Metric Learning from Poses for Temporal Clustering of Human Motion

Adolfo López-Méndez[1]
adolf.lopez@upc.edu

Juergen Gall[2]
juergen.gall@tue.mpg.de

Josep R. Casas[1]
josep.ramon.casas@upc.edu

Luc van Gool[3]
vangool@vision.ee.ethz.ch

[1] Technical University of Catalonia (UPC)
Barcelona, Spain

[2] MPI for Intelligent Systems
Tuebingen, Germany

[3] ETH Zurich
Switzerland

Segmenting human motion into distinct actions is a highly challenging problem. From the motion analysis perspective, segmentation is difficult due to large stylistic variations, temporal scaling, changes in physical appearance, irregularity in the periodicity of human motions and the huge number of actions and their combinations. From a semantic viewpoint, segmentation is inherently elusive and difficult because in the vast majority of cases it is not clear when a set of poses describes an action. For instance, punching with the left hand and punching with right hand can be different actions, but it might be also regarded as punching or even more general as boxing.

We propose to learn what makes a sequence of poses different from others such that it should be annotated as an action, as illustrated in Fig. 1.

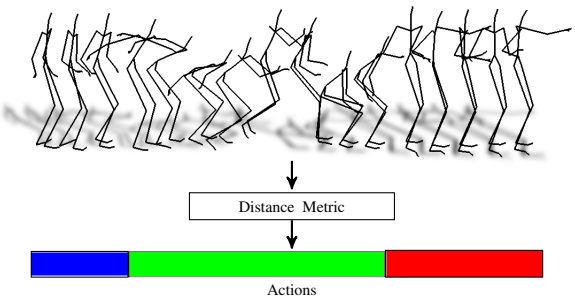

Figure 1: System Overview: Human motion sequences are clustered into different actions using a learned distance metric. We use annotations available in a mocap dataset to learn a distance metric that captures the semantic similarity between skeleton motion.

We make use of already annotated motion capture datasets and formulate action segmentation as a weakly supervised temporal clustering problem for an unknown number of clusters. Since publicly available datasets might contain different motions and action labels than the test sequences, we can not use the annotation directly for action segmentation. Instead, we use the annotations to learn a distance metric for skeleton motion using relative comparisons in the form of *samples of the same action are more similar than they are to a different action*. This is very intuitive since the sequences of a single database are usually labeled based on a semantic similarity.

We obtain a set of 14 relevant joint positions $\{\mathbf{q}_1,\dots,\mathbf{q}_{14}\}$ that can be easily obtained in different datasets [1]; see Fig. 2. We define a feature vector using these joint positions, and their velocity and acceleration:

$$\mathbf{x} = \{\mathbf{q}_1,\dots,\mathbf{q}_{14},\dot{\mathbf{q}}_1,\dots,\dot{\mathbf{q}}_{14},\ddot{\mathbf{q}}_1,\dots,\ddot{\mathbf{q}}_{14}\} \quad (1)$$

In the paper, we propose a set of relative constraints for pose features in order to capture the semantic similarity between poses given action labels of mocap datasets. We then rely on Information Theoretic Metric Learning (ITML) [3] in order to find the distance metric d_A, parameterized by a positive semi-definite matrix $\mathbf{A}$:

$$d_A(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{A}(\mathbf{x}_i-\mathbf{x}_j) \quad (2)$$

Since for each feature $\mathbf{x}_i$ we have only an action label y_i, we define the constraints based on triplets of points $(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)$ with class labels (y_i,y_j,y_k), where feature vectors with the same label should be closer to each other than to feature vectors with different labels. As an example, if $y_i = y_j \wedge y_j \neq y_k$ then the learned metric should hold $d_A(\mathbf{x}_i,\mathbf{x}_j) \leq \min(d(\mathbf{x}_i,\mathbf{x}_k),d(\mathbf{x}_j,\mathbf{x}_k))$.

The learned distance metric is then used to cluster the feature vectors in a test sequence into k motion primitives. The obtained primitives are provided to a hierarchical Dirichlet process (HDP)[5] that clusters the motion sequence into distinct behaviors (see Fig. 2). Provided that HDPs are non-parametric Bayesian models for infinite component mixtures, the number of actions (clusters) in a motion sequence is automatically estimated.

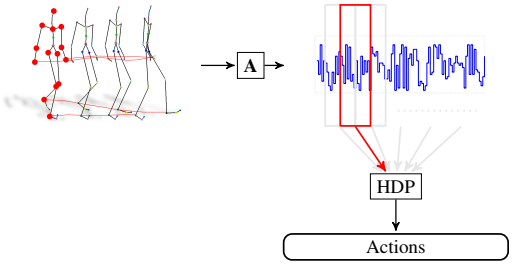

Figure 2: Detailed overview of our approach. A set of pose-based features are extracted using 14 relevant joints (marked with red spheres). These features are subsequently clustered into primitives using a metric ($\mathbf{A}$) learned on related action sequences. In order to infer the different actions in a sequence, we first group the primitives using a sliding window. Then, we provide the resulting sets of primitives to a hierarchical Dirichlet process.

We conduct experiments on two publicly available mocap datasets: the CMU dataset [2], and the HDM05 dataset [4]. Specifically, we carry cross-dataset experiments in order to validate that the learned metric can be used for unseen actions and across datasets.

Details about the proposed constraints, implementation and evaluation methodology are given in the paper. Our conclusion, supported on the experimental results, is that the learned metrics improve the clustering results even across datasets and do not require that the actions of the test sequences are present in the training data. Furthermore, the method does not require to know the actual number of actions in a sequence. This makes our semi-supervised temporal clustering approach a compelling alternative to other unsupervised methods.

[1] J. Barbič, A. Safonova, J. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard. Segmenting motion capture data into distinct behaviors. In *Proceedings of Graphics Interface 2004*, GI '04, pages 185–194, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2004. Canadian Human-Computer Communications Society.

[2] Carnegie Mellon University Motion Capture Database. http://mocap.cs.cmu.edu. URL `http://mocap.cs.cmu.edu`.

[3] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 209–216, 2007.

[4] HDM05 Mocap Dataset. http://www.mpi-inf.mpg.de/resources/hdm05/index.html. URL `http://www.mpi-inf.mpg.de/resources/HDM05/index.html`.

[5] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.