SCALE-INDEPENDENT SPATIO-TEMPORAL STATISTICAL SHAPE REPRESENTATIONS FOR 3D HUMAN ACTION RECOGNITION

Marco Körner, Daniel Haase and Joachim Denzler
Department of Mathematics and Computer Sciences, Friedrich Schiller University of Jena
Ernst-Abbe-Platz 2, 07743 Jena, Germany

Keywords: Human action recognition, Manifold learning, PCA, Shape model.

Abstract: Since depth measuring devices for real-world scenarios became available in the recent past, the use of 3d data now comes more in focus of human action recognition. We propose a scheme for representing human actions in 3d, which is designed to be invariant with respect to the actor’s scale, rotation, and translation. Our approach employs Principal Component Analysis (PCA) as an exemplary technique from the domain of manifold learning. To distinguish actions regarding their execution speed, we include temporal information into our modeling scheme. Experiments performed on the CMU Motion Capture dataset shows promising recognition rates as well as its robustness with respect to noise and incorrect detection of landmarks.

1 INTRODUCTION AND RELATED WORK

In the last decades the recognition and analysis of actions and motions performed by humans have become one of the most promising fields in computer vision research and lead to a wide variety for research topics in computer vision. This family of problems aims to determine human activities automatically based on several sensor observations. A wide range of industrial as well as academic applications are based on this research, e.g. the interaction between humans and machines, surveillance and security, entertainment, video content retrieval as well as the research in medical and life sciences.

In early years of scientific interest those methods concentrated on the evaluation of 2d image sequences delivered by gray value or color cameras (Gavrila, 1999; Turaga et al., 2008; Poppe, 2010). Due to the massive amount of research those methods now achieve very good results on the standard Weizmann 2d action recognition dataset (Gorelick et al., 2007). Several of those approaches are based on the evaluation of changes in silhouettes or the extraction of interest point features in space-time volumes created by subsequent video frames (Laptev, 2005; Dollar et al., 2005). Furthermore the combination of shape and optical flow is used for action recognition (Ke et al., 2007).

In contrast to this huge amount of scientific work concerning 2d images, 3d data was not yet used in a remarkable quantity. However, the recent development of depth measuring devices such as Time-of-Flight (ToF) sensors or sensors based on the projection and capturing of structured light patterns make 3d data available in a fast and inexpensive way.

In this paper we present a spatio-temporal representation scheme for human actions given as sequences of 3d landmark positions which models the spatial variations in a contextual way and takes into account the temporal coherence between subsequent frames based on manifold learning techniques. After presenting our approach in Sec. 2 we show numerous experiments evaluated on the CMU Motion Capture (MoCap) dataset in Sec. 3. A summary and a brief outlook in Sec. 4 conclude this paper.

2 STATISTICAL SHAPE REPRESENTATION

Manifold learning techniques are widely used for classification tasks like face detection and emotion recognition (Zhang et al., 2005). For action recognition from 2d video streams the usability of Principal Component Analysis (PCA), and Independent Component Analysis (ICA) on motion silhouettes have been compared (Yamazaki et al., 2007). Locality Preserving Projections (LPP) were utilized in combination with a special Hausdorff distance measure on silhouettes (Wang and Suter, 2007). A comparison of further techniques for dimensionality reduction like Locality Sensitive Discriminative Analysis (LSDA) and Local...
Spatio-Temporal Discriminant Embedding (LSTDE) was presented in (Jia and Yeung, 2008). Tensor PCA was presented in (Jia and Yeung, 2008). Conditional Random Fields (CRF) was employed to (ASM) (Cootes et al., 1995) are mass- 

To preserve scale invariance of our modeling, a predefined skeleton model is used for projection each time.

Combining all zero-mean skeleton configurations at every available time step yields the matrix of landmarks

\[
L = \begin{pmatrix}
I_1 - I_\mu \\
... \\
I_N - I_\mu
\end{pmatrix} \in \mathbb{R}^{N_l \times N_l}, \quad I_\mu = \frac{1}{N_l} \sum_{i=1}^{N_l} I^i.
\]

Performing Principle Component Analysis (PCA) on \(L\) will return its matrix

\[
P_L = (v_1^L | \cdots | v_{N_l}^L) \in \mathbb{R}^{N_l \times N_l}
\]

of eigenvectors sorted according to their corresponding eigenvalues \(\lambda^L_k\) descendingly representing the importance of each data space direction. Using these eigenvectors as basis vectors, every arbitrary skeleton configuration represented by a 3d landmark coordinate set can be expressed as a linear combination

\[
I^i = I_\mu + P_L^i b^f 
\]

of the data matrix columns and the frame-specific shape parameter vector \(b^f\) added to the constant mean shape \(I_\mu\).

