouette-based action sequences.

The seminal works of Isomap and LLE, together with Laplacian eigenmap, are difficult to map new data points to the low-dimensional embedding space. This limitation makes them unsuitable for classification tasks formulated under the inductive setting. LPP addresses this problem by finding the optimal linear approximations to the eigenfunctions of the Laplace-Beltrami operator on the manifold. As a result, LPP is linear and defined for both training and test data. In particular, LPP constructs a nearest neighbor graph. By using the Laplacian of the graph, LPP can find a mapping which optimally preserves the local neighborhood information. However, LPP is in general unsupervised and hence its mapping function is not necessarily optimal for maximizing class separability. Recently, some discriminant manifold embedding methods have been proposed. In particular, LDE constructs two nearest neighbor graphs, a within-class graph and a between-class graph, to model the local discriminant structure of the data manifold. MFA is essentially the same. However, these two methods consider the within-class and between-class relations as equally important, which may cause the discovered manifold structure of the within-class data points to be biased. On the other hand, LSDA first constructs one nearest neighbor graph, and then splits it into the within-class graph and the between-class graph. Consequently, LSDA is more flexible for data manifold analysis. The proposed method in this paper is more related to LSDA. An illustration of 2-dimensional manifold embedding using LPP, LSDA and our LSTDE method for human silhouette sequences of multiple action classes from [11] is shown in Fig. 1.

3. Local spatio-temporal discriminant embedding

In this section, we present our LSTDE algorithm specially designed for human action recognition. We choose silhouettes as feature representations for human action video sequences. Given h training human action sequences, we assume the associated sequences of moving silhouettes $\{X^q\}_{q=1}^h$ can be extracted from the original videos. The extracted silhouette images of each sequence contain foreground human body shape information, the size and position of which may vary across the whole video sequence. We thus center and normalize them so that the resulting nor-



LPP LSDA LSTDE

Figure 1. An illustration of 2-dimensional manifold embedding using LPP, LSDA and our LSTDE method for human silhouette sequences of multiple action classes from [11].

malized silhouette images are of equal image size and contain as much foreground body shape information as possible. We represent each frame of these silhouette sequences as a vector-based data point $x_i \in R^n$. Then any human action silhouette sequence X^q can be represented as $X^q = (x_1^q, x_2^q, \ldots, x_{N_q}^q)$, where N_q is the length of sequence X^q.

3.1. Objective functions

Suppose there are totally m data points $\{x_1, \ldots, x_m\}$ in all the h training silhouette sequences. We assume these data points to be living on or close to some dynamic shape manifold $\mathcal{M}$. The goal is then to derive a low-dimensional embedding that characterizes the local geometrical and discriminant properties of the data manifold. Using these m data points, we can build a nearest neighbor graph G to model the local geometrical structure of $\mathcal{M}$. For each data point x_i, we find its k nearest neighbors in G and introduce an edge between x_i and each of its neighbors. Then we get an affinity matrix W for G, which is defined as:

$$W_{ij} = \begin{cases} 1, & \text{if } x_i \text{ and } x_j \text{ are nearest neighbors} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

In order to discover the local spatial discriminant structure of the data manifold, we partition the set of k nearest neighbors of x_i into two subsets, one for data points with the same action class label as x_i and another for data points with different class labels. Accordingly, the nearest neighbor graph G yields two graphs, the within-class graph G_w and the between-class graph G_b. Let W_w and W_b be the affinity matrices of G_w and G_b respectively. Naturally we have $W = W_w + W_b$.

