Human Motion Recognition Using Isomap and Dynamic Time Warping

Jaron Blackburn and Eraldo Ribeiro

Computer Vision and Bio-Inspired Computing Laboratory
Department of Computer Sciences
Florida Institute of Technology
Melbourne, FL 32901, USA
{jblackburn,eribeiro}@fit.edu
http://www.cs.fit.edu/~eribeiro

Abstract. In this paper, we address the problem of recognizing human motion from videos. Human motion recognition is a challenging computer vision problem. In the past ten years, a number of successful approaches based on nonlinear manifold learning have been proposed. However, little attention has been given to the use of isometric feature mapping (Isomap) for human motion recognition. Our contribution in this paper is twofold. First, we demonstrate the applicability of Isomap for dimensionality reduction in human motion recognition. Secondly, we show how an adapted dynamic time warping algorithm (DTW) can be successfully used for matching motion patterns of embedded manifolds. We compare our method to previous works on human motion recognition. Evaluation is performed utilizing an established baseline data set from the web for direct comparison. Finally, our results show that our Isomap-DTW method performs very well for human motion recognition.

Keywords: human motion recognition, non-linear manifold learning, dynamic time warping.

1 Introduction

The automatic recognition of human motion from videos is a challenging research problem in computer vision. The interest in obtaining effective solutions to this problem has increased significantly in the past ten years motivated by both the rise of security concerns and increased affordability of digital video hardware. Recent works in the computer vision literature have proposed a number of successful motion recognition approaches based on nonlinear manifold learning techniques [17,8,23]. Nonlinear manifold learning techniques aim at addressing simultaneously the inherent high-dimensionality and non-linearity of representing human motion patterns. However, within this category of methods, little attention has been given to the use of isometric feature mapping (Isomap) [20]. In this paper, we bridge this gap by proposing a new method for automatic recognition of human motion and actions from single-view videos.

A. Elgammal et al. (Eds.): Human Motion 2007, LNCS 4814, pp. 285–298, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Our approach uses non-linear manifold learning of human silhouettes in motion. The approach is similar to the ones proposed by [17,8,23]. However, we cast the problem of recognizing human motion as the one of matching motion manifolds. Our matching procedure is based on an adapted multidimensional dynamic time warping (DTW) matching measurement [22,2].

Our contribution in this paper is twofold. First, we demonstrate the applicability of Isomap for dimensionality reduction in human motion recognition. Secondly, we show how an adapted dynamic time warping algorithm can be successfully used for matching motion patterns in the Isomap embedded manifold. To accomplish our goals, we commence by assuming that the observed human motion patterns can be represented by point-wise trajectories in a lower dimensional space using isometric non-linear manifold mapping. Our proposed algorithm starts by learning Isomap representations of known motion patterns from a set of training images. The learning of the manifold projection mapping is accomplished by means of an invertible radial basis function (RBF) mapping as described in [8]. The initial Isomap projection does not encode any temporal relationship between image frames. Temporal information is introduced into the learned manifold after the projection to the manifold space. The nonlinear manifold augmented with temporal information will then form the learned motion pattern to be used for the recognition of novel motion sequences. Finally, recognition is accomplished by means of a nearest-neighbor classification scheme based on a dynamic time warping score. Figure 1 illustrates sample output from each of the three main steps of the method (i.e., Preprocessing, Model Generation, Recognition). The process in the figure is briefly described as follows. A single video-frame post preprocessing is provided as an example of the functionality performed in this step. In the model generation step, the Isomap projection and the addition of time are shown. Additionally, a comparison of the Isomap projection ($\circ$) to the inverse RBF learned projection ($\times$) is illustrated. During the recognition step, the learned projection is used to map the test sequence ($\bullet$) into the lower dimensional space. Finally, the DTW moves the projected data (solid line) to the temporally aligned data (thin dotted line) to perform the match to the template (thick dashed line).

Our experiments show that the use of Isomap with DTW performs very well for human motion recognition. We test our method on a set of standard human motion sequences widely used in the literature. Finally, we provide a comparison between our approach and recently published methods [4,17,3,23]. Specifically, we apply our algorithm to the data set created by [3] for direct comparison to both [3] and [23]. The data set is also similar enough in nature to compare our results to the approaches presented in [17] and the single-view case in [4]. We show that our method obtains superior results to [4,17,3], and obtains the same 100% recognition rate as the Hidden Markov Model method proposed by [23].

The remainder of this paper is organized as follows. In Section 2, we commence by providing a brief survey of the related literature on human motion recognition. Section 3 describes the details of our motion recognition method.

