Int J Comput Vis (2013) 101:482–497
DOI 10.1007/s11263-012-0557-0

Using Segmented 3D Point Clouds for Accurate Likelihood Approximation in Human Pose Tracking

Nicolas Lehment · Moritz Kaiser · Gerhard Rigoll

Received: 13 November 2011 / Accepted: 8 August 2012 / Published online: 30 August 2012

Abstract The observation likelihood approximation is a central problem in stochastic human pose tracking. In this article we present a new approach to quantify the correspondence between hypothetical and observed human poses in depth images. Our approach is based on segmented point clouds, enabling accurate approximations even under conditions of self-occlusion and in the absence of color or texture cues. The segmentation step extracts small regions of high saliency such as hands or arms and ensures that the information contained in these regions is not marginalized by larger, less salient regions such as the chest. To enable the rapid, parallel evaluation of many poses, a fast ellipsoid body model is used which handles occlusion and intersection detection in an integrated manner. The proposed approximation function is evaluated on both synthetic and real camera data. In addition, we compare our approximation function against the corresponding function used by a state-of-the-art pose tracker. The approach is suitable for parallelization on GPUs or multicore CPUs.

Keywords Human pose tracking · Depth image · Observation likelihood approximation · Stochastic tracking · Parallel computing

N. Lehment (✉) · M. Kaiser · G. Rigoll
Institute for Human-Machine-Communication, Technische Universität München, Arcisstr. 21, 80333 München, Germany
e-mail: lehment@tum.de

M. Kaiser
e-mail: moritz.kaiser@tum.de

G. Rigoll
e-mail: rigoll@tum.de

1 Introduction

Human pose tracking is a highly complex task which has undergone rapid development over the last decade. Applications range from entertainment systems to professional HCI applications, e.g. in sterile environments of hospitals.

In recent years, the two classical approaches of monocular and multi-camera pose tracking have been joined by depth image based pose tracking (as summarized by Poppe 2007). The majority of depth data based pose tracking systems employ stochastic methods. The observation likelihood function lies at the heart of these stochastic methods, providing a measure of confidence that a given pose hypothesis is supported by the observed data. Typically, an observation likelihood function is derived from edge or feature matching between a deformable body model and the current observation (Isard and Blake 1998; Deutscher and Reid 2005). Azad et al. (2008) use edge cues in combination with a separate hand and head tracker, Bernier et al. (2008) consider a combination of 3D contour points and a separate hand tracker, Darby et al. (2008) work only with the 3D contour points while Fontmarty et al. (2007) use edge points and a number of other cues.

Most methods considered so far use a stereo camera or a multi-camera setup of four or more calibrated cameras. The growing availability of affordable and precise depth sensing cameras, such as Microsoft's Kinect system or ASUS's Xtion camera, have sparked an increased interest in the use of pure depth data for pose estimation. Most recent approaches to pose tracking, such as presented by Zhu and Fujimura (2009), Ganapathi et al. (2010) and Baak et al. (2011) work directly on the depth data obtained from the sensor supported by key-point detection. The work of Shotton et al. (2011) and Girshick et al. (2011) extends methods of image classification to tackle the body part detection and