Since the amount of represented variances of landmark sequences captured by the eigenvectors decreases
massively according to the evolution of their corresponding eigenvalues, the number of columns in the eigenvector matrix $P_L$ can be restricted to achieve a substantial reduction of dimensionality.

Fig. 2(a) shows the first three action components, while Fig. 2(b) depicts the corresponding eigenvectors of an action from class walking.

2.2 Integration of Temporal Context

While the previously described representation solely models linear variations of skeleton joints, the temporal evolution of configurations might contribute helpful information for the recognition and analysis of articulated actions. For this reason, our model is extended to include this temporal component.

In (Bosch et al., 2002) such a temporal modeling of periodical actions was already used to model a beating heart. This was pointed out to be a generalization of the multi-view integration approach of (Lelieveldt et al., 2003) and (Oost et al., 2006). Instead of considering a skeleton configuration at a single time step $t_0$ to obtain the model parameters, they regard a series of sequential time steps $t_0 < t_1 < \ldots < t_k$ or alternatively multiple views $(\theta_1, \theta_2, \ldots, \theta_k)$ at the same time step as a single configuration.

Applied to our problem, the provided model models the temporal evolution of skeleton configurations by appending subsequent input matrices horizontally:

$$L^{k_{\text{hist}}}_{t_0} = \begin{pmatrix} L^{(t_0)}_{\text{hist}} & \ldots & L^{(t_{k_{\text{hist}}})}_{\text{hist}} \end{pmatrix} \in \mathbb{R}^{(k_{\text{hist}}+1) \times N_t},$$

$$L^{k_{\text{hist}}}_{t_{t_0}} = \begin{pmatrix} \begin{pmatrix} L^{(t_{t_0})}_{\text{hist}} & \ldots & L^{(t_{k_{\text{hist}}})}_{\text{hist}} \end{pmatrix} \end{pmatrix} \in \mathbb{R}^{(k_{\text{hist}}+1) \times (N_t - (k_{\text{hist}} + 1))}.$$

This approach allows us to distinguish between an action and its reverse counterpart as well as to classify the speed of execution.

3 EXPERIMENTS

3.1 Dataset

In order to evaluate the proposed methods, we have chosen eight different actions performed by different actors from the CMU MoCap dataset, as shown in more detail in Tab. 1. While we have selected common actions with slightly different executions like walking, running, marching and sneaking or hopping and jumping, we also took complex motions—salsa and golfwring—into account.

3.2 Discriminability of Eigenvector Representation

When performing PCA on sequential data $L$, the result shows the most important directions of variance in the data. For this reason, the eigenvectors $v_L^k$ corresponding to the largest eigenvalues $\lambda_L^k$ are supposed to encode most of the information, while the eigenvectors corresponding to the lower eigenvalues model only minor changes in the data as well as noise.

Fig. 3 depicts the evolution of the eigenvalues for all action classes in our dataset. As can be seen, after a strong descent up to the third principal component, the eigenvalues converged strongly towards zero. After a certain component, there was no substantial contribution to the data, which became apparent at the 12th eigenvalue, as indicated by the vertical line in Fig. 3. As depicted in Tab. 2, in most the cases two to three eigenvectors were sufficient to cover 90% of the variances occurred while execution of an action. Solely the action classes with high variances in all directions need more discriminability, which can be handled by increasing the number of eigenvectors. This fact can also be seen in Tab. 3, where the first three eigenvectors $v_L^k$ are shown together with their mean shapes $I_\mu$.

The back projection error $\epsilon_{\text{action}}(t') = \| (t' - P_{L_{\text{action}}}^T P_{L_{\text{action}}}^T) - t' \|_2$ obtained by trans-
Table 2: Comparison of variance covering facilities of our representation scheme. While the usage of two to three eigenvectors allows to achieve 90% of the variances obtained during simple actions, more dimensions are needed to represent more complex actions.

<table>
<thead>
<tr>
<th>Action Class</th>
<th>Amount of variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90%</td>
</tr>
<tr>
<td>walking</td>
<td>2</td>
</tr>
<tr>
<td>running</td>
<td>2</td>
</tr>
<tr>
<td>marching</td>
<td>3</td>
</tr>
<tr>
<td>sneaking</td>
<td>3</td>
</tr>
<tr>
<td>hopping</td>
<td>2</td>
</tr>
<tr>
<td>jumping</td>
<td>3</td>
</tr>
<tr>
<td>golfing</td>
<td>3</td>
</tr>
<tr>
<td>salsa</td>
<td>8</td>
</tr>
</tbody>
</table>

Figure 3: Evolution of eigenvalues for different action classes. Eigenvalues are decreasing massively up to the third component, while they remain static for higher-order components.

