3D GAIT ESTIMATION FROM MONOSCOPIC VIDEO

Angel D. Sappa† Niki Aifanti‡ Sotiris Malassiotis‡ Michael G. Strintzis‡

Computer Vision Center†
Edifici O, Campus UAB
08193 Bellaterra - Barcelona, Spain
angel.sappa@cvc.uab.es

Informatics & Telematics Institute‡
1st Km Thermi-Panorama Road
Thermi-Thessaloniki, Greece
{naif, malasiot}@iti.gr strintzi@eng.auth.gr

ABSTRACT

This paper presents a new approach for 3D gait estimation from monocular image sequences, using both a kinematics and a walking motion models as sources of prior knowledge. The proposed technique consists of two major stages. Firstly, the motion trajectory and the pedestrian's footprints are detected throughout the segmented video sequence. Secondly, as the 3D human model, driven by the prior motion model, walks over this trajectory, the joints' angles are locally adjusted to the pedestrian's walking style. This tuning process is performed once per walking cycle and not per frame, saving considerable CPU time. In addition, local tuning allows handling displacements at different speeds or directions. The target application is the augmentation of 2D television sequences with depth information that may be used in future 3D-TV systems.

1. INTRODUCTION

3D-TV opens a new and attractive field of applications, from more realistic movies to interactive environments. However, in order to fully exploit these new 3D-TV systems all the existing 2D video material should be converted into 3D. Theoretically, it is not possible to completely recover 3D information from 2D video sequences when no other extra information is given or can be estimated. Since television sequences are populated with objects with known structure and motion such as humans, cars, etc, prior knowledge would arguably aid the recovery of the scene. Prior knowledge in the form of kinematics constraints (average size of an articulated structure, degrees of freedom (DOFs) for each articulation), or motion dynamics (physical laws ruling the objects' movements), is a commonplace solution to handle the aforementioned problem.

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 Ramón y Cajal Program*.

In real world conditions, 3D human motion modeling using monocular image sequences constitutes a complex and challenging problem, which involves difficulties such as: self-occlusions, depth ambiguities of the body parts, walking direction estimation, erroneous background segmentation, etc. (see [1] for more details).

In order to avoid some of the aforementioned problems, 3D human walking modeling has been usually tackled by making simplifying assumptions (e.g. [2], [3], [4]) or by imposing constraints on the motion (e.g. walking in a plane orthogonal to the camera with a constant speed [5], [6]). Moreover, in order to register the projection of the computed 3D model with the given image, several features have been combined [7], such as skin color, edges, skeleton, optical flow, etc.

The proposed approach consists in dividing the given walking sequence into separate walking cycles, which are independently processed. An explicit motion model, defined by a set of motion curves driving each articulation, is used as initial approximation of the motion. These curves, obtained from anthropometric studies [6], are individually tuned by the algorithm according to the walking attitude of each pedestrian (Fig. 1). The main advantage comparing with previous approaches is that matching between the projection of the 3D model and the image features is performed once per walking cycle and not per frame. A brief description of the 3D body modeling, together with depth estimation is given below. The proposed technique is presented in section 4 and section 5. Section 6 shows experimental results and finally conclusions and future work are introduced in section 7.

2. 3D BODY MODELING

In the current work, similarly than in [8], an articulated structure defined by 16 links (superquadrics) and 22 DOF, 4 for each arm and leg and 6 for the torso (3 for orientation and 3 for position) was chosen (Fig. 2(*left*)). However, in order to reduce the complexity, it was assumed that while walking, the legs' and arms' movements are contained in parallel planes and that the body's orientation is always or-

0-7803-8554-3/04/$20.00 ©2004 IEEE.

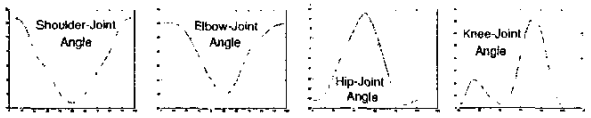

Figure 1. Motion curves computed according to [6].

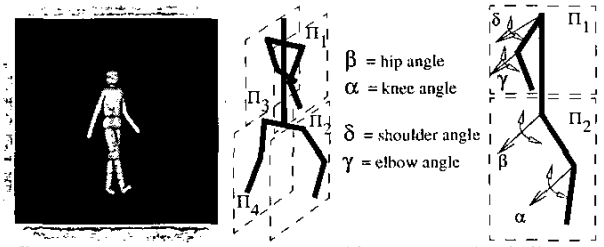

Figure 2. (*left*) Illustration of a 22 DOF model built with superquadrics. (*right*) Simplified articulated structure (12 DOFs).

thogonal to the floor. Hence, the final model is defined by 2 DOF for each arm and leg and 4 for the torso (3 for position plus 1 for orientation). The movements of the limbs are based on a hierarchical approach using Euler angles. The body posture is synthesized by concatenating the transformation matrices associated with the joints, starting from the torso.

