Computer Vision and Image Understanding **83,** 79–95 (2001)
doi:10.1006/cviu.2001.0918, available online at http://www.idealibrary.com on IDEAL®

NOTE

A Silhouette Based Technique for the Reconstruction of Human Movement

Andrea Bottino and Aldo Laurentini

Dipartimento di Automatica ed Informatica, Politecnico di Torino,
Corso Duca degli Abruzzi, 24, 10129 Torino, Italy
E-mail: bottino@polito.it, laurentini@polito.it

Received June 27, 2000; revised November 30, 2000

A number of promising applications have renewed researchers' interest in the analysis of human movements. In general, motion capture could play an important role in many areas that require storing, analyzing, or reproducing the motion of human beings. Current motion capture techniques are based on intrusive sensory systems, which might be disturbing or impossible to apply in several application areas. In this paper we present a novel non-intrusive technique able to reconstruct unconstrained motion. From multiple-viewpoint images taken with an ordinary camera a 3D reconstruction is computed with a technique known as volume intersection. Motion data are acquired by fitting a model of the performer to the reconstructed volume. Data about the reconstruction accuracy achievable with our technique in a virtual environment are also provided.

Key Words: silhouettes; volume intersection; model based recognition; motion capture.

1. INTRODUCTION

The interest in the analysis of human motion is motivated both by the recent technical improvements in real-time signal processing hardware and by a variety of new promising application areas. Motion capture (MC) plays an important role since it enables one to store, reproduce, and analyze the motion of human beings. Several applications involving MC already exist, and many others are foreseen. Among them are virtual reality (character animation, games, interactive virtual worlds), sport performance analysis and athlete training, the clinical study of orthopedic patients, computer-driven rehabilitation environments, choreography, smart surveillance systems, gesture-driven user interfaces, and video annotation (see [1, 9]).

1077-3142/01 $35.00

The commercial MC equipment existing at present is based on intrusive sensory systems which exploit either of two technologies: magnetic and optical tracking. Both techniques require more or less bulky objects to be attached to the body of the performer which might disturb the subject and more or less affect his or her gestures. For some applications, such as analyzing sport performances, this could be a serious drawback. Also observe that both techniques supply precise position data only for a limited number of points. These points, in the most favorable case, are located on the moving skin of the subject several centimeters from the feature being tracked, usually joints. These data are suitable for driving the realistic motion of a 3D character, but could be insufficient for working out with precision the 3D posture and motion of the body, as required for biomechanics and sport studies. Last, placing objects on the body is out of the question for applications such as area surveillance.

A number of nonintrusive MC techniques have been reported in the literature. These approaches can be model based or not. In the latter case, it is typically more difficult to establish feature correspondence between consecutive frames. On the contrary, explicit shape modeling can improve the recovery of the body structure, that is labeling and tracking of body parts. Also, one can take advantage of the a priori knowledge about human motion which can be exploited to enhance the motion analysis process.

In model-based approaches the human body is represented with some kind of model whose 3D posture and motion are matched with the physical data. Stick articulated models, as in [13], idealize the human skeleton. Ellipsoidal blobs [5, 6], cylinders and generalized cylinders [15], deformed superquadrics [10], geons [3], and parametric solids and finite elements [14] have been used to build models which mimic more or less closely the human body.

The proposed approaches also differ on the dimensionality of the analyzed space (2D or 3D), on the sources of information, and on the approach to motion recovery. In [13] a system which is able to recover the posture of a human body performing gymnastic movements from a monocular view is presented. In [15] the type of motion analyzed is restricted to the walking cycle. Moving parts are detected by applying a change detection algorithm followed by binary image operations. A similar approach is presented in [17]. Both edge and region information are used to determine the posture of the model together with camera orientation by means of an iterated Kalman filter. The Pfinder system [18] tracks body features such as head, hands, and body. Different body parts and the background scene are described in statistical terms by a spatial and color distribution. Analysis is reinforced by means of a predictive Kalman filter. The approach reported in [2], which exploits 2D projection of blobs to develop a real-time system to track arms and head from two different views using nonlinear estimation techniques, is similar. The paper [8] describes a batch framework based on 2D measurements from a single view. The 3D pose reconstruction process exploits several constraints, including kinematic constraints, joint angle limits, dynamic smoothing, and 3D key frames specified by the user. In [10] four orthogonal views are used to track the whole body of a person. Pose recovery and tracking is obtained by applying a prediction, synthesis, image analysis, and state estimation chain. In [5] the performer is tracked using a region-based motion estimation framework and the model is fitted to the images by means of a Newton–Raphson style minimization. Another recent approach, based on annealed particle filtering, which exploits both edge and silhouette information extracted from multiple cameras, is presented in [7]. Comprehensive references on many other techniques can be found in [1, 9, 15].

The purpose of this paper is to develop a sufficiently simple and robust alternative approach to allow the implementation of practical equipment. Our approach is based on multiple 2D silhouettes of the body extracted from 2D images and can be outlined as follows:

- Different cameras are used to obtain views of a human body. From each image a 2D silhouette of the performer is extracted.
- A volumetric description of the object is recovered by intersecting the solid cones obtained by back-projecting from each viewpoint the corresponding silhouette (*volume intersection, VI*). The final voxel representation can be obtained at different resolutions.
- The posture is recovered by fitting a model of the human body to the reconstructed volume. This is obtained by minimizing a suitable distance function between the volume and the model with a search through the space of pose parameters.

