3D Human Walking Modeling

Angel D. Sappa[1], Niki Aifanti[2], Sotiris Malassiotis[2], and Michael G. Strintzis[2]

[1] Computer Vision Center, Edifici O, Campus UAB,
08193 Bellaterra - Barcelona, Spain
angel.sappa@cvc.uab.es
[2] Informatics & Telematics Institute, 1st Km Thermi-Panorama Road,
57001 Thermi-Thessaloniki, Greece
{naif,malasiot}@iti.gr
strintzi@eng.auth.gr

Abstract. This paper presents a new approach for 3D human walking modeling from monocular image sequences. An efficient feature point selection and tracking approach has been used to compute feature points' trajectories. Peaks and valleys of these trajectories are used to detect key frames—frames where both legs are in contact with the floor. These frames, together with prior knowledge of body kinematics and a motion model, are the basis for the 3D reconstruction of human walking. The legs' configuration at each key frame contributes to tune the amplitude of the motion model. Differently than previous approaches, this tuning process is not performed at every frame, reducing CPU time. In addition, the movement's frequency is defined by the elapsed time between two consecutive key frames, which allows handling walking displacement at different speed. Experimental results with different video sequences are presented.[1]

1 Introduction

Walking motion is the most common type of human locomotion. 3D motion models are required for applications such as: intelligent video surveillance, pedestrian detection for traffic applications, gait recognition, medical diagnosis and rehabilitation, human-machine interface [1]. Due to the wide interest it has generated, 3D human motion modeling is one of the most active areas within the computer vision community.

Vision-based human motion modeling approaches usually combine several computer vision processing techniques (e.g. video sequence segmentation, object tracking, motion prediction, 3D object representation, model fitting, etc.). Different techniques have been proposed to find a model that matches a walking displacement. These approaches can be broadly classified into monocular or multi camera approaches.

[1] 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*.

F.J. Perales and B.A. Draper (Eds.): AMDO 2004, LNCS 3179, pp. 111-122, 2004.

A multicamera system was proposed by [2]. It consists of a stereoscopic technique able to cope not only with self-occlusions but also with fast movements and poor quality images. This approach incorporates physical forces to each rigid part of a kinematics 3D human body model consisting of truncated cones. These forces guide each 3D model's part towards a convergence with the body posture in the image. The model's projections are compared with the silhouettes extracted from the image by means of a novel approach, which combines the Maxwell's demons algorithm with the classical ICP algorithm. Although stereoscopic systems provide us with more information for the scanned scenes, 3D human motion systems with only one camera-view available is the most frequent case.

Motion modeling using monocular image sequences constitutes a complex and challenging problem. Similarly to approach [2], but in a 2D space and assuming a segmented video sequence is given as an input, [3] proposes a system that fits a projected body model with the contour of a segmented image. This boundary matching technique consists of an error minimization between the pose of the projected model and the pose of the real body—all in a 2D space. The main disadvantage of this technique is that it finds the correspondence between the projected body parts and the silhouette contour, before starting the matching approach. This means that it looks for the point of the silhouette contour that corresponds to a given projected body part, assuming that the model posture is not initialized. This problem is still more difficult to handle in those frames where self-occlusions appear or edges cannot be properly computed.

Differently than the previous approaches, the aspect ratio of the bounding box of the moving silhouette has been used in [4]. This approach is able to cope with both lateral and frontal views. In this case the contour is studied as a whole and body parts do not need to be detected. The aspect ratio is used to encode the pedestrian's walking way. However, although shapes are one of the most important semantic attributes of an image, problems appear in those cases where the pedestrian wears clothes not so tight or carries objects such as a suitcase, handbag or backpack. Carried objects distort the human body silhouette and therefore the aspect ratio of the corresponding bounding box.

In order to be able to tackle some of the problems mentioned above, some authors propose simplifying assumptions. In [5] for example, tight-fitting clothes with sleeves of contrasting colors have been used. Thus, the right arm is depicted with a different color than the left arm and edge detection is simplified especially in case of self-occlusions. [6] proposes an approach where the user selects some points on the image, which mainly correspond to the joints of the human body. Points of interest are also marked in [7] using infrared diode markers. The authors present a physics-based framework for 3D shape and non-rigid motion estimation based on the use of a non-contact 3D motion digitizing system. Unfortunately, when a 2D video sequence is given, it is not likely to affect its content afterwards in such a way. Therefore, the usefulness of these approaches is restricted to cases where access in making the sequence is possible.

