Analyzing Human Movements from Silhouettes using Manifold Learning

Liang Wang and David Suter
ARC Centre for Perceptive and Intelligent Machines in Complex Environments
Monash University, Clayton, VIC, 3800, Australia
{liang.wang, d.suter}@eng.monash.edu.au

Abstract*

A novel method for learning and recognizing sequential image data is proposed, and promising applications to vision-based human movement analysis are demonstrated. To find more compact representations of high-dimensional silhouette data, we exploit locality preserving projections (LPP) to achieve low-dimensional manifold embedding. Further, we present two kinds of methods to analyze and recognize learned motion manifolds. One is correlation matching based on the Hausdorrf distance, and the other is a probabilistic method using continuous hidden Markov models (HMM). Encouraging results are obtained in two representative experiments in the areas of human activity recognition and gait-based human identification.

1. Introduction

Visual analysis of human movements [2] aims to detect, track and recognize people, and more generally, to understand human behaviours. Interest in this is strongly driven by a wide spectrum of promising application areas such as smart surveillance, perceptual interface, etc.

Previous studies extract various features from raw video data for human motion analysis, e.g., optical flow [10], spatiotemporal gradients [11], local descriptors [12], the tracked trajectories [8], etc. However, tracking is complex due to the large variability in the shape and articulation of the human body. When using image measurements in terms of spatiotemporal gradients, optical flow or other intensity-based features, the recognition results depend greatly on the image recording conditions. In contrast, human silhouette extraction from videos is easier and more feasible for current vision techniques, especially in the environments with stationary cameras.

Human motion can be regarded as temporal variations of human silhouettes. Thus the method that we present prefers to directly analyze moving silhouettes for human movement analysis. Since all images collected during movements generally lie on a low dimensional manifold embedded in the high dimensional image space, it will be ideal to analyze human motions in a more compact low dimensional space. Recently, some promising frameworks for dimensionality reduction have been introduced, e.g., isometric feature mapping (Isomap) [17], local linear embedding (LLE) [16] and locality preserving projections (LPP) [18]. Accordingly, some researchers are exploring these newer methods for different vision applications, e.g., Elgammal and Lee [19] proposed an approach to inferring 3D body pose from silhouettes using gait manifold learned by LLE. Wang *et al*. [20] learned the intrinsic object structure by Isomap to enhance tracking of parameterized contours. However, research on the manifold learning for more complex human movement analysis and recognition is still very limited.

Based on the above considerations, this paper proposes an effective framework to analyze human movements from silhouettes, in which we explore *LPP* to achieve the low-dimensional embedding of dynamic silhouette data. Two kinds of methods are then presented to recognize the learned manifolds, one of which is the nearest manifold method using the mean *Hausdorrf distance* metric, and the other is a probabilistic modelling and recognition method based on *HMM*. We demonstrate real applications of the proposed method to *human gait and activity analysis*.

The main purpose and contributions of this paper are summarized as follows. 1) Our aim is to examine the feasibility of using the features available directly from (probably imperfect) space-time silhouettes for analyzing human motions. 2) We propose a general framework for visual learning and recognition of sequential silhouette data. 3) We successfully exploit real application of LPP to discover intrinsic structure of dynamic data manifolds. 4) Two kinds of recognition methods are presented in the manifold subspace. Their good performance for human gait and activity recognition is examined. 5) Relatively, the proposed method is easy to understand and implement. The use of only binary silhouette cue make our method free from some problems arising in most previous studies, e.g., imperfect 2D or 3D feature tracking, expensive and noise-sensitive optical flow computation, etc.

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

2. Related work

In this paper, two areas of interest are human activity recognition and gait-based human identification which we now briefly survey.

