Quasi View-Independent Human Motion Recognition in Subspaces

Pattreeya Tanisaro, Florian Mahner and Gunther Heidemann

Institute of Cognitive Science, Osnabrück University
ICMLC 2017
Singapore, 24-26 Feb 2017
**Motivation**

We are interested in dimensionality reduction techniques for complex and nonlinear multidimensional time-series data. -> Human motions!

**Benefits of dimensionality reduction**

- **Speed.** Reduced computational cost and memory requirement.
- **Interpretation of data by means of visualization.**

However, encapsulating data to a lower-dimensional has also a disadvantage, namely, the classification is much more difficult to solve.

**Research Questions:**

1. How to apply traditional manifold learning approaches with time-series problems?
2. What is the suitable method to normalize the time-series data i.e., human motions?
We test our results for both 3D and 2D projection for 10 actions, which are bending, boxing, golfing swing, jumping forward, marching, running, walking, standing body cross-crunch exercise, side-twist exercise and salsa dance.
Feature Vectors

We benchmark the feature vectors from the following methods:

1. Data in space by way of naive approach
2. Subspaces: i) traditional manifold learning ii) our proposed scheme.

\[
V_1 \in N_{f,1} \times (N_j \cdot N_d) \\
\vdots \\
V_m \in N_{f,m} \times (N_j \cdot N_d) \\
\vdots \\
V_M \in N_{f,M} \times (N_j \cdot N_d) \\
N_M \times (N_j \cdot N_d)
\]

**Figure:** Feature vectors using the naive approach

- \( N_{f,m} \) is a number of frames of video \( m \), where \( m \) is the video index in \( \{1..M\} \).
- \( M \) be the number of training videos, and \( N_j \) the number of skeleton joints fixed at \( N_j = 15 \).
- \( N_d \) is the number of the dimension in \( \{2, 3\} \) for 2D and 3D projection.
- A total number of frames \( N_M \) is computed from \( N_M = \sum_{m=1}^{M} N_{f,m} \).
We follow the technique suggested by Körner and Denzler to normalize data points by computing a zero-mean skeleton configuration at each frame.

Let \( \ell_i = ((x_1, y_1), \ldots, (x_{N_j}, y_{N_j}))_i \) is a joint matrix at frame \( i \) for 2D projection \( (N_d = 2) \), \( d \in \{x, y\} \).

For each video \( m \), we compute the mean average of the joints from all frames for each dimension \( d \) of \( L_{m,d} \) as \( L_{m,d} = \frac{1}{N_{f,m}} \sum_{i=1}^{N_{f,m}} \ell_i \).

We normalize the data in order to compute the eigenvectors \( V_{e,c,m} \). The normalized vector of the video \( m \) is defined as \( V_{\mu,m} \) using \( V_{\mu,m,d} = V_{m,d} - L_{m,d} \), and \( L_m \) is applied frame-wise.

The feature vectors of video \( m \) can be obtained from the mean average vector and the normalized basis vectors as \([L_m; V_{e,c,m}]\).

---

\(^2\)M. Körner and J. Denzler, Analyzing the Subspaces Obtained by Dimensionality Reduction for Human Action Recognition from 3d Data, AVSS 2012
The first benefit of the proposed method is: the extracted features are fixed and very small no matter how long the videos are ($M \ll N_M$).
Experimental Setup

Figure: Tracking of each skeletal joint location for $N_j = 15$ on 2D projection. A subject performs *jumping forward* from different angles. From left to right: $-45^\circ$ at frame 150, $-45^\circ$ at frame 250, $0^\circ$ at frame 250, and $-90^\circ$ at frame 250.

- Pick five subjects for training and test set each for ten actions.
- Subjects in the training set are excluded from the test set.
- For the training set, we apply five camera angles $\{-90, -45, 0, 45, 90\}$ to each video. We have $10 \times 5 \times 5$ videos ($M = 250, N_M = 122645$) for training data.
- For the test set, we use twenty-one angles in $\{-100, -90, ..., 90, 100\}$. We have $10 \times 5 \times 21$ videos ($M = 1050, N_M = 570759$) for testing recognition of untrained subjects and from untrained camera angles.
Our Contributions

Our contributions:

1. We show the robustness of applying the proposed feature extraction technique for various dimensionality reduction approaches under two conditions:
   1. untrained camera angles in combination with untrained subjects
   2. data loss; by further extension of 1) by subsampling the original test data.

2. We propose an anatomically in-frame normalization using a center of torso which can improve the recognition rate by at least 10%.

Figure: The largest areas of distribution of motions from the basis vector with the largest eigenvalue using PCA. Left: Sphere distributions of 3D skeletons at 0°. Right: at 90°.
Results: Recognition Rates From 3D and 2D Projection

Figure: Recognition rate from 3D (blue) and 2D (red) projection using the naive approach with five classifiers and, using in-frame referenced by the center of torso vs. raw spatial data.

Figure: Recognition rates from 3D using dimensionality reduction from spatial vs. in-frame reference data. The results are from several manifold learning techniques using 3 principal components in combination with RF and 1-NN.
Results: Recognition Rates of 2D Data in Subspaces

**Figure:** Recognition rates from 2D projection using dimensionality reduction with 3 principal components.

**Figure:** Comparison of recognition rates from resampling test data with different subsampling factors using various feature extraction techniques.
Results: Confusion Matrix

Figure: Two actions from two camera angles that often misclassified in subspaces: walking and running.

<table>
<thead>
<tr>
<th></th>
<th>bend</th>
<th>box</th>
<th>golf</th>
<th>jump</th>
<th>march</th>
<th>run</th>
<th>salsa</th>
<th>crunch</th>
<th>twist</th>
<th>walk</th>
</tr>
</thead>
<tbody>
<tr>
<td>bend</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>box</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>golf</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>jump</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>march</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>93.4</td>
<td>0</td>
<td>0</td>
<td>2.8</td>
<td>0</td>
<td>3.8</td>
</tr>
<tr>
<td>run</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>96.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3.8</td>
</tr>
<tr>
<td>salsa</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>crunch</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>twist</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>walk</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.8</td>
<td>0</td>
<td>0</td>
<td>95.2</td>
</tr>
</tbody>
</table>

Tabelle: Confusion matrix of classifying 1050 test samples of the 2D skeleton using Isomap-3N+RF.
Conclusion

We demonstrate the robustness of using the proposed scheme of representing human motion sequences in subspaces.

- The presented algorithm is quite robust against data sequence loss and yields high recognition rates even with untrained subjects and untrained angles. Moreover, the training and testing are fast.
- The normalization with in-frame reference data can improve the recognition by more than 10%.
- We also show that the proposed using PCA and 1-NN for the case of data in sequence loss is not suitable.
- The proposed algorithm and the normalization technique could be further extended for other time-series applications.
Thank you for your attention!