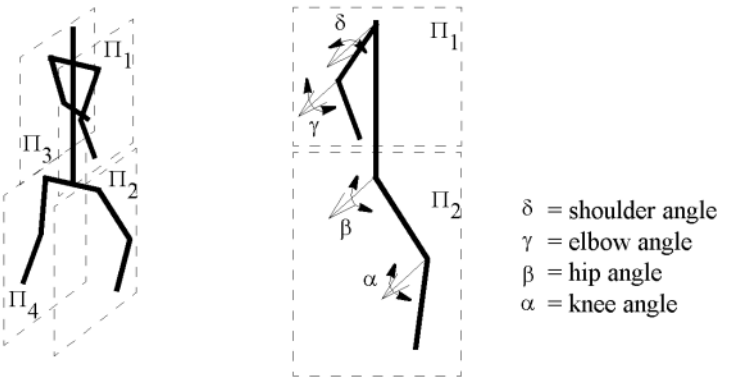

Fig. 1. Simplified articulated structure defined by 12 DOFs, arms and legs rotations are contained in planes parallel to the walking's direction

Recently, a novel approach based on feature point selection and tracking was proposed in [8]. This approach is closely related to the technique proposed in this work. However, a main difference is that in [8] feature points are triangulated together and similarity between triangles and body parts is studied, while in the current work, feature point's trajectories are plotted on the image plane and used to detect key frames. Robustness in feature point based approaches is considerably better than in those techniques based on silhouette, since silhouette does not only depend on walking style or direction but also on other external factors such as those mentioned above. Walking attitude is easier captured by studying the spatio-temporal motion of feature points.

In this paper a new approach to cope with the problem of human walking modeling is presented. The main idea is to search for a particular kinematics configuration throughout the frames of the given video sequence, and then to use the extracted information in order to tune a general motion model. Walking displacement involves the synchronized movements of each body part—the same is valid for any cyclic human body displacement (e.g., running, jogging). In this work, a set of curves, obtained from anthropometric studies [9], is used as a coarse walking model. These curves need to be individually tuned according to the walking attitude of each pedestrian. This tuning process is based on the observation that although each person walks with a particular style, there is an instant in which every human body structure achieves the same configuration. This instant happens when both legs are in contact with the floor. Then, the open articulated structure becomes a closed structure. This closed structure is a rich source of information useful to tune most of the motion model's parameters. The outline of this work is as follows. The proposed technique is described in section 2. Experimental results using different video sequences are presented in section 3. Conclusions and further improvements are given in section 4.

2 The Proposed Approach

The proposed approach consists of two stages. In the first stage feature points are selected and tracked throughout the whole video sequence in order to find key frames' positions. In the second stage a generic motion model is locally tuned by using kinematics information extracted from the key frames. The main advantage comparing with previous approaches is that matching between the projection of the 3D model and the body silhouette image features is not performed at every frame (e.g., hip tuning is performed twice per walking cycle). The algorithm's stages are fully described below together with a brief description of the 3D representation used to model the human body.

2.1 Body Modeling

Similarly to [10], an articulated structure defined by 16 links is used. However, in order to reduce the complexity, motion model is simplified and consists of 12 DOF. This simplification assumes that in walking, legs' and arms' movements are contained in parallel planes (see illustration in Fig. 1). In addition, the body orientation is always orthogonal to the floor. Thus the orientation is described by only one DOF (this degree of freedom allows us to model trajectories that are not orthogonal to the camera direction, see Fig. 11). Hence, the final model is defined by two DOF for each arm and leg and four for the torso (three for the position plus one for the orientation).

The articulated structure is represented by 16 superquadrics (Fig. 2); a complete mathematical description of this volumetric model can be found in [10].

2.2 Feature Point Selection and Tracking

Feature point selection and tracking approaches were chosen because they allow capturing the motion's parameters by using as prior knowledge the kinematics of the body structure. In addition, point-based approaches seem to be more robust in comparison with silhouette based approaches. Next, a brief description of the techniques used is given.