Human activity recognition: Traditional methods of human activity analysis are based on tracking models in either 2D or 3D spaces [8,15]. In the work of Yaccob and Black [8], an action was represented by 40 curves derived from the tracking results of five body parts of a cardboard people model. Other work obtains intensity or gradient based features for motion recognition. Zelnik-Manor and Irani [11] used marginal histograms of spatiotemporal gradients at a few temporal scales to cluster and recognize video events. The work of Efros *et al*. [10] adopted a spatiotemporal descriptor based on blurred optical flow measurements to recognize actions on ballet, tennis and football datasets. There has also been significant interest in approaches that exploit local descriptors on interest points in images or videos. Schuldt *et al*. [12] constructed video representations in terms of local space-time features for action recognition. Silhouette-based methods are becoming popular. Bobick and Davis [13] proposed view-based temporal templates for representation and recognition of aerobics actions. Blank *et al*. [14] performed action recognition by utilizing the properties of the solution to the Poisson equation to extract features from the space-time silhouettes.

Gait recognition: Recently, some methods have been suggested for the task of human identification by using gait [3-7], based on the observation that people can recognize others by simply observing their gaits. Most of the existing methods extract features from the silhouettes of the person and identify individuals based on those features or their temporal variations. Collins *et al*. [4] established a method based on template matching of body silhouettes in key frames for human identification. Lee *et al.* [6] described a moment-based representation of gait appearance for the purposes of person identification and gender classification. Phillips *et al*. [5] proposed a baseline algorithm for human identification using direct spatiotemporal correlation of silhouette images.

3. Learning motion manifolds

It is a formidable task to learn the complete structure of the motion manifold in the high dimensional image space. Our idea is to embed the nonlinear manifold of human motions in a low dimensional subspace for more compact feature extraction and representation.

3.1. Visual silhouette inputs

Our basic assumption is that an associated sequence of foreground silhouettes of a moving person can be obtained from the original video. Each silhouette image is then centred and normalized on the basis of keeping the aspect ratio property of the silhouette so that the resulting images contain as much foreground as possible, do not distort the motion shape, and are of equal dimensions for all input frames. We directly use the normalized silhouette images as visual inputs for manifold learning. Figure 1 shows two examples of visual silhouette inputs.

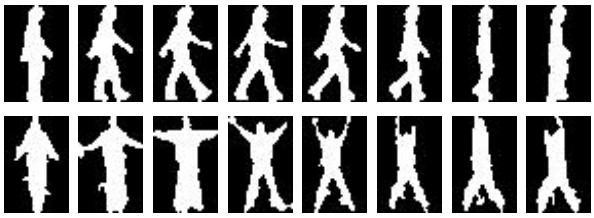

Figure 1. Examples of visual inputs from the actions of walking (*top*) and jumping jack (*bottom*)

3.2. Manifold learning using LPP

We choose LPP to find manifold subspaces based on a few reasons: a) LPP can explicitly model and discover the intrinsically nonlinear manifold structure of motions by the use of an adjacency graph; b) Like LLE, LPP has locality preserving characteristic, which makes it less sensitive to outliers; c) LPP is a linear embedding, thus it is computationally more efficient than nonlinear methods; and d) Many nonlinear methods (e.g., Isomap and LLE) are defined only on the training data points and how to evaluate the maps on new test data points remains unclear. However LPP can be easily applied to any new data points.

According to [18], the major procedure of manifold learning using LPP is described as follows.

Construct the data matrix. Given m different classes of motions and each class represents a sequence of input silhouettes. Each silhouette image with the resolution of $r \times c$ is represented by an h-dimensional ($h=r \times c$) vector $\boldsymbol{f}$ in a raster-scan manner. Let $\boldsymbol{f}_{i,j}$ be the jth input frame in the ith class and n_i the number of such inputs in the ith class. The total number of training samples is $n=n_1+n_2+\cdots+n_m$, and the whole training data set can be represented by $\boldsymbol{X}=[\boldsymbol{f}_{1,1}, \boldsymbol{f}_{1,2}, \cdots, \boldsymbol{f}_{1,n1}, \boldsymbol{f}_{2,1}, \cdots, \boldsymbol{f}_{m,nm}]=[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$.

