Unsupervised Analysis of Human Behavior Based on Manifold Learning

Yu-Ming Liang, Sheng-Wen Shih, Arthur Chun-Chieh Shih, Hong-Yuan Mark Liao, and Cheng-Chung Lin

Abstract—In this paper, we propose a framework for unsupervised analysis of human behavior based on manifold learning. First, a pairwise human posture distance matrix is calculated from a training action sequence. Then, the isometric feature mapping (Isomap) algorithm is applied to construct a low-dimensional structure from the distance matrix. The data points in the Isomap space are consequently represented as a time-series of low-dimensional points. A temporal segmentation technique is then applied to segment the time series into subseries corresponding to atomic actions. Next, a dynamic time warping (DTW) approach is applied for clustering atomic action sequences. Finally, we use the clustering results to learn and classify atomic actions using the nearest neighbor rule. Experiments conducted on real data demonstrate the efficacy of the proposed method.

I. INTRODUCTION

Understanding human behavior involves two key issues: 1) how to segment the input action sequence into atomic actions; and 2) how to recognize each segmented atomic action. Since the atomic actions are unknown beforehand, a large number of manually labeled training examples must be collected when using a supervised learning approach. Therefore, unsupervised learning approaches are always preferable for human behavior analysis.

In general, unsupervised learning is more difficult than supervised learning. Therefore, the number of published unsupervised learning methods is much less than that of supervised ones [1-3]. A general unsupervised system for human action analysis usually includes three stages: temporal segmentation, atomic action clustering, and atomic action learning and classification. Since the human shape can be modeled as an articulated object with high degrees of freedom, the dimension of a human shape descriptor is usually very high. Under these circumstances, the computation required for temporal segmentation, clustering and classification in a high dimensional feature space is not intuitive and may be very time consuming. Theoretically, a continuous human action sequence can be considered as the variation of human postures lying on a low-dimensional manifold embedding in a high-dimensional space which can be effectively learnt from a set of training data. In this paper, we propose a framework of unsupervised analysis of human behavior based on manifold learning.

The main contribution of this study is to propose a framework of unsupervised analysis of human behavior based on the Isomap algorithm. First, we adopt convex-hull shape context (CSC) [4] to represent a human posture. As to a training action sequence, it is consequently mapped into a manifold trajectory in the Isomap space during the training process. For properly separating an action sequence into atomic actions, the break points between any two consecutive atomic actions have to be identified. To do the job, a manifold trajectory is represented as a time series of low-dimensional points. Then, we devise a temporal segmentation technique to correctly segment the manifold trajectory into atomic actions. Next, we apply a dynamic time warping (DTW) algorithm to perform atomic action sequence clustering. Finally, we use the clustered results to represent each cluster by an exemplar. For an input atomic action, we use the nearest neighbor rule to classify it into a correct category.

II. MANIFOLD LEARNING: THE ISOMAP ALGORITHM

The goal of the Isomap algorithm is to seek a Euclidean space that best preserves the geodesic distances between any two data points in the original high-dimensional space [5]. The Isomap algorithm comprises the following three steps: (1) Construct neighborhood graph; (2) Compute the pairwise geodesic distances; and (3) Construct d-dimensional embedding. An important issue of the Isomap algorithm is to determine the dimension d of the Isomap space. The residual variance, the following is used to evaluate the error of dimensionality reduction

\[ R_d = 1 - r^2(G, D_d) \]  

where G is the geodesic distance matrix; \( D_d \) is the Euclidean distance matrix in the d-dimensional space; and \( r(G, D_d) \) is the correlation coefficient of G and \( D_d \). The value of d is determined using a trial and error approach to reduce the
is close to eigenvectors. The coordinate matrix follows:

\[ X \]