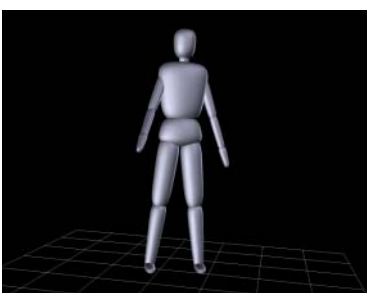

Fig. 2. Illustration of a 22 DOF model built with superquadric

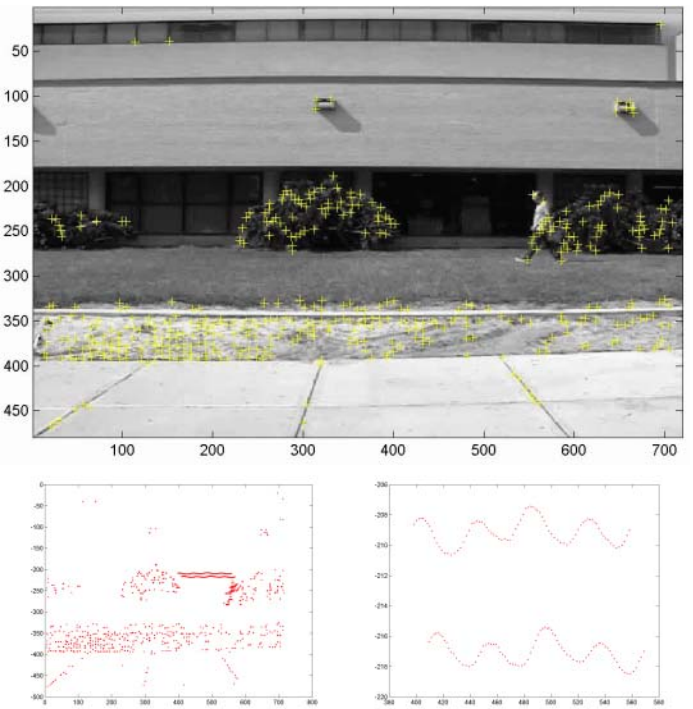

Fig. 3. (*top*) Feature points from the first frame of the video sequence used in [13]. (*bottom-left*) Feature points' trajectories. (*bottom-right*) Feature points' trajectories after removing static points

Feature Point Selection. In this work, the feature points are used to capture human body movements and are selected by using a corner detector algorithm. Let *I(x,y)* be the first frame of a given video sequence. Then, a pixel *(x,y)* is a corner feature if at all pixels in a window W_S around *(x,y)* the smallest singular value of G is bigger than a predefined σ; in the current implementation W_S was set to *5x5* and $\sigma = 0.05$. G is defined as:

$$G = \begin{bmatrix} \sum I_x^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^2 \end{bmatrix}$$

and (I_x, I_y) are the gradients obtained by convolving the image I with the derivatives of a pair of Gaussian filters. More details about corner detection can be found in [11]. Assuming that at the beginning there is no information about the pedestrian's position in the given frame, and in order to enforce a homogeneous feature sampling, input frames are partitioned into 4 regular tiles (*2x2* regions of *240x360* pixels each in the illustration presented in Fig. 3).

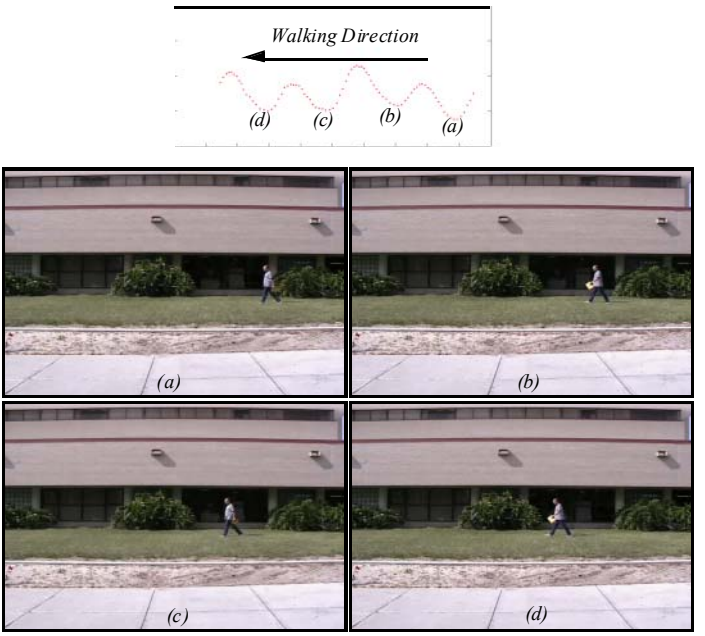

Fig. 4. (*top*) A single feature point's trajectory. (*middle* and *bottom*) Key frames associated with the valleys of a feature point's trajectory

Feature Point Tracking. After selecting a set of feature points and setting a tracking window W_T (*3x3* in the current implementation) an iterative feature tracking algorithm has been used [11]. Assuming a small interframe motion, feature points are tracked by minimizing the sum of squared differences between two consecutive frames.