Construct the adjacency graph. Let $\boldsymbol{G}$ be a graph with n nodes. An edge will be put between nodes i and j if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are close, where the 'close' can be defined by ε-neighbourhoods ($\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 < \varepsilon, \varepsilon \in R$) or the K-nearest neighbours [18]. We choose the K-nearest neighbours to construct the adjacency graph, i.e., $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ will be connected if $\boldsymbol{x}_i$ is among the K-nearest neighbours of $\boldsymbol{x}_j$ or $\boldsymbol{x}_j$ is among the K-nearest neighbours of $\boldsymbol{x}_i$. To measure the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, we use the cosine similarity

$$cos(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{(\left|\boldsymbol{x}_i\right| \cdot \left|\boldsymbol{x}_j\right|)} \quad (1)$$

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

We also use the supervised form of LPP (named SLPP) by integrating the class information when constructing the affinity graph. That is, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ will be directly connected if they belong to the same class.

Choose the weights. The weight matrix $\boldsymbol{W}$ is a sparse symmetric $n \times n$ matrix with w_{ij} representing the weight of the edge joining vertices i and j, and 0 if there is no such edge. There are two kinds of variations for weighting, i.e., heat kernel ($w_{ij} = e^{-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 / t}$, $t \in R$) and simple-minded 0-1 weighting [18]. We choose the 0-1 weighting rule.

Eigenmaps: Compute the eigenvectors and eigenvalues for the generalized eigenvector problem [18]

$$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{e} = \gamma \boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{e} \tag{2}$$

where $\boldsymbol{D}$ is a diagonal matrix whose entries are column (or row) sums of $\boldsymbol{W}$, i.e., $D_{ii}=\sum_j w_{ji}$, $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$ is the Laplacian matrix. Let the column vectors $\boldsymbol{e}_0$, ⋯, $\boldsymbol{e}_{l-1}$ be the solutions of (2), ordered according to their eigenvalues $\lambda_0<\lambda_1<\cdots<\lambda_{l-1}$. The embedding is represented by

$$\boldsymbol{y}_i = \boldsymbol{E}^T \boldsymbol{x}_i , \quad \boldsymbol{E} = \left[\boldsymbol{e}_0, \boldsymbol{e}_1, \cdots, \boldsymbol{e}_{l-1}\right] \tag{3}$$

Each data point is embedded into a point in the low dimensional feature space, thus a movement is mapped into a curve with temporal order in such subspace.

4. Recognizing motion manifolds

4.1. Matching-based method

The manifold curves of movements can be themselves used in a naive way for matching-based recognition. Since the computed manifold of each motion sequence depends on its duration and temporal shift, an ideal distance metric should be able to handle such changes. The Hausdorff distance provides an elegant solution by determining the resemblance of one point set to another. The manner in which it is computed implicitly includes temporal constraints between observation vectors.

Motion similarity measure: Assume that two motion sequences are respectively projected into $\boldsymbol{M}_1$ ($l \times T_1$) and $\boldsymbol{M}_2$ ($l \times T_2$), where l is the reduced dimensionality, and T_1 and T_2 are the durations of these two motions, respectively. A variant of the Hausdorff metric, i.e., the mean value of the minimums, is used here.

$$S\left(\boldsymbol{M}_1, \boldsymbol{M}_2\right) = \underset{1 \le i \le T_1}{mean}\left(\underset{1 \le j \le T_2}{min}\left(\left\| \frac{\boldsymbol{M}_1(i)}{\left\|\boldsymbol{M}_1(i)\right\|} - \frac{\boldsymbol{M}_2(j)}{\left\|\boldsymbol{M}_2(j)\right\|} \right\| \right) \right) \tag{4}$$

Since the Hausdorff distance is oriented, the similarity measure is thus modified to ensure symmetry

$$d = S\left(\boldsymbol{M}_1, \boldsymbol{M}_2\right) + S\left(\boldsymbol{M}_2, \boldsymbol{M}_1\right) \tag{5}$$

Nearest-manifold classification: Motion classification is performed in a nearest neighbour framework

$$c_1 = arg\ \underset{i}{min}\, d\left(TM, RM_i\right) \tag{6}$$

where RM_i represent the ith reference motion pattern, i=1, 2, …, m, and TM is a test sequence.