On the other hand, each x_i belongs to a silhouette video sequence X^q. There is a temporally continuous short video segment $S_i = (x_{i-t}, \ldots, x_i, \ldots, x_{i+t})$ associated with x_i, which has a segment length of $2t+1$ frames. In the following, we abuse the notation of $S_i = (x_{i-t}, \ldots, x_i, \ldots, x_{i+t}) \in R^{n\times(2t+1)}$, to describe a data matrix of a short video segment which contains vector-based data points in its columns. Obviously, all data points in S_i have the same class label as x_i. Suppose x_i and x_j are connected in graph G_b (i.e., they are nearest neighbors with different class labels), and S_i and S_j are their associated short video segments. In order to discover the local temporal discriminant structure of the data manifold, it is desirable that the principal angles between the linear subspaces formed by S_i and S_j be maximized after manifold embedding.

The aim of our algorithm is to find a transformation matrix $A \in R^{n\times l}$, with $l \ll n$, such that after embedding via the transformation $y_i = A^\top x_i$ has the following properties: (1) neighboring data points with the same class label are close in the low-dimensional embedding space; (2) neighboring data points with different class labels are far apart in the embedding space; (3) linear subspaces spanned by local short video segments S_i and S_j of different classes have large principal angles. To satisfy the above criteria, we need to optimize the following objective functions with respect to A:

$$\max_A \sum_{ij} \|A^\top x_i - A^\top x_j\|^2 W_{b,ij} \quad (2)$$

$$\min_A \sum_{ij} \|A^\top x_i - A^\top x_j\|^2 W_{w,ij} \quad (3)$$

$$\min_A \sum_{ij} F(A^\top S_i, A^\top S_j) W_{b,ij} \quad (4)$$

where $F(A^\top S_i, A^\top S_j)$ represents the similarity between the local short video segments S_i and S_j in the embedding space. When the similarity value is minimized, the principal angles between their corresponding linear subspaces are maximized. For notational simplicity, we rewrite the objective function (2) in the trace form as:

$$\begin{aligned} & \frac{1}{2}\sum_{ij} \|A^\top x_i - A^\top x_j\|^2 W_{b,ij} \\ = \; & \frac{1}{2}\sum_{ij} \mathrm{Tr}\{A^\top (x_i - x_j)(x_i - x_j)^\top A\} W_{b,ij} \\ = \; & \frac{1}{2}\mathrm{Tr}\Big\{A^\top \sum_{ij} \big((x_i - x_j) W_{b,ij} (x_i - x_j)^\top\big) A\Big\} \\ = \; & \mathrm{Tr}\{A^\top (X D_b X^\top - X W_b X^\top) A\} \\ = \; & \mathrm{Tr}\{A^\top X L_b X^\top A\}, \end{aligned} \quad (5)$$

where $X = (x_1, \ldots, x_m) \in R^{n \times m}$ is the data matrix, D_b is a diagonal matrix whose entries are row or column sums of W_b (since W_b is symmetric), and $L_b = D_b - W_b$ is the Laplacian matrix of G_b. Similarly, the objective function (3) can be rewritten as:

$$\min_A \mathrm{Tr}\{A^\top X(D_w - W_w)X^\top A\}, \tag{6}$$

where D_w is a diagonal matrix whose entries are row or column sums of W_w.

3.2. Computation of principal angles

Principal angles, also known as canonical correlations [30], have been used for recognizing human actions [31]. In [28], Kim *et al.* proposed a discriminative canonical correlation analysis method for image set classification. In general, there are many ways to compute the principal angles between the linear subspaces spanned by S_i and S_j. The singular value decomposition (SVD) solution is chosen here for its numerical stability [29]. Assume $P_i \in R^{n \times d}$ is an orthonormal basis matrix for S_i with $S_i S_i^\top \simeq P_i \Lambda_i P_i^\top$ and d is the dimensionality of the reduced subspace. Similarly we get P_j for S_j. We can compute the SVD of $P_i^\top P_j$ as:

$$P_i^\top P_j = \tilde{Q}_{ij} \Lambda \tilde{Q}_{ji}^\top, \tag{7}$$

where $\Lambda = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$. Canonical correlations, which are cosines of the principal angles, are the singular values. The orthonormal basis matrices of the subspaces of the data on the embedding space can be obtained as:

$$Y_i Y_i^\top = (A^\top S_i)(A^\top S_i)^\top \simeq (A^\top P_i)\Lambda_i (A^\top P_i)^\top. \tag{8}$$