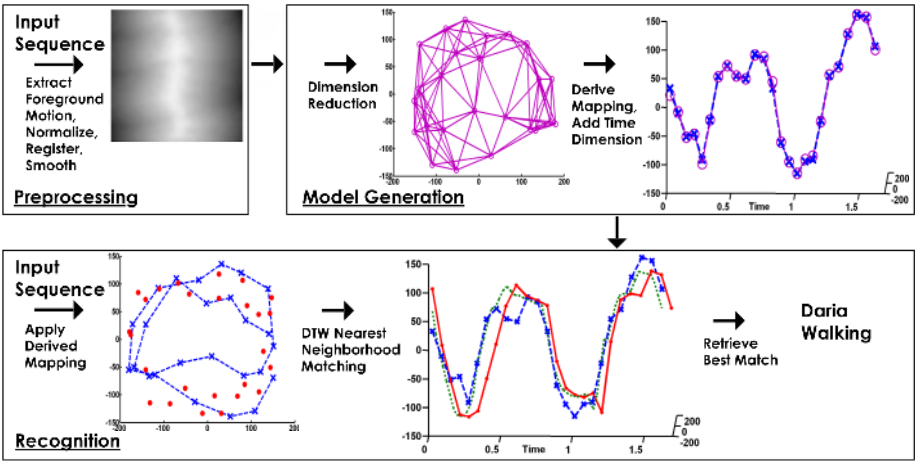

Fig. 1. Motion manifold creation and recognition using DTW. The purple ○ with solid lines denote a projected training sequence. The blue × with dashed lines denote a learned motion template. The red ● with and without solid lines denote a projected test sequence. The green dotted lines denote DTW aligned test data.

Then, in Section 4, we show our preliminary results using the proposed method. Finally, in Section 5, we present our conclusions and directions for future work.

2 Related Literature

The literature on the problem of recognizing human motion from videos sequences is extensive [1,11,17,8,23]. In this paper, we focus ourselves on the methods addressing the specific problem of recognizing human motion from image sequences without the use of markers, tracking devices, or special body suits. In general, such methods can be broadly classified into multiple-view and single-view methods. Multiple-view methods address the motion recognition problem using image sequences obtained from multiple cameras placed at different spatial locations [4,10,18]. The strength of these methods is their power to resolve ambiguous human motion patterns that may result from self-occlusion and viewpoint-driven appearance changes. However, multiple view approaches usually require the availability of synchronized camera systems and controlled camera environments. On the other hand, single-view methods rely only on information provided by a single video camera [3,4,6,8,9,14,15,17,18,23]. Under the single-view assumption, human motion recognition becomes a significantly more challenging and ill-posed problem.

In general, single-view motion recognition is performed using three main steps. The first of these consists of an image processing step in which the image is filtered to reduce the presence of noise (e.g., background, acquisition noise) and to enhance the presence of useful features (e.g., contours, textures, skeletons).

The second step aims at representing the motion information obtained from a sequence of extracted features. The motion information from human activities is inherently both highly non-linear and high-dimensional. As a result, this step will usually try to obtain relevant (i.e., discriminative) motion information using a reduced dimensional space-time representation. The representation can be accomplished, for instance, by making use of explicit measurements on the image to which a pre-determined model is fitted (i.e., skeleton-based methods [18,10], part-segmentation-based methods [15,6,14,9]).

More recently, research in human motion recognition has shifted toward the concept of identifying a motion directly from appearance rather than fitting the visual input to a physical model [4,17,3,23]. Indeed, most of these works have avoided direct feature extraction techniques as they tend to be sensitive to variations such as color, texture, and clothing. Instead, recent work has focused on the use of silhouettes or other high-level abstractions from the raw input data. In this paper, we propose an approach that falls under this later category. Work in silhouette-based human motion recognition can be grouped in terms of the main steps used to approach the problem: image preprocessing, motion pattern representation, and recognition or matching approach.

We begin by discussing the image preprocessing step. This is usually the first step of any recent approach to human motion activity identification. Here, the image foreground (i.e., moving object) is extracted by means of motion segmentation techniques. Standard techniques include the ones based on Motion-History Images (MHI) [4,17,3]. Motion history -based representations allow for simultaneous description of both the dynamics of the motion and the shape of objects. However, as pointed out by Bobick and Davis [4], MHI-based methods are not suited for representing the underlying motion when the observed object returns to similar positions (e.g., cyclic motion patterns). Alternatively, object's silhouette information alone can be used as an input for recognition systems. Wang and Suter [23] used silhouettes as the input to their recognition method. Elgammal and Lee [8] also used silhouettes without motion history. In this paper, we use a similar smoothing technique as the one presented in [8]. However, our distance function representation places a higher weight on the moving object's medial-axis. This reduces the influence of variations in silhouette's contours.