4.2. State-space method

Although the Hausdorff distance can reflect temporal association of motions, it is not explicit in modelling such temporal constraints. Also, the matching-based method is subject to individual-frame noise in input data. State-space models are more ideal to explicitly represent temporal transition process of the movement. In particular, HMMs [1] have been demonstrated to a potent tool for analyzing time-varying data, and sophisticated algorithms for the HMM-based learning and recognition are available.

Parameter training of HMM: In the training stage, we specify the number of states for each class of motion empirically, and use the data-driven design of HMM with no restriction of the topology. In detail, the model parameters describing an HMM is represented by the triplet $\gamma=\{\pi_j, a_{ij}, b_j\}$, where π_j is the initial probability of the jth state being the first state, a_{ij} denotes the transition probability of the jth state occurring immediately after the ith one, and b_j is the probability for a feature vector $\boldsymbol{O}_t$ conditioned on the jth state. Assume that the state set is $q_t \in \{s_1, \cdots, s_{Ns}\}$ with the number of states N_s, and the state-conditional observation density is simply modelled as a multivariate Gaussian model, we have

$$\begin{aligned} &\pi_i = \mathrm{P}\left(q_1 = s_i\right), \quad 1 \le i \le N_s \\ &a_{ij} = \mathrm{P}\left(q_{t+1} = s_j \mid q_t = s_i\right), \quad 1 \le i, j \le N_s \\ &b_j\left(\boldsymbol{O}_t\right) = \mathrm{P}\left(\boldsymbol{O}_t \mid q_t = s_j\right) = \mathrm{N}\left(\boldsymbol{O}_t; \boldsymbol{u}_j, \Sigma_j\right) \end{aligned} \tag{7}$$

The parameters of the HMM are initialized to random values and the Baum-Welch algorithm is used to estimate the parameters iteratively using the forward-backward procedure [1]. Given a set of motion manifolds from the same class, we may extend the training to include multiple sequences. At each time of iteration, the contribution from individual sequences is summed up in the procedure of forward-backward parameter estimation.

HMM-based recognition: Once the separate HMMs are trained for all classes of motions, recognition of a new test sequence can be performed based on the likelihood computed for the input in terms of individual HMMs. Given m classes of HMMs γ_1, γ_2, ⋯, γ_m, and the associated manifold $\boldsymbol{Y}=[\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T]$ of a test sequence, this test is declared to belong to the class c_2 represented by the HMM with the maximum likelihood, i.e.,

$$c_2 = arg\ \underset{i}{max}\, P\left(\boldsymbol{Y} \mid \gamma_i\right) \tag{8}$$

5. Experiments

5.1. Experiment I: Human activity recognition

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

Due to the lack of a common evaluation database in the domain of human activity recognition, we use a recent database reported in [14][*]. To the best of our knowledge, this database is one of few concurrent action databases available in the public domain, and is appreciably sized in terms of the number of subjects, actions and videos. It consists of 81 low-resolution videos (180×144, 25fps) from 9 people, each performing 9 natural activities, i.e., bending (bend), jumping-forward-on-two-legs (jump), jumping-in-place-on-two-legs (pjump), jumping jack (jack), running (run), walking (walk), galloping-sideways (side), waving-one-hand (wave1), and waving-two-hands (wave2). Together with one more recently added activity of skipping (skip), this dataset in total includes 10 activities and 90 videos. The sample images are shown in Figure 2. Different people have different physical sizes and perform activities differently both in styles and speeds. This dataset asks different people to perform the same activities, thus providing more realistic data for the test of the method's versatility.

