ARTICLE IN PRESS

Robotics and Autonomous Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

Semantic parametric body shape estimation from noisy depth sequences

Alexandru Eugen Ichim [a], Federico Tombari [b,c,*]

[a] École Polytechnique Fédérale de Lausanne, Switzerland
[b] University of Bologna, Italy
[c] TU Munich, Germany

HIGHLIGHTS

- A framework for tracking and modeling of human bodies from sequences of depth maps.
- Modular and extensible energy cost optimization, with depth and prior constraints.
- Compact semantic tags associated to the estimated body shape using L1 relaxation.
- Relies on the tools and algorithms provided by the Point Cloud Library (PCL).
- 3 fps performance for continuous tracking and modeling on the CPU.

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
3D body modeling
3D body tracking
Depth data
Point Cloud Library

ABSTRACT

The paper proposes a complete framework for tracking and modeling articulated human bodies from sequences of range maps acquired from off-the-shelf depth cameras. In particular, we propose an original approach for fitting a pre-defined parametric shape model to depth data by exploiting the 3D body pose tracked through a sequence of range maps. To this goal, we make use of multiple types of constraints and cues embedded into a unique cost function, which is then efficiently minimized. Our framework is able to yield compact semantic tags associated to the estimated body shape by leveraging on semantic body modeling from MakeHuman and L1 relaxation, and relies on the tools and algorithms provided by the open source Point Cloud Library (PCL), representing a good integration of the functionalities available therein.

1. Introduction and related work

The task of 3D body modeling aims at automatically obtaining an accurate 3D model of a person's body. The possibility of having at disposal an accurate 3D model adapted to the body characteristics of a subject opens up new directions in a variety of applications, such as in the fields of entertainment (e.g. 3D avatar creation for videogaming and movie special effects), fitness (e.g., for automatic estimation of the body mass), apparel (e.g., for virtual changing room applications), interactive design, and security (people detection and identification).

The output of this task is generally represented by a parametric 3D body model, with the parameters estimated so that the model adapts to the specific characteristics of the subject being scanned. It is often the case that these parametric models are open sourced and available to the community so to favor interchange and standardization. While earlier parametric models [1] were based on simple Principal Component Analysis (PCA) of standard human poses (e.g., T/A poses), more recent approaches also model minute body deformations such as muscle bulging under complex poses, e.g., the SCAPE models [2,3]. Another possibility of sourcing parametric body models is from semantic models, i.e., models built by artists, where each body shape modifier has an associated semantic tag, such as it is the case of MakeHuman [4].

Accurately estimating the 3D body model traditionally requires dedicated and expensive hardware to acquire high resolution scans of the body, generally by means of 3D laser scanners or high frame-rate structured light sensors. In addition, this procedure is characterized by high processing time due to the re-positioning of the scanner from different view points, the acquisition and the joint 3D registration of the different scans. To overcome

* Corresponding author at: University of Bologna, Italy.
E-mail address: federico.tombari@unibo.it (F. Tombari).

http://dx.doi.org/10.1016/j.robot.2015.09.029
0921-8890/

Please cite this article in press as: A.E. Ichim, F. Tombari, Semantic parametric body shape estimation from noisy depth sequences, Robotics and Autonomous Systems (2015), http://dx.doi.org/10.1016/j.robot.2015.09.029

such limitations, the work of [5] proposed to fit a 3D parametric model to a frame acquired by means of a monocular RGB camera. Although not fully automatic due to the need of user interaction as well as limited in the modeling accuracy due to the 2D to 3D fitting, this work introduced the concept of using low-cost hardware for the task of 3D body modeling.

Successively, thanks to the popularity of consumer depth cameras originated by the development of the Microsoft Kinect, other works [6–9] have tackled 3D body modeling by means of the noisy range data acquired from such low-cost 3D sensors. Initially, [6] proposed to fit each parametric 3D model obtained from SCAPE [2] on a certain number of range depth maps (e.g. 4) by optimizing an objective cost function relying on 3D data fitting as well as silhouette fitting. The main limitations of such a method are represented by the constraints imposed by the system, in the form of a specific pose (*T-pose*) that the subject has to assume throughout the sequence, and by the overall efficiency (more than one hour is reported to process one subject). Successively, in [8], simplified SCAPE shape models are estimated from two depth maps of the subject (one frontal, one from the back) in real time by optimizing a cost function composed of two terms, respectively taking into account point-to-point and point-to-plane fitting. Analogously to [6], this method carries out the modeling by relying on a small number of slightly overlapping frames, hence might suffer from the presence of noise in the data.

Differently, non-parametric shape modeling approaches have been also proposed. This is the case of [7], where a moving voxel grid is used for each body part to integrate together surface measurements obtained from a depth map sequence within a Truncated Signed Distance Function (TSDF) representation, thus allowing to build volumetric models for both the background and each piecewise body part. Due to the TSDF fusion, the output is not a parametric body model, but a piecewise smooth 3D mesh reconstruction of the body. Another non-parametric approach is the one proposed in [9], where real-time pose and shape estimation is obtained via a probabilistic approach based on a Gaussian Mixture Model (GMM). Also in this case, the input is represented by a sequence of RGB-D frames. A purely point-based technique is proposed by [10] for the people re-identification task; the authors use the Microsoft SDK to track and segment the body, and then the points are accumulated by transforming each limb to the standard A-pose.