Fitting a model of the human body to the volume reconstructed by VI is in principle equal to fitting in 2D the projections of the model to the various silhouettes. However, fitting in 3D allows a better understanding of the problems due to incorrect volumetric reconstruction (see Section 2). In addition, fitting in 2D would require one to project many times the boundary of the 3D model, while reconstruction should be computed only once for each frame.

The organization of the paper is as follows: Section 2 describes the multiple silhouettes approach, Section 3 outlines the different components of our MC system, Section 4 covers the posture reconstruction, Section 5 reports the accuracy of the overall process, and Section 6 contains concluding remarks and outlines future developments.

2. THE MULTIPLE SILHOUETTE APPROACH

Reconstructing 3D shapes from 2D silhouettes is a popular approach in computer vision. A two dimensional silhouette is the contour of the projection on the view plane of a 3D object. The VI technique (Fig. 1) recovers a volumetric description R of the object O from different silhouettes by intersecting the solid cones obtained by back-projecting from each viewpoint the corresponding silhouette [12, 20]. R is a bounding volume which more or less closely approximates O, depending on the viewpoints and the object itself.

The rationale of the VI approach's quality is that silhouettes can usually be obtained with simple and robust algorithms from intensity images. In addition, VI does not compel us to find correspondences between multiple images and above all is a not-intrusive technique.

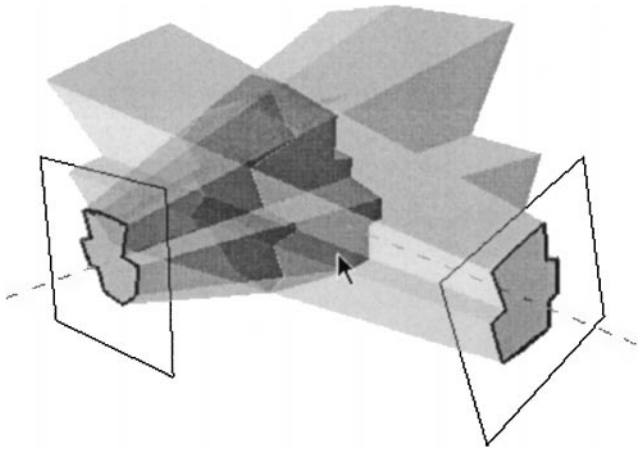

FIG. 1. The VI technique.

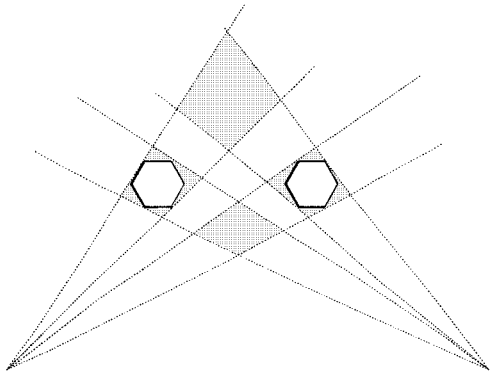

FIG. 2. Phantom volumes with two cameras.

However, using VI for reconstructing the human body requires us to face several difficulties. Bad placement and an insufficient number of cameras could produce bulges which affect the correct placement of the model. In addition, because of the complex shape of the human body, this technique can produce phantom volumes, that is unconnected volumes or protrusions not corresponding to real parts of the body, as can be seen in Fig. 2. For each view, two silhouettes are generated by two objects but VI reconstructs two more objects and we are not able to tell the phantoms from the real objects without further information.

Therefore, due to the complex structure of the human body, an adequate number of cameras must be used and a careful positioning of the view points must be performed. However, in model-based motion capture the "phantom" problem is by far less severe, since exploiting continuity is a powerful tool for fitting the "true" volumes in ambiguous cases.

3. THE MOTION CAPTURE SYSTEM

This section outlines the different components of our MC system. First we describe the model of the human body used for pose recovery. The Cameras and Silhouettes section covers the problems of modeling the cameras and of extracting silhouettes from image planes. Then the VI algorithm is described in detail together with the posture reconstruction algorithm.

3.1. The Model

Our model consists of two components: a representation of the skeleton and a representation of the body surrounding it. The following sections describe those components in details.

3.1.1. The skeleton. The skeleton has 15 segments which are connected by spherical joints. The model is composed by the following body parts: head, trunk, pelvis, upper arms, forearms, hands, thighs, shins, and feet.

Skeleton segments are organized in a tree whose root is located in the pelvis (see Fig. 3). Each segment inherits the transformation of its parent. Some constraints have been introduced to model the structure of human motion according to the anatomy and physics

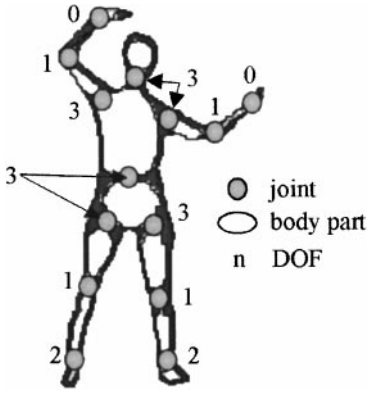

FIG. 3. The human body model.

of human body motion. Elbows and knees provide only one degree of freedom (DOF), ankles cannot roll, and, considering as an approximation that the forearm and the hand are rigidly connected, wrists have no DOFs. The range of values spanned by the DOFs is also constrained by reasonable bounds. The total number of DOFs of the model, including the *(x, y, z)* position of the radix of the tree, is 32.