Human motion information is inherently both highly dimensional and complex. Therefore, dimensionality reduction is a standard procedure in the preprocessing of motion data for recognition. Here, the key idea is to find a suitable reduced representation of the motion while maintaining sufficient discriminating data for performing the recognition. To accomplish these goals, past works have used simple data reduction techniques such as principal component analysis (PCA) [17] and Locality Preserving Projections (LPP) [13,23]. The main advantage of these linear approaches is their ability to produce a direct mapping to the embedding space. Nevertheless, the nature of human motion is highly non-linear. Indeed, for complex motions of long duration, recent advances in non-linear dimensionality reduction techniques provide significant improvements of human motion recognition. Techniques in this group include the Isometric feature

mapping (Isomap) [20] and the Local Linear Embedding (LLE) [19]. Evidence of the effectiveness of these non-linear manifold learning methods for human motion recognition has been widely reported in the computer vision literature [23,3,8].

Finally, the recognition step in most motion recognition methods aim at determining the maximum similarity between an unobserved test sequence and pre-learned motion models. Some methods use distance measurements such as the Mahalanobis distance [4] or the Hausdorff distance to establish matches between the learned templates and test sequences [17,3,23]. Methods using the Hausdorff distance are sensitive to non-isometrically similar datasets (i.e., the Hausdorff distance compares each point from one set to every point in the second set regardless of temporal sequence). In order to address this limitation, Wang and Suter [23] propose the use of the Hausdorff distance only as a baseline for a Hidden Markov Model (HMM) matching procedure. Additional important works using HMM for human motion analysis and recognition include [12,5,24]. HMM allows for a principled probabilistic modeling of the temporal sequential information. An alternative way to approach the matching of data sequences is to use Dynamic Time Warping (DTW) [22,2]. DTW has been used in the context of matching data sequences in several applications such as speech recognition, economics, and bio-informatics. DTW provides an approximate similarity measurement while allowing for matching partially identical sequences.

The method proposed in this paper uses an adapted DTW algorithm to perform recognition by matching trajectories on a non-linear manifold space representation. Our paper aims at demonstrating the effectiveness of the Isomap-DTW combination. To the best of our knowledge, this combination has not yet been explored in the human motion recognition literature. In several cases Isomap has been dismissed in favor of Local Linear Embedding or other algorithms mostly due to the greater focus on the local relationship perservation. In other cases Isomap has been dismissed due to the lack of an inverse mapping which other algorithms readily elucidate. The inverse mapping issue has been solved by Elgammal and Lee [8]. Additionally, DTW has also been dismissed in favor of HMM. Our work demonstrates the potential of using Isomap and DTW for matching motion manifolds to accomplish accurate human motion recognition. Next, we describe the details of our motion recognition method.

3 Our Method

In this section, we describe the details of the steps of our method.

3.1 Data Preprocessing

The selection of Isomap for our algorithm imposes a restriction on the input data set. Isomap asymptotically converges for a large class of nonlinear manifolds. The convergence is achieved when the input data has a large enough frequency of coverage within the high dimension space. Consequently, Isomap must be supplied an input data set sufficiently representative to create a meaningful

embedded manifold space. This is a reasonable restriction for any machine learning problem and, given a specific domain, the required amount of input needed for adequate characterization can be obtained via experimentation. Given that a large enough representative set is required, there are two actions that can be taken to aid in preconditioning of the data. The first action is to select a more constrained search space and the second is to generalize the hypothesis sets remaining within the reduced search space. These two techniques aid in decreasing the amount of training data that may be required.

Constraining the Search Space. In many cases the search space can be preconditioned to a much smaller set. One common preconditioning used for images is the reduction of color representation to gray scale representation. Thoughtful manipulation of the search space not only aids in reducing the representative data set needed for learning the manifold space, but can also increase the robustness of the learned mapping.

In the particular case of human motion several recent works established successful results by reducing the space to the silhouette of subjects [3,8]. This discards much of the data associated with internal clothing details, and removes all background data from the search space. The end result focuses the observed dimensionality to strictly the motion performed.

For this particular problem domain, the registration of the silhouettes in the image frames also limits the size of the search space. This preprocessing discards the motion caused by translation and further constrains the space to the motions relative to the internal deformation of the shape. Nevertheless, a simplistic resizing alteration could change the aspect ratio of the subject, and result in an undesirable change to the internal deformation. In our implementation, the registration is performed by isolating the foreground silhouette using a simple background subtraction operation. A bounding box is then constructed for each frame that encompasses the foreground pixels. The largest frame size is chosen to represent the standard frame size for the entire sequence. Finally, all remaining frames are aligned (center pixel) to the center of the standard selected frame.