The objective of MDS [6] is to find a Euclidean distance reconstruction that best preserves the inter-point distances. Given a distance matrix \( D = [d_{ij}] \in \mathbb{R}^{n \times n} \), where \( d_{ij} \) is the distance between points \( i \) and \( j \), MDS can construct a set of \( n \) points in the \( d \)-dimensional Euclidean space whose inter-point distances are close to those in \( G \). Let \( x_i = (x_{ij})^T \) denote the coordinates of the \( i \)th point in the Isomap Euclidean space. The Euclidean distance between the \( i \)th and \( j \)th points can be computed as follows:

\[
d_i^2 = (x_i - x_j)^T (x_i - x_j) = x_i^T x_i + x_j^T x_j - 2x_i^T x_j.
\]

In order to overcome the indeterminacy of the solution due to arbitrary translation, the following zero-mean assumption is imposed:

\[
\sum_{i=1}^{n} x_i = 0.
\]

From Equations (2) and (3), the inner product between \( x_i \) and \( x_j \) can be derived as follows:

\[
b_{ij} = x_i^T x_j = -\frac{1}{2} (d_i^2 - \frac{1}{n} \sum_{j=1}^{n} d_j^2) = \frac{1}{n} \sum_{j=1}^{n} d_{ij}.
\]

Let \( D = [d_{ij}] \) denote the distance matrix computed in the Isomap space. Since the Isomap space is to be determined such that \( D \) is close to \( G \), the inner product matrix \( B = [b_{ij}] \) can be obtained by

\[
B = -\frac{1}{2} HG H^T,
\]

where \( H = I - \frac{1}{n} 11^T \) is the centering matrix with \( I = [1, \ldots, 1]^T \), a vector of \( n \) ones. Let \( X = [x_1, \ldots, x_n]^T \) be the \( n \times d \) matrix of the unknown coordinates of the \( n \) points in the Isomap space. The inner product matrix can be expressed as \( B = XX^T \). To compute \( X \) from \( B \), we can decompose \( B \) into \( VAV^T \), where \( A = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \), \( \lambda_i \geq \lambda_{i+1} \geq \cdots \geq \lambda_n \geq 0 \), is the diagonal matrix of eigenvalues and \( V = [v_1, v_2, \ldots, v_n] \) is the matrix of corresponding eigenvectors. The coordinate matrix \( X \) can be calculated as follows:

\[
X = V \Lambda^{\frac{1}{2}},
\]

where \( \Lambda^{\frac{1}{2}} = \text{diag}(\lambda_{1}^{\frac{1}{2}}, \lambda_{2}^{\frac{1}{2}}, \ldots, \lambda_{n}^{\frac{1}{2}}) \) and \( v^* = [v_1, v_2, \ldots, v_n] \).

### III. THE PROPOSED APPROACH

For simplicity, it is assumed that the input video sequence has been processed to obtain the human silhouette sequence, namely the action sequence. The proposed approach comprises four steps which will be described in the following four subsections, respectively.

#### A. Isomap Learning for Human Behavior

To construct the low-dimensional structure of human behavior from a training action sequence, the human posture must be effectively represented in high-dimensional space. In this work, we adopt the convex-hull shape contexts (CSC) [4] to represent the shape of a human silhouette. After each silhouette in the training action sequence is represented as a CSC descriptor, a pairwise shape distance matrix can be calculated based on shape matching. The computed distance matrix is used to compute an Isomap using the method described in Section 2. Therefore, each human silhouette is transformed into a low-dimensional point in the Isomap space. Figure 1 shows the residual variance of Isomap on the training data computed with different value of \( d \), from which the dimension of the Isomap space can be selected as four. Figure 2 shows the constructed 4-D Isomap space.

#### B. Temporal Segmentation

The purpose of temporal segmentation is to identify suitable break points to partition a continuous action sequence into atomic actions. In Figure 2, it is obvious that different atomic actions that can be distinguished using the CSC descriptors will have different trajectories. Therefore, the segmentation process is to identify break points between any two successive atomic action trajectories. To deal with this problem, we first represent the manifold trajectory as a time series of \( d \)-D data points and then calculate the magnitude (i.e., the two norm) of each point, as shown in Figure 3. In general, the human motion slows down at the boundary of an atomic action. Therefore, the local minima and the local maxima of the magnitude series can be regarded as candidates for break points. Furthermore, since humans usually return to a rest posture after completing an atomic action, we define the points possessing low action speeds and the postures nearby the rest posture as the break points of atomic actions. Notably, since the rest postures appear at almost all the atomic actions,
they are the most common postures which will usually be mapped into the neighborhood of the origin of the Isomap space due to the zero-mean assumption formulated in Equation (3). Therefore, we only use the local minima as the break points for obtaining atomic action sequences. In the magnitude series shown in Figure 3, ten local minima can be detected which divide the action trajectory into ten atomic actions.