3.1.2. The surface. The surface is defined through a triangular mesh consisting of more than 600 triangles depicted in Fig. 4. The complete set of shape parameters can be arranged to match the characteristics of the real performer. This surface representation has a medium level of accuracy.

3.2. Cameras and Silhouettes

The system must be accurately calibrated to ensure correct correspondence between the visual cone of each camera and the 3D common world; this is done using Tsai's method [16], which requires an accurate identification of 3D reference points to obtain all the camera parameters. Reference points are obtained by means of a particular calibration object (Fig. 5) that is a not-deformable structure containing a grid of squares of known

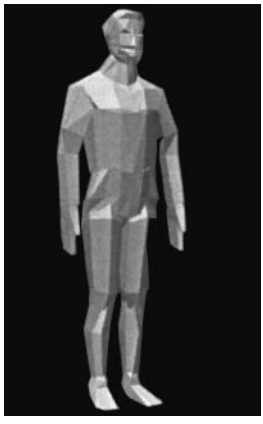

FIG. 4. Model surface.

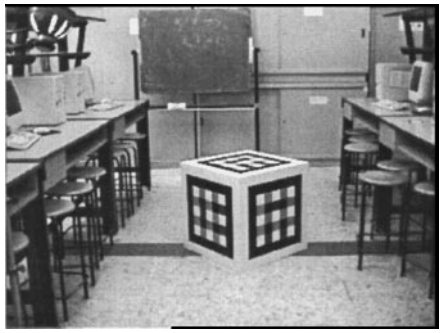

FIG. 5. Image of the calibration object.

position and dimension, whose centers are the reference points. The squares have different colors and are arranged on patterns that allow the calibration process to clearly identify the face of the object they belong to and thus to obtain the complete 3D coordinates of the reference points.

Our approach to silhouette extraction is mainly based on the ideas presented in [18, 19]. Since we use stationary cameras, the silhouette extraction system processes a scene that consists of a static background and one single moving person. Thus we define a model of the background scene and compare it with the current frame. The scene on the background is modeled as a texture surface, every point of which contains a mean value color and a distribution around that mean. The rationale of this approach is to reduce the effect of noise of the images acquired with a CCD camera. The initial model can be computed with a short sequence (about 200 frames) of the empty scene.

The color data associated to each pixel of the scene model are represented in *YUV* format. The advantage of this representation is that *UV* are less sensitive to changes in light intensity and differences between shadowed and not shadowed areas appear almost only in the *Y* component.

To extract the silhouette of the performer from the current frame, each pixel of the frame is thresholded against the expected value, given by the corresponding pixel of the scene model. Since noise can be different for each pixel, a fixed threshold would be an exceedingly rough approximation. Hence, a different threshold for each pixel is evaluated in the preprocessing step and two values are associated to each pixel of the scene model: the mean μ color and the threshold for each component given by

$$T_c(x, y) = \alpha_{\text{tol}} \cdot \max(n_{c,\max} - \mu_c, \mu_c - n_{c,\min}) \qquad c = Y, U, V,$$

where $n_{c,\max}$ and $n_{c,\min}$ are the minimal and maximal value of the component c at the point (x, y) and α_{tol} is a tolerance factor with $\alpha_{\text{tol}} < 1$. For each component of pixel p we evaluate the inequality:

$$|p_c - \mu_c| < T_c(x, y). \tag{1}$$

If the inequality is false for both the U and V components or false for Y and at least one of the *UV* components, the pixel is assigned to the silhouette since its color is significantly different from the background. To compensate for changes in lighting, if the pixel belongs

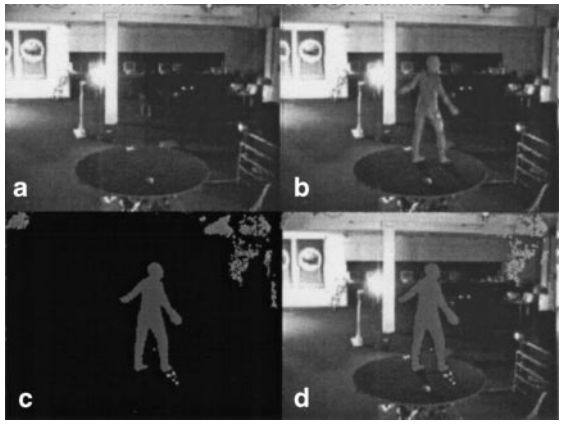

FIG. 6. (a–d) The scene model, a frame of the sequence, and the extracted silhouette.

to the background the pixel statistic is updated using a simple adaptive filter,

$$\mu_t = \alpha \cdot p + (1 - \alpha) \cdot \mu_{t-1},$$

where t refers to the current frame and $t-1$ to the previous one. In order to avoid the identification of cast shadows as part of the silhouette we observe that there is a potential shadow only if the pixel has a similar color but is darker than the expected value. In this case, we consider a second threshold for Y, given by

$$T_s(x, y) = \alpha_{\text{shadow}} \cdot T_Y(x, y),$$

where $\alpha_{\text{shadow}} > 1$, usually set as 1.2; if inequality (1) for the Y component computed using T_s is again false the pixel is assigned to the silhouette.

After the silhouette identification process we apply a postprocessing phase to remove spurious features or to fill undesired holes in the silhouette.