Forming an arbitrary skeleton configuration $l'$ from Euclidian space $\mathbb{R}^3$ into the reduced eigenspace $\mathbb{V}_{\text{action}}$ of a certain action class and back to Euclidian space, where $P_{l'\text{action}} = (v_1^{\text{action}}, \ldots, v_{k_{cv}}^{\text{action}})$ is a matrix containing the eigenvectors corresponding to the first $k_{cv}$ largest eigenvalues of $L_{\text{action}}$, give a quantitative justification for this postulation, as can be seen in Fig. 4. As a result, the ordering of remaining eigenvectors is no longer meaningful. Therefore, they are not considered in the following classification purposes.

3.3 Feature Vector Design and Classification

In order to distinguish action classes, features have to be derived from the sequence of skeleton configurations. Using the representation described before, feature vectors $v_L = (l_1^{\mu}, v_1^{l'}, \ldots, v_{k_{cv}}^{l'})$ are extracted from a series $L'$ of landmark vectors $l'$ by concatenating its mean shape $l_1^{\mu}$ and its eigenvectors corresponding to the first $k_{cv}$ eigenvalues.

In Fig. 5(a) one can observe that the recognition rate during classification had a maximum peak at $k_{cv} = 3$, which argues for a high degree of discriminability. This is emphasized by the vertical line in Fig. 5(a). Without using any eigenvectors, only the mean shape is taken into account during feature extraction, which leads to lower discriminability. Using more eigenvectors would cause a more exact reconstruction of the skeleton configuration and therefore a smaller discriminability due to the increased coverage of variability.

For simplicity, we used the $k$ Nearest Neighbor ($k$-NN) framework for classification, which assigns a class label to a feature vector employing an arbitrary

Table 3: Comparison of the mean shapes and the first three eigenvectors of the action classes in our dataset. Note that similar actions have similar first eigenvectors and different second or third eigenvectors while different actions can already be distinguished by their first eigenvectors.

<table>
<thead>
<tr>
<th>Action Class</th>
<th>Mean Shape</th>
<th>Eigenvectors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$l_\mu$</td>
<td>$v_1^{l'}$</td>
</tr>
<tr>
<td>walking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>running</td>
<td></td>
<td></td>
</tr>
<tr>
<td>marching</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sneaking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hopping</td>
<td></td>
<td></td>
</tr>
<tr>
<td>jumping</td>
<td></td>
<td></td>
</tr>
<tr>
<td>golfing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>salsa</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: Back projection errors obtained by transforming skeleton configurations in every time step of a hopping sequence (thick line) from Euclidean space into action-specific eigenspaces and back to Euclidean space. Small errors suggest that the mapping is appropriate for the given action representation, while high errors are indicating poor mapping facilities.
Table 4: Confusion matrix with overall recognition rates obtained by exhausting leave-one-out test on our dataset.

<table>
<thead>
<tr>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>walking</td>
<td>100 0 0 0 0 0 0</td>
</tr>
<tr>
<td>running</td>
<td>33 56 0 11 0 0 0</td>
</tr>
<tr>
<td>marching</td>
<td>4 0 93 0 0 4 0</td>
</tr>
<tr>
<td>sneaking</td>
<td>0 0 0 100 0 0 0</td>
</tr>
<tr>
<td>hopping</td>
<td>13 7 0 0 80 0 0</td>
</tr>
<tr>
<td>jumping</td>
<td>0 0 7 0 86 0 0</td>
</tr>
<tr>
<td>golfing</td>
<td>0 0 0 0 0 100 0</td>
</tr>
<tr>
<td>salsa</td>
<td>0 0 3 0 0 0 97</td>
</tr>
</tbody>
</table>

distance measure $d(y_{\text{test}}, y_{\text{target}})$ between the feature vector $y_{\text{test}}$ and a representative prototype vector $y_{\text{target}}$. In our experiments, we chose the Euclidean distance $d(y_{\text{test}}, y_{\text{target}}) = \| y_{\text{test}} - y_{\text{target}} \|_2$.

As can be seen in Fig. 5(b), using $k = 1$ gave the best recognition rate, while increasing the number of neighbors caused apparent worse results as well as higher computational time for classification.

Using this feature extraction scheme and the 1-NN classifier, we were able to achieve results as shown in the confusion matrix obtained by exhaustive leave-one-out test in Tab. 4. As one can see, most of the action classes in our dataset were recognized correctly in more than 80% of the cases, while 4 classes gave recognition rates of nearly 100%. Solely the action class running has been confused with the semantically related classes walking and sneaking due to their similar variations during execution.