C. Atomic Action Clustering

After segmenting the training action sequence, the segmented actions are clustered in order to identify and to model each atomic action. Since the time duration of each segmented action sequence is different, a dynamic time warping (DTW) [7] approach is used to cluster the segmented action sequences. The purpose of DTW is to align and compare two sequences \( A=(a_1,a_2,...,a_n) \) and \( B=(b_1,b_2,...,b_n) \) by finding an optimal warping path \( W=(w(1),w(2),...,w(m)) \) between them. Therefore, the DTW matching score between \( A \) and \( B \) can be calculated by

\[
DTW(A,B) = \min_{W} \left\{ \sum_{i=1}^{n} \| a_i - b_{w(i)} \| \right\} \tag{7}
\]

Since the definition of DTW matching score is not symmetric, we define the DTW action distance between two segmented action sequences by

\[
AD(A,B) = \frac{1}{2} (DTW(A,B) + DTW(B,A)) \tag{8}
\]

After the pairwise action distances between all segmented actions are calculated, we can group similar atomic actions into one cluster using the hierarchical clustering algorithm [8].

D. Atomic Action Learning and Classification

Suppose that the coordinates of the training data points are \( \{x_1,x_2,...,x_n\} \) and the coordinate vector of a new data point is denoted by \( \boldsymbol{x}_{n+1} \); summation over \( i=1,...,n+1 \) and \( j=1,...,n+1 \) for \( d_{ij} \) in Equation (2) leads to

\[
\boldsymbol{x}_{n+1}^T \boldsymbol{x}_j = \frac{1}{n} \sum_{i=1}^{n} d_{ij}^2 - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_j \, . \tag{9}
\]

The inner product of \( \boldsymbol{x}_{n+1} \) and \( \boldsymbol{x}_j \) can also be calculated using Equation (2) as follows

\[
\boldsymbol{x}_{n+1}^T \boldsymbol{x}_j = -\frac{1}{2} (d_{n+1}^2 - \boldsymbol{x}_{n+1}^T \boldsymbol{x}_{n+1} - \boldsymbol{x}_j^T \boldsymbol{x}_j) \tag{10}
\]

From Equations (2), (9) and (10), the value of \( \boldsymbol{x}_{n+1}^T \boldsymbol{x}_j \) can be calculated as follows:

\[
\begin{align*}
\boldsymbol{x}_{n+1}^T \boldsymbol{x}_j &= -\frac{1}{2} (d_{n+1}^2 - \frac{1}{n} \sum_{i=1}^{n+1} d_{i+1}^2 - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n} \sum_{j=1}^{n} d_{ij}^2) \\
&= f_j
\end{align*} \tag{11}
\]

Therefore, \( \sum_{j=1}^{n} f_j = \mathbf{f} = [f_1,...,f_n]^T \) \tag{12}

Substituting Equation (6) into Equation (12), we have

\[
(\sqrt{\lambda_1} \mathbf{v}_1,...,\sqrt{\lambda_d} \mathbf{v}_d)^T \boldsymbol{x}_{n+1} = \mathbf{f} \, . \tag{13}
\]

The least-square solution of \( \boldsymbol{x}_{n+1} \) can be derived as follows:

\[
\boldsymbol{x}_{n+1} = (\frac{1}{\sqrt{\lambda_1}} \mathbf{v}_1,...,\frac{1}{\sqrt{\lambda_d}} \mathbf{v}_d)^T \mathbf{f} \tag{14}
\]

After each human posture in the test action sequence is projected into the Isomap space, the test sequence can be classified into a cluster using the nearest-neighbor approach based on the DTW distance. Furthermore, a threshold is used to judge whether an unknown action sequence belongs to one of the learnt clusters.