In Fig. 6a the scene model built for one of the cameras used in the test sequence is depicted, while Fig. 6b shows one of the frames of the sequence in which a person moves into the active area. Figures 6c and 6d show the result of the silhouette extraction process.

3.3. The Volume Intersection Algorithm

The VI algorithm works at various resolutions and outputs the boundary voxels of the reconstructed volume R. The running time of the algorithm depends on the number of boundary voxels and thus approximately on the square of the linear resolution.

The outline of the algorithm is as follows:

- a 3D point **P** is an *internal point*, belonging to R, if each projection of **P** in an image plane (according to the camera model) belongs to the corresponding silhouette;
- a voxel is a *boundary voxel* if some, but not all, of its vertices belong to R;
- after finding with a simple heuristic one boundary voxel, the algorithms checks the six adjacent voxels and selects as boundary voxel those which share with the first voxel a *boundary face*, that is a face whose vertices are not all interior or all exterior.

By recursively applying these rules, all the boundary voxels are found.

4. DETERMINING THE POSTURE OF THE MODEL

Pose recovery is based on a search through the 32 dimensional space of pose parameters and implies finding the pose of the model which more closely approximates the actual appearance of the moving subject. The approximation accuracy is given by a similarity function between the current model pose and the volume R obtained by VI. This function is obtained by summing the squared distance between each voxel center C_i to the closest segment of the model.

Let $\wp$ be a vector with 32 parameters required to specify a posture and $d_j(C_i)$ be the distance between the voxel center C_i and the surface of the segment j. Let S_j, $j = 1, \ldots, 15$, be the set of voxel centers closest to segment j. We define the distance function as:

$$D(\wp, R) = \sum_{j=1}^{15} w_j \cdot \sum_{\forall C_i \in S_j} d_j^2(C_i).$$

The contribution of each segment to $D(\wp, R)$ depends on the number of voxels assigned to the segment and on the dimension of the corresponding part of the body. The purpose of the weights w_i is to enhance the contribution of the smallest parts of the model in order to obtain similar posture errors for trunk and limbs. Convenient values of the weights have been experimentally found.

To minimize $D(\wp, R)$ we use the gradient method. The process is stopped when $\Delta D(\wp, R)$ is lower than a predefined threshold.

In order to reduce the number of computations required, each segment is approximated at the first stage of the reconstruction algorithm by an oriented bounding ellipsoid (OBE; see Fig. 7). The size of the axes of each OBE is equal to the dimensions of the boundary box of the corresponding segment. The posture recovery process is a two-stage process: a coarser first step in which the OBEs are fitted to the reconstructed volume and a finer second step in which fitting is applied to the real model.

To recover the motion of the model, the above procedure is applied to each frame of the motion sequence. An exception made for the first time, the starting position of the model is that obtained for the previous frame, since each time the model is close to its final position, the computation of the new posture requires relatively few steps. In addition, some sort of

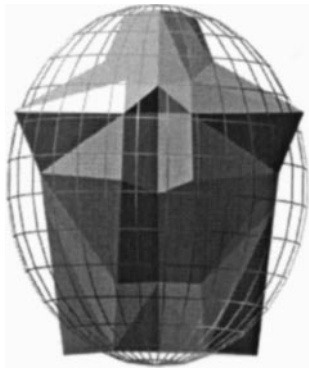

FIG. 7. Oriented bounding ellipsoid (OBE) of the trunk.

implicit filtering takes place, since possible local minima of the distance function due to phantom volumes are avoided.

5. EXPERIMENTAL RESULTS

The experimental work has been divided into two phases.

First, the system has been tested in a virtual environment in order to investigate the precision of both 3D direct reconstruction and model-based posture and motion identification for several postures of the body and various resolutions. Obviously, evaluating the precision of reconstruction is much easier in the virtual world than in the real world. In fact we know a priori the exact posture of the body, and the model used for fitting the reconstructed volume is the same model which produces the silhouettes.

Second, we applied the proposed approach to real image sequences.

5.1. Accuracy in a Virtual Environment

We define the posture error as the average of the distance between corresponding vertices of the reference model and the reconstructed model. In order to cover many significant postures and typical movements, we evaluated the reconstruction accuracy for several different image sequences (see Fig. 8):

- a straight walk, in which the dummy performs a full gait cycle (two steps of 1 m each) recorded in 42 frames
- a circular walk on a path 2 m across (80 frames)
- a run (42 frames)
- a gymnastic movement (40 frames)