3.4 Robustness to Noise

In real-world applications the input data for action classification are not ideal. Hence we modeled the influence of additive, zero-mean normally distributed, and uncorrelated Salt-and-Pepper noise to quantitatively evaluate the robustness of our approach.

As can be seen in Fig. 6(a), adding Gaussian noise to the input data did not negatively affect the classification results. This might be explained by the mean subtraction on the one hand and the usage of PCA on the other hand during modeling. In order to find the principal components, noise added to the data will only affect the eigenvectors corresponding to the smaller eigenvalues, while the inherent and consistent information of movement over time is still captured by the eigenvectors corresponding to the larger eigenvalues.

A similar behavior can be observed in the case of adding uncorrelated Salt-and-Pepper noise to input data. As can be seen in Fig. 6(b), while the recognition rates were decreasing with the amount of added noise, simple median filters applied to the single channels along the time dimension were able to drastically reduce these effects. It can be seen that an amount of 70% Salt-and-Pepper noise can be handled by applying a 15-frame temporal median filter which only results in a small decrease in the recognition rates.
As mentioned in Sec. 2.2, we not only model the variations of landmark transitions during a fixed time period, but also integrate the evolution of these movements by incorporating the temporal context during an action execution.

Tab. 5 shows that the integration of temporal information into the action model affects the recognition rates slightly.

<table>
<thead>
<tr>
<th>Historical Offset $\Delta_{hist}$</th>
<th>Number of History Frames $h_{hist}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>92.45</td>
</tr>
<tr>
<td>10</td>
<td>93.08</td>
</tr>
<tr>
<td>15</td>
<td>91.82</td>
</tr>
<tr>
<td>30</td>
<td>90.57</td>
</tr>
</tbody>
</table>

### 3.5 Comparison to Other Work

Although human action recognition was widely investigated for 2d data, there is less work available concerning the case of having access to 3d data. A similar approach to classify human actions in 3d data was taken in (Han et al., 2010), but they selected less action classes from the CMU MoCap dataset. While they distinguish only 3 action classes with small variations in execution, recognition rates of 98.29% were obtained without taking the presence of noise into account. In (Junejo et al., 2011), the same database has been used to create artificial 2d views and evaluating several distance metrics on the landmark points without modeling the shape at all. They observed recognition rates of about 90.5% in average when combining all their camera views for training and testing. The approach of (Shen et al., 2008) employed homography constraints and lead to an overall recognition rate of about 92%.

Compared to those results, our approach performs similarly (92.45%) on the same data even in the presence of noise.

### 3.6 Use of Temporal Context

As mentioned in Sec. 2.2, we not only model the variations of landmark transitions during a fixed time period, but also integrate the evolution of these movements by incorporating the temporal context during an action execution.

Tab. 5 shows that the integration of temporal information into the action model affects the recognition rates slightly. We tested several values for the number of history frames $h_{hist}$ integrated to the model as well as the temporal offset $\Delta_{hist} = (t_i - t_{i-1})$, $1 \leq i \leq h_{hist}$ of these frames. The observed behavior can be explained by taking into account the variability of action executions within the dataset, where, for example, one actor performs slower while another performs faster.

Although this fact is not requested in the given scenario, it would allow us to distinguish actions regarding the execution speed which can be of interest in further applications. For instance, the confusion between action classes running and walking or sneaking could be dissolved exploiting these temporal information.

### 4 SUMMARY AND OUTLOOK

We proposed a method for representing sequences of human actions while integrating spatial and temporal information into a combined model. This representation scheme was shown to be suitable for human action classification applications. Experiments performed on the CMU motion capturing dataset gave promising results which are able to compete with existing state of the art approaches.

To overcome certain false classifications, a hierarchy of single binary classifiers can be built. One can observe that similar motions are grouped into closer subtrees, while diverging actions are located in distinct subtrees.

Another field of research is the design of features used for classification. Since closely related classes tend to be confused, more sophisticated features should help to overcome this behavior.

The parameter vector $b_l$ could be used to build a self-similarity matrix instead of using the Euclidian landmark distances as proposed by (Junejo et al., 2011). More sophisticated distance measures like the angular distance in the manifold space could benefit the discriminability of the action classes.

For feeding real-world data to our approach, skeleton configurations can be extracted from frames provided by depth measuring camera devices such as Microsoft Kinect or PMD, which was recently shown to be possible in real-time (Li et al., 2010; Shotton et al., 2011). The combination of Active Shape Model based landmark detection and our proposed action representation could also be promising.

### ACKNOWLEDGEMENTS

The data used in our experiments was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

### REFERENCES