Generalize the Hypothesis Sets. After the initial search space reduction, generalization is performed by converting the silhouette to a gray level gradient using a distance transform similar to [8]. In our method, we perform the distance transform so that the highest values are assigned to the silhouette's most medial axis points. Once the smoothing is completed, the intensity range in all images is re-scaled to a predefined maximum value (e.g., 255). The result of this preprocessing step is illustrated in Figure 2. Gray scale images are used, however, the color versions illustrate effect on the silhouette's medial axis. The smoothing decreases the variance between subtle differences of similar images, such as those caused by clothing and hair variance. Data sets containing both large volumes and small volumes with significant amount of discriminative features for recognition in smaller volumes may be sensitive to this preprocessing. For human motion, this does not seem to be an issue, and we believe this preprocessing increases the overall robustness of recognition.

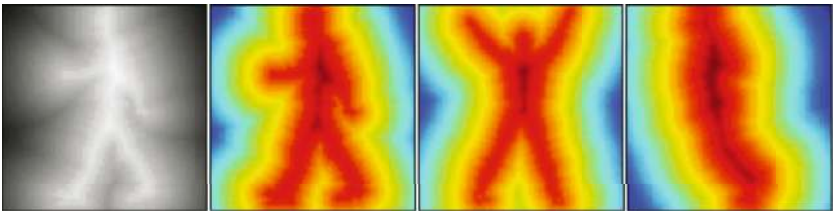

Fig. 2. Sample preprocessed data: Walk (gray scale), Walk, Jack, Jump (color)

3.2 Motion Pattern Learning Using Isomap

In this part of our algorithm, we use the isometric feature mapping or Isomap [20] to obtain template models of the observed motions. Here, our goal is to build a model representation of each motion pattern in our training set. These models will later be used in the matching step to accomplish recognition of unknown motion patterns. The key idea here is to use Isomap as a means of representing the actual intrinsic dimensionality of the analyzed data. Elgammal and Lee [8] used Locally Linear Embedding (LLE) as a manifold learning technique in their motion recognition work. However, Isomap manifolds have been reported to retain more global relationships than its LLE counterpart [7]. This part of our method is divided into two main steps. First, an Isomap manifold is created for each motion available in the training dataset. Secondly, radial basis function mappings are estimated for mapping the learned Isomap manifold space back to the template images. These functions admit an inverse map that allows for the extraction of the manifold embedding for new images. These steps are detailed as follows.

Isometric mapping of silhouette patterns. In this step, we will use Isomap to build a manifold representation of our motion sequence. The input data used by this step is a set of smoothed silhouette images obtained by the preprocessing step of our method. Let $Y = \{y_i \in \mathbb{R}^d, i = 1, \ldots, N\}$ be the set of preprocessed image data (i.e., smoothed silhouette images), and $X = \{x_i \in \mathbb{R}^m, i = 1, \ldots, N\}$ be the corresponding embedding points. The embedded points X are determined using the following three-step Isomap algorithm: **(1)** Create a weighted graph G of points in Y with weights $d_Y(i, j)$ representing the pairwise distance between neighbors. In our algorithm, a neighborhood is defined by the k-nearest neighbors; **(2)** Estimate the pairwise geodesic distances $d_X(i, j)$ between all manifold points by finding the shortest path distances in the graph G. These shortest path distances are denoted by $d_G(i, j)$; **(3)** Finally, apply classical multidimensional scaling (MDS) on D_G to map the data onto an m-dimensional Euclidean space X. It is worth pointing out that $d_X(i, j)$ and $d_Y(i, j)$ are Euclidean pairwise distances within manifold space while $d_G(i, j)$ represents the actual geodesic distances. The coordinate vectors $\mathbf{x}_i$ in X are chosen by minimizing the following L^2 cost function:

$$E = \sqrt{\sum_{ij} \left[\tau\left(d_G(i,j)\right) - \tau\left(d_X\left(i,j\right)\right)\right]^2} \tag{1}$$

where τ in an operator that converts distances to inner products as described in [20]. The use of this operator supports efficiency in the optimization process.

However, the above embedding procedure does not directly allow for the mapping of new images onto the same manifold. In order to address this issue, Elgammal and Lee [8] proposed the use of an approximate invertible mapping from the embedded space to the image space. This mapping is based on radial basis functions. For completeness, this mapping is briefly described next. Further details of this method can be found in [8].

