MONOCULAR 3D HUMAN BODY RECONSTRUCTION TOWARDS DEPTH AUGMENTATION OF TELEVISION SEQUENCES

Angel Sappa Niki Aifanti Sotiris Malassiotis Michael G. Strintzis

Informatics & Telematics Institute
1st Km Thermi-Panorama Road
Thermi-Thessaloniki, Greece
{angel.sappa@iti.gr}

ABSTRACT

This paper addresses the reconstruction of 3D human body models from 2D video sequences. Considering that the input frames are already segmented, the proposed technique consists of three stages. These stages are independently applied over each segmented frame. Firstly, a skeleton of a human figure obtained from the segmented image is extracted by means of a fast algorithm based on a Voronoi diagram of the boundary points. Afterwards, the skeleton is labelled according to the human body parts (e.g. head, upper arm, lower arm, torso, etc). Secondly, an initial 3D model posture is estimated from the labelled skeleton. Finally, an iterative closest point (ICP) implementation is used to refine the initial model posture by maximizing the similarity between the projected 3D model and the segmented image. Experimental results with video sequences are presented.

1. INTRODUCTION

The use of 3D Human Body Models (HBM) is experiencing a continuous and accelerated growth. This is partly due to the increasing demand of more realistic representations from computer graphics and computer vision communities. Computer graphics pursue a realistic modelling of both the human body geometry and its associated motion. Applications such as: games, virtual reality or animations demand highly realistic models. On the contrary, computer vision seeks for an efficient and accurate model for applications such as: intelligent video surveillance, motion analysis, telepresence, human-machine interface.

The current work is focused on the generation of 3D HBMs within the computer vision field. In other words, the objective is to extract 3D HBMs from the information provided by a steady single camera. The target application is depth augmentation of common television sequences for future 3D-Displays [1].

Due to its widespread interest, there has been an abundance of work on the vision-based human body model reconstruction in recent years; however, in spite of all the effort it is still an open research area with a lot of work to be done. Recovering the shape and the pose of the human body, with only one point of view, is an ill-posed problem due to self-occlusions and motion ambiguities. In spite of the aforementioned difficulties, 3D human body reconstruction from 2D images has been addressed by many researchers. In the early eighty [2] proposes a model-based technique to compute a synthetic 3D model by using monocular images. This technique extracts a pairs of parallel lines in a segmented real image and matches them with the legs of a projected 3D model.

Other model-based approaches, using monocular perception systems, have been recently proposed by [3] and [4]. In [3] the problem of human arms modelling is addressed, while [4] tackles the full body modelling. It is based on maximizing the joint probability density function of the position and velocity of the body parts. The drawback of this approach is the requirement of markers (light bulbs strapped to the body joints) for facilitating the image analysis. In [5] a probabilistic approach is introduced for modelling 3D human motion for synthesis and tracking. The goal of this technique is to predict the 3D pose by using the observed motion history. Although the obtained results are quite promising, the aforementioned techniques are computationally expensive or need some kind of learning/training process. In [5], for example, a large data base with different body postures is required.

Unlike the previous approaches, in the current work 3D human body postures are estimated by using explicitly the information extracted from 2D video sequence instead of relying on probabilistic methods. Assuming a segmented image is given as an input, the proposed technique consists of three stages. Firstly, a human body skeleton of the given segmented image is extracted. Sec-

This work has been carried out as part of the ATTEST project (Advanced Three-dimensional TElevision System Technologies, IST-2001-34396). The first author has been supported by the EU under a Marie Curie Post-Doctoral Fellowship (HPMD-GH-01-00086-01).

0-7803-7750-8/03/$17.00 ©2003 IEEE

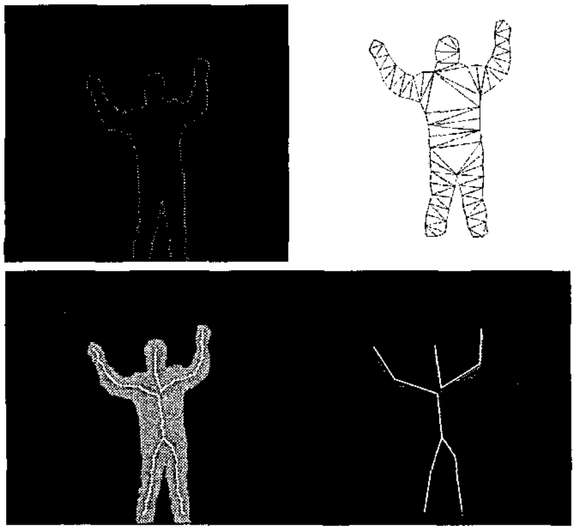

Figure 1. (*top-left*) Boundary points extracted from the segmented input image. (*top-right*) Constrained Delaunay triangulation of a subset of boundary points—only one out of twenty points were considered. (*bottom-left*) Skeleton computed from the Voronoi diagram. (*bottom-right*) Labelled body parts.

ondly, the skeleton posture is used to initialize a 3D model of a human body. Finally, a 2D projection of the initial 3D model is registered with the original image by using the iterative closest point algorithm.

