3D Human Body-Part Tracking and Action Classification Using a Hierarchical Body Model

Leonid Raskin
raskinl@cs.technion.ac.il
Michael Rudzsky
rudzsky@cs.technion.ac.il
Ehud Rivlin
rivlin@cs.technion.ac.il

Computer Science Department
Technion -Israel Institute of Technology
Haifa, Israel, 3200

Human body pose tracking is a challenging task for several reasons. The large variety of poses and high dimensionality of the human model complicates the examination of the entire subject and makes it harder to detect each body part separately. However, the poses can be presented in a low dimensional space using the dimensionality reduction techniques. Such a reduction improves the trackers robustness, ability to recover from temporary target loss, and the computational effectiveness.

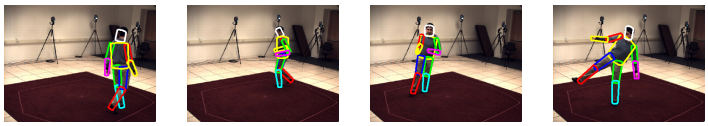
In this paper we introduce a Hierarchical Annealing Particle Filter (H-APF) tracker which exploits the Hierarchical Human Body Model in order to achieve accurate body part estimates. We apply a nonlinear dimensionality reduction using the Hierarchical Gaussian Process Latent Variable Model (HGPLVM) [1] and the Annealing Particle Filter (APF) [2] is used for the propagation between the sequential frames. A hierarchical model of the human body expresses conditional dependencies between the body parts and allows us to capture properties of separate parts. The human body model consists of two independent parts: one containing information about 3D location and orientation of the body and the other describing the articulation of the body. The articulation is represented as a hierarchy of body parts. The method uses previously observed poses from various motion types to generate mapping functions from the low-dimensional latent spaces to the data space that describes the poses. The tracking algorithm consists of two stages. First, the particles are generated in the latent space and are transformed to the data space using the learned mapping functions. Second, rotation and translation parameters are added to obtain valid poses. Finally, the likelihood function is calculated in order to evaluate how well these poses match the image. The resulting tracker estimates the locations in the latent spaces that represent poses with the highest likelihood. We demonstrate that our tracker is capable of robust tracking the human poses even in the scenarios that envolves several different motions.

Our method performs action classification in the latent spaces, produced by HGPLVM. A pose estimated on each frame corresponds to a coordinate in the latent space. Therefore, an action is represented by a curve in this latent space. The classification of the motion is based on the comparison of the sequences of latent coordinates that the tracker produces to the sequences that represent different actions (we denote such sequences as models). The modified Frèchet distance [3] is used in order to perform the comparison. This approach allows for the introduction of actions different from those used for the learning of the latent spaces by exploiting the model that represents it. We also show that the action classification, when performed in the latent space, is robust and has a high accuracy rate.

We tested H-APF tracker using the HumanEvaI and HumanEvaII [4] and some home-made datasets. The sequences contain different activities, such as *walking*, *boxing*,and *jogging*, which were captured by several (four) synchronized and calibrated cameras. The sequences were captured using the MoCap system that provides the correct 3D locations of the body joints, such as shoulders and knees. This information is used for an evaluation of the results and a comparison to other tracking algorithms. The error is calculated, based on a comparison of the tracker's output to the ground truth, using the average distance in millimeters between 3-D joint locations [5].

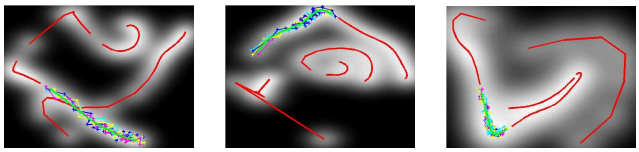
We trained HGPLVM with several different motion types and used this latent space in order to track the body parts on the videos from the HumanEvaI and HumanEvaII datasets. Figure 1 shows the result of the tracking of the HumanEvaII(S2) dataset which combines three different behaviors: walking, jogging and balancing.

Do demonstrade the ability of the classification algorithm we used several datasets. One of them includes five different activities: (1) *hand waving*, (2) *lifting an object*, (3) *kicking*, (4) *sitting down*, and (5) *punching*. For each activity five different sequences were captured performed by different actor. A cross-validation procedure was applied: for each motion type one sequence was used in order to construct the latent space and define the model of the motion type and the rest were used for evaluation. Figure 2 shows the trajectories produced by the trackers for the *sitting down* action projected on different latent spaces from different hierarchy levels. The average correct classification rate for these sequences is 90%.



frame 230 frame 640 frame 700 frame 800

Figure 1: Tracking results of H-APF tracker. Sample frames from the combo1 sequence from HumanEvaII(S2) dataset.



(a) (b) (c)

Figure 2: Tracking trajectories of *sitting down* action projected to different latent spaces: (a)hierarchy level 1, latent space 1; (b) hierarchy level 2, latent space 3; (c)hierarchy level 3, latent space 5. The green line represents the correct model (a model of *sitting down* action), the red lines represent the incorrect models (models of *waving, punching, kicking and picking an object* actions), and the colored lines represent the trajectories produced by the tracker. The crosses on the colored lines represent the actual latent locations, that were estimated by the H-APF tracker.

[1] N. D. Lawrence and A. J. Moore. Hierarchical gaussian process latent variable models. *Proc. International Conference on Machine Learning (ICML)*, 2007.

[2] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision (IJCV)*, 61(2):185–205, 2004.

[3] H. Alt, C. Knauer, and C. Wenk. Matching polygonal curves with respect to the Frèchet distance. *Proc. 18th International Symposium on Theoretical Aspect of Computer Science (STACS)*, 2:63–74, 2001.

[4] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2:2041–2048, 2006.

[5] A. Balan, L. Sigal, and M. Black. A quantitative evaluation of video-based 3d person tracking. *IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 349–356, 2005.