The Visual Computer

Experimenting with nonintrusive motion capture in a virtual environment

Andrea Bottino,
Aldo Laurentini

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

A growing number of promising applications requires recognizing human posture and motion. Conventional techniques require us to attach foreign objects to the body, which in some applications is disturbing or even impossible. New, nonintrusive motion capture approaches are called for. The well-known shape-from-silhouette technique for understanding 3D shapes could also be effective for human bodies. We present a novel technique for model-based motion capture that uses silhouettes extracted from multiple views. A 3D reconstruction of the performer can be computed from a silhouette with a technique known as volume intersection. We can recover the posture by fitting a model of the human body to the reconstructed volume. The purpose of this work is to test the effectiveness of this approach in a virtual environment by investigating the precision of the posture and motion obtained with various numbers and arrangements of stationary cameras. An average 1% position error has been obtained with five cameras.

Key words: Silhouettes – Volume intersection – Model-based recognition – Motion capture

Correspondence to: A. Bottino

1 Introduction

Capturing human motion is an important practical issue. Several applications already exist, and many others are foreseen. Among them are virtual reality (character animation, games, interactive virtual worlds), motion analysis (sport performance analysis, athletes training, clinical study of orthopedic patients, choreography of dance and ballet), smart surveillance systems, gesture-driven user interfaces, and telerobotics (Gavrila 1999).

Understanding human posture and motion is a challenging task since the human body is a complex, articulated and flexible object, and many postures produce self-occlusion.

Several kinds of commercial motion capture equipment exist at present. They are based on the idea of tracking key points of the subject (usually joints) (Lee and Chen 1985; Rashid 1980; Webb and Aggarwal 1982). This can be done with one of two technologies: magnetic and optical tracking. Both require more or less bulky objects to be attached to the body, which might disturb the subject and affect his gestures. In many of the present and future applications of motion analysis, nonintrusive sensory methods are preferable, and in some cases, they are necessary.

Several authors have dealt with motion capture or systems that analyze human movement. These are based on vision, but many approaches are model based. The human body is usually represented by some kind of model whose 3D posture and motion are matched with physical data. Stick-articulated models, as those of Leung and Yang (1995), idealize the human skeleton. Ellipsoidal blobs (Bregler and Malik 1998), cylinders and generalized cylinders (Marr and Nashihara 1985; Mohan and Nevatia 1989; Rohr 1994), deformed superquadrics (Solina and Bajcsy 1990; Gavrila and Davis 1996), geons (Biederman 1987), parametric solids and finite elements (Pentland and Horowitz 1991; Pentland and Scarloff 1991), have been used to build models that mimic the human body more or less closely. Motion estimation has been improved by predictive Kalman filtering (Pentland and Horowitz 1991; Pentland and Scarloff 1991; Rohr 1994; Wachter and Nagel 1999). The proposed approaches also differ for the dimensionality of the analyzed space (2D or 3D), for the sources of information, and for the approach to motion recovery. Leung and Yang (1995) present a system that can recover the posture of a human gymnast from a monocular view. Rohr (1994) restricts the type of motion analyzed

to the walking cycle. He detects moving parts with a change-detection algorithm followed by binary image operations. Wachter and Nagel (1999) use a similar approach that exploits both edge and region information to determine the posture of the model and camera orientation by means of an iterated Kalman filter. The Pfinder system (Wren et al. 1997) tracks features such as those of the head, hands, and body. Various body parts and the background scene are described in statistical terms by spatial and color distributions. Azarbayejani and Pentland (1996) report a similar approach, which exploits 2D projection of blobs to track the arms and head from two views with nonlinear estimation techniques. Gavrila and Davis (1996) use four orthogonal views to recover the posture. The pose is recovered and tracked by a prediction, synthesis, image analysis, and state estimation chain. Bregler and Malik (1998) track the performer in a region-based, motion-estimation framework, and fit the model to the images by means of a Newton-Raphson-style minimization. Kakadiaris and Metaxas (1998) present another approach. They do not use a model defined a priori, but recover the body parts from the image sequence.

Comprehensive references to many techniques in this area can be found in the recent surveys of Aggarwal and Cai (1999) and Gavrila (1999).

