3D Reconstruction of Human Motion from Video

Hashim Yasin, Björn Krüger, and Andreas Weber

Department of Computer Science II,
University of Bonn, Germany.
{yasin,kruegerb,weber}@cs.uni-bonn.de

Abstract. This paper presents a novel framework for 3D full body reconstruction of human motion from uncalibrated monocular video data. We first detect and track feature sets from video sequences by employing MSER and SURF feature detection techniques together with prior information obtained from the motion capture database. By deriving suitable feature sets from both video and motion capture data, we are able to employ fast motion retrieval techniques to find the best relevant prior poses. The resulting 3D motion sequences are reconstructed by an energy minimization process that takes multiple prior terms into account.

Keywords: Video based motion retrieval, 3D motion reconstruction.

1 Introduction

Motion reconstruction from video has been remained a major research topic from the last decade. In this paper, we reconstruct 3D motion from uncalibrated monocular video stream by employing data-driven approach. First, we detect and track feature sets from monocular video. We utilize Maximally Stable Extremal Regions (MSER) and Speeded Up Robust Features (SURF) feature detection techniques combined with some prior knowledge retrieved from the motion capture (MoCap) library, to make feature detection and tracking more robust. We employ online lazy neighbourhood graph (OLNG) for efficient similarity search in the line of Tautges et al. [1] to work with 2D feature sets. These 2D feature sets, extracted either from video or MoCap data, are used as control input signal.

2 Feature Design and Retrieval

In this section, we demonstrate how we retrieve motion with the help of suitable feature sets both from the motion capture library and video input. For motion retrieval from MoCap library, we employ 3D feature sets $\mathcal{F}_{3D}^{15}$ based on positions of hands, feet and head similar to Krüger et al. [2]. We extract 2D feature sets $\mathcal{F}_{2D}^{10}$ as done in the paper of Yasin et al. [3] by projection of 3D feature sets onto 2D plane at different viewing directions like azimuth angles (0-10-350) degrees with 10 degree step size and elevation angles (0-15-90) degrees with step size 15 degree. These 2D feature sets are further normalized by translating mean (center

of mass) to its origin of the coordinate system. For video data, we detect and track positions of hands, feet and head to develop video based feature sets $\mathcal{F}_{\text{vid}}^{10}$ for input query. We employ together low-level image feature detection techniques (MSER and SURF) and high-level 3D prior information available in MoCap database. This 3D prior data are back projected into image plane and used to make detection and tracking more robust and precise. For similarity search, we employ kd-tree, build upon 2D feature sets $\mathcal{F}_{\text{2D}}^{10}$, and a graph structure OLNG like in [1,2]. As a result, we obtain k-nearest neighbours ($k$nn) from MoCap database. We select the best N poses from these retrieved nearest poses through OLNG by considering step sizes and minimum cost of the paths of $k$nn. In our experiments, the value of k is fixed to be 2^{12} and the value of N is set to be 256.

3 Online Motion Reconstruction

We have developed our reconstruction algorithm on the basis of retrieved N similar poses from MoCap database with joint angle configurations $Q^t = \{\boldsymbol{q}_1^t, \ldots, \boldsymbol{q}_N^t\}$ and positional information $X^t = \{\boldsymbol{x}_1^t, \ldots, \boldsymbol{x}_N^t\}$ at current frame t. We formulate our reconstruction methodology as energy minimization problem as,

$$P_{\text{rec}} = argmin(w_p E_p + w_j E_j + w_s E_s + w_c E_c). \tag{1}$$

with user defined weights w_p, w_j, w_s and w_c. The energy terms E_p, E_j, E_s and E_c are optimized by gradient descent based method and are explained as below;

Pose Energy Term, E_p, measures a-priori likelihood of the synthesized joints angle configurations, called as synthesized pose, from the MoCap library. We employ symmetric square root kernel function to estimate probability density similar to [1], which is formulated into energy term as,

$$E_p = \sum_{n=1}^{N} w_n^t \cdot \sqrt{|\tilde{\boldsymbol{q}}_n^t - \boldsymbol{q}^t|}. \tag{2}$$

where $\tilde{\boldsymbol{q}}_n^t$ is the joints angle configurations of the best retrieved N poses and $\boldsymbol{q}^t$ is the joint angel configurations obtained in a PCA space at frame t.