Points, lying on the head or shoulders, are the best candidates to satisfy the aforementioned assumption. Most of the other points (e.g. points over the legs, arms or hands, are missed after a couple of frames). Fig. 3(*top*) illustrates feature points detected in the first frame of the video sequence used in [13]. Fig. 3(*bottom-left*) depicts the trajectories of the feature points when all frames are considered. On the contrary, Fig. 3(*bottom-right*) shows the trajectories after removing static points. In the current implementation we only use one feature point's trajectory. Further improvements could be to merge feature points' trajectories in order to generate a more robust approach.

2.3 Motion Model Tuning

The outcome of the previous stage is the trajectory of a feature point (Fig. 4(*top*)) consisting of peaks and valleys. Firstly, the first-order derivative of the curve is computed to find peaks' and valleys' positions by seeking the positive-to-negative zero-crossing points. Peaks correspond to those frames where the pedestrian reaches the maximum height, which happens in that moment of the half walking cycle when the hip angles are minimum. On the contrary, the valleys correspond to those frames

where the two legs are in contact with the floor and then, the hip angles are maximum. So, the valleys are used to find key frames, while the peaks are used for footprint detection. The frames corresponding to each valley of Fig. 4(*top*) are presented in Fig. 4(*middle*) and (*bottom*). An interesting point of the proposed approach is that in this video sequence, in spite of the fact that the pedestrian is carrying a folder, key frames are correctly detected; as a result, legs and torso correspondence are easily computed. On the contrary, an approach based on shape, such as [3], will face difficulties, since it will try to minimize the matching error based on the whole shape (including folder).

After detecting key frames corresponding to the valleys of a trajectory, it is necessary to define also the footprints of the pedestrian throughout the sequence. In order to achieve this, body silhouettes were computed by an image segmentation algorithm [12]. Additionally, other segmented video sequences were provided by [14]. Footprint positions are computed as follow.

Throughout a walking displacement sequence, there is always, at least, one foot in contact with the ground, with null velocity (pivot foot). In addition, there is one instant per walking cycle in which both feet are in contact with the floor (both with null velocity). The foot that is in contact with the floor can be easily detected by extracting its defining *static points*. A point is considered as a static point $spt^{F}_{(i,j)}$ in frame F, if it remains as a boundary point $bp^{F}_{(i,j)}$ (silhouette point, Fig. 5(*left*)) in at least three consecutive frames—value computed experimentally $stp^{F}_{(i,j)} \Rightarrow (bp^{F-1}_{(i,j)}, bp^{F}_{(i,j)}, bp^{F+1}_{(i,j)})$.

The result of the previous stage is a set of static points distributed along the pedestrian's path. Now, the problem is to cluster those points belonging to the same foot. Static points defining a single footprint are easily clustered by studying the peaks' positions in the feature point's trajectory. All those static points in a neighborhood of $F \pm 3$ from the frame corresponding to a peak position (F) will be clustered together and will define the same footprint (fp_i). Fig. 5(*right*) shows the footprints detected after processing the video sequence of Fig. 4.

Fig. 5. (*left*) Five consecutive frames used to detect static points. (*right*) Footprints computed after clustering static points generated by the same foot (picks in a feature point's trajectory (Fig. 4(*top*))