In this work, we propose a framework aimed at efficient 3D parametric body modeling from noisy depth sequences acquired with consumer depth cameras. Conversely to [6,8], one main contribution of this work is to leverage on the temporal cue by explicitly tracking the 3D subject and estimating its 3D pose through a sequence of frames. This allows us to integrate the noisy body shape of the subject over many temporally correlated frames, effectively averaging out noise. The modeling procedure is carried out by minimization of an energy cost which includes, as a second contribution of our approach, additional set of cues with respect to those used in previous works, based on silhouette and 3D surface fitting, as well as skeleton similarity, PCA and smoothness. We show that the combinations of these terms induce a more robust estimation of the body model. Finally, and differently to [6,8], we propose to use MakeHuman models [4] due to their better integration with semantic information associated to each body part. In conjunction with this, a third contribution is a specific L1 minimization of the energy cost term associated with our modeling scheme, so as to induce sparsity in the semantic tags and automatically yield a compact semantic description of each acquired body.

In Sections 2 and 3 we illustrate the entire proposed pipeline, which tracks the body motion of the subject and estimates the pose of its body joints over time, then, from the 3D estimated body pose at each frame, it refines the parameters of a MakeHuman body model by cost function optimization. A graphical overview of the proposed pipeline is shown in Fig. 1. Our framework relies on open-source computer vision and full body modeling libraries such as the Point Cloud Library (PCL) and MakeHuman, and it is easily customizable for different tasks requiring different precision and performance due to the modularity of its nature. In our initial implementation it is able to process frames at a speed of 3 fps and does not require specific constraints on the pose of the subject. To demonstrate the effectiveness of our approach, in Section 4 we show some qualitative examples of body models estimated and tracked from real data acquired from consumer depth cameras, as well as measured accuracy of the estimated body model with respect to specific body parts. We also demonstrate the usefulness of compact semantic body tags associated to our estimated body models.

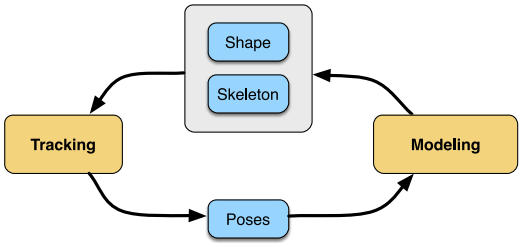

Fig. 1. Overview of the proposed tracking and modeling pipeline.

2. Proposed methodology

2.1. Data representation

In our system, the articulated human bodies are represented as quad and/or triangle meshes. The bodies can be articulated via the underlying skeleton. Skeletons are composed of multiple joints disposed in a tree hierarchy (see Fig. 2 for an example), based on which local node transformations are propagated: $\boldsymbol{T}^j_{abs} = \boldsymbol{T}^{parent(j)}_{abs}\boldsymbol{T}^j_{local}$, where $\boldsymbol{T}^j_{abs}$ is the world transformation of joint j, and $\boldsymbol{T}^j_{local}$ is its local transformation with respect to its parent node in the skeleton tree. Each joint influences a number of mesh vertices in its vicinity, as defined by the linear blend skinning model: $\boldsymbol{v}^{pose}_i = \sum_j w^j_i \boldsymbol{T}^{pose}_j * (\boldsymbol{T}^{rest}_j)^{-1} * \boldsymbol{v}^{rest}_i$, where each joint j in the skeleton has $\boldsymbol{T}^{rest}_j$ as the transformation corresponding to the rest pose (A or T-pose) and $\boldsymbol{T}^{pose}_j$ the transformation of the joint in the posed skeleton configuration, and w^j_i is the blend skinning weight of joint j over mesh vertex $\boldsymbol{v}_i$. The joints are modeled by their rest transformation (expressed using rotation matrix $\boldsymbol{R}^{rest}_j$ and translation vector $\boldsymbol{t}^{rest}_j$) and the pose rotation parametrized using Euler angles $\boldsymbol{\beta}$: $\boldsymbol{T}^{pose}_j = \boldsymbol{R}^{rest}_j * \boldsymbol{R}^x(\beta^x_j) * \boldsymbol{R}^y(\beta^y_j) * \boldsymbol{R}^z(\beta^z_j) + \boldsymbol{t}^{rest}_j$.

The skeleton model deforms the mesh based on the pose of the body, but does not take into account the deformations that define the identity of a person. To this end, we employ a global linear deformation model in which vertices $\boldsymbol{v}^{rest}_i$ are expressed as linear combinations $\boldsymbol{s}$ of bases stacked as columns into matrix $\boldsymbol{B}$: $\boldsymbol{v}^{rest}_i = \boldsymbol{m}_i + \boldsymbol{B}_i\boldsymbol{s}$. Previous work such as [1–3] uses statistical models derived from a set of registered scans of people. The framework we propose allows for such models to be used (they use the same linear system), but in our implementation we employed blendshapes exported from the popular human body modeling software MakeHuman [4]. These blendshapes correspond to the sliders in the MakeHuman application, that is used by numerous artists and game developers to generate realistic character assets. As a result, our model is based on a set of non-orthogonal bases

Please cite this article in press as: A.E. Ichim, F. Tombari, Semantic parametric body shape estimation from noisy depth sequences, Robotics and Autonomous Systems (2015), http://dx.doi.org/10.1016/j.robot.2015.09.029