Learning embedded-space-image mappings. The main goal in this step is to obtain an invertible approximate mapping between the embedded manifold space and the image space. Let $t_j \in \mathbb{R}^m, j = 1, \ldots, N_t$ be a set of N_t cluster centers in the embedding space obtained by using a K-Means clustering algorithm. In this paper, we choose N_t such that $N_t = \frac{3}{4}N$. The radial basis function interpolants $f^k : \mathbb{R}^m \rightarrow \mathbb{R}^d$ can be found and satisfy the condition $y_i^k = f^k(x_i)$. Here, k is the k^{th} dimension (pixel) in the image space. More specifically, the interpolant is given by:

$$f^k\left(x\right) = p^k\left(x\right) + \sum_{i=1}^{N_t} w_i^k \phi\left(\left|x - t_i\right|\right) \tag{2}$$

Equation 2 can also be written in matrix form as:

$$f\left(x\right) = B \cdot \psi\left(x\right) \tag{3}$$

where B is a $d \times (N_t + m + 1)$ dimensional matrix, and ψ is given by:

$$\psi = \left[\phi\left(\left|x - t_1\right|\right) \ldots \phi\left(\left|x - t_{N_t}\right|\right)\ 1\ x^\mathsf{T}\right]^\mathsf{T} \tag{4}$$

Finally, B can be obtained by solving the linear system:

$$\begin{pmatrix} A & P_x \\ P_t^\mathsf{T} & 0_{(m+1)\times(m+1)} \end{pmatrix} B^\mathsf{T} = \begin{pmatrix} Y \\ 0_{(m+1)\times d} \end{pmatrix} \tag{5}$$

where A is $N \times N_t$ matrix with $A_{ij} = \phi\left(\left|x_i - t_j\right|\right)$, $i = 1 \ldots N$, $j = 1 \ldots N_t$, ϕ is the thin-plate spline $\phi\left(u\right) = u^2 \log\left(u\right)$, P_x is a $N \times (m+1)$ matrix with i^{th} row $\left[1\ x_i^\mathsf{T}\right]$, and P_t is a $N_t \times (m+1)$ matrix with i^{th} row $\left[1\ t_i^\mathsf{T}\right]$.

The mapping in Equation 5 can be inverted by calculating the Moore-Penrose pseudo-inverse of the matrix B:

$$\psi\left(x\right) = (B^\mathsf{T}B)^{-1}B^\mathsf{T}y \tag{6}$$

This function can be used to map each training image-frame to the embedded template space. The final motion model manifold is then created by reintroducing the time dimension into the manifold representation. This is accomplished by

assigning each frame its corresponding time from the original sequence. The motion manifold construction is now complete, and test frames can be efficiently converted to each of the template manifold spaces before entering the recognition phase. The recognition step is described as follows.

3.3 Recognition

We perform recognition by means of a matching function based on dynamic time warping [22,2]. We adapted the original DTW framework to allow for the matching of motion patterns in manifold space. The key modifications in the DTW algorithm are the following. First, we interpolate both the model template and test manifolds to have the same number of points. Secondly, we use a multi-dimensional version of the DTW with an adapted scoring system using the basic Sakoe-Chiba band constraint. These few modifications permit the DTW algorithm to adjust to nonlinear variations in the input motion patterns. Our main modifications to the DTW algorithm are described as follows.

Interpolation of Inputs. This is a preprocessing step used to improve the quality of the input data before proceeding with the actual DTW alignment. There are several sources of spatial and temporal variations that need to be considered. First, temporal synchronization of video frames cannot be guaranteed (e.g., cameras of various frame rates). A second source of noise is related to the spatial and temporal variations that occur whenever humans perform the same motion repeatedly. The original DTW algorithms does not require same size sequences. Also, uniformity in the sampling rate of the manifolds' time-series is not required. However, results tend to improve when sequences are of similar sampling rates. This interpolation step allows the time aligning properties of DTW to more accurately compensate for the nonlinear variants by matching to anticipated intermediary missing frames.

Adapted Distance Measure. The standard DTW distance measurement is obtained by integrating the values along a path of a distance matrix relating the final manifold points to the initial manifold points. This path search is performed in a dynamic programming manner. In the standard DTW algorithm, all visited nodes contribute to the final distance reported. However, our distance measure only aggregates distances associated with transitions to the next state of the template into the final distance measure. We have modified this distance function slightly to remove additions which simply indicate the time warping is keeping the test manifold in the same state for a longer duration to remain synchronous with the template.