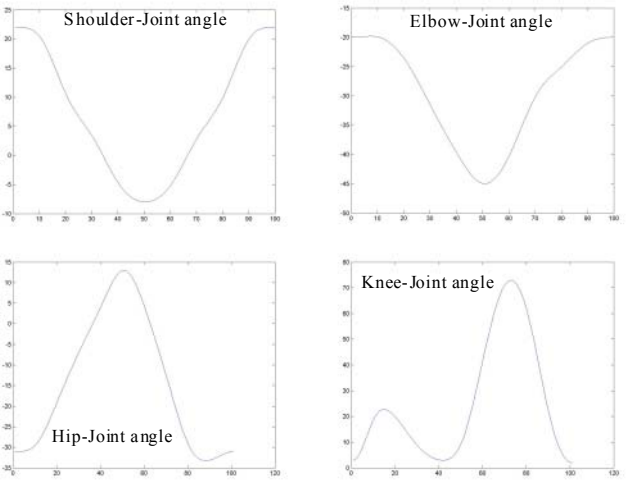

Fig. 6. Motion curves of the joints at the shoulder, elbow, hip and knee (computed from [9])

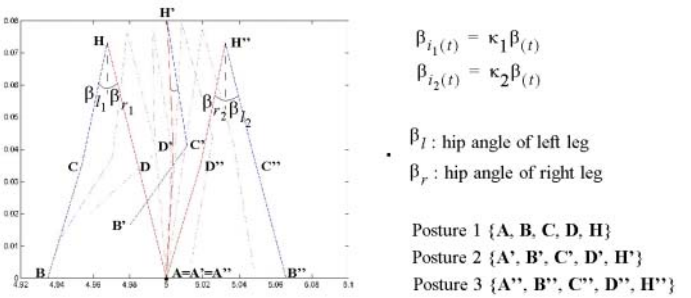

Fig. 7. Half walking cycle executed by using scale factors (κ_1, κ_2) over the hip motion curve presented in Fig. 6 (knee motion curve is not tuned at this stage). Spatial positions of points (**D**, **H**, **C** and **B**) are computed by using angles from the motion curves and trigonometric relationships

As it was introduced above, key frames are defined as those frames where both feet are in contact with the floor. At every key 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, together with the maximum value of the hip motion model (Fig. 6) are used to compute a scale factor κ. This factor is utilized to adjust the hip motion model to the current pedestrian's walking. Actually, it is used for half the walking cycle, which does not start from the current key frame but from a quarter of the walking cycle before the current key frame until halfway to the next one. The maximum hip angle in the next key frame is used to update this scale factor.

This local tuning, within a half walking cycle, is illustrated with the 2D articulated structure shown in Fig. 7, from Posture 1 to Posture 3. A 2D articulated structure was

chosen in order to make the understanding easier, however the tuning process is carried out in a 3D space, estimated by using an average pedestrian's height. The two footprints of the first key frame are represented by the points **A** and **B**, while the footprints of the next key frame are the corresponding points **A"** and **B"**. During this half walking cycle one foot is always in contact with the floor (so points **A** = **A'**= **A"**), while the other leg is moving from point **B** to point **B"**. In halfway to **B"**, the moving leg crosses the other one (null hip angle values). Points **C**, **C'**, **C"** and **D**, **D'**, **D"** represent the left and right knee, while the points **H**, **H'**, **H"** represent the hip joints.

Given the first key frame, the scale factor κ_1 is computed and used to perform the motion ($\beta_{i_1(t)}$) through the first quarter of the walking cycle. The second key frame (**A"**, **B"**) is used to compute the scale factor κ_2 . At each iteration of this half walking cycle, the spatial positions of the points **B**, **C**, **D** and **H** are calculated using the position of point **A**, which remains static, the hip angles of Fig. 6 scaled by the corresponding factor κ_i is and the knee angles of Fig. 6. The number of frames in between the two key frames defines the sampling rate of the motion curves presented on Fig. 6. This allows handling variations in the walking speed.

As aforementioned, the computed factors κ_i are used to scale the hip angles. The difference 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 human silhouette. 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 Fig. 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 computed 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. They are computed simultaneously using an efficient search method.

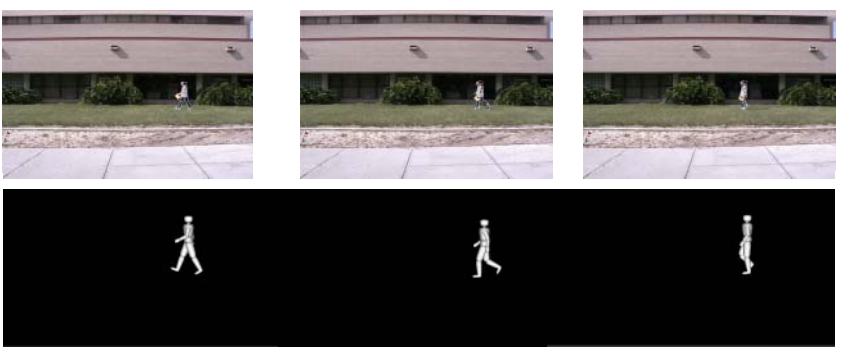

Fig. 8. (*top*) Three different frames of the video sequence used in [13]. (*bottom*) The corresponding 3D walking models