3. DEPTH ESTIMATION

The transformation matrices defining the 3D orientation (R_c) and position (T_c) of the camera coordinate system have been assumed known. The mapping between the world-coordinate reference frame and the camera-coordinate reference frame is given by:

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = R_c \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} + T_c \qquad (1)$$

where a 3D point in the camera-coordinate reference frame is represented by $P_c(x_c, y_c, z_c)$, while in the world-coordinate reference frame by $P_w(x_w, y_w, z_w)$. The pedestrian's height is assumed to be equal to an average human height. Finally, the perspective projection of a point to the image plane is defined as:

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} = \frac{f_c}{z_c} \begin{bmatrix} x_c \\ y_c \end{bmatrix} - \begin{bmatrix} col/2 \\ row/2 \end{bmatrix} \qquad (2)$$

where f_c is the focal length.

Using these equations, the 3D world coordinates of the center of the segmented image are computed (Fig. 3 (*top-right*)) for every input frame. Fig. 3(*bottom-left*) illustrates the center point's path resulting from the estimated depth values. The result is not smooth enough since variations in the segmented figure affect depth estimation values. In order to generate a smoother path, the estimated depth values are filtered by using a spline curve. These depth values will be used to compute the footprints' position. In addition, since the person's body is oriented towards the direction of movement, the gradient of the path is used to determine the model's orientation.

4. FOOTPRINT DETECTION

We may safely assume that the center of gravity of a walking person is continuously in movement. However, throughout walking displacements there is at least a foot with null velocity (pivot foot) and one instant per walking cycle in which both feet are in contact with the floor (both with null velocity). The latter happens when the pedestrian changes from one pivot foot to the other. Frames containing these configurations will be called *anchor frames* and can be easily detected by extracting *static points* through the given video sequence. A point is considered as a static point $spt^F_{(i,j)}$ in frame F, if it remains a boundary point $bp^F_{(i,j)}$ in at least three consecutive frames—value computed experimentally $stp^F_{(i,j)} \Rightarrow \left(bp^{F-1}_{(i,j)}, bp^F_{(i,j)}, bp^{F+1}_{(i,j)}\right)$. Fig. 4(*top*) shows an illustration of static points (black points) detected after processing consecutive frames.

Static points defining a single footprint do not belong all to the same frame but to a sequence of several consecutive frames, in which the foot was in contact with the floor. Considering that foot's sole is not a rigid surface, a single foot generates a set of static points during the time in which different parts of it are in contact with the floor. Hence, points belonging to the same footprint should be clustered.

The implemented clustering technique is similar to a region growing using both spatial and temporal information. The label associated with a static point (footprint index), computed in a frame F, $stp^F_{(i,j)}$, is propagated to a neighbor static point $stp^N_{(i\pm r, j\pm c)}$, $r \in \{0,1\}, c \in \{0,1\}$ if this spatial neighbor point has been computed in a temporal neighborhood of a maximum of Γ frames from the considered frame F (i.e. $|F-N| \leq \Gamma$). In the current implementation Γ has been experimentally set to six. The proposed footprint labeling technique starts labeling those static points contained in the first frame and then propagates these labels to consecutive frames. At each stage, new static points are labeled either with new index values or with footprint indexes propagated from previous frames. An example of the results obtained after applying this technique is presented in Fig. 4(*bottom-right*). Notice that using temporal information, spatial ambiguities generated by trajectories parallel to the camera direction or by path self-crossing are easily avoided Fig. 4(*bottom-left*).

5. MOTION MODEL TUNING

The result from the previous clustering stage is a list of footprints with a corresponding set of points defining each one. In addition to the set of points, each footprint has an

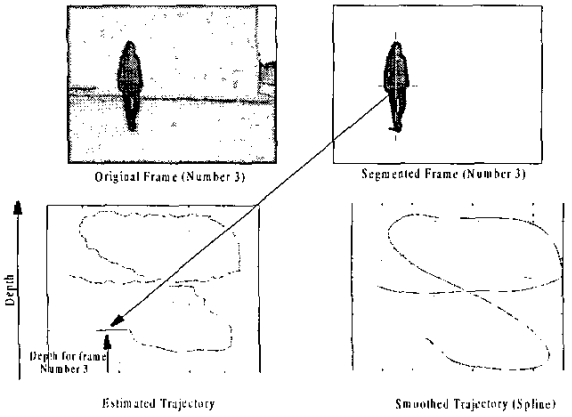

Fig. 3. Illustration of the center point's path computed from a video sequence of 540 frames