In this paper, we present and discuss a new, nonintrusive motion-capture technique based on multiple video images. The approach is based on 2D silhouettes of the body extracted from 2D images, and could be sufficiently simple and robust for practical implementations. The main features of our approach are:

1. Silhouettes can usually be obtained with a simple and robust algorithm from intensity images.
2. A direct reconstruction of the 3D shape of the body can be computed from the silhouettes with a technique known as *volume intersection* (Laurentini 1994; Zheng 1994), and the 3D posture and motion of models of the human body can be obtained by fitting a model to the reconstructed volume

However, the number and position of the cameras, subject to practical constraints, strongly affect the accuracy of the 3D reconstruction, and, thus, also the estimation of the posture and motion of a model. Studying this problem is a necessary step towards effective motion capture from multiple silhouettes. The purpose of this paper is twofold:

1. To demonstrate a model-based motion capture based on this approach in a virtual environment.
2. To investigate the precision of our approach for both direct 3D reconstruction and model-based motion identification with various arrangements of cameras and various resolutions, models, and motions.

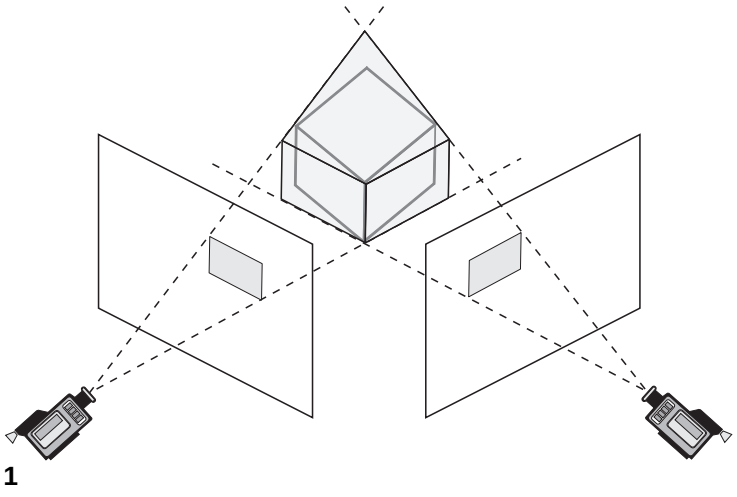
Obviously, the second task is much easier in the virtual world than in the real world since we know the exact posture of the body a priori, and the model used for fitting the reconstructed volume is the same as the model that produces the silhouettes.

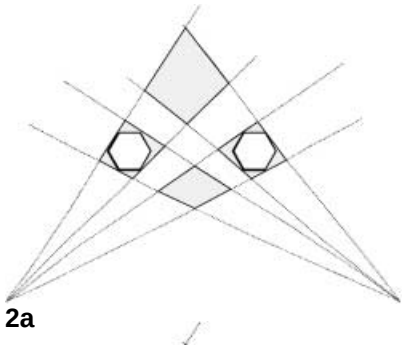
The content of this paper is as follows. In Sect. 2, we discuss the problems connected to the reconstruction of 3D shapes from silhouettes. In Sect. 3, we describe the virtual environment and the algorithms for determining posture and motion. In Sect. 4 we present and discuss the experimental results obtained.

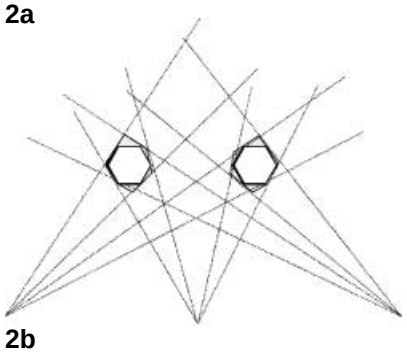
2 The multiple silhouette approach to motion capture

Reconstructing 3D shapes from 2D silhouettes is a popular approach in computer vision. A 2D silhouette of a 3D object is the region occupied by the projection of the object on the view plane. The volume intersection (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 (Martin and Aggarwal 1983; Potemesil 1987; Noborio et al. 1988; Ahuja and Veenstra 1989; Chian and Aggarwal 1989; Srinivasan et al. 1990; Zheng 1994). R is a volume that approximates O; how closely depends on the viewpoints and the object itself. The rationale of the VI approach is that silhouettes can usually be obtained with simple and robust algorithms from intensity images. In addition, VI does not compel us to find correspondence between multiple images.