Note that canonical correlations are only defined for the orthonormal basis matrices of subspaces as in equation (7). However, $A^\top P_i$ is generally not orthonormal, but we can normalize it using QR factorization. Specifically, $A^\top P_i$ can be decomposed as:

$$A^\top P_i = \Psi_i \Delta_i, \tag{9}$$

where $\Psi_i \in R^{n \times d}$ is the orthonormal matrix formed by the first d columns and $\Delta_i \in R^{d \times d}$ is an invertible upper triangular matrix. Then the normalized P_i' can be computed as:

$$P_i' = P_i \Delta_i^{-1}. \tag{10}$$

We use the normalized $A^\top P_i'$ to represent the orthonormal basis matrix of the embedded data. Given the SVD computation:

$$(A^\top P_i')^\top (A^\top P_j') = Q_{ij} \Lambda Q_{ji}^\top, \tag{11}$$

the similarity of the local short video segments S_i and S_j in the embedding space is defined as the sum of the canonical correlations:

$$\begin{aligned} F(A^\top S_i, A^\top S_j) &= \max_{Q_{ij}, Q_{ji}} \mathrm{Tr}\{Q_{ij}^\top P_i'^\top A A^\top P_j' Q_{ji}\} \\ &= \max_{Q_{ij}, Q_{ji}} \mathrm{Tr}\{A^\top P_j' Q_{ji} Q_{ij}^\top P_i'^\top A\}. \end{aligned} \tag{12}$$

By simple linear algebra, we can show that $2\,\mathrm{Tr}\{A^\top A - A^\top P_j' Q_{ji} Q_{ij}^\top P_i'^\top A\}$ is equivalent to $\mathrm{Tr}\{A^\top (P_j' Q_{ji} - P_i' Q_{ij})(P_j' Q_{ji} - P_i' Q_{ij})^\top A\}$. So the objective function (4) can be reformulated as:

$$\max_A \mathrm{Tr}\{A^\top C_b A\}, \tag{13}$$

where

$$C_b = \sum_{ij} (P_j' Q_{ji} - P_i' Q_{ij})(P_j' Q_{ji} - P_i' Q_{ij})^\top W_{b,ij}. \tag{14}$$

3.3. Optimal embedding

Since each entry of D_w in (6) is a row or column sum of the within-class affinity matrix W_w, it provides a natural measure for the data points. Specifically, if $D_{w,ii}$ is large, then the class containing x_i has a high density around x_i. We therefore impose $\mathrm{Tr}\{A^\top X D_w X^\top A\} = 1$ and hence the objective function in (6) becomes:

$$\min_A 1 - \mathrm{Tr}\{A^\top X W_w X^\top A\}, \tag{15}$$

or

$$\max_A \mathrm{Tr}\{A^\top X W_w X^\top A\}. \tag{16}$$

Given the reformulated objective functions (5), (16) and (13), the overall optimization problem for LSTDE becomes:

$$\begin{aligned} \max_A \big\{ \beta \left(\alpha \mathrm{Tr}\{A^\top X L_b X^\top A\} + (1-\alpha)\mathrm{Tr}\{A^\top C_b A\} \right) \\ + (1-\beta)\mathrm{Tr}\{A^\top X W_w X^\top A\} \big\} \\ s.t. \quad \mathrm{Tr}\{A^\top X D_w X^\top A\} = 1, \end{aligned} \tag{17}$$

where α, β are parameters with $0 \le \alpha \le 1$, $0 \le \beta \le 1$. Thus, the columns of an optimal A can be obtained as the generalized eigenvectors corresponding to the l largest eigenvalues of the following generalized eigenvalue equation:

$$\begin{aligned} \left(\beta \left(\alpha X L_b X^\top + (1-\alpha) C_b \right) + (1-\beta) X W_w X^\top \right) \mathbf{a} \\ = \lambda X D_w X^\top \mathbf{a}. \end{aligned} \tag{18}$$