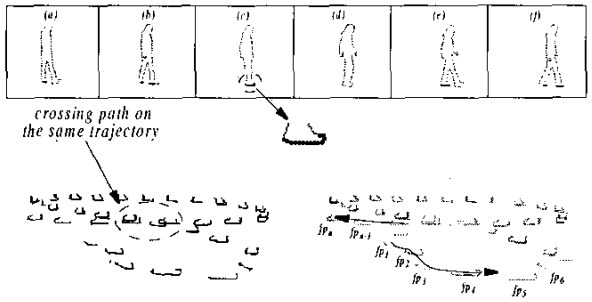

Fig. 4. (*top*) Static points detected for each frame. (*bottom-left*) All the computed static points represented in a single image. (*bottom-right*) Footprints labelled by the spatial and temporal clustering of the static points.

associated list containing the numbers of the frames where it was detected, $fp_q\{F_a, F_b, \ldots, F_f\}$. Intermediate frames, where footprints have not been detected at all (e.g. Fig. 4(*top-d*)), are also included in that list for continuity.

The objective is to find those frames where both feet are in contact with the floor (Fig. 4(*top*) (*a*)-(*b*) and (*e*)-(*f*)). This happens in every half walking cycle in several consecutive frames (Fig. 4(*top*) shows one out of four of the original frames). The frame lying in the center of those consecutive frames is defined as anchor frame. At every anchor frame, the articulated human body structure reaches a posture with maximum hip angles. In the current implementation, hip angles are defined by the legs and the vertical axis containing the hip joints. This maximum value is used to compute a scale factor κ_i, which adjusts the hip motion model (Fig. 1) to the pedestrian's walking. This local tuning, within a half walking cycle, is illustrated in Fig. 5, where the computed scale factor is actually used for a quarter of the walking cycle, from the current anchor frame until halfway to the next one. During the next quarter of the walking cycle an updated scale factor, calculated from the maximum hip angle in the

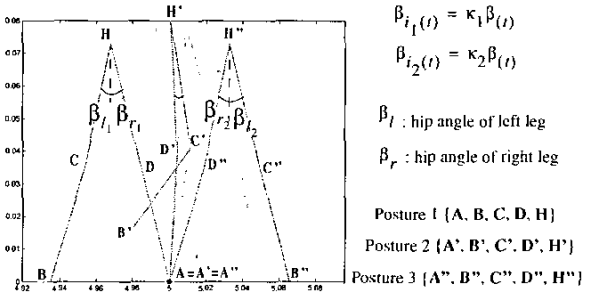

Figure 5. Half walking cycle executed by using scale factors (κ_1, κ_2) over the hip motion curve presented in Fig. 1. Spatial position of points (D, H, C and B) are computed by using angles from the motion curves and trigonometric relationships.

next anchor frame, is used. A 2D articulated structure is depicted in Fig. 5 in order to make understanding easier. However the tuning process is carried out in 3D space. Depth values of footprints (**A**, **B**) or (**A"**, **B"**) are computed as an average of the center point's depth in the frames where these footprints appear. The number of frames in between two anchor frames defines the sampling rate of the motion curves.

The differences in walking between people implies that all the motion curves should be modified by using an appropriate scale factor for each one. In order to estimate these factors an error measurement (registration quality index: *RQI*) is introduced. The proposed *RQI* measures the quality of the matching between the projected 3D model and the corresponding walking human figure. It is defined as: $RQI = overlappedArea/totalArea$, where total area consists of the surface of the projected 3D model plus the surface of the walking human figure less the overlapped area, while the overlapped area is defined by the overlap of these two surfaces. Firstly, the algorithm computes the knee scale factor that maximizes the *RQI* values. In every iteration, an average *RQI* is calculated for all the sequence. In order to speed up the process the number of frames was subsampled. Afterwards, the elbow and shoulder scale factors are estimated similarly.

6. EXPERIMENTAL RESULTS

The proposed technique has been tested with different outdoor video sequences. The video sequence used as an illustration throughout this work consists of 540 frames of *240x320* pixels each, which have been segmented using the technique presented in [9]. In this sequence, the pedestrian follows the trajectory depicted in Fig. 3 and her walking speed is variable. Results of the proposed algorithm are presented in Fig. 6. Notice that the model is able to follow the pedestrian independently of the walking direction, in particular when the pedestrian changes direction (see last two frames in Fig. 6(*top*)). The corresponding 3D models