4 Experimental Results

In this section, we evaluate the effectiveness of our motion recognition method. Our main goal here is to show that our method is able to recognize a number of motion patterns acquired by a single camera. To accomplish this goal we provide

a comparison between our method and two recently published motion recognition methods [3,23]. For this comparative study, we use the same dataset used by the methods in [3,23]. The data set contains a collection of nine individuals performing ten distinct actions. The actions and the corresponding labels used in our experiments are the following: bending over (Bend), jumping jack (Jack), hopping across the screen (Jump), jumping up and down in place (Pjump), running (Run), stepping sideways to one direction (Side), hopping on one foot across the screen (Skip), walking (Walk), waving one arm (Wave1), and waving both arms (Wave2). We divided the data set into training subset and testing subset. These two subsets cover all individuals performing all actions. However, in the case of the Bend action, the data set did not contain enough frames to allow for the creation of two distinct action subsets. We addressed this problem by sampling every other frame in the Bend sequence to create the training and testing subsets. Additionally, in some cases, the starting point of the motion was significantly different (e.g., half-cycle sequence). This was addressed by manually stitching the two halves of incomplete motions into a single test motion. The resulting datasets were then used in the experiments described in this section.

We began by preprocessing each sequence to extract the foreground motion information. For simplicity, we used a background subtraction method to facilitate the extraction of the moving foreground silhouette. For cases where a clean background is not available, a more robust foreground segmentation method can be used [21]. The resulting silhouette images were both normalized and registered as described in Section 3.1. In our experiments, we evaluated the performance of the proposed method for images of varying sizes. The sizes used were 16×16, 24×24, and 32×32 pixels. Once the processed sequences were at hand, we compared our Isomap-based method against both the LLE and the LPP dimensionality reduction techniques. For all methods, the local manifold similarity was based on the K-nearest neighbors. Here, the K neighborhood was chosen as suggested in [23]. Accordingly, we used values of K ranging from 5 to 15 to ensure at least an overlap ranging from 10 to 15, respectively. Each motion manifold space created by these embeddings contained two dimensions and were generated from the images without taking any temporal information into consideration. Temporal information was subsequently reintroduced creating manifolds such as those illustrated in Figure 3. The manifolds in Figure 3 also illustrate the use of linear extrapolation between subsequent data points to define the motion manifold. A sampling of 64 evenly-spaced data points were taken from both the learned motion manifold and the test motion manifold for input to the DTW algorithm. A standard sequence size of 64 was chosen to represent approximately twice the size of the largest number of frames for any of the motions in the experiment's dataset. This sampling rate allows the DTW algorithm to perform alignment to interpolated frames that are missing in the learned models due to temporal misalignment in the frame sequence. The algorithm's power to extract meaningful intermediate frames is illustrated in Figure 4. With the exception of a few degraded cases each motion sequence is recognizable despite only the first and last

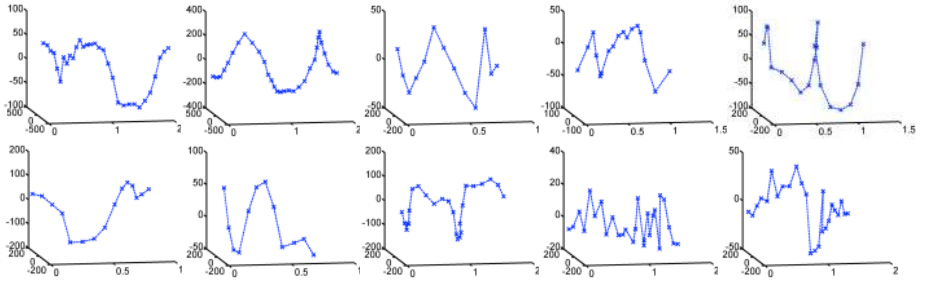

Fig. 3. Motion manifolds for Daria. Top row: Bend, Jack, Jump, Pjump, Run. Bottom row: Side, Skip, Walk, Wave1, Wave2.

silhouettes of each sequence falling exactly on a projected data point. Temporal misalignment and missing frames are common issues in many of the analyzed videos. The DTW was constrained using a Sakoe-Chiba band of 25%. Figure 5 illustrates that our proposed method using Isomap-DTW achieved almost exact recognition rates for the tested activities.

