Journal of Visual Languages and Computing 48 (2018) 10–18

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Data-driven human model estimation for realtime motion capture

Le Su[1,b], Lianjun Liao[1,a,c,d], Wenpeng Zhai[b], Shihong Xia[*,a]

[a] Institute of Computing Technology, CAS, Beijing 100190, China
[b] Civil Aviation University of China, Tianjin 300300, China
[c] North China University of Technology, Beijing 100144, China
[d] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Keywords:
Human model estimation
Data driven
Human motion capture
Depth image

ABSTRACT

In this paper, we present a practicable method to estimate individual 3D human model in a low cost multi-view realtime 3D human motion capture system. The key idea is: using human geometric model database and human motion database to establish geometric priors and pose prior model; when given the geometric prior, pose prior and a standard template geometry model, the individual human body model and its embedded skeleton can be estimated from the 3D point cloud captured from multiple depth cameras. Because of the introduction of the global prior model of body pose and shapes into a unified nonlinear optimization problem, the accuracy of geometric model estimation is significantly improved. The experiments on the synthesized data set with noise or without noise and the real data set captured from multiple depth cameras show that the estimation results of our method are more reasonable and accurate than the classical methods, and our method is better noise-immunity. The proposed new individual 3D geometric model estimation method is suitable for online realtime human motion tracking system.

1. Introduction

With the rapid development of computing technology, three-dimensional (3D) human body models and their dynamic motions are widely used in the digital entertainment industry. Human performance mainly involves human body shapes and motions. Human motion capture and tracking is a hot issue in computer vision and graphics [1]. It mainly studies how to quickly reconstruct accurate human geometry model and human motion sequence from the input depth data stream. The motion capture technology has important application value in the movie stunt, animation games, sports training and other fields, for example, the captured human motion sequence can be used to guide sports training, or to improve the sense of reality of game characters. Unlike traditional marker-based motion capture method [2], RGB-D motion capture system has no need of body-worn sensors [3], which make uncomfortable movements, and difficult tasks as marker labeling or missing-marker estimation. However, up to now, the existing low-cost commercial RGB-D human capture system such as Kinect, suffered from ill-pose problem caused by limbs occlusions or self-occlusions, and could not robustly reconstruct reasonable accurate 3D human motion sequence. In contrast with single-view-based system such as [4], the multi-view based methods such as [5], could achieve even more accuracy by minimizing the influence of ill-pose problem caused by limbs occlusion or self-occlusion. Realtime human motion capture systems are also reported. However, these methods need a pre-established human model, when human body size changed significantly or a pose was not in the database, it would fail.

To address this issue, we present a practicable method to estimate individual 3D human model in a low-cost multi-view realtime 3D human motion capture system. The key idea is: based on human geometric model database and motion database, establish geometric priors and pose prior model; then with the established geometric priors, pose prior and a standard template geometry model, the individual human body model and its embedded skeleton can be estimated from the captured 3D point clouds from multiple depth cameras. The main contributions in this work are as follows:

- We proposed a new individual 3D geometric model estimation method suitable for online realtime human motion tracking system.
- Successfully introduced the global prior model of human body pose and shapes into the nonlinear optimization problem of human geometry model estimation, and consequently the accuracy of geometric model estimation is significantly improved.

* Corresponding author.
E-mail addresses: liaolianjun@ict.ac.cn (L. Liao), xsh@ict.ac.cn (S. Xia).
[1] These authors are co-first authors.

https://doi.org/10.1016/j.jvlc.2018.05.001
Received 17 December 2017; Received in revised form 12 February 2018; Accepted 2 May 2018

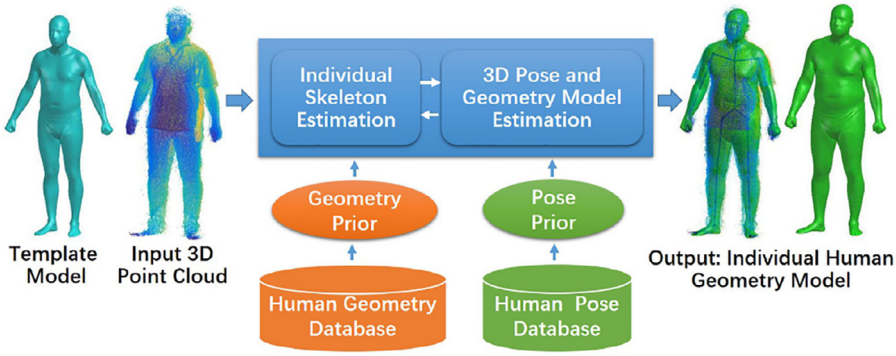

Fig. 1. System overview.