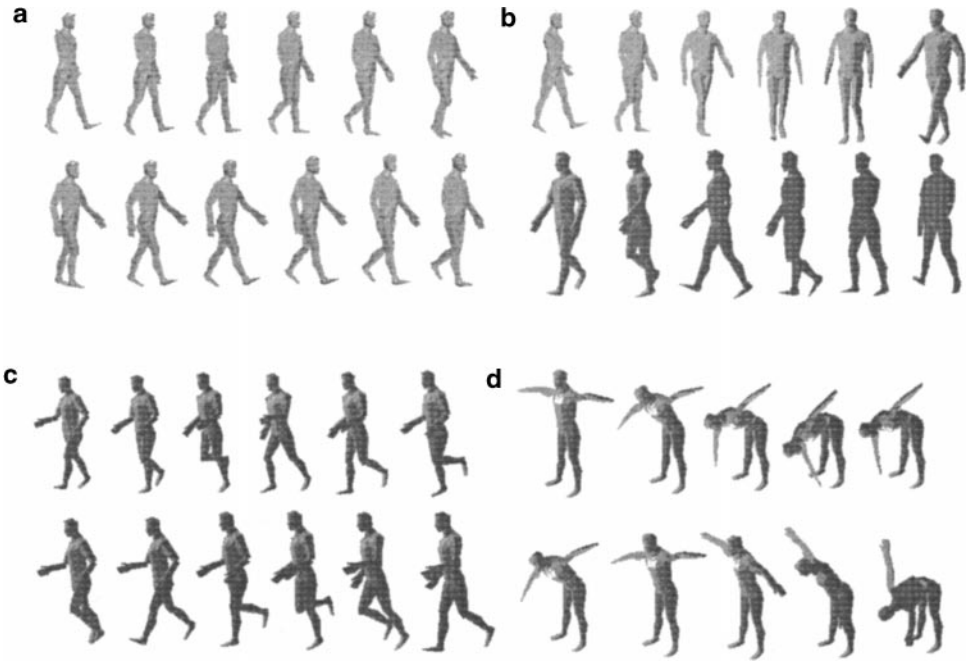

FIG. 8. (a–d) Linear walk, circular walk, run, and gymnastic sequence.

TABLE 1
Summary Results for Linear Walk Sequence

Voxel	Mean error	Max error	Min. error	St. Dev.
25	17.05	21.54	12.61	2.37
35	16.31	23.25	9.91	3.23
45	18.69	23.93	11.60	3.36

TABLE 2
Summary Results for Circular Walk Sequence

Voxel	Mean error	Max error	Min. error	St. Dev.
25	22.54	34.18	13.51	3.99
35	21.67	29.68	12.07	3.91
45	22.90	33.55	12.64	3.69

TABLE 3
Summary Results for Run Sequence

Voxel	Mean error	Max error	Min. error	St. Dev.
25	24.34	37.22	16.37	5.03
35	18.44	25.57	9.20	3.79
45	22.10	31.61	12.32	4.55

TABLE 4
Summary Results for Gymnastic Sequence

Voxel	Mean error	Max error	Min. error	St. Dev.
25	18.57	29.42	12.22	1.28
35	17.93	32.65	11.65	3.97
45	18.57	30.90	9.52	4.42

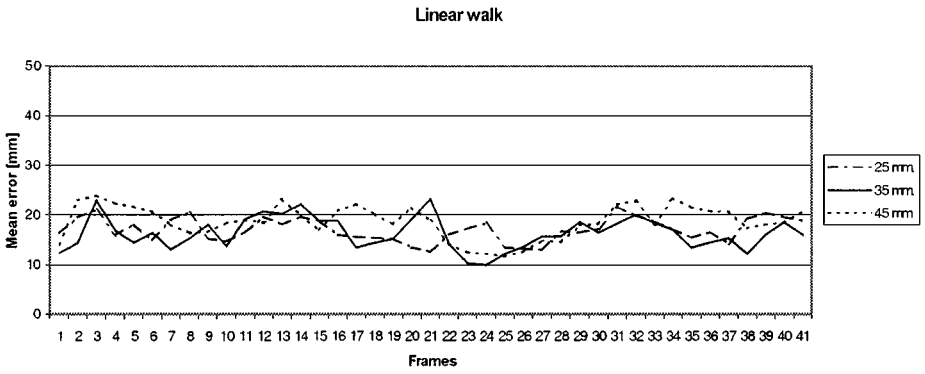

DIAGRAM 1. Average posture error in mm for linear walk sequence.

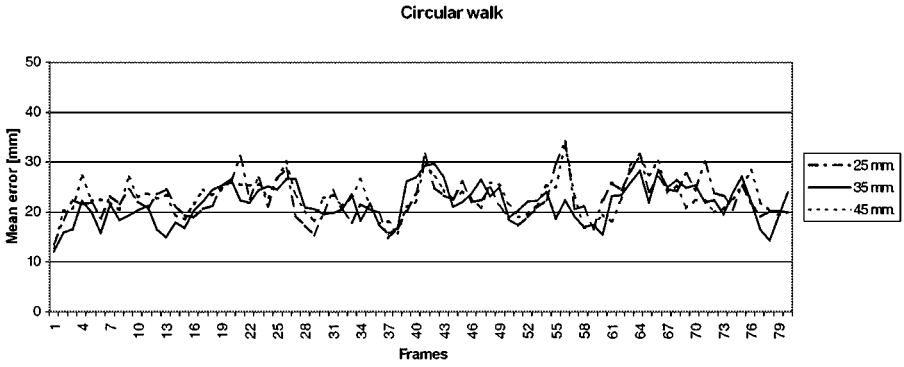

DIAGRAM 2. Average posture error in mm for circular walk sequence.

To evaluate how resolution affects posture precision we have reconstructed the volume using three different voxel sizes (45, 35, and 25 mm). Five cameras have been used for all the tests (four cameras located in a horizontal plane, 1.5 m above the floor, and the fifth shooting the dummy from above). The active area is 4×4 m wide.

The model used to create the motion sequences is 1.80 m high. The results obtained are summarized in Tables 1 to 3, where we report the posture errors averaged over all the frames of the sequences. Diagrams 1 to 4 report the average posture errors for each frame of the sequences, expressed in mm, for decreasing voxel size. The best average error obtained for the different sequences is between 16 and 21 mm, that is almost 1% of the body size. The best reconstruction has been achieved for all the sequences using voxels of 35 mm. The diagrams also show that the accuracy of the reconstruction is relatively unaffected by the voxel size. For completeness, in Table 5 we also present the angular errors averaged over all the frames for the test sequences.