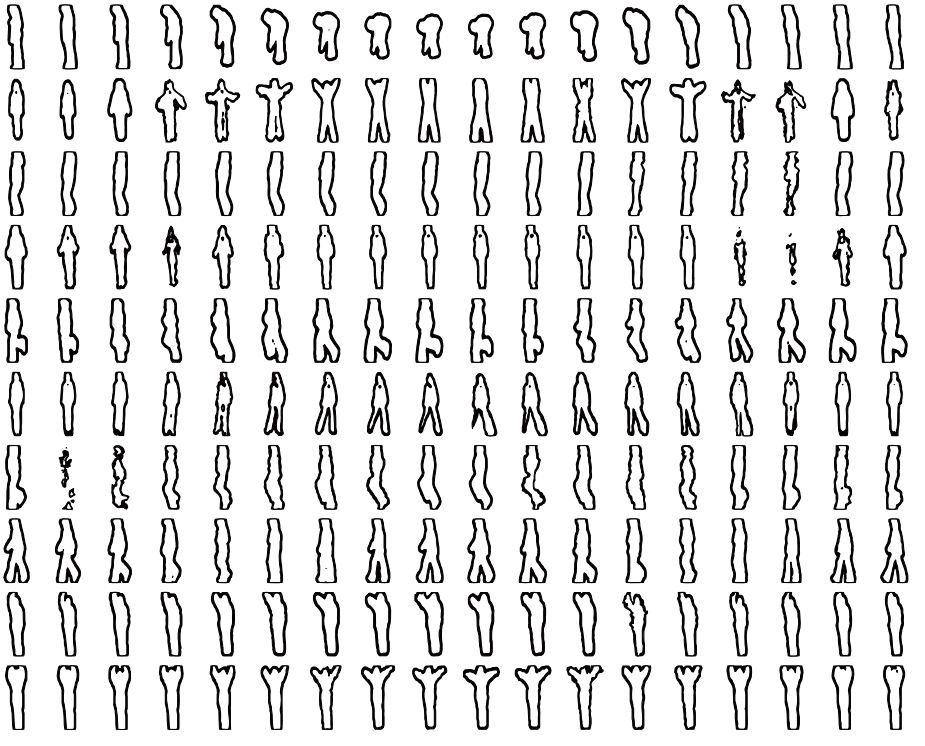

Fig. 4. Silhouette contour of the projection from manifold space to image space

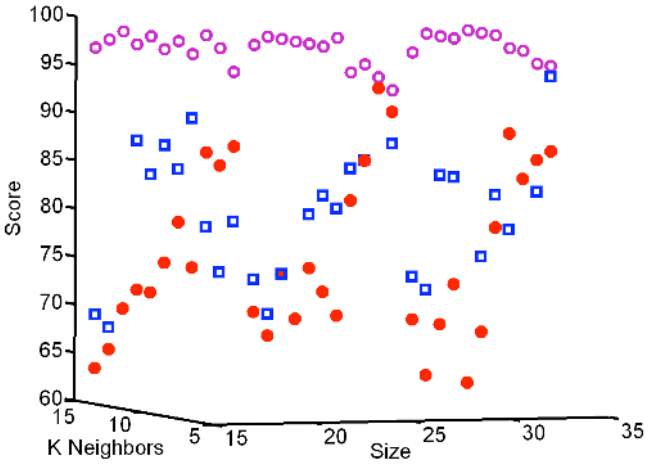

Fig. 5. Isomap(○), LLE(●) and LPP(box) Overall Activity Recognition with a Sakoe-Chiba's band of 25%. The image size is the width and height of the images after preprocessing which is also equivalent to $\sqrt{d}$.

The recognition scores in Figure 5 represent the percentage of motions correctly identified. The size of the images, the k-neighborhood sizes and the dimensionality reduction techniques used were varied for comparison. The results in Figure 5 provide evidence to support our claim that the global perservation of the Isomap data reduction technique can elucidate more meaningful manifolds for recognition via DTW. The recognition results shown in Figure 5 using Isomap with DTW are superior to those reported in [4,3]. Moreover, our results were equivalent to the ones obtained using supervised LPP-Hausdorrf-distance, unsupervised-LPP-HMM, and supervised-LPP-HMM [23]. However, our algorithm achieves this same high recognition rate with smaller image size, smaller neighborhood size, and no supervision. It is worth pointing out that, although Masoud *et al.* [17] utilized a different action database, the motions performed were comparable to the ones used in our experiments. Additionally, the best results reported in [17] were only in the lower 90% range, while our algorithm achieved 100% at several occasions. Also, although our experiments utilized periodic sequences, our method does not require motion periodicity. The specific dataset was used for comparison purposes only.

The subjects used for training are identical to the subjects used for testing. As a result, we are currently unable to infer the generalization capabilities of the proposed method with respect to recognizing unseen subjects. While we are not covering this specific issue in this paper, it is expected that models for one individual may be able to elucidate matches to similar motions performed by other individuals not captured for a particular model.

Finally, our results for all other tested Isomap configurations consistently achieved activity recognition rates above 95%. This demonstrates that, without

any experimental tuning, our technique performs very well in comparison to other established human motion recognition methods.