Joint Energy Term, E_j, minimizes unwanted artifacts arises due to 2D-3D transformation and compels joints positions $\boldsymbol{x}^t$, resulted from forward kinematic of synthesized pose, according to the prior joints positions of the N best poses,

$$E_j = \sum_{n=1}^{N} w_n^t \cdot \sqrt{|\tilde{\boldsymbol{x}}_n^t - \boldsymbol{x}^t|}. \tag{3}$$

Smooth Energy Term, E_s, keeps away the jittering and jerkiness effects and imposes smoothness in a way that newly reconstructed pose is bound to be according to the previously two reconstructed poses and the prior knowledge about smoothness between neighbouring candidates exists in MoCap database,

$$E_s = \sum_{n=1}^{N} w_n^t \cdot \sqrt{|\tilde{\boldsymbol{S}}_n^t - \boldsymbol{S}^t|}. \tag{4}$$

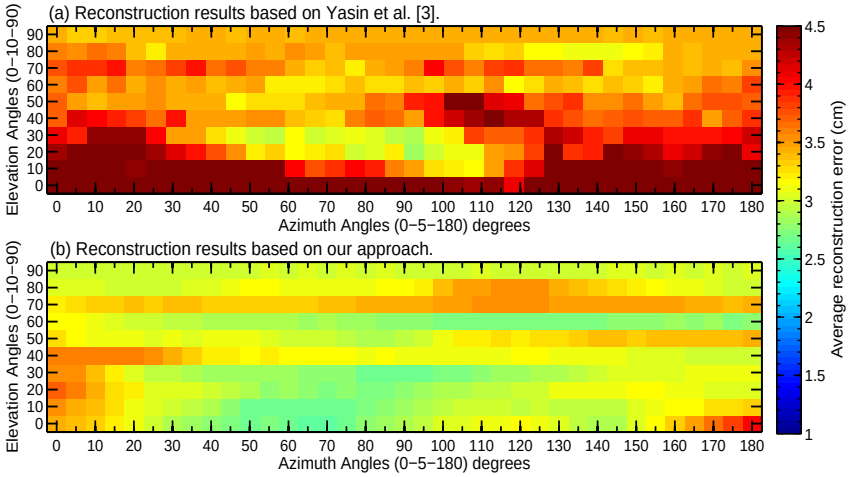

Fig. 1. Average reconstruction error graphs for walking in straight motion with azimuth angles (0-5-180) degrees and elevation angles (0-10-90) degrees.

where $\tilde{\boldsymbol{S}} = \tilde{\boldsymbol{x}}^t - 2\tilde{\boldsymbol{x}}^{t-1} + \tilde{\boldsymbol{x}}^{t-2}$ with position vectors $\tilde{\boldsymbol{x}}^t$, $\tilde{\boldsymbol{x}}^{t-1}$ and $\tilde{\boldsymbol{x}}^{t-2}$ of the N best retrieved poses; and $\boldsymbol{S} = \boldsymbol{x}^t - 2\boldsymbol{x}^{t-1} + \boldsymbol{x}^{t-2}$ with position vectors $\boldsymbol{x}^t$, $\boldsymbol{x}^{t-1}$ and $\boldsymbol{x}^{t-2}$ of the reconstructed poses at frames t, $t-1$ and $t-2$ respectively.

Control Energy Term, E_c, minimizes the distance between feature sets of the performed input motion and the synthesized motion. We project 3D feature sets of hands, feet and head of the current synthesized pose into 2D plane at specific view and normalize these 2D feature sets. The estimated 2D feature sets $\hat{\boldsymbol{x}}^t_{est,2d}$ extracted from input motion at current frame t are then get subtracted from these normalized 2D feature sets $\hat{\boldsymbol{x}}^t_{2d}$ of current frame t as,

$$E_c = \sqrt{|\hat{\boldsymbol{x}}^t_{2d} - \hat{\boldsymbol{x}}^t_{est,2d}|}. \tag{5}$$

4 Results and Conclusion

We employ HDM05 [4] MoCap library and record video input motions using Kinect RGB camera with resolution 587×440 pixels and frame rate 30 frames per seconds. We testify our approach on variety of motions like straight walking, side walking, walking in a circle, jumping jack and cartwheel motions etc.