The outline of this paper is as follows. Section 2 addresses the skeleton extraction stage. Section 3 introduces the used 3D model and describes the initialization of the posture parameters. Section 4 presents the approach used for registering the computed model with the original image. Finally, section 5 presents experimental results by using a video sequence. Conclusions and further improvements are given in section 6.

2. SKELETON EXTRACTION

Given a segmented image as an input (segmented images were computed by using the techniques proposed in [6] and [7]), the skeleton of the contained human figure is extracted. After implementing and comparing different options an algorithm based on a Voronoi diagram has been chosen. The proposed technique consists of the following steps. Firstly, the boundary points of the segmented image are extracted (see Fig. 1(*top-left*)) and triangulated by means of a constrained Delaunay algorithm [8]. The constraint is used to enforce a triangulation inside the polyline defined by linking consecutive boundary points. In order to reduce the CPU time, not all the boundary points are considered but only a subset (taking into account that the points are arranged in a list, in the current implementation only one out of twenty points were used). Fig. 1(*top-right*) illustrates the triangular mesh obtained by using a subset of the points presented in Fig. 1(*top-left*). Afterwards, the corresponding Voronoi diagram is extracted from the obtained triangular mesh and used to define the skeleton [9] (see Fig. 1(*bottom-left*)). Finally, the computed skeleton is labelled according to the different parts of the human body (i.e. legs, torso, arms and head). The implemented heuristic consists in labelling the lowest point as a member of one of the legs and going up until a bifurcation is reached. Every time a bifurcation is reached, a new member of the body part is labelled. Fig. 1(*bottom-right*) presents the skeleton extracted by using the Voronoi diagram and the corresponding set of segments defining the body parts. The torso and head were represented by a single segment, while the legs and arms were represented by two segments in order to preserve the human body anatomy.

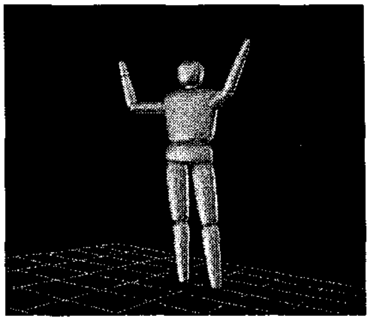

Figure 2. Illustration of a 22 DOF model built with superquadric

3. 3D BODY MODELLING

Superquadrics are a family of parametric shapes capable of modelling a large set of blob-like objects, such as spheres, cylinders, parallelepipeds and shapes in between [10]. A superquadric surface is given by the following parametric equation:

$$\mathbf{x}(\theta, \phi) = \begin{bmatrix} \alpha_1 \cos^{\varepsilon_1}(\theta) \cos^{\varepsilon_2}(\phi) \\ \alpha_2 \cos^{\varepsilon_1}(\theta) \sin^{\varepsilon_2}(\phi) \\ \alpha_3 \sin^{\varepsilon_1}(\theta) \end{bmatrix} \qquad (1)$$

where $-\pi/2 \leq \theta \leq \pi/2$, $-\pi \leq \phi < \pi$. The parameters α_1, α_2 and α_3 define the size of the superquadric along the *x*, *y* and *z* axis respectively, while ε_1 is the squareness parameter in the latitude plane and ε_2 is the squareness parameter in the longitudinal plane. Furthermore, superquadric shapes can be deformed with tapering, bending and cavities. In our model, the different body parts are represented with superquadrics tapered along the y-axis. The parametric equation is then written as:

$$\mathbf{x}'(\theta, \phi) = \begin{bmatrix} \left(\frac{t_1}{\alpha_2}x_2 + 1\right)x_1 \\ x_2 \\ \left(\frac{t_3}{\alpha_2}x_2 + 1\right)x_3 \end{bmatrix} \quad (2)$$

where $-1 \le t_1$, $t_3 \le 1$ are the tapering parameters and x_1, x_2 and x_3 are the elements of the vector in equation (1). The parameters α_1, α_2 and α_3 were defined in each body part according to anthropometric measurements. An example can be seen in Fig. 2.