5 Conclusions and Future Work

In this paper, we presented a method for recognizing human action and motion patterns. Our method works by matching motion projections in Isomap non-linear manifold space using dynamic time warping (DTW). Dynamic time warping has been used in the past in many sequence alignment applications. However, the application of DTW to matching human motion manifolds has been somewhat unexplored. Moreover, we showed that Isomap manifold learning combined with DTW can be an effective way to both represent and match human motion patterns.

Our algorithm achieved accurate activity recognition results using an adapted implementation of DTW with a basic Sakoe-Chiba band optimization. Our experiments established the potential of the method for human motion recognition.

Future work includes the improvement of the computational efficiency of our recognition method by introducing indexing mechanisms such as the one suggested in [16]. Additionally, we plan to investigate the use of statistical neighborhood approach in our adapted DTW to help improve the classification results for both LLE and LPP.

References

1. Aggarwal, J.K., Cai, Q.: Human motion analysis: a review. Computer Vision and Image Understanding 73(3), 428–440 (1999)
2. Berndt, D.J., Clifford, J.: Finding patterns in time series: a dynamic programming approach. Advances in Knowledge Discovery and Data Mining, 229–248 (1996)
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: IEEE International Conference on Computer Vision, Washington, DC, USA, pp. 1395–1402 (2005)
4. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 23(3), 257–267 (2001)
5. Bregler, C.: Learning and recognizing human dynamics in video sequences. In: CVPR 1997. Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition, p. 568. IEEE Computer Society Press, Los Alamitos (1997)
6. Bregler, C., Malik, J.: Learning appearance based models: Mixtures of second moment experts. In: Advances in Neural Information Processing Systems, vol. 9, p. 845. The MIT Press, Cambridge (1997)
7. de Silva, V., Tenenbaum, J.B.: Global versus local methods in nonlinear dimensionality reduction. In: Becker, S., Thrun, S., Obermayer, K. (eds.) NIPS, pp. 705–712. MIT Press, Cambridge (2002)
8. Elgammal, A., Lee, C.-S.: Inferring 3D body pose from silhouettes using activity manifold learning. In: IEEE International Conference on Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, vol. 02, pp. 681–688 (2004)

9. Fanti, C., Zelnik-Manor, L., Perona, P.: Hybrid models for human motion recognition. In: CVPR. IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, vol. 1, pp. 1166–1173 (2005)
10. Gavrila, D., Davis, L.: 3D model-based tracking of humans in action: A multi-view approach. In: CVPR. Conference on Computer Vision and Pattern Recognition, June 18–20, 1996, San Francisco, CA, USA (1996)
11. Gavrila, D.M.: The visual analysis of human movement: a survey. Computer Vision and Image Understanding 73(1), 82–98 (1999)
12. Green, R.D., Guan, L.: Quantifying and recognizing human movement patterns from monocular video images-part i: a new framework for modeling human motion. IEEE Trans. Circuits Syst. Video Techn. 14(2), 179–190 (2004)
13. He, X., Niyogi, P.: Locality preserving projections. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems 16, MIT Press, Cambridge (2004)
14. Heisele, C., Woehler, B.: Motion-based recognition of pedestrians. In: IEEE International Conference on Pattern Recognition, Washington, DC, USA, vol. 2, p. 1325. IEEE Computer Society Press, Los Alamitos (1998)
15. Ju, S.X., Black, M.J., Yacoob, Y.: Cardboard people: A parameterized model of articulated motion. In: International Conference on Automatic Face and Gesture Recognition, Killington, Vermont, pp. 38–44 (1996)
16. Keogh, E.J., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. Knowledge and Information Systems 7(3), 358–386 (2005)
17. Masoud, O., Papanikolopoulos, N.: A method for human action recognition. Image Vision Computing 21(8), 729–743 (2003)
18. Rehg, J., Kanade, T.: Digiteyes: Vision-based human hand tracking. Technical Report CMU-CS-93-220, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA (December 1993)
19. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
20. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)
21. Tian, Y.-L., Lu, M., Hampapur, A.: Robust and efficient foreground analysis for real-time video surveillance. In: CVPR. IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, vol. 1, pp. 1182–1187 (2005)
22. Vintsyuk, T.K.: Speech discrimination by dynamic programming. Cybernetics and Systems Analysis 4(1), 52–57 (1968)
23. Wang, L., Suter, D.: Analyzing human movements from silhouettes using manifold learning. In: IEEE International Conference on Video and Signal Based Surveillance, Washington, DC, USA, p. 7 (2006)
24. Wilson, A.D., Bobick, A.F.: Parametric hidden markov models for gesture recognition. IEEE Trans. Pattern Anal. Mach. Intell. 21(9), 884–900 (1999)