Two-level GMM Clustering of Human Poses for Automatic Human Behavior Analysis

Víctor Ponce, Miguel Reyes, Xavier Baró, Mario Gorga, and Sergio Escalera

Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Spain
Centre de Visió per Computador, Universitat Autònoma de Barcelona, Spain
E-mail: v88ponce@gmail.com, mreyese@gmail.com, xbaro@cvc.uab.es, mario.gorga@gmail.com, sergio@maia.ub.es

Abstract

Detect human poses is a main step in the study of human behavior analysis. Our achievement is to find a non supervised method to determine key poses of a human gesture for a posterior analysis of a human behavior. We use Kinect system for body feature extraction and then we apply a global and local feature clustering using a two-level Gaussian Mixture Model approach.

Keywords: Human Pose, Human Behavior Analysis, Clustering.

1 Introduction

Detect human poses is a main step in the study of human behavior analysis. However, human pose detection is a challenging task because of the huge inter/intra-limb feature variability in both still images and image sequences. From the point of view of data acquisition, many methodologies treat images captured by visible-light cameras. Computer Vision are then used to detect, describe, and learn visual features. The problem of limb detection and pose recovery becomes even more difficult because of the difficulties of uncontrolled environments: illumination changes, different points of view or occlusions, just to mention a few. In this work, we deal with the problem of human pose detection using a depth representation of image data using the Microsoft Kinect. Using this sensor, Shotton et al. [2] present one of the greatest advances in the extraction of the human body pose from depth images, representing the body as a skeletal form comprised by a set of joints. Based on the human skeleton representation, we propose a two-level Gaussian Mixture Model process to perform clustering of the space of human poses. Using this approach, poses are grouped in coherent sets which can be useful to train posterior general purpose human behavior systems. We test our approach on a novel data set of human behavior, showing higher coherency of human pose clusters in comparison with classical one-level GMM approach.

The rest of the paper is organized as follows: Section 2 describes the data acquisition process and feature description of human poses. Section 3 presents the two-level GMM approach for human pose clustering. Section 4 shows preliminary qualitative results, and finally, Section 5 concludes the paper.

2 Human Pose Representation

This section describes the processing of depth data in order to perform the segmentation of the human body, obtaining its skeletal model, and computing its feature vector.

For the acquisition of depth maps we use the

public API OpenNI software [1]. This middleware is able to provide sequences of images at rate of 30 frames per second. The depth images obtained are 340×280 pixels resolution. These features are able to detect and track people to a maximum distance of six meters from multi-sensor device.

We use the method of [2] to detect the human body and its skeletal model. The approach of [2] uses a huge set of human samples to infer pixel labels through Random Forest estimation, and skeletal model is defined as the centroid of mass of the different dense regions using mean shift algorithm. Experimental results demonstrated that it is efficient and effective for reconstructing 3D human body poses, even against partial occlusions, different points of view or no light conditions. The articulated human model is defined by the set of 15 reference points shown in Figure 1. This model has the advantage of being highly deformable, and thus, able to fit to complex human poses.

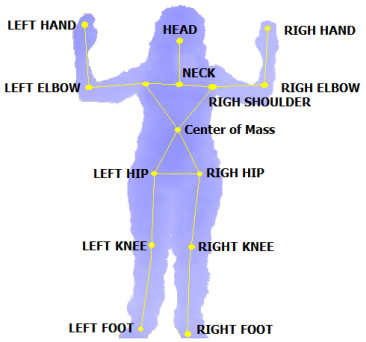

Figure 1: The 3D articulated human model consisting of 15 distinctive points.