The choices of free parameters α and β in equation (18) may have different effects on the optimal A. When a larger β value is used, the derived low-dimensional embedding space using A favors more the local between-class data discrimination. On the other hand, when a smaller β value is used, the derived embedding space can preserve more local within-class neighborhood relations. At the same time, the value of α can balance the local spatial and local temporal discriminating capabilities.

3.3.1 Iterative learning

The generalized eigenvalue problem in equation (18) requires the prior knowledge of C_b, while the computation of C_b in (14) involves the variables P' and Q which can only be obtained given an estimate of A. We thus propose an iterative method to optimize A. More specifically, given a current estimate of A, the normalized P' can be computed using equations (9) and (10), then Q can be computed by SVD as in equation (11). After obtaining the variables P' and Q, we can compute C_b based on equation (14) and then update A by solving the generalized eigenvalue problem (18). The pseudocode for the iterative optimization of LSTDE is given in **Algorithm** 1.

Algorithm 1: Local Spatio-Temporal Discriminant Embedding (LSTDE)

input : All m data points $X = (x_1, \ldots, x_m) \in R^{n\times m}$ with class labels
output: $A \in R^{n\times l}$
Preprocessing:
1. Construct W_w, W_b from m data points
2. Compute D_w, D_b, L_b from W_w, W_b
3. Get S_i, S_j for all $W_{b,ij} \neq 0$
4. Compute $P_i \in R^{n\times d}$ for all S_i: $S_i S_i^\top \simeq P_i \Lambda_i P_i^\top$
Initialize:
5. $A = I \in R^{n\times n}$
Iterate:
6. For all P_i, do: $A^\top P_i = \Psi_i \Delta_i \rightarrow P'_i = P_i \Delta_i^{-1}$
7. For all P'_i, P'_j pairs, do: $(A^\top P'_i)^\top (A^\top P'_j) = Q_{ij} \Lambda Q_{ji}^\top$
8. Compute
$C_b = \sum_{ij} (P'_j Q_{ji} - P'_i Q_{ij})(P'_j Q_{ji} - P'_i Q_{ij})^\top W_{b,ij}$
9. Compute eigenvectors $\{\mathbf{a}_i\}_{i=1}^n$ by (18), $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$
End
10. $A = [\mathbf{a}_1, \ldots, \mathbf{a}_l]$

4. Experiments

4.1. Data settings

We evaluate our manifold embedding method using the human action dataset from [11]. This dataset contains 10 action classes performed by nine different human subjects. The actions include bending (bend), jumping jack (jack), jumping-forward-on-two-legs (jump), jumping-in-place-on-two-legs (pjump), running (run), galloping-sideways (side), skipping (skip), walking (walk), waving-one-hand (wave1), and waving-two-hands (wave2). There are totally 93 sequences since some types of actions are performed twice by some subjects. We do not intend to address the foreground detection issue in this work, so the silhouette masks obtained in [11] are directly used in our experiments. We center and normalize all silhouette frames into the same 64×48 dimension, and convert them into 3072-dimensional vectors in a raster-scan manner. Some example silhouette frames of different actions are shown in Fig. 2. To avoid the singular matrix problem and also for computational efficiency, we preprocess the data using PCA so that 98% information is kept in the sense of low rank approximation.

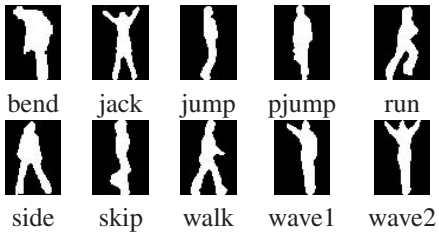


Figure 2. Normalized example silhouette frames of different actions.