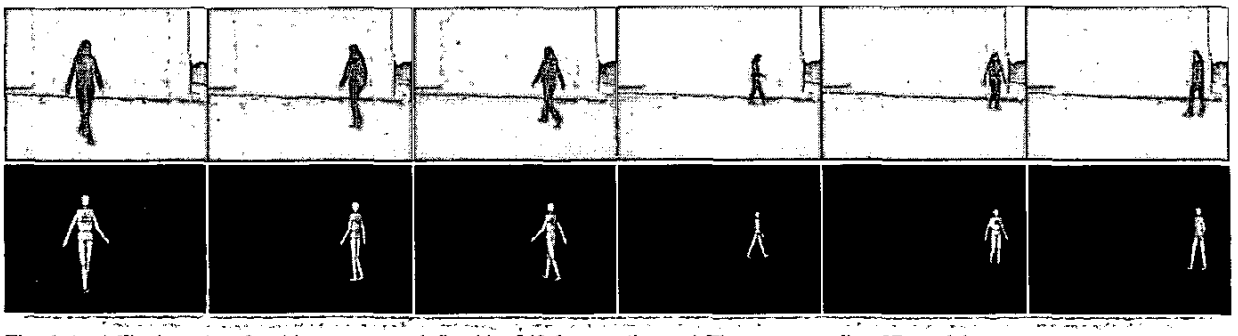

Fig. 6. (*top*) Final results of a video sequence defined by 540 frames. (*bottom*) The corresponding 3D models.

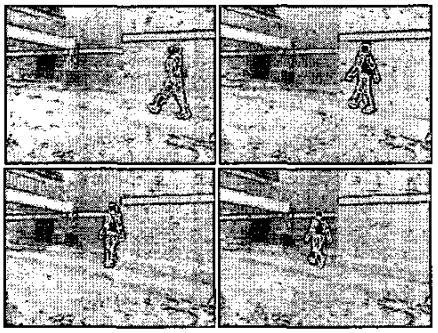

Fig. 7. Final result of a short video sequence (in white the projected model).

are presented in Fig. 6(*bottom*). The *RQI* for the total video sequences has been 0.59.

Fig. 7 shows some results of a video sequence defined by 70 frames of *240x320* pixels each. The segmented input frames have been provided by the authors of [10]. In this case the *RQI* value was 0.5. The main reason for this poor performance is that the pedestrian is carrying a backpack and he is not wearing so tight clothes. The average CPU time for the different stages of the proposed algorithm was 0.37 seconds per frame. This time includes depth estimation, static point and footprint detection and finally local tuning of the motion model parameters.

7. CONCLUSIONS

A new approach towards human motion modeling and recovery has been presented. It exploits prior knowledge regarding a person's movement as well as human body kinematics constraints. At this paper only walking has been modeled. No constraints about the direction or speed are imposed. Experimental results with different pedestrian, speeds and walking directions demonstrate robustness of the algorithm with no compromise in computational complexity. Further work will include the tuning of not only motion model's parameters but also geometric model's parameters in order to find a better fitting. In this way, external objects attached to the body (like a handbag or backpack) could be added to the body and considered as a part of it.

8. REFERENCES

[1] A. Sappa, N. Aifanti, N. Grammalidis and S. Malassiotis, "Advances in Vision-Based Human Body Modeling", *Chapter book in: 3D Modeling and Animation: Synthesis and Analysis Techniques for the Human Body*, N. Sarris and M.G. Strintzis (Eds.), Idea-Group Inc., 2004 (*in press*).

[2] D. Gavrila and L. Davis, "3-D Model-Based Tracking of Humans in Action: a Multi-View Approach", *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, San Francisco, USA, 1996.

[3] C. Barron and I. Kakadiaris, "Estimating Anthropometry and Pose from a Single Camera", *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Hilton Head Island, USA, 2000.

[4] D. Metaxas and D. Terzopoulos, "Shape and Nonrigid Motion Estimation through Physics-Based Synthesis", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 6, June 1993, pp. 580-591.

[5] S. Wachter and H. Nagel, "Tracking of Persons in Monocular Image Sequences", *IEEE Computer Vision and Image Understanding*, Vol. 74, No. 3, pp. 174-192, 1999.

[6] K. Rohr, "Human Movement Analysis Based on Explicit Motion Models", *Chapter 8 in Motion-Based Recognition*, M. Shah and R. Jain (Eds.), Kluwer Academic Publisher, Dordrecht Boston 1997, pp. 171-198.

[7] H. Sidenbladh, M. Black and D. Fleet, "Stochastic Tracking of 3D Human Figures Using 2D Image Motion", *European Conference on Computer Vision*, Dublin, Ireland, 2000.

[8] A. Sappa, N. Aifanti, S. Malassiotis and M. Strintzis, "Monocular 3D Human Body Reconstruction Towards Depth Augmentation of Television Sequences", *IEEE Int. Conf. on Image Processing*, Barcelona, Spain, Sep. 2003.

[9] C. Kim and J. Hwang, "Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications", *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 12, No. 2, Feb. 2002, pp. 122-129.

[10] 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*, Barcelona, Spain, Sep. 2000.