In order to subsequently make comparisons and analyze the different extracted skeletal models, we need to normalize them. In this sense, we use the neck joint of the skeletal model as the origin of coordinates (OC). Then, the neck is not used in the frame descriptor, and the remaining 14 joints are used in the frame descriptor computing their 3D coordinates with respect to the OC. This transformation allows us to relate pose models that are at different depths, being invariant to translation, scale, and tolerant to corporal differences of subjects. Thus, the final feature vector $\mathbf{V_j}$ at frame j that defines the human pose is described by 42 elements (14 joints × three spatial coordinates),

$$V_j = \{\{v^1_{j,x}, v^1_{j,y}, v^1_{j,z}\}, ..., \{v^{14}_{j,x}, v^{14}_{j,y}, v^{14}_{j,z}\}\}$$

3 Human Pose Clustering

In order to group the previous pose descriptions in pose clusters, we use standard Gaussian Mixture Model. Our goal is to group the set of frame pose descriptions in clusters so that posterior learning algorithms can improve generalization in Human Behavior Analysis systems. We use a full covariance GMM of K components parameterized as follows,

$$\boldsymbol{\theta} = \{\pi(k), \mu(k), \Sigma(k), k = 1..K\}, \quad (1)$$

Then, a likelihood value based on the probability distributions $p(\cdot)$ of the GMM is obtained as follows,

$$\mathbf{GMM}(\mathbf{V}, \mathbf{k}, \boldsymbol{\theta}) = \sum_i -\log p(V|k_i, \boldsymbol{\theta}) - \log \pi(k_i) \quad (2)$$

Based on this standard probabilistic GMM model, our two-level clustering procedure is defined as follows,

1) First level: Use three spatial components of descriptor V for each joint $i, i \in [1, .., 14]$ and perform GMM of k^1 clusters, namely GMM^i_1

2) Second level:

2.1) Define for each pose a new feature vector,

$$V^* = \{v^1_1, .., v^1_{k^1}, .., v^1_1 4, .., v^1_{k^1} 4\}$$

of size $14 \cdot k^1$, where v^j_i is the probability result of applying GMM model GMM^i_1 at features from V corresponding to spatial coordinates of j-th joint.

2.2) Use components of descriptor V^* and perform GMM of k^2 clusters, namely GMM_2.