We use nine-fold cross validation for evaluation. Each time we take silhouette frames of 10 action sequences of one subject as test data, and use those of the remaining eight subjects for LSTDE learning. The recognition accuracy is averaged over nine runs of cross validation.

4.2. Design of a two-stage recognition scheme

The data point of each test silhouette frame is first projected into the embedding space using the optimal projection matrix A learned, which results in a low-dimensional representation of the test data point. To make better use of the temporal shape variation information between different actions, we design a two-stage recognition scheme.

In the first stage, the test action sequence is recognized on a frame-by-frame basis. For simplicity, we use a variant of k-nearest neighbor classifier as the baseline classification method in this stage, where the value of k can be set to be the same as that of the nearest neighbor graph G. However, similar body shapes may occur during the execution of different actions, which essentially correspond to the manifold intersection areas of different actions in Fig. 1. Simple nearest neighbor classification based on individual frames may fail in these highly confusing areas. We then consider using more frames in the test sequence, e.g., a short video segment, in the second stage.

In particular, k nearest neighbors are first found using Euclidean distance for the low-dimensional embedding $y_i = Ax_i$ of each test frame x_i. If all these k nearest neighbors (rather than the majority of the k nearest neighbors for a standard k-nearest neighbor classifier) have the same class label, then the test frame x_i is recognized as that class. Otherwise x_i probably falls inside some confusing area. We then use a short segment $(x_{i-t}, \ldots, x_i, \ldots, x_{i+t})$ which temporally centers at x_i in the test sequence. After projection using A, these $2t+1$ data points $(y_{i-t}, \ldots, y_i, \ldots, y_{i+t})$ in the embedding space form a local temporal linear subspace. Similarly, those k nearest neighbors have their associated short video seg-

Classifier	Recognition accuracy (%) of different methods					
	PCA	LDA	LPP	LDE	LSDA	LSTDE
k-nearest neighbor	81.21 (58)	84.11 (8)	83.35 (31)	84.34 (40)	83.50 (34)	83.46 (28)
two-stage recognition scheme	84.03 (58)	82.83 (8)	85.69 (31)	84.18 (40)	85.86 (34)	90.91 (28)

Table 1. Recognition accuracy using different manifold embedding methods. The number in parentheses is the optimal dimensionality l of the embedding space.

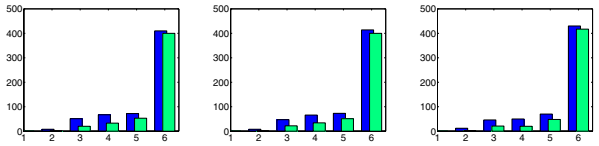


Figure 3. Histograms of both the correctly classified and total test data points for different majority numbers $2, 3, 4, 5, 6$ in 6 nearest neighbors when using a standard k-nearest neighbor classifier, with respect to LPP, LSDA and our method. The green bars stand for correctly classified data points from all the test data points which are represented by the blue bars.

ments which form their corresponding local temporal subspaces in the embedding space. We compute the principal angles between the test local temporal subspace and those subspaces associated with the k nearest neighbors, and then the test frame x_i is recognized as the class of the nearest neighbor with the smallest principal angles between them. We expect better performance can be achieved by this two-stage recognition scheme, based on three justifications: (1) We choose a variant of k-nearest neighbor classifier in the first stage to distinguish the test silhouette frames in the confusing areas of different actions. These test frames are likely to be the major source of misclassification, which will be verified in Fig. 3. (2) In the second recognition stage, we use short video segments rather than single frames in these highly confusing areas. (3) Our proposed LSTDE method finds optimal embedding in the sense of both local spatial and local temporal discriminating capabilities.