These results are similar to those obtained with a simpler model, consisting of cylinders of various widths [4].

5.2. Recovering Model Postures from Real Image Sequences

The video sequences used in our tests have been acquired in two different environments with different background and lighting conditions. This also helped to improve the robustness of the various components by providing them with a wide range of input data. We

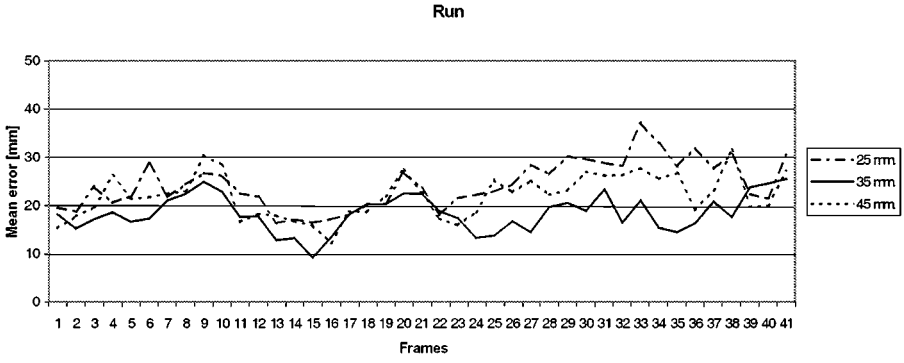

DIAGRAM 3. Average posture error in mm for run sequence.

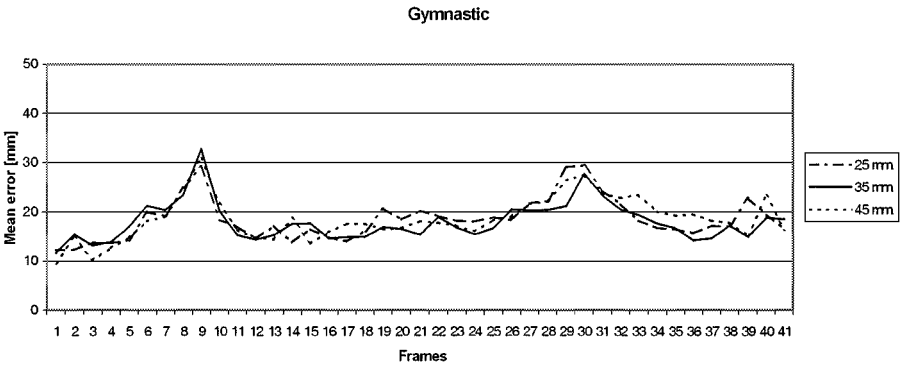

DIAGRAM 4. Average posture error in mm for gymnastic sequence.

have used five video cameras to record five different views of the performer. The sequences have been synchronized by flashing a light at the outset and detecting for each sequence the starting video frame containing the flash. The actor performed freely in the work area since we wanted to test our approach for real unconstrained motion.

To avoid the burden of measuring the characteristics of the real performer, we developed an automatic measurement process to calibrate the model by means of an initialization stage which exploits both known poses and known movements. Similar approaches can be found in [10, 11, 18].

In the virtual environment we found that placing a camera over the head of the performer improved the quality of the volume reconstruction. However, in both the test environments this was not possible due to practical problems. Despite this drawback, the reconstructed sequences look satisfactory when seen at real frame rate (25 frame/s). Results of the reconstruction process can be seen in Fig. 10, which shows the composition of the real and virtual models (Fig. 9 contains the same frames with only the real performer). In Fig. 11 we show a comparison between the original image, the reconstructed voxel model (with voxels of 50 mm), and the parametrical shape model for several frames of a bow sequence reconstructed using five cameras.

TABLE 5
Angular Errors in Degrees Averaged over all the Frames for Different Sequences and Different Voxel Sizes

Sequence	Mean	Min	Max	St. Dev
Walk Vox.25	1.56	0.46	4.97	1.01
Walk Vox.35	1.54	0.72	5.11	0.97
Walk Vox.45	1.75	0.54	5.72	1.13
Circular Vox.25	1.66	0.90	2.58	0.58
Circular Vox.35	1.58	0.80	2.53	0.51
Circular Vox.45	1.75	0.81	2.62	0.57
Run Vox.25	1.63	0.86	3.24	0.67
Run Vox.35	1.26	0.66	2.31	0.48
Run Vox.45	1.44	0.91	2.31	0.40
Gym Vox.25	3.96	0.73	10.91	3.71
Gym Vox.35	4.01	0.83	9.85	3.37
Gym Vox.45	4.29	0.76	10.52	3.70

FIG. 9. Outtakes from camera 2.

FIG. 10. Reconstructed postures.