Figure 2. Example images of each kind of activity. From top left to bottom right: bend, jack, jump, pjump, run, side, skip, walk, wave1, and wave2, respectively

We directly adopt the masks from [14] for subsequent processing. Whether the other activities in this dataset are in essence periodic or not, people are asked to perform them multiple times in a repetitive manner (except for bending). We extract 198 sequences from the original videos by periodicity detection and segmentation, each of which includes a complete action. The numbers of each kind of activity sequences are respectively 9, 23, 24, 27, 14, 22, 25, 16, 19, and 19 for bend, jack, jump, pjump, run, side, skip, walk, wave1, and wave2. We normalize all silhouette images into the same dimension (i.e., 48×32 pixels). Each image is denoted by a 1536-dimensional vector, and a considerable number of such visual inputs are used to learn activity manifolds. Figure 3 shows spatiotemporal projections of activities, where the points with same colours are from the same activity (Bend Jack Jump Pjump Run Side Skip Walk Wave1 Wave2). From Figure 3, we can see that the SLPP has better visual clustering effect for each class of activity than LPP, but both of them have compact clustering within the same activities. Note that distributions of jump, run and skip are relatively closer due to their high similarities.

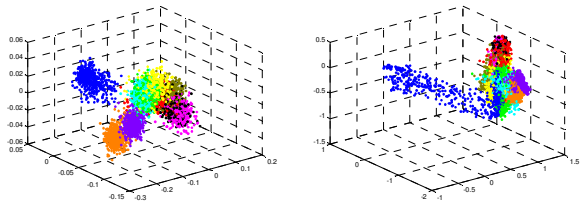

Figure 3. Activity manifolds: SLPP (left) and LPP with K=20 (right)

To compute an overall unbiased estimate of the true recognition rates, we use the leave-one-out rule. Each time, we first leave one sequence out (the sequences taken from the same original video is removed, while other activities of the same subject remain), then train on all the remaining sequences, and finally classify this left-out sequence according to its differences with respect to the rest examples. If this left-out sequence is classified correctly, it must exhibit a high similarity to a sequence from a different person performing the same activity.

After obtaining activity manifolds, we use the methods described in Section 4 to perform activity recognition. Figure 4 shows pair-wise similarities (198×198) using the mean Hausdorff distance, in which the darker the pixel is, the more similar two activity sequences are. From each squared sub-matrix along the diagonal line (i.e., 9×9, 23×23, …, 19×19), we can see there are lower similarity values within the sequences with the same activity, and higher values between the sequences with different activities. For the training of HMM parameters, multiple sequences including the same activity are used to estimate individual HMM parameters. We specify the number of state in the HMM design in a range of 3-5 empirically.

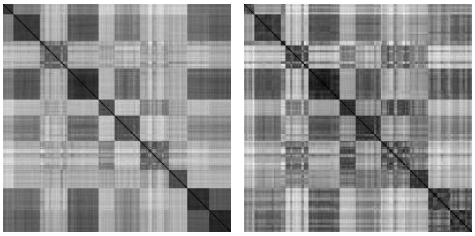

Figure 4. Pairwise similarity: SLPP (*left*) and LPP with K=20 (*right*)

Figure 5 shows correct classification rates (CCR) of activity recognition, from which the following conclusions can be drawn: 1) dynamic silhouette manifolds are indeed informative for classifying human activities; 2) generally, the supervised LPP performs better than the unsupervised.

[*] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

This is because it integrates class label information in training, thus increasing the discrimination ability; and 3) the HMM-based method performs somewhat better than the Hausdorrf-distance based method. This is probably because the statistical nature of the HMM renders overall robustness to representation and recognition.

Figure 5 gives a confusion matrix, in which the element of each row represents the probability that certain kind of activity is classified as other kinds of activities, from which it can be seen that most activities have perfect classification, and only a few skip activities are confused. High similarities among silhouettes in these motions with similar moving patterns may contribute to the confusion.