The model used through the current implementation has 22 DOFs—four for each arm and leg, three for the torso and three for the head. We did not assign any DOFs to the palms or the feet for simplicity. The movements of the limbs are based on a hierarchical approach (the torso is considered the root) using Euler angles. The body posture is synthesized by concatenating the transformation matrices associated with the joints, starting from the root. Kinematic constraints have also been introduced in order to generate a realistic 3D model.

The unknown model parameters, which are the DOFs, are initially estimated using information extracted from the labelled skeleton. More precisely, the orientation and the length of the skeleton's labelled parts define the initial values of the DOFs, providing an initial human body posture. In case that some skeleton parts are missed, due to occlusions or segmentation failure, the algorithm uses temporal continuity. It is based on the posture computed in the previous frame updated with a displacement estimated by using the three precedent frames.

4. 3D POSTURE ESTIMATION

This section describes the technique used to compute a 3D human body posture from a 2D projection. This technique is based on the Iterative Closest Point—ICP—algorithm (originally proposed in [11]), which starts by using the initial human body posture as described in the previous section. The aim of the algorithm is to establish registration between the edge points—boundary points in Section 2—extracted from the segmentation and the projected occluding boundary points of the 3D model.

In order to obtain the projected occluding boundary points, the normals of each point on the superquadric surfaces are calculated. The dot product between the normal vectors and the viewing direction is then obtained. The sign of the dot product indicates whether this point lies on the front or the back surface. After eliminating the back-facing polygons, the occluding boundary vertices can be specified. The projection of the occluding boundary is subsequently obtained by projecting the boundary points onto the image plane.

The relative position and scale of the human figure is obtained from the segmented image. Consequently, the parameters computed by the ICP algorithm are the DOFs of the model $\omega = [\omega_1, \omega_2, \dots \omega_{22}]$. The objective is to estimate the model parameters that align the projected occluding boundary points with the extracted edge points. This is achieved by means of an iterative energy minimization technique. The computation of the distance between the projected occluding boundary points $B_i(\omega)$ $(i = 1, \dots, N)$ and the edge points E_j $(j = 1, \dots, M)$ is based on an approximate though efficient technique. According to this technique, the distance of an edge point E_j from an occluding boundary point may be approximated by $\|E_j - B_i(\omega)\|$, where $B_i(\omega)$ is the closest occluding boundary point, which was obtained by a KD-tree search algorithm. Thus, the estimation of the model parameters is achieved by the minimization of the following function:

$$D(\omega) = \sum_{j=1}^{M} w_j \|E_j - B_{E_j}(\omega)\|^2 \quad (3)$$

where w_j is a weighting factor and $B_{E_j}(\omega)$ denotes the occluding boundary point which was found to be the closest to E_j. The minimization process is then performed by means of the Levenberg-Marquard non-linear least squares technique.

The use of the weighting factors w_j aims at limiting the effect of outliers on the estimation process. The weighting factors are calculated according to the following:

$$w_j = \begin{cases} 0 & d_j > 3\sigma \\ \sigma/d_j & \sigma < d_j < 3\sigma \\ 1 & d_j < \sigma \end{cases} \quad (4)$$

where d_j is the residual fitting error for point E_j and σ the error variance.

Since the number of DOFs is significant, if the algorithm is applied to the whole body at the same time, it may be easily trapped in a local minima. For this reason, several groups of body parts are separately processed. At the beginning, head's and torso's abduction parameters are estimated. In the next step, the edges close to the head's and torso's boundary projection are excluded. Thus, only the edges associated with the arms and legs are left. We use prior knowledge of the human body geometry to assign the remaining edge pixels to arms or legs. Then, the ICP algorithm is applied independently for each labelled group of points (e.g. left arm, right arm).