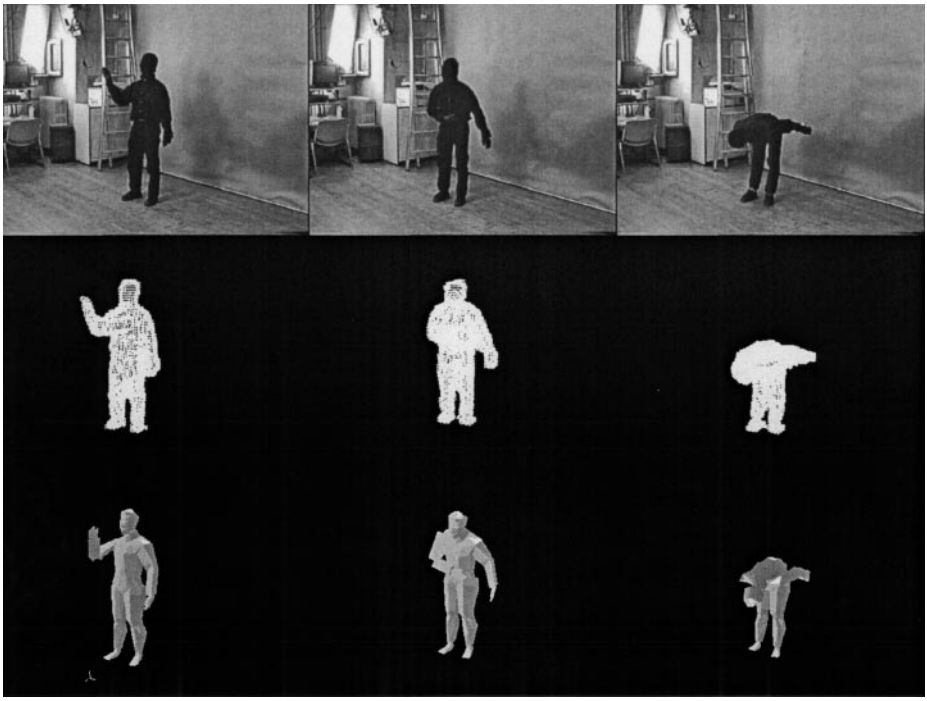

FIG. 11. Original images, reconstructed voxel models, and parameterized shape models for a bow sequence.

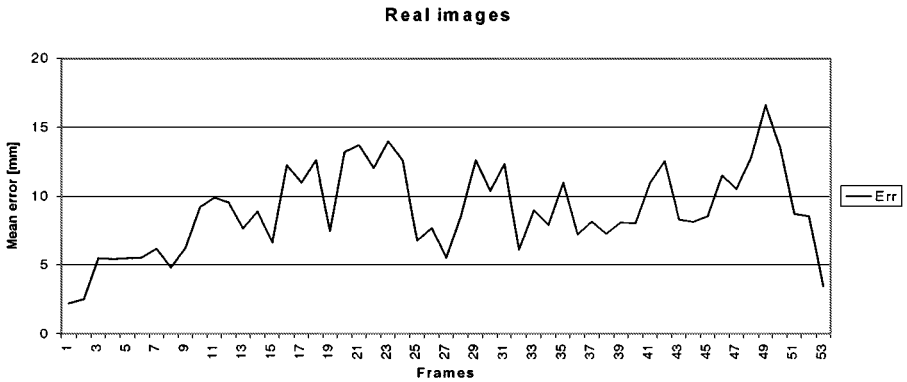

DIAGRAM 5. Average posture difference for reconstruction of real image sequences using 30 and 50 mm voxels.

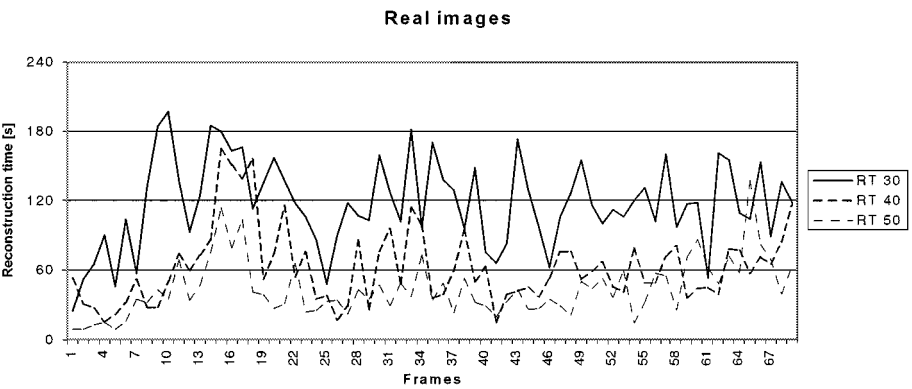

DIAGRAM 6. Average reconstruction time of real image sequences using 30, 40, and 50 mm voxels.

TABLE 6
Summary Results of Reconstruction Time

Voxel	Mean rec. time (s)
30	117.32
40	62.46
50	44.35

To evaluate how different resolutions affect pose reconstruction, we present the difference between the postures obtained using voxels of 30 and 50 mm. As can be seen in Diagram 5, the mean difference is relatively low (its mean value for the first 55 frames is 8.9 mm).

Although real time is not one of the goals of this work, we also computed the mean reconstruction time per frame for different resolutions. The system has been tested on a 500 MHz Pentium. Results are reported in Diagram 6 and summarized in Table 6. As can be seen the mean reconstruction time varies approximately according to the square of the linear resolution, going from 117 s for voxels of 30 mm to 44 s for voxels of 50 mm. Also, it should be noted that the reconstruction time varies greatly from frame to frame, since the number of iterations of the positioning algorithm is not constant.

6. CONCLUSIONS AND FUTURE WORKS

We have demonstrated an approach able to reconstruct unconstrained human motion in realistic situations without markers or external devices attached to the body of the subject. The approach presented is based on multiple 2D silhouettes of the body extracted from 2D images. From each set of silhouettes the performer can be reconstructed with a technique known as volume intersection. Posture recovery is then obtained by fitting a model of the human body to the reconstructed volume.