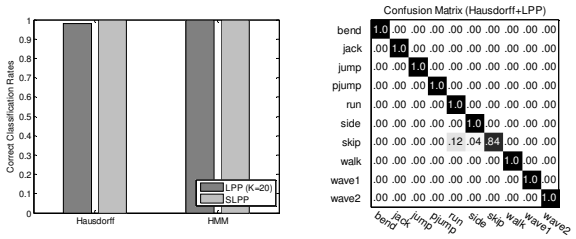

Figure 5. CCRs (*left*) and confusion matrix (*right*)

Two important parameters in the LPP-based manifold learning are the number of the nearest neighbours and the reduced dimension. All experiments have shown the choice of K in a range of 10-25 have similar recognition results, which suggests that K is easily selected to obtain stable results. From the relationship between the reduced dimensions and the recognition rates, we find that SLPP generally needs a lower dimension than LPP to obtain the best results; and our method generally does not need a high dimension to obtain good results. For consistency, here we report all the results with respect to l=20 and K=20.

We also compare the proposed method with a related method described in [9], which uses linear PCA on the filtered images using an IIR (infinite impulse response) filter for obtaining low-dimensional activity description. A best recognition rate of 92.8% using the nearest centroid manifold distance was reported on a test dataset of 8 actions and 168 sequences. We re-implement and evaluate this method on our dataset, and the best recognition rate is 85.86%, which is lower than any of our methods. This is probably because that, on the one hand, our method just uses binary silhouettes as inputs, thus being insensitive to the low colour contrast and texture changes of clothes, and on the other hand, compared with PCA, the LPP is less sensitive to outliers and noise, and more suitable to find intrinsic structures of activity manifolds. The work in [14] reported an almost 100% recognition rate on 549 test cubes derived from the same dataset (without skipping there, but the skipping is easily confused here). Our results are comparable to those of Blank *et al*., but our feature extraction seems simpler than theirs.

5.2. Experiment II: Gait recognition

Most of gait recognition algorithms evaluate their performance on datasets with lateral view because more apparent gait motions can be examined and captured in such a viewing angle. Here we select the NLPR dataset with lateral view [7] for this experiment. It includes 20 subjects, 4 sequences per subject, thus a total of 80 gait sequences (20×4). These sequence images are captured at a rate of 25 fps with the resolution 352×240. The length of each sequence varies with the pace of the walker, but is generally above 2 gait periods. Figure 6 shows example images. Relatively, this dataset is more challenging for human movement analysis because these sequences belong to the same walking activity. But they are performed by different subjects with different physical structures and motion manners.

Figure 6. Example images in the NLPR gait dataset

We directly use the silhouette data obtained in [7] for algorithm evaluation. Similarly, each silhouette image is normalized into a 48×32 resolution, and the supervised or unsupervised LPP methods are used to learn walking manifolds. Figure 7 shows 3D visualization including only 6 subjects, in which the same colour represents the distributions of walking sequences from the same subject.

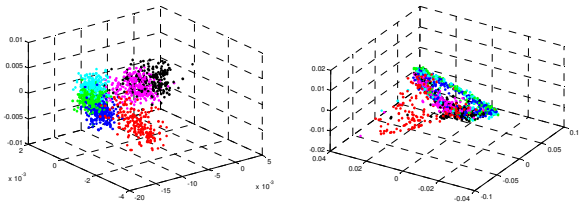

Figure 7. Walking manifolds: SLPP (*left*) and LPP with K=15 (*right*)

We realize human identification using the leave-one-out rule. Note that 'class' in this experiment means human ID (i.e., labels 1-20). Figure 8 shows pair-wise similarities (80×80) using the mean Hausdorff distance, in each of which each squared sub-matrix along the diagonal line (i.e., 4×4) has relatively higher similarity, especially for SLPP, which suggests distinguishable abilities among different subjects of walking gaits. For the HMM-based method, the numbers of states are set to 5 for all sequences. Figure 9 shows CCRs of human identification. Note that all results reported here are with respect to l=20 and K=15. From Figure 9, we can draw some similar conclusions to Experiment I. Although the visual effects of the sequences from different subjects in the manifold

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

subspace may not be as apparent as in Experiment I (with relatively bigger variations between different 'classes'), the classification results are satisfactory because of the introduction of temporal relation during recognition.