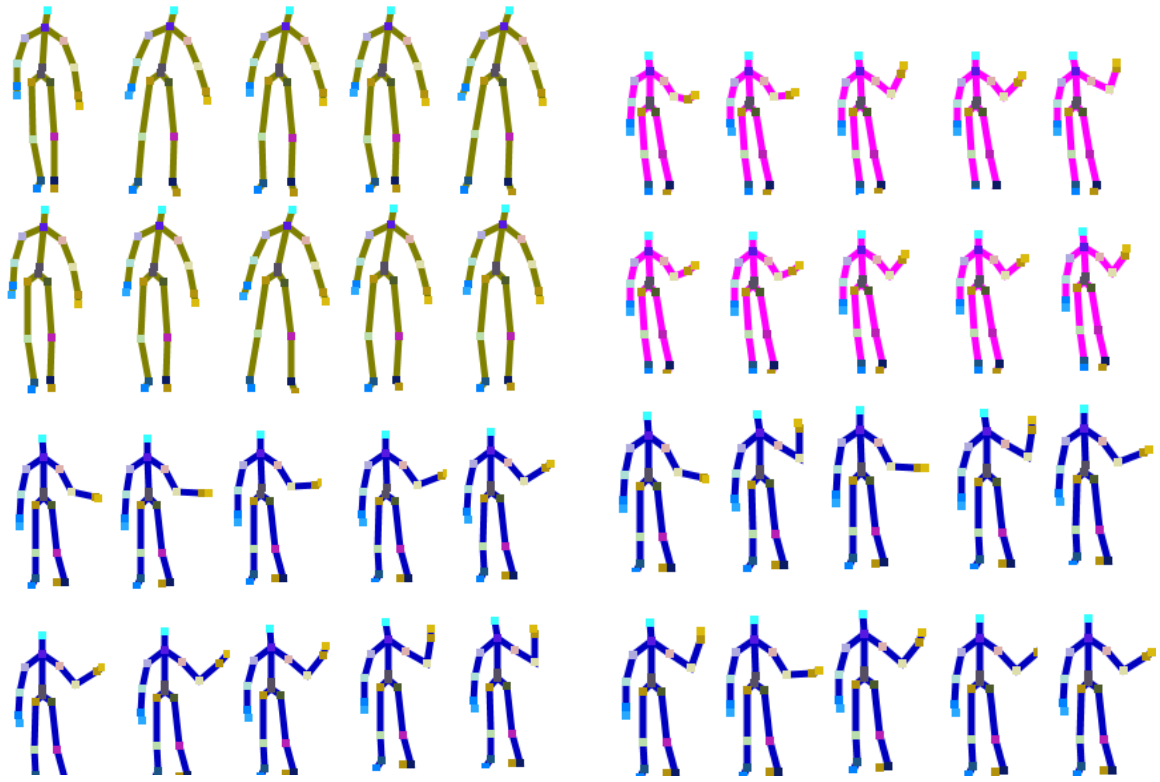

Figure 2: Sample examples from clusters defined using one-level GMM (left) and two-level GMM (right), respectively.

Given a new frame, then, human skeleton is obtained as described before, and feature vector V is computed and tested using two-level GMM description, obtaining a final probability for the most likely cluster from the set of k^2 possible pose clusters.

4 Results

Before the presentation of the results, first, we discuss the data, methods and parameters, and validation protocol of the experiments.

Data: We designed a new data set of gestures using the Kinect device consisting of seven different categories. It has been considered 10 different actors and different environments, having a total of 130 data sequences with 32 frame gestures. Thus, the data set contains the high variability from uncontrolled environments. The resolution of the video depth sequences is 340×280.

Methods and parameters: The people detection system used is provided by the public library OpenNI. This library has a high accuracy in people detection, allowing multiple detection even in cases of partial occlusions. The detection is accurate as people remain at a minimum of 60cm from the camera and up to 4m, but can reach up to 6m but with less robust and reliable detection. We perform classical one-level GMM and the proposed two-level GMM in Matlab programming, using $k^1 = 5$

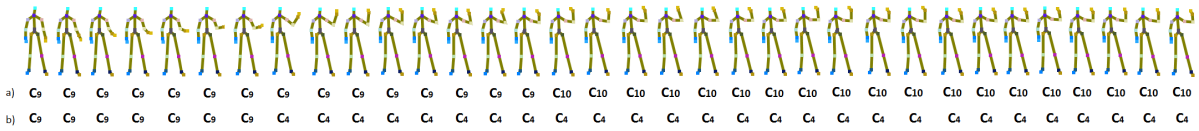

Figure 3: Cluster assignments for some gesture pose sequences.

and $k^2 = [10, 20, 40]$. k^2 is the number of clusters used for standard one-level GMM.

We tested the one and two-level GMM procedures on the designed data set. We show two qualitative results. First, in Figure 2, we show some examples of samples that fall in a particular cluster using one-level GMM and some samples that fall in a two-level GMM cluster. Up and down results are poses from different subjects. From this qualitative results we can observe that in the case of one-level cluster samples have more visual variability, grouping different pose from different subjects in the same cluster. This is mainly because all joints and spatial coordinates are considered independent in the one-level GMM procedure, and large movement in a particular joint affects global clustering for a particular pose. On the other hand, in the proposed two-level GMM clustering, all joints are first independently clustered, and grouped with equal probability in the second GMM level. As a result, we can observe that samples from the proposed clustering have more visual similarity, offering more discriminative information for better generalization of Human Behavior recognition techniques.

As an example of application, in the second qualitative result shown in Figure 3, we show consecutive visual descriptions of some data set gestures. At the bottom of the sequences, we show a first row that represents the cluster number assigned by a one-level GMM, and a second row with the assigned cluster using two-level GMM. One can see that in most cases both grouping techniques assigns consecutive poses to same clusters, but as shown earlier, the clusters assigned by the one-level GMM have more visual variability, being inefficient for human behavior generalization purposes.

5 Conclusion

In this paper, we designed a data set of human actions and described individual frames using pose skeleton models from depth map information. We proposed a two-level GMM clustering algorithm in order to group similar poses so that posterior Human Behavior analysis techniques can improve generalization. We showed some preliminary qualitative results comparing our approach with the classical one-level GMM clustering strategy, showing a more visual coherent grouping of poses.

References

[1] Open natural interface. November 2010. Last viewed 14-07-2011 13:00.

[2] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. *CVPR*, 2011.