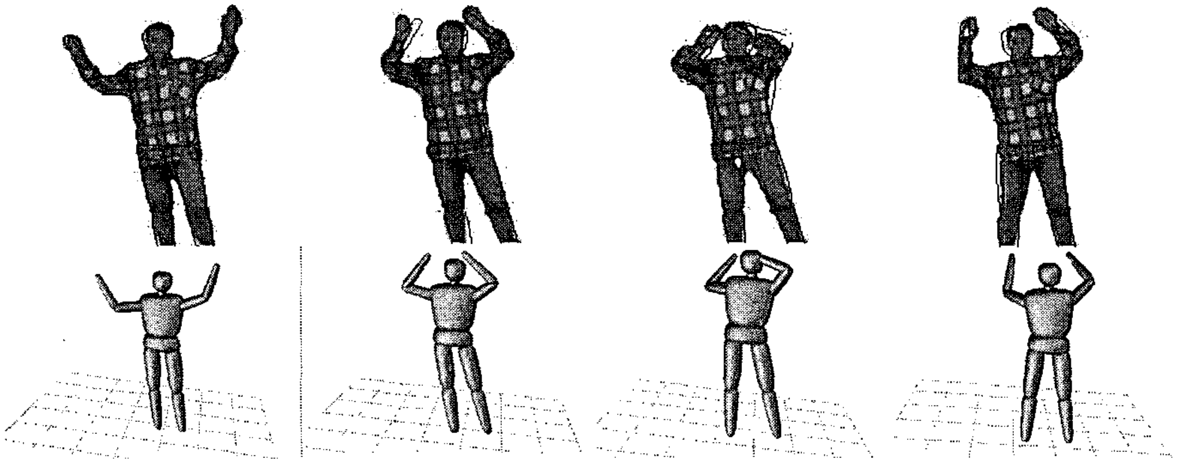

Figure 3. (*top*) Original segmented images used as an input for the algorithm, in black the projected occluding boundary points of the 3D model are represented. (*bottom*) The corresponding 3D models generated by using the proposed technique. Minor misalignments are due to posture ambiguities.

5. EXPERIMENTAL RESULTS

The proposed technique has been tested with several segmented video sequences generated using [6] and [7]. The average CPU (Pentium III, 1GHz processor) time to compute and label the skeleton of the human body was 0.1 sec. per frame. This labelled skeleton is used to estimate a coarse 3D posture which is improved by the registration of the projected occluding boundary points with the edge points extracted from the segmented image. The average CPU time to register the different projected body parts was 2 sec. per frame. Fig. 3(*top*) presents a set of four frames used as an input for the proposed technique. The final projected occluding boundary points, computed with ICP, are represented in black. The obtained 3D models are illustrated in Fig. 3(*bottom*).

6. CONCLUSIONS AND FURTHER WORK

A new technique to generate 3D models of human bodies from 2D video sequences has been presented. It is based on the processing of single frames, avoiding expensive probabilistic approaches and learning problems. The proposed technique consists of three stages. Firstly, a skeleton of a human body figure is extracted and labelled. Next, the posture of a 3D model is estimated by using the aforementioned skeleton, and finally a registration algorithm is used to tune the parameters of the model (DOFs).

Further work will include the prediction of human body posture using consecutive frames. The initial human body posture estimation could be improved and the CPU time reduced, by using the posture of a previous frame and by studying the history of the articulations' movement.

7. REFERENCES

[1] A. Redert, et al., ATTEST- Advanced Three-dimensional TElevision System Technologies, *3DPVT'02*, Padova, Italy, June, 2002.

[2] D. Hogg, Model-Based Vision: A Program to See a Walking Person, *Image and Vision Computing*, 1(1), February 1983.

[3] L. Goncalves, E. Di Bernardo, E. Ursella and P. Perona, Monocular tracking of the human arm in 3D, *IEEE Int. Conf. on Computer Vision*, 1995.

[4] Y. Song, L. Goncalves, E. Di Bernardo and P. Perona., Monocular Perception of Biological Motion—Detection and Labeling, *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Fort Collings, USA, 1999.

[5] H. Sidenbladh, M. Black and L. Sigal, Implicit Probabilistic Models of Human Motion for Synthesis and Tracking, *European Conf. on Computer Vision*, Copenhagen, Denmark 2002.

[6] F. Ernst, P. Wilinski and K. Van Overveld, Dense Structure-from-Motion: An Approach Based on Segment Matching, *ECCV 2001*.

[7] S. Jabri, Z. Duric, H. Wechsler and A. Rosenfeld, Detection and Location of People in Video Images Using Adaptive Fusion of Color and Edge Information, *15th. Int. Conf. on Pattern Recognition 2000*.

[8] O. Faugeras, *Three-Dimensional Computer Vision. The MIT Press*, 1993.

[9] Sing-Tze Bow, Pattern Recognition and Image Preprocessing, *Marcel Dekker, Inc*, 1992.

[10] F. Solina and R. Bajcsy, Recovery of Parametric Models from Range Images: The Case for Superquadrics with Global Deformations, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 2, February 1990.

[11] P. Besl and N. McKay, A Method for Registration of 3-D Shapes, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 14, no. 2, February 1992.