Manifold Learning for ToF-based Human Body Tracking and Activity Recognition

Loren Arthur Schwarz
schwarz@cs.tum.edu
Diana Mateus
mateus@cs.tum.edu
Victor Castañeda
castaned@cs.tum.edu
Nassir Navab
navab@cs.tum.edu

Chair for Computer Aided Medical
Procedures & Augmented Reality
Technische Universität München
Garching bei München, Germany

Recent technological advances have lead to the development of cameras that measure depth by means of the time-of-flight (ToF) principle [5]. ToF cameras allow capturing an entire scene instantaneously, and thus provide depth images in real-time. Despite the relatively low resolution, this type of data offers a clear advantage over conventional cameras for specific applications, such as human-machine interaction. In this paper, we propose a method that allows simultaneously *recognizing the performed activity* and *tracking the full-body pose* of a person observed by a a single ToF camera. Our method removes the need for identifying body parts in sparse and noisy ToF images [4] or for fitting a skeleton using expensive optimisation techniques [1].

The proposed method consists of learning a prior model of human motion and using an efficient, sampling-based inference approach for activity recognition and body tracking (Figure 1). The prior motion model is comprised of a set of low-dimensional manifold embeddings for each activity of interest. We generate the embeddings from full-body pose training data using a manifold learning technique [2]. Each of the embeddings acts as a low-dimensional parametrisation of feasible body poses [3] that we use to constrain the problem of body tracking only from depth cues. In a generative tracking framework, we sample the low-dimensional manifold embedding space by means of a particle filter and thus avoid exhaustively searching the full-body pose space. This way, we are able to track multiple pose hypotheses for different activities and to select one that is most consistent with the observed depth cues. Our depth feature descriptor, intuitively a sparse 3D human silhouette representation, can easily be extracted from ToF images.

The overall method combines the distinctiveness of multiple local, activity-specific motion models into a global model capable of recognising and tracking multiple activities from simple observations.

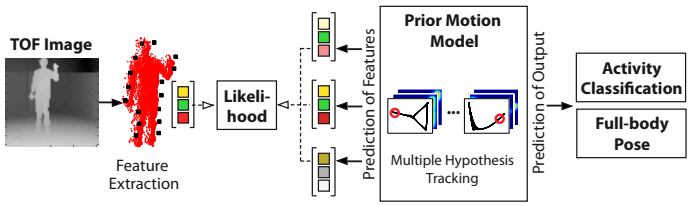

Figure 1: The proposed full-body tracking and activity recognition method is based on a learned motion model containing activity-specific manifolds of feasible poses. Tracking is achieved using a particle filter.

Let $\mathbf{y} \in \mathbb{R}^{d_y}$ denote a full-body pose, consisting of the joint angles of a simple skeleton body model and let $\mathbf{s} \in \mathbb{R}^{d_s}$ be a feature vector representing a ToF depth image. We are given a training dataset of labelled full-body poses and ToF feature vectors $\{\mathbf{Y}^\alpha, \mathbf{S}^\alpha\}$, $\alpha \in \{1,\dots,M\}$, for M activities of interest. The dataset is acquired with a synchronized motion capture and ToF camera system. Each activity α contains N_α training poses, i.e. $\mathbf{Y}^\alpha = [\mathbf{y}_1^\alpha,\dots,\mathbf{y}_{N_\alpha}^\alpha]$ and $\mathbf{S}^\alpha = [\mathbf{s}_1^\alpha,\dots,\mathbf{s}_{N_\alpha}^\alpha]$. During the training phase, we learn a prior motion model that consists of the following activity-specific components:

1. *Manifold embeddings* $\mathbf{X}^\alpha = [\mathbf{x}_1^\alpha,\dots,\mathbf{x}_{N_\alpha}^\alpha]$ generated from the full-body pose training data $\mathbf{Y}^\alpha$ using Laplacian Eigenmaps [2], such that each embedding point $\mathbf{x}_i^\alpha$ corresponds to a full-body pose $\mathbf{y}_i^\alpha$.
2. *Regression mappings* $f_{xs}^\alpha(\mathbf{x})$ and $f_{xy}^\alpha(\mathbf{x})$, learned from training data, that allow predicting feature vectors and full-body poses, respectively, from embedding locations $\mathbf{x}$.

Figure 2: Pose estimation examples for several frames of different activities. *Top:* Segmented input 3D ToF data of a person. *Bottom:* Corresponding estimated (red) and ground truth (blue) full-body poses.

3. *Pose likelihood priors* $p_{\text{pose}}(\alpha,\mathbf{x})$ that model the feasibility of a pose of activity α, given by its low-dimensional representation $\mathbf{x}$.
4. *Activity switching priors* $p_{\text{switch}}(\alpha,\mathbf{x})$ that model the likelihood of switching to another activity from any embedding location $\mathbf{x}$.

In the testing phase, we recognize the performed activity $\hat{\alpha}_t$ and predict the full-body pose $\hat{\mathbf{y}}_t$ at every time step t, given only observed feature vectors $\mathbf{s}_t$. We model the state of our dynamic system as a pair $(\hat{\alpha}_t, \hat{\mathbf{x}}_t)$ of an activity index and a position in the corresponding manifold embedding. For state inference, we employ a particle filter that efficiently samples the embedding space and tracks multiple pose hypotheses.

We recorded a training and testing dataset using a ToF camera synchronized with an optical motion capture system. Depth features were extracted from the ToF images according to the procedure described in the full paper. The descriptor has $d_s = 48$ dimensions and the manifold embeddings have $d_x = 2$ dimensions. We considered 10 activities: clapping, golfing, hurrah (arms up), jumping jack, knee bends, picking something up, punching, scratching head, playing the violin and waving. Each of the movements was recorded 6 times with 10 actors. Only the depth features were used for testing, the motion capture data served as ground truth. Our experiments were performed in a cross-validation scheme. Over all testing sequences, 92% of all non-idle frames were classified as the correct activity. Misclassification mainly occurred between activities with similar poses, such as *waving* and *scratching head*. The predicted full-body poses deviate from the ground truth poses by 4.21° per joint, or alternatively, by 29.1mm in 3D space (see Figure 2). The detailed evaluation in the full paper shows that our method can reliably recognize movements of multiple activities and precisely estimate full-body pose.

[1] A O Balan, L Sigal, M J Black, J E Davis, and H W Haussecker. Detailed human shape and pose from images. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, May 2007.

[2] M Belkin and P Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373 – 1396, Feb 2003.

[3] A Elgammal and C Lee. The role of manifold learning in human motion analysis. *Human Motion Understanding, Modeling, Capture and Animation*, pages 1–29, 2008.

[4] V Ganapathi, C Plagemann, D Koller, and S Thrun. Real time motion capture using a single time-of-flight camera. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[5] A Kolb, E Barth, R Koch, and R Larsen. Time-of-flight sensors in computer graphics. *EUROGRAPHICS*, pages 119–134, 2009.