Marker-less 3D Human Motion Capture with Monocular Image Sequence and Height-Maps

Yu Du[1], Yongkang Wong[2], Yonghao Liu[1], Feilin Han[1], Yilin Gui[1],
Zhen Wang[1], Mohan Kankanhalli[2,3], Weidong Geng[1⋆]

[1]College of Computer Science, Zhejiang University
[2]Interactive & Digital Media Institute, National University of Singapore
[3]School of Computing, National University of Singapore
{answeror,yonghaoliu,hanfeilin,ylgui,wangzh_cs,gengwd}@zju.edu.cn,
yongkang.wong@nus.edu.sg, mohan@comp.nus.edu.sg

Abstract. The recovery of 3D human pose with monocular camera is an inherently ill-posed problem due to the large number of possible projections from the same 2D image to 3D space. Aimed at improving the accuracy of 3D motion reconstruction, we introduce the additional built-in knowledge, namely height-map, into the algorithmic scheme of reconstructing the 3D pose/motion under a single-view calibrated camera. Our novel proposed framework consists of two major contributions. Firstly, the RGB image and its calculated height-map are combined to detect the landmarks of 2D joints with a dual-stream deep convolution network. Secondly, we formulate a new objective function to estimate 3D motion from the detected 2D joints in the monocular image sequence, which reinforces the temporal coherence constraints on both the camera and 3D poses. Experiments with HumanEva, Human3.6M, and MCAD dataset validate that our method outperforms the state-of-the-art algorithms on both 2D joints localization and 3D motion recovery. Moreover, the evaluation results on HumanEva indicates that the performance of our proposed single-view approach is comparable to that of the multi-view deep learning counterpart.

Keywords: Human Pose Estimation, Height-map

1 Introduction

Marker-less motion capture is an active field of research in computer vision and graphics with applications in computer animation, video surveillance, biomedical research, and sports science. According to the recent study on world population aging [1], the life expectancy at age 60 and above is expected to grow in the next few decades. This anticipates an emerging need in video-based analysis systems to monitor the elderly in nursing home as an event alert system.

Existing motion capture approaches can be broadly divided into two categories: (1) methods based on monocular camera [2–5], and (2) methods that rely

[⋆] Corresponding author.