A quantitative comparison between estimated and true pose is important to evaluate the proposed system. Experiments in a virtual environment proved that the reconstruction accuracy for different motion sequences is between 1.6 and 2.1 cm. (about 1% of the reference object). Although no firm statements about the accuracy of reconstruction can be made for real sequences, the perceived accuracy looks satisfactory for most of the target applications of the system. Another interesting result is that the precision is relatively unaffected by reconstructing the 3D volumes at low resolution for real images also. This benefits the amount of computation required and could be important in cases where a wide area is observed.

It would be interesting to compare the reconstruction precision of our technique (even if obtained in a highly artificial condition) with that of other motion capture techniques. However, this does not appear to be an easy task. One reason is that, as far as we know, no comparable data are available. For intrusive MC approaches, optical markers are tracked with millimetric precision, and similar data are claimed for magnetic tracking. However, no precision data are supplied about the body of the performer.

As far as non-intrusive approaches are concerned, several have been presented and demonstrated with real images, but usually no precision data are available. Clearly, the reason is

that this would require knowing the true posture. The only attempt to perform precise error analysis on computer vision based motion capture studies known to the authors is described in [2]. However, their measurements only refer to the position of a hand moving along a straight trajectory of known dimension and are not easily comparable with our results.

In order to improve our technique, we are planning to consider several issues:

- Constrained motion does not actually include self-intersection avoidance, which might be useful for pruning incorrect poses during the pose reconstruction process.
- Dynamic filtering (such as Kalman filtering) can remove noisy components of the recovered sequence and prediction can be used to boost the reconstruction process.
- Finally, we plan to avoid the limitation of using stationary cameras and having a static background.

REFERENCES

1. J. K. Aggarwal and Q. Cai, Human motion analysis: a review, *Comput. Vision Image Understanding* **73**, 1999, 428–440, doi: cviu.1998.0744.
2. A. Azarbayejani and A. Pentland, Real-time self-calibrating stereo person tracking using 3D shape estimation from blob features, in *Proc. of International Conf. on PR, Vienna, 1996*.
3. I. Biederman, Recognition-by-components: A theory of human image understanding, *Psych. Rev.* **94**, 1987, 115–147.
4. A. Bottino, A. Laurentini, and P. Zuccone, Toward non-intrusive motion capture, in *Proc. of Third Asian Conf. on Computer Vision, Hong Kong, China (1), 1998*, pp. 417–423.
5. C. Bregler and J. Malik, Tracking people with twists and exponential maps, in *Proc. IEEE Conf. on CVPR, 1998*, pp. 8–15.
6. G. K. M. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler, A real time system for robust 3D voxel reconstruction of human motions, in *Proc. IEEE Conf on CVPR, 2000*, pp. 714–720.
7. J. Deutscher, A. Blake, and I. Reid, Articulated body motion capture by annealed particle filtering, *Proc. IEEE Conf. on CVPR, 2000*, pp. 126–133.
8. D. E. Di Franco, T. Cham, and J. M. Rehg, *Recovery of 3D Articulated Motion from 2D Correspondences*, Technical Report Series, CRL 99/7, Cambridge Research Laboratory, 1999.
9. D. M. Gavrila, The visual analysis of human movement: A survey, *Comput. Vision Image Understanding* **73**, 1999, 82–98, doi: cviu.1998.0716.
10. D. M. Gavrila and L. S. Davis, 3D Model-based tracking and recognition of human movement: a multi-view approach, in *Proc. IEEE CS Conf. on CVPR, San Francisco, CA, 1996*, pp. 73–80.
11. I. Kakadiaris and D. Metaxas, 3D Human body model acquisition from multiple views, in *Proc. of the Fifth International Conference on Computer Vision, Boston, 1995*, pp. 618–623.
12. A. Laurentini, How far 3D shapes can be understood from 2D silhouettes, *IEEE Trans. Pattern Anal. Mach. Intell.* **17**, 1995, 188–195.
13. M. K. Leung and Y. Yang, First sight: a human body outline labeling system, *IEEE Trans. Pattern Anal. Mach. Intell.* **17**, 1995, 359–377.
14. A. Pentland and S. Sclaroff, Closed-form solutions for physically based shape modeling and recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* **13**, 1991, 715–729.
15. K. Rohr, Toward model-based recognition of human movements in image sequences, *CVGIP: Image Understanding* **59**, 1994, 94–115.

16. R. Y. Tsai, A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses, *IEEE J. Robotics Automation* **3**, 1987, 323–344.
17. S. Wachter and H.-H. Nagel, Tracking persons in monocular image sequences, *Comput. Vision Image Understanding* **74**, 1999, 174–192, doi: cviu.1999.0758.
18. C. Wren and A. Pentland, Dynamic models of human motion, in "*Proc. of Third IEEE International Conf. on Automatic Face and Gesture Recognition*," *Nara, Japan, 1998*, pp. 22–27.
19. M. Yamada, E. Kazuyuki, and J. Ohya, A new robust real-time method for extracting human silhouettes from color images, in *Proc. of Third IEEE International Conf. on Automatic Face and Gesture Recognition, Nara, Japan, 1998*, pp. 528–553.
20. J. Zheng, Acquiring 3D models from sequences of contours. *IEEE Trans. Pattern Anal. Mach. Intell.* **16**, 1994, 163–177.