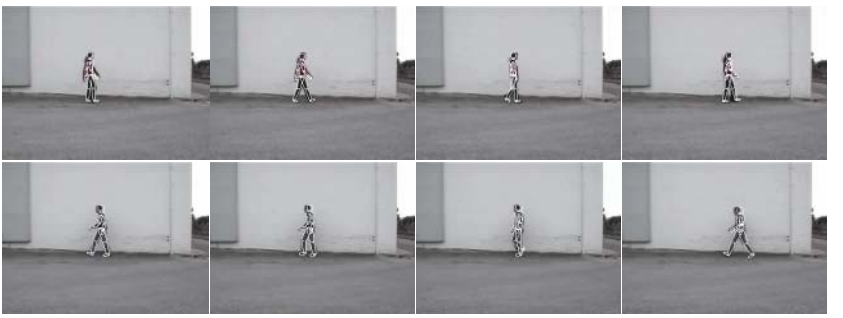

Fig. 9. Input frames of two video sequences (*240x320* each)

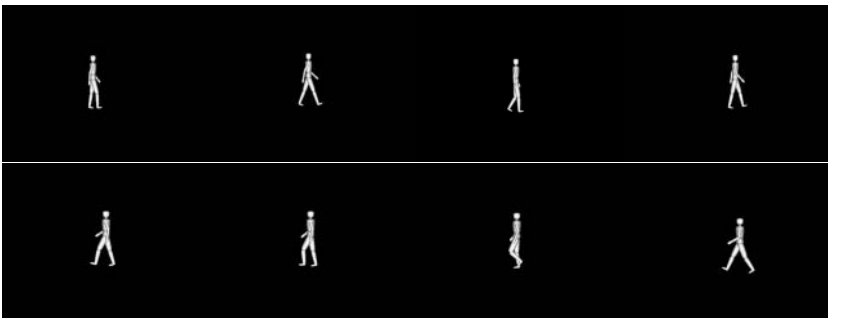

Fig. 10. 3D models corresponding to the frames presented in Fig. 9 (*top*) and (*bottom*) respectively

3 Experimental Results

The proposed technique has been tested with video sequences used in [13] and [14], together with our own video sequences. In spite that the current approach has been developed to handle sequences with a pedestrian walking over a planar surface, in a plane orthogonal to the camera direction, the technique has been also tested with an oblique walking direction (see Fig. 11) showing encouraging results. The video sequence used as an illustration throughout this work consists of 85 frames of *480x720* pixels each, which have been segmented using the technique presented in [15]. Some of the computed 3D walking models are presented in Fig. 8(*bottom*), while the original frames together with the projected boundaries are presented in Fig. 8(*top*).

Fig. 9(*top*) presents frames of a video sequence defined by 103 frames (*240x320* pixels each), while Fig. 9(*bottom*) correspond to a video sequence defined by 70 frames (*240x320* pixels each). Although the speed and walking style is considerably different, the proposed technique can handle both situations. The corresponding 3D models are presented in Fig. 10 (*top*) and (*bottom*) respectively.

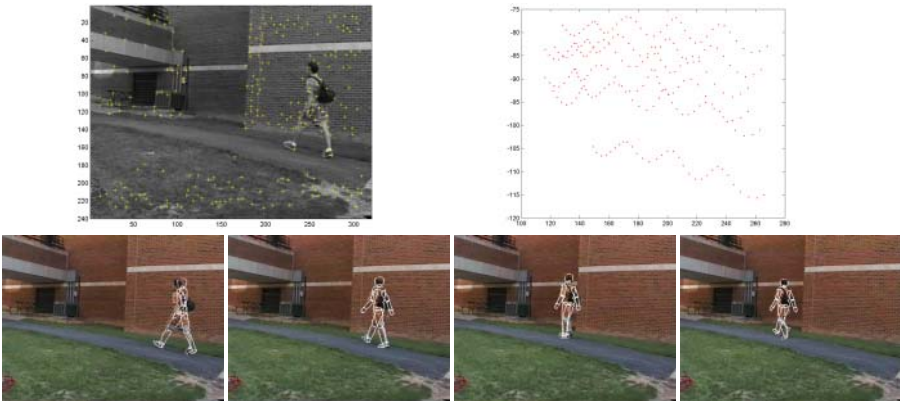

Fig. 11. (*top-left*) Feature points of the first frame. (*top-right*) Feature points' trajectories. (*bottom*) Some frames illustrating the final result (segmented input has been provided by [14])