Despite the simplicity of the basic idea, the VI approach raises a number of questions, such as: which objects are exactly reconstructable; which is the closest approximation that can be obtained for nonreconstructable objects; what can be inferred about the unknown object from the reconstructed object? The geometric concept of the visual hull (Laurentini 1994, 1995, 1997) provides answers to these questions, which are relevant for all VI applications. For







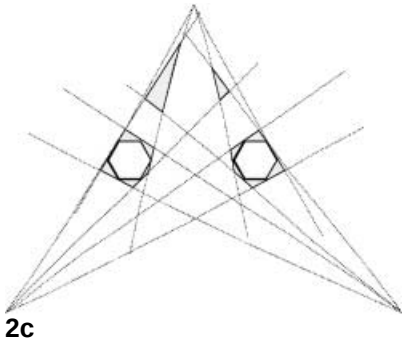


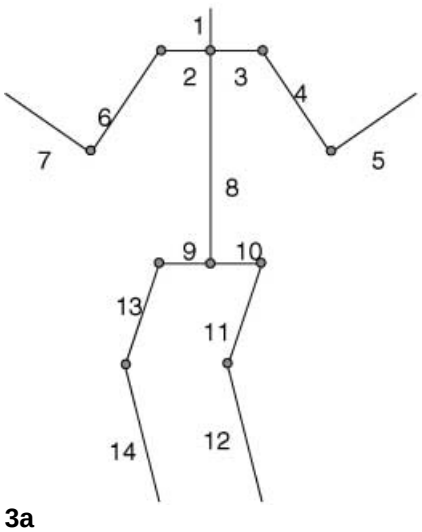
Fig. 1. The volume intersection (VI) technique
Fig. 2a–c. Phantom volumes

instance, when the arms are close to the trunk or the legs are close to each other, the visual hull of the body is larger than the body itself, which, therefore, cannot be exactly reconstructed from the silhouettes alone.

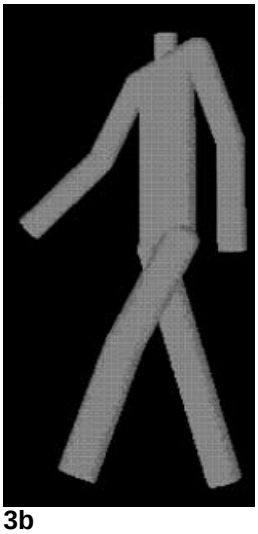
Using VI for reconstructing the human body leads to certain difficulties. Poor placement and an insufficient number of cameras could produce bulges that affect the correct placement of the model. In addition, because of the complex shape of the human body, this technique can produce "phantom" volumes; that is, unconnected volumes or protrusions that do not correspond to real parts of the body.

Let us consider the 2D example of Fig. 2a. For each view, the two objects generate two silhouettes and, without further information, we cannot distinguish the phantom objects from the real ones. In this case, a third viewpoint can easily overcome the problem (Fig. 2b). However, three viewpoints would not be sufficient to avoid phantoms. For instance, in Fig. 2c, three evenly spaced viewing directions still produce two phantom objects.

It is not difficult to understand that phantom volumes are likely to be produced when a human body is reconstructed with few silhouettes. Situations as those shown before in 2D can occur in many planes sectioning a human body. Thus, investigating the numbers and positions of cameras suitable for reconstructing 3D bodies with sufficient precision is among the goals of this paper.



3a



3b

4

Fig. 3. The skeleton (**a**); the dummy (**b**)
Fig. 4. The walking dummy

Model-based posture understanding based on VI consists in fitting a model of the human body to the volume reconstructed by VI. In principle, this is exactly the same as fitting the projections of the model to the various silhouettes in two dimensions (at least for objects of unknown shape), since the information provided by the set of silhouettes or by the reconstructed volume is the same. However, fitting in three dimensions makes it easy to visualize the bulges that affect the recovered posture.
In model-based motion capture, the phantom problem is much less severe, since exploiting continuity is a powerful tool for fitting the "true" volumes in ambiguous cases.

3 The virtual environment and motion capture system

In this section, we describe the various software components developed for reconstructing a 3D body from its silhouettes and for capturing its motion in a virtual environment. The VI, posture determination, and motion capture software can also deal with silhouettes extracted from real-world images, and will be used in our future work. Two models of the human body have been used in our experiments. In this section, we present the simpler model. The second is described in Sect. 4.