4.3. Experimental results and comparison

We compare our method with some representative manifold embedding methods, namely, PCA, LDA [25], LPP [17], LDE [21], and LSDA [22]. For each method, both a standard k-nearest neighbor classifier and our newly designed two-stage recognition scheme are used for comparison. In particular, we choose $t = 3$, so 7-frame local temporal video segments are used in the two-stage recognition scheme. If longer segments are used, higher recognition accuracy may be achieved. For LPP, LSDA and our method, the number of nearest neighbors in the graph G is empirically chosen to be $k = 6$. The parameters α, β in equation (18) are also optimized as $\alpha = 0.1$ and $\beta = 0.5$. Table 1 summarizes the recognition results of different methods, with the optimal reduced dimensionality l shown in parentheses. From Table 1, we can draw the following conclusions:

1. Manifold learning of human actions represented by silhouette sequences is indeed very effective for classifying different types of actions. Recognition even based on individual frames only can give fairly good results.
2. Our two-stage recognition scheme is very effective, especially for PCA, LPP, LSDA, and our method. This is also consistent with what we have expected in Section 4.2.
3. When using the standard k-nearest neighbor classifier for recognition, LDA and LDE are slightly better than the other methods. It may be because they put more emphasis on globally separating data points of different classes. However, exploiting less local geometrical structure of each within-class data manifold makes them not appealing, especially for human action sequences with intrinsic dynamic shape manifolds.
4. Based on the two-stage recognition scheme, our LSTDE method greatly outperforms all other methods, which essentially benefits from both the local spatial and local temporal discriminating capabilities.

To investigate what test data points are the major source of misclassification when using the standard k-nearest neighbor classifier, we plot in Fig. 3 the histograms of both the correctly classified and total test data points for different majority numbers in k nearest neighbors, e.g. the majority numbers $2, 3, 4, 5, 6$ in 6 nearest neighbors, with respect to LPP, LSDA and LSTDE. Fig. 3 suggests that most of the misclassified data points come from the majority numbers $2, 3, 4, 5$ when using a standard 6-nearest neighbor classifier. Essentially among these data points our LSTDE method outperforms the other methods in the second recognition stage. From Table 1 and Fig. 3, we can conclude that our two-stage recognition scheme is an effective design for human action recognition.

We also investigate which actions are misclassified by showing confusion tables in Fig. 4 with respect to the best three methods LPP, LSDA and LSTDE. The elements in each row of the confusion tables represent the probabilities that data points of certain action are classified as some other actions. Fig. 4 shows that there exist highly similar space-time human body shapes between the actions *jump*,

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	**1.0**									
jack		**.99**		.01						
jump	.03		**.79**		.10	.05	.03			
pjump		.02		**.93**		.05				
run			.02		**.62**	.04	.22	.10		
side				.23	.02	**.65**	.05	.05		
skip			.20		.16		**.62**	.02		
walk					.02	.03		**.95**		
wave1				.21					**.79**	
wave2		.03		.02					.05	**.90**

LPP

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	**1.0**									
jack		**.98**		.01						.01
jump	.03		**.84**		.03	.05	.05			
pjump		.05		**.95**						
run					**.60**	.07	.22	.11		
side				.23		**.72**	.02	.03		
skip			.24		.27		**.49**			
walk						.02		**.98**		
wave1				.24					**.76**	
wave2		.03							.03	**.94**

LSDA

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	**1.0**									
jack		**1.0**								
jump	.03		**.89**		.03	.05				
pjump		.02		**.98**						
run					**.76**	.04	.13	.07		
side				.19		**.79**		.02		
skip			.20		.16		**.64**			
walk								**1.0**		
wave1				.20					**.80**	
wave2									.02	**.98**

LSTDE

Figure 4. Confusion tables with respect to LPP, LSDA and LSTDE. The elements in each row represent the probabilities that data points of certain action are classified as some other actions.

skip and *run*, between *side* and *pjump*, and between *jack*, *pjump* and *wave*. While other methods perform worse, our method gives better results especially among these confusing actions.