To map a new human posture into the existing Isomap space is very time consuming, especially when the number of training postures is large. To deal with this problem, we propose a simplified mapping approach based on some chosen key data points. First, we select some key data points from the training set based on the k-means algorithm. These selected key data points will form a simplified Isomap space. Then, these selected key data points are used as the training data to estimate the coordinate vector of a new data point (repeat Equations (9)-(14)). Suppose that the coordinates of the key data points are \( \{ \boldsymbol{x}_{q_1}, \boldsymbol{x}_{q_2},...,\boldsymbol{x}_{q_k} \} \). Because the estimation of the coordinate vector of a new data point is based on the zero-mean assumption of training data points, we transform the coordinate system as follows:

\[
\boldsymbol{x}_q = \boldsymbol{x}_q - \bar{x}_q , \tag{15}
\]

where \( \bar{x}_q = \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{x}_q \). Therefore, the inner product of \( \boldsymbol{x}_{n+1} \) and all of the key data points can be calculated using Equation (12):

\[
\sum_{q=1}^{k} \boldsymbol{x}_{n+1}^T \boldsymbol{x}_{q} = \boldsymbol{x}_{n+1}^T \mathbf{x}_q = \mathbf{f} = [f_{q_1},...,f_{q_k}]^T \tag{16}
\]

The least-square solution of \( \boldsymbol{x}_{n+1} \) can be derived as follows:

\[
\boldsymbol{x}_{n+1} = (\mathbf{X}_q \mathbf{X}_q)^{-1} \mathbf{X}_q \mathbf{f} . \tag{17}
\]

Finally, the coordinate vector of a new data point \( \boldsymbol{x}_{n+1} \) can be calculated by \( \boldsymbol{x}_{n+1} + \bar{x}_q \).

IV. EXPERIMENTS

We conducted a series of experiments to evaluate the...
effectiveness of the proposed method. The data used in the experiments included one training sequence and two test sequences performed by two human subjects. The training data contained 25 atomic action sequences, approximately 1983 frames, belonging to five different action classes that were performed by the human subject 1. Some typical image frames for each atomic action class are shown in Figure 4. Using the Isomap algorithm, a 4-D Isomap space was constructed from the training data, as shown in Figure 5. Next, we represent the manifold trajectory in Figure 5 as a time series of data points, and then applied temporal segmentation to this time series. The results of temporal segmentation are shown in Figure 6.

Next, two test action sequences were used to evaluate the performance of the proposed temporal segmentation and action classifier. The first test sequence was obtained from the human subject 2 and this sequence contained 46 atomic actions all belonged to the five learnt action clusters. Figure 7 shows the constructed atomic action trajectories by mapping the new test data into the Isomap space using the method described in Section 3.4. In this experiment, all atomic actions were correctly segmented and classified. The second test sequence was obtained by asking the human subject 1 to perform new actions that were different from all the trained atomic actions. The constructed atomic action trajectories and five mean trajectories are shown in Figure 8. Using the

proposed action classification method, these atomic actions were all successfully classified as unknown actions.

In the final experiment, we evaluated the performance of the proposed action classifier based on the simplified mapping approach. Figure 9 shows the selected key data points and the reconstructed Isomap. In this experiment, it is obvious that the reconstructed Isomap is very similar to the original Isomap (see Figure 5). We conducted experiments on the two test action sequences. Figure 10 shows the reconstructed atomic action trajectories by mapping the new test data into the simplified Isomap space.

V. CONCLUSIONS

In this paper, we have proposed a framework for unsupervised analysis of human behavior based on Isomap learning. A pairwise CSC distance matrix is first calculated from the training action sequence. The Isomap algorithm is then applied to construct the low-dimensional structure from the distance matrix. Next, the data points in the Isomap space are represented as a time series of low-dimensional points, and a temporal segmentation process is applied to segment this sequence into atomic actions. A DTW approach is then applied for atomic action clustering. Finally, the clustering results are used to learn and classify atomic actions.

REFERENCES