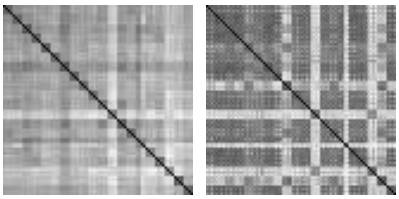

Figure 8. Pairwise similarity: SLPP (*left*) and LPP with K=15 (*right*)

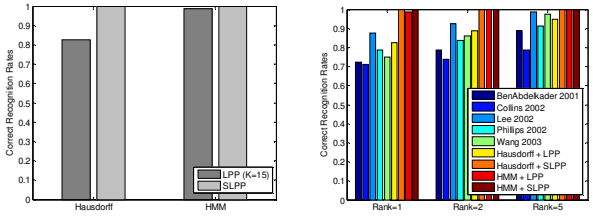

Figure 9. CCRs (*left*) and algorithm comparison (*right*)

We also compare the performance of the proposed method with those of a few silhouette-based methods described in [3-7] on the same silhouette data, as shown in Figure 9. We find that 1) the HMM-based method always performs better than all other algorithms; and 2) For SLPP, the Hausdorff-based method outperforms all other algorithms; for the unsupervised LPP, it performs worse than [6], but superior to [3,4,5,7].

6. Summary and future work

In this paper, our emphasis has been placed on human activity and gait analysis. To this end, we have proposed a general framework to learn and recognize sequential silhouette data in low-dimensional manifold space, and demonstrated encouraging applications of this technique.

Although the proposed framework performs well, much work still remains open, e.g., further algorithm evaluation on a larger database, fusion of shape and kinematics cues, view-invariant feature extraction, etc.

References

[1] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, 77: 257-286, 1989.

[2] D. Gavrila, The visual analysis of human movement: a survey, *CVIU*, 73 (1): 82-98, 1999.

[3] C. BenAbdelkader *et al*., EigenGait: motion-based recognition of people using image self-similarity, *AVBPA*, pp. 284-294, 2001.

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

[5] P. Phillips *et al*., The gait identification challenge problem: data sets and baseline algorithm, *ICPR*, 2002.

[6] L. Lee and W. Grimson, Gait analysis for recognition and classification, *AFG*, pp. 155-162, 2002.

[7] L. Wang *et al*., Silhouette analysis based gait recognition for human identification, *PAMI*, 25 (12): 1505-1518, 2003.

[8] Y. Yacoob and M. Black, Parameterized modelling and recognition of activities, *CVIU*, 73 (2): 232-247, 1999.

[9] O. Masoud and N. Papanikolopoulos, Recognizing human activities, *AVSS*, pp. 157-162, 2003.

[10] A. Efros *et al*., Recognizing action at a distance, *ICCV*, 2: 726-733, 2003.

[11] L. Zelnik-Manor and M. Irani, Event-based analysis of video, *CVPR*, 2: 123–130, 2001.

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

[13] A. Bobick and J. Davis, The recognition of human movement using temporal templates, *PAMI*, 23 (3): 257–267, 2001.

[14] M. Blank *et al*., Action as space-time shapes, *ICCV*, 2: 1395-1402, 2005.

[15] R. Green and L. Guan, Quantifying and recognizing human movement patterns from monocular video images, *TCSVT*, 14 (2): 179-190, 2004.

[16] S. Roweis and L. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science*, 290: 2323–2326, 2000.

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

[18] X. He and P. Niyogi, Locality preserving projections, *NIPS*, 2003.

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

[20] Q. Wang, G. Xu, and H. Ai, Learning object intrinsic structure for robust visual tracking, *CVPR*, 2: 227-233, 2003.

* This work is supported by ARC Centre for Perceptive and Intelligent Machines in Complex Environments, Monash University, Australia