In case of **synthetic data**, 2D data extracted from the MoCap motion clip, named as synthetic data, are given as input to the system for 3D reconstruction. We evaluate our approach with our previous contribution Yasin et al. [3] and construct average reconstruction error graphs where azimuth angles are given along x-axis, elevation angles are drawn along y-axis and reconstruction error is color-coded. From the results shown in Figure 1, it is quite obvious that the reconstruction has been improved in all viewing directions when we employ our

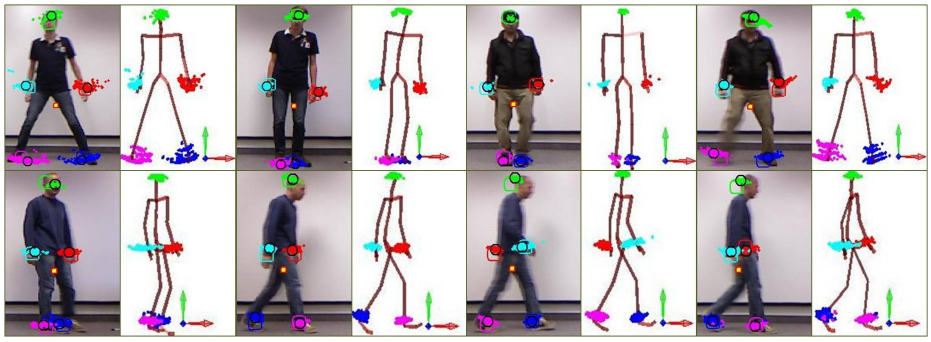

Fig. 2. Tracking and reconstruction results of our approach with $k$nn for different types of motions. The further results can be seen in supplementary material available at [5].

proposed approach where we use symmetric square root kernel function instead of multivariate normal distribution model to estimate probability density.

For **video data**, our proposed approach comparatively outperforms as well. We have improved the process of detection and tracking through 3D prior existing knowledge and ultimately reconstruction results as compared to Yasin et al. [3]. From experiments, we observe that mistracking has been reduced from 20-25% to 8-10% on an average, with back projection of 3D prior knowledge.

As **conclusive remarks**, we have proposed an efficient data driven full body reconstruction approach for video data by using the positions of just hands, feet and head. We have competently utilized the 3D prior knowledge to make low level image based feature detection and tracking process more robust. Our system performs reconstruction with frame rate roughly 5-6 frames per second.

References

1. Tautges, J., Zinke, A., Krüger, B., Baumann, J., Weber, A., Helten, T., Müller, M., Seidel, H.P., Eberhardt, B.: Motion reconstruction using sparse accelerometer data. ACM Trans. Graph. **30** (2011) 18:1–18:12
2. Krüger, B., Tautges, J., Weber, A., Zinke, A.: Fast local and global similarity searches in large motion capture databases. In: 2010 ACM SIGGRAPH / Eurographics Symposium on Computer Animation. SCA '10, Aire-la-Ville, Switzerland, Switzerland, Eurographics Association (2010) 1–10
3. Yasin, H., Krüger, B., Weber, A.: Model based full body human motion reconstruction from video data. In: 6th International Conference on Computer Vision / Computer Graphics Collaboration Techniques and Applications (MIRAGE 2013). (2013)
4. Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., Weber, A.: Documentation Mocap Database HDM05. Technical Report CG-2007-2, Universität Bonn (2007)
5. Department of Computer Graphics, University of Bonn: Motion Tracking, Retrieval and 3D Reconstruction from Video (2013) `http://cg.cs.uni-bonn.de/vmotrec`.