In the above experiments, 7-frame local temporal video segments were used in the second recognition stage. One may expect to get higher recognition accuracy if longer segments are used. We also perform experiments to investigate the influence of different video segment lengths on the recognition performance. In particular, we choose $t \in \{1, 3, 5, 7, 9, 12\}$. The average recognition results using cross validation for different methods are plotted in Fig. 5. The results suggest that longer video segments do improve the performance for all methods. The best results can be achieved when t is between $7 \sim 10$, i.e., when video segments of $15 \sim 21$ frames in length are used. When more than 21 frames are used, the recognition accuracy cannot be increased further. Since the human action sequences in the dataset [11] contain periodic body motions, a video segment of 21 frames may already contain one whole action cycle.

4.4. Robustness test

The recognition of human actions can be further challenged when action sequences are captured in front of non-uniform backgrounds, with partial occlusions and non-rigid deformations, at changing viewpoints, etc. To investigate the robustness of our method to these high irregularities of real-world actions, we perform further experiments using 10 video sequences of people walking in various difficult scenarios [11]. In particular, these videos include diagonal walking (changing scale and viewpoint), walking with a dog (non-rigid deformation), walking when swinging a bag (rigid deformation), walking in a skirt (changing clothes), walking with partially occluded legs (partial occlusion), sleepwalking (different style), limping (different style), walking with knees up (different style), walking when carrying a briefcase (carrying objects), and normal walking (background change). Some example frames and their associated segmented silhouettes are shown in Fig. 6. Table 2 reports the frame-based recognition accuracy of these 10 walking sequences using our LSTDE method. From Table 2, we can see that except for four sequences (walk when swinging a bag, walk with partially occluded legs, sleepwalk, and walk with knees up), the recognition results for all the other walking sequences with high irregularities are comparable with those results in Fig. 4 for sequences captured under normal conditions, which demonstrates that our method has low sensitivity to these challenging factors.

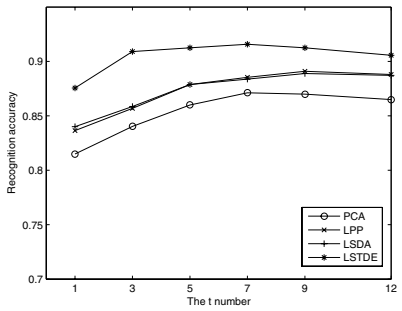


Figure 5. Recognition performance using different video segment lengths.

Test sequence	Recognition accuracy (%)
Diagonal walk	82.14
Walk with a dog	90.74
Walk when swinging a bag	74.58
Walk in a skirt	92.16
Walk with partially occluded legs	71.19
Sleepwalk	56.06
Limpwalk	83.96
Walk with knees up	66.02
Walk when carrying a briefcase	92.86
Normal walk	95.16

Table 2. Recognition results of robustness test.

5. Conclusion

In this paper, we address human action recognition using manifold learning methods. In particular, we propose a novel local spatio-temporal discriminant embedding (LSTDE) method. LSTDE can find an optimal embedding in the sense of both local spatial and local temporal discriminating capabilities. Our method is able to perform recog-

Figure 6. Example frames of walking sequences for robustness test (from left to right and top to bottom): walk with a dog, walk when swinging a bag, walk in a skirt, walk with partially occluded legs, sleepwalk, limp, walk with knees up, and walk when carrying a briefcase.

nition on a frame-by-frame or short video segment basis. We use silhouette sequences of human actions as feature representations. Experimental results demonstrate that our method can accurately recognize human actions, and outperforms some representative manifold embedding methods.