Finally, the proposed algorithm was also tested on a video sequence, consisting of 70 frames of *240x320* pixels each, containing a diagonal walking displacement (Fig. 11). The segmented input frames have been provided by the authors of [14]. Although the trajectory was not on a plane orthogonal to the camera direction, feature point information was enough to capture the pedestrian attitude.

4 Conclusions and Future Work

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. Although constraints about walking direction and planar surfaces have been imposed, we expect to find results similar to the ones presented in [17] (frontal and oblique walking direction).

The extension of the proposed technique to model other kinds of human body cyclic movements (such as running or going up/down stairs) will be studied by using the same technique (point features detection merged with the corresponding motion model). In addition, the use of a similar approach to model the displacement of other articulated bodies (animals in general [16]) will be studied. Animal motion modeling (i.e. cyclic movement) can be understood as an open articulated structure, however, when more than one extremity is in contact with the floor, that structure becomes a closed kinematics chain with a reduced set of DOFs. Therefore, a motion model could be computed by exploiting these particular features.

Further work will also 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.

References

[1] Sappa, A, Aifanti, N., Grammalidis, N and Malassiotis, S: 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] Delamarre, Q and Faugeras, O.: 3D Articulated Models and Multi-View Tracking with Physical Forces, Special Issue on Modelling People, Computer Vision and Image Understanding, Vol. 81, 328-357, 2001.

[3] Ning, H. Tan, T., Wang, L. and Hu, W.: Kinematics-Based Tracking of Human Walking in Monocular Video Sequences, Image and Vision Computing, Vol. 22, 2004, 429-441.

[4] Wang, L., Tan, T., Hu, W. and Ning, H.: Automatic Gait Recognition Based on Statistical Shape Analysis, IEEE Trans. on Image Processing, Vol. 12 (9), September 2003, 1-13.

[5] Gavrila, D. and Davis, L.: 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.

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

[7] Metaxas, D. and Terzopoulos, D.: Shape and Nonrigid Motion Estimation through Physics-Based Synthesis, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 10 (6), June 1993, 580-591.

[8] Song, Y., Goncalves, L. and Perona, P.: Unsupervised Learning of Human Motion, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 25 (7), July 2003, 1-14.

[9] Rohr, K.: 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.

[10] Sappa, A., Aifanti, N., Malassiotis, S. and Strintzis, M.: Monocular 3D Human Body Reconstruction Towards Depth Augmentation of Telefvision Sequences, IEEE Int. Conf. on Image Processing, Barcelona, Spain, September 2003.

[11] Ma, Y., Soatto, S., Kosecká, J. and Sastry, S.: An Invitation to 3-D Vision: From Images to Geometric Models, Springer-Verlag New York, 2004.

[12] Kim, C. and Hwang, J.: Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications, IEEE Trans. on Circuits and Systems for Video Technology, Vol. 12 (2), February 2002, 122-129.

[13] Phillips, P., Sarkar, S., Robledo, I., Grother, P. and Bowyer, K.: Baseline Results for the Challenge Problem of Human ID Using Gait Analysis, IEEE Int. Conf. on Automatic Face and Gesture Recognition, Washington, USA, May 2002.

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

[15] Kim, C. and Hwang, J.: Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications, IEEE Trans. on Circuits and Systems for Video Technology, Vol. 12 (2), February 2002, 122-129.

[16] Schneider, P. and Wilhelms, J.: Hybrid Anatomically Based Modeling of Animals, IEEE Computer Animation'98, Philadelphia, USA, June 1998.

[17] Wang, L, Tan, T., Ning, H. and Hu, W.: Silhouette Analysis-Based Gait Recognition for Human Identification, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 25 (12), December 2003, 781-796.