3.1 The simpler model

The skeleton of the human body has been modeled as a tree of 14 rigid segments connected by ball joints (Fig. 3a). The lengths and the widths of the segments agree with the average measurements of the male population (Fantin and Dias 1994) and the body, without the head, is about 1.50 m high. The body is modeled by cylinders of various widths centered about the segments (Fig. 3b), except for the trunk, whose section is a rectangle with smoothed angles. To specify a posture, one must specify a vector $\wp$ with 32 parameters: the (x, y, z) position of the radix of the tree and two angles for each segment, except for the trunk. An extra parameter is required to specify its rotation, since it is not symmetrical with respect to its axis. The model is constrained to avoid physically impossible positions, and simple motion algorithms can be used to drive it in various motion sequences (Fig. 4), which are sufficiently realistic for our purposes. For details, see Blunno and Falcione (1996).

3.2 Cameras and silhouettes

Any number of stationary cameras can be located anywhere in the virtual 3D environment (Fig. 5). The position and orientation of each camera are specified with a view center, an optical axis, and an up vector. Each virtual camera provides a frame of 512×512

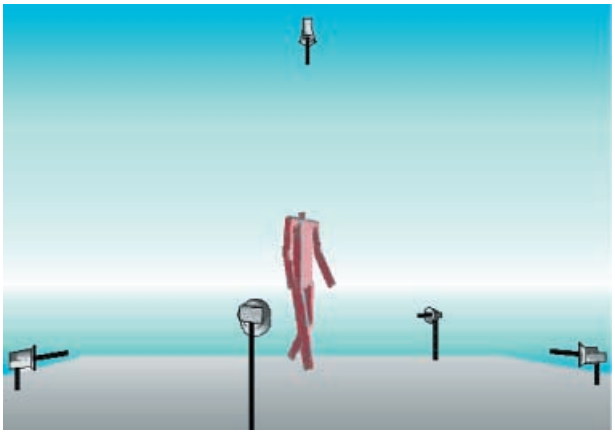


Fig. 5. The virtual environment

two-level pixels in the virtual image plane. OpenGl has been used to model the virtual environment and to obtain the silhouettes.

3.3 The volume intersection algorithm

Provision has also been made for images produced by a real camera, and the silhouettes can be back-projected with Tsai's camera model and calibration data, consisting of a set of 3D world coordinates of feature points and the corresponding 2D coordinates of the feature points in the image. The reader is referred to the original paper for further details (Tsai 1987).

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

Let a 3D point P be an *internal point* if it belongs to R; otherwise, an *external point*. Clearly, each projection of an internal point in an image plane belongs to the corresponding silhouette. The vertices of a boundary voxel of R cannot be all internal or all external points. The VI algorithm is as follows. After finding an initial boundary voxel, the algorithms checks the six adjacent voxels and selects those that share *a boundary face* with the first voxel as the boundary voxels. The vertices of such a boundary face cannot be all internal points or all external points.

All the boundary voxels are found by applying these rules recursively. The initial boundary voxel can be found as follows:

- For the first frame, a random internal point is chosen as the center of the initial voxel; if the selected voxel is not a boundary voxel, then a boundary voxel can be found by exploring the voxel space along the coordinate axis.
- For the following frames, we look for boundary voxels of the previous frame that are also on the boundary at the current frame.

In Fig. 6, we show a 2D example of the VI algorithm. In Fig. 6c and d we see how boundary voxels are recursively found: the white dotted voxels are discarded since they do not share a boundary face with the boundary voxel previously found, while the rule is applied recursively to the black dotted ones. In Fig. 7 we show some outputs of the VI algorithm at various resolutions.