References

[1] S. Carlsson and J. Sullivan, Action recognition by shape matching to key frames, *Workshop on Models versus Exemplars in Computer Vision*, 2001. 1

[2] R. Collins, R. Gross, and J. Shi, Silhouette-based human identification from body shape and gait, *AFG*, 2002. 1

[3] A. Efros, A. Berg, G. Mori, and J. Malik, Recognizing action at a distance, *ICCV*, 2003. 1

[4] C. Schuldt, I. Laptev, and B. Caputo, Recognizing human actions: a local SVM approach, *ICPR*, 3, pp. 32-36, 2004. 1

[5] C. Bregler, Learning and recognizing human dynamics in video sequences, *CVPR*, 1997. 1

[6] Y. Yacoob and M.J. Black, Parametrized modeling and recognition of activities, *CVIU*, vol.73, no.2, pp.232-247,1999. 1

[7] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, Behavior recognition via sparse spatio-temporal features, *ICCV Workshop: VSPETS*, 2005. 1

[8] Y. Ke, R. Sukthankar, and M. Hebert, Efficient visual event detection using volumetric features, *ICCV*, pp.166-173, 2005. 1

[9] I. Laptev, On space-time interest points, *IJCV*, 64(2-3), pp. 107-123, 2005. 1

[10] A. Bobick and J. Davis, The recognition of human movement using temporal templates, *PAMI*, 23(3) pp. 257-267, 2001. 1

[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, Action as space-time shapes, *PAMI*, vol. 29, no. 12, pp. 2247-2253, 2007. 1, 2, 3, 5, 7

[12] L. Wang and D. Suter, Learning and matching of dynamic shape manifolds for human action recognition, *TIP*, vol. 16, no. 6, pp. 1646-1661, 2007. 1, 2

[13] L. Wang and D. Suter, Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model, *CVPR*, 2007. 1

[14] J.C. Niebles, H. Wang, and L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, *BMVC*, 2006. 1

[15] F. Lv and R Nevatia, Single view human action recognition using key pose matching and Viterbi patch searching, *CVPR*, 2007. 1

[16] S. Ali, A. Basharat, and M. Shah, Chaotic invariants for human action recognition, *ICCV*, 2007. 1

[17] X. He and P. Niyogi, Locality preserving projections,*NIPS*, 2003. 1, 6

[18] J. B. Tenenbaum, V. de Silva, and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science*, vol. 290, pp. 2319-2323, 2000. 2

[19] S. Roweis and L. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science*, vol. 290, pp. 2323-2326, 2000. 2

[20] M. Belkin and P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, *NIPS*, pp. 585-591, 2001. 2

[21] H.-T. Chen, H.-W. Chang, and T.-L. Liu, Local discriminant embedding and its variants, *CVPR*, 2005. 2, 6

[22] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, Locality sensitive discriminant analysis, *IJCAI*, pp. 708-713, 2007. 2, 6

[23] M. Sugiyama, Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis, *Journal of Machine Learning Research*, 8, pp. 1027-1061, 2007. 2

[24] S. Yan, D. Xu, B. Zhang, and H.-J. Zhang, Graph embedding: A general framework for dimensionality reduction, *CVPR*, 2005. 2

[25] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, Wiley-Interscience, Hoboken, NJ, 2nd edition, 2000. 6

[26] A. Elgammal and C-S. Lee, Inferring 3D body pose from silhouettes using activity manifold learning, *CVPR*, 2 pp. 681-688, 2004. 2

[27] C. Sminchisescu and A. Jepson, Generative modelling for continuous non-linearly embedded visual inference, *ICML*, pp. 140-147, 2004. 2

[28] T.-K. Kim, J. Kittler, and R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, *PAMI*, vol. 29, no. 6, 2007. 4

[29] A. Bjorch and G.H. Golub, Numerical methods for computing angles between linear subspaces, *Math. Computation*, vol. 27, no. 123, pp. 579-594, 1973. 4

[30] H. Hotelling, Relations between two sets of variates, *Biometrika*, vol. 28, no. 34, pp. 321-372, 1936. 4

[31] Y. Sheikh, M. Sheikh, and M. Shah, Exploring the space of a human action, *ICCV*, 2005. 4