2. Related work

Model-free methods such as [6–8], considering no prior information of human body, identified human pose in a frame through feature point detection of image or mesh. The drawback is that it neglected the influence of previous frames on the pose of the current frame, that is, ignoring the essence that human motion is a continuous process of spatial and temporal variation. Model-based methods need a 3D model scanned in advance, such as methods based on point cloud ICP [9–11], or methods based on multi view depth camera [12,13]. The cost of the 3D scanner is high, and it is time-consuming to process the scanned data. It also suffers from error accumulation and tracking the long time movement.

Data driven methods, with the help of a 3D pose database constructed in advance from captured motion data, can usually achieve compelling results. Siddiqui and Medioni [14] and Baak et al. [15] estimated human pose in each frame by detecting feature points from depth image. Since the models in the database are the standard 3D human body models, it could not always get reasonable results when the size of actors was much different from the standard model in database. Ye et al. [16] retrieved the optimal match between the 3D point cloud captured from multi depth camera and the 3D human pose in the database, and then estimated the full-body pose by deforming the retrieved pose back to the captured 3D point cloud through non-rigid registration. The human pose database was composed of the 3D point clouds calculated from the depth images which were generated by projecting a standard 3D body model driven by an embedded skeleton. Zhang et al. [5] presented an efficient physics-based motion reconstruction algorithm, which integrated the input depth data from 3 Kinect cameras, foot pressure data from wearable pressure sensors and detailed full-body geometry, and then reconstructed offline full-body motion (i.e. kinematic data) and human dynamic data. When tracking the 3D skeletal poses, the global PCA for features dimension reduction process was done firstly on the 3D body pose set in the CMU motion database, and thus the reconstructed 3D pose prior was equivalent to imposing additional range constraint on human joint angle. Zhu et al. [17] succeeded in tracking human motion by combining the semantic feature detection of human limbs based on Bayesian estimation and inverse kinematical optimization calculation with constraints such as joint limit avoidance. However with the assumption that human head in the image is always located above its waist, therefore, it could not deal with the pose that did not meet this condition. Wei et al. [4] provided a fast, automatic method for capturing full-body motion data using a single depth camera. It described the realtime 3D human posture reconstruction from monocular depth image as formulating the registration problem in a Maximum A Posteriori (MAP) framework and iteratively registered a 3D articulated human body model with monocular depth image. Xu et al. [18] presented a novel markerless motion capture system (FlyCap), using multiple cooperative flying cameras to store the captured RGBD sensor data and the recorded VO data in the onboard NUC for later off-line reconstruction.

In general, from the above human motion capture system, we can see that the model-based method often is more better in accuracy and the human geometry model estimation plays an important role in these methods. Therefore, our motivation of this paper is how to reasonably and accurately estimate the individual geometry model suitable for realtime human motion system. The most related work to ours in this paper is that proposed by Anguelov et al. [19] which introduce the SCAPE method, a data-driven method, for building a human shape model that spans variation in both subject shape and pose. Generally, the SCAPE method can reconstruct a reasonable and accurate 3D human geometric model and pose. However, the SCAPE method required accurate point correspondence between non-rigid model and the target 3D point cloud to ensure the 3D human pose accuracy estimation and large-scale deformation, and when the target 3D human pose differs largely from the template model, the results of SCAPE would be obviously unreasonable.

3. Overview

We propose an effective approach to online estimate individual human geometric model for a low-cost realtime 3D human motion capture system. Fig. 1 gives the pipeline of our system. The input data is the 3D point cloud captured by multi-view calibrated depth camera and a standard template human geometric model. Then the individual 3D geometric models and embedded skeletons consistent with input point clouds are accurately constructed by making full use of human geometric model database and human pose database. We will demonstrate that our method can quickly reconstruct reasonable and accurate individual 3D geometric models and its embedded skeletons, and it is suitable for realtime human motion capture system.

4. Data acquisition

This section focuses on the data acquisition method from multiple depth cameras and the representation form of pose database.

4.1. Spatial and temporal alignment

There are four deep cameras (Microsoft Kinect v2.0) connected to one PC, which extended three PCI card to obtain three additional USB3.0 ports (only one USB3.0 hub on my mainboard). Since Microsoft's Kinect driver does not support multiple Kinect cameras connecting simultaneously, we use the open source device driver libfreenect2 [20].

Temporal alignment. The frame rate of Kinect is 30 fps, that is, the acquisition cycle is about 33 ms [21]. Attaching a timestamp for every depth frame when capturing, the synchronous group comprises of depth frames with time difference of less than half a period ($\leq$ 15 ms).

Camera parameters. There are four Kinect cameras located at the