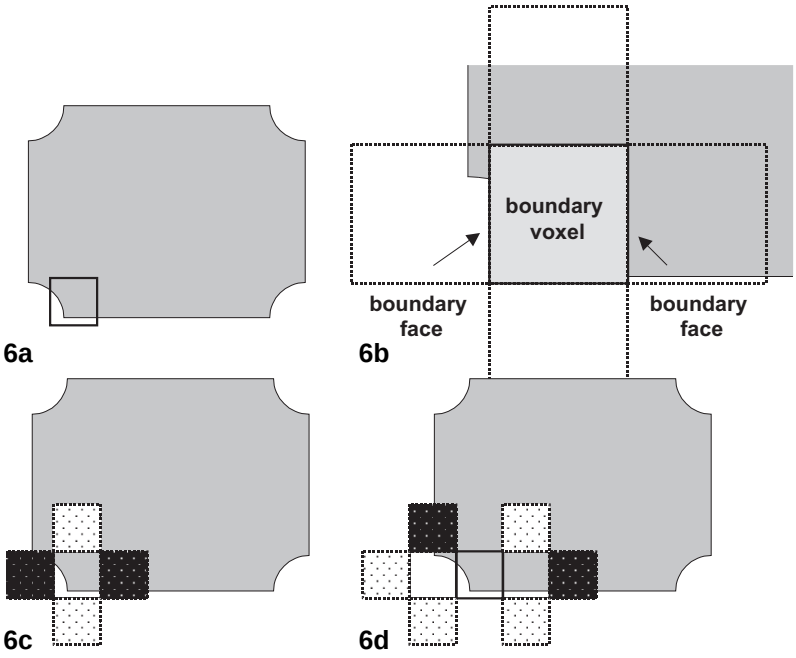
3.4 Determining the posture of the model

Fitting the model to volume R obtained by VI requires minimization of some measure of distance between the dummy and the reconstructed volume. A rather straightforward choice is to minimize the distance between the boundary voxels of R and the surface of the dummy. In more detail, let C_i be a voxel center. Its distance from the surface of the dummy is assumed to be the distance between C_i and the closest surface of a segment of the dummy. Let this be segment j, and let the distance be $d_j(C_i)$. Let S_j, $j = 1, \ldots, 14$ be the set of centers closest to each segment. We define the distance function as:

$$D(\wp, R) = \sum_{j=1}^{14} w_j \cdot \sum_{\forall C_i \in S_j} d_j^2(C_i).$$

The contribution of each segment to $D(\wp, R)$ depends on the number of voxels assigned to the segment, which can change at each iteration, and it also depends 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 found experimentally.

For minimizing $D(\wp, R)$, we use the gradient method in the 32D space of the position parameters. The process is stopped when $\Delta D(\wp, R)$ becomes less than a predefined threshold.



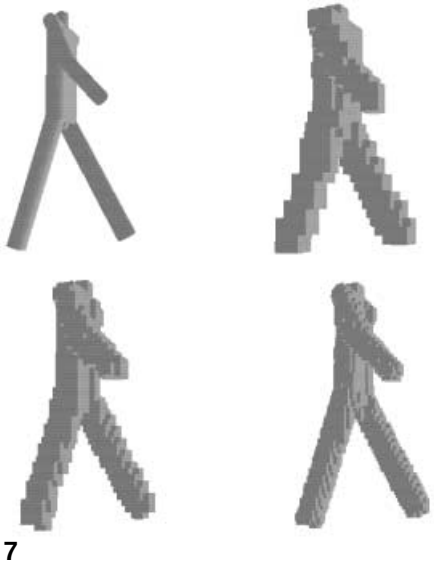


Fig. 6a–d. Example of the VI algorithm running in two dimensions

Fig. 7. The original posture and the volumes reconstructed at increasing resolutions (voxel sizes of 50 mm, 35 mm, and 20 mm)

As for the VI algorithm, the reconstruction time depends mainly on the number of voxels, and thus approximately on the square of the linear resolution.

It is worth noting that the distance measure chosen is computationally expensive, even though some simplification is possible for a sequence of contiguous postures. We also experimented with a simpler distance function, in which the distances were computed between the centers of the boundary voxels and the skeleton segments. However, this choice resulted in a less accurate posture determination, particularly for the cases in which the bulges produced

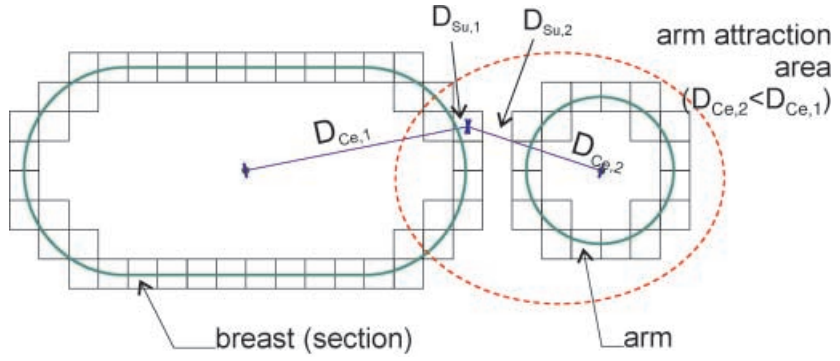


Fig. 8. Horizontal section of the dummy

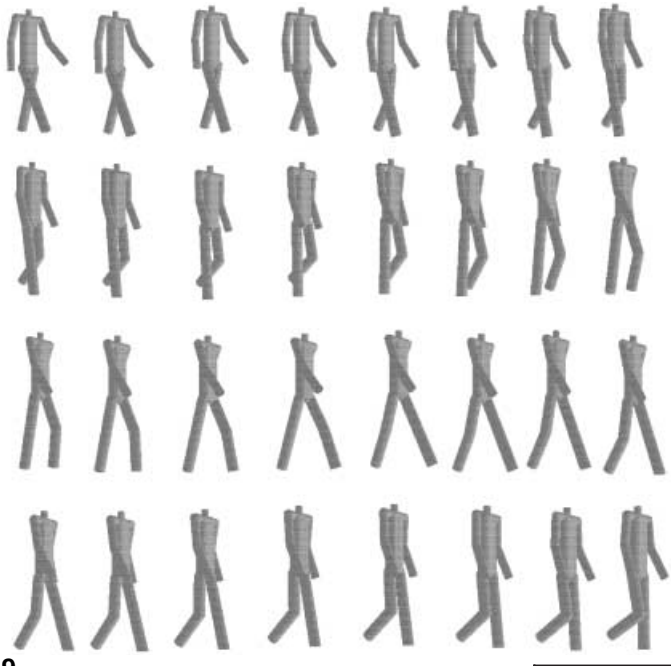
by the VI algorithm were more conspicuous. The reason for this can be understood from the example in Fig. 8, which shows a horizontal section of the dummy. In this case, the correct assignment of the voxel marked with a small cross is to segment 1, since the distance $D_{Su,1}$ of its center from the surface of segment 1 is less than $D_{Su,2}$, the distance from the surface of segment 2. However, if we consider the distances from the center of the segments, then it is $D_{Ce,2} < D_{Ce,1}$, and the point would be assigned to segment 2.

3.5 Recovering the motion of the model

In order to recover the motion of the dummy, the procedure just described is applied to each frame of the motion sequence. Except for the first frame, the starting position of the model is the one obtained from the previous set of silhouettes. Since the dummy is close to its final position at each iteration, the computation of the new posture requires relatively few steps. In addition, some sort of implicit filtering takes place, since possible local minima of the distance function due to phantom volumes are avoided.
The dummy easily fits the first volume obtained, provided that the start-up position has almost the same global orientation as the moving dummy.

4 Experimental results

In this section, we present and discuss the precision performances of our algorithm.
In the first test, we consider a full gait cycle (two steps of about 1 m each), recorded in 42 frames, with which we can consider several postures. The initial frames of the sequence are depicted in Fig. 9.
To evaluate how resolution affects posture precision, we use three voxel sizes for the VI algorithm (50 mm, 35 mm, and 20 mm).
The number and position of the cameras strongly affects the VI reconstruction, and then the posture precision. To reduce the degrees of freedom (DOFs) of our search, we explored mainly arrangements of three, four, and five cameras. It is worth noting that optical motion capture systems typically use six to eight cameras for capturing the motion of the full body (Webb and Aggarwall 1982; Sabel 1996; Wells and Tutt 1998). As another practical constraint, we locate most cameras in a horizontal plane, 1 m above the floor.
First, we experimented with various arrangements of three cameras in this plane, and we found no satisfactory positioning. In all the cases tested, the position of the arms of the dummy was completely wrong for many frames in the gate cycle, with errors of more than 10 cm.
The accuracy was markedly improved by a fourth camera in the same plane, and by a fifth camera, located at 4 m from the ground level, with a vertical optical axis (that is, above the head of the dummy). However, adding more cameras on the plane seemed to improve the accuracy only marginally.
We experimented with several arrangements of the four cameras in the horizontal plane. Four of them are shown in Fig. 10, where the dotted lines of the grid are 1 m apart. The corresponding positioning errors are reported in Fig. 11 for all the frames of the gait cycle. They are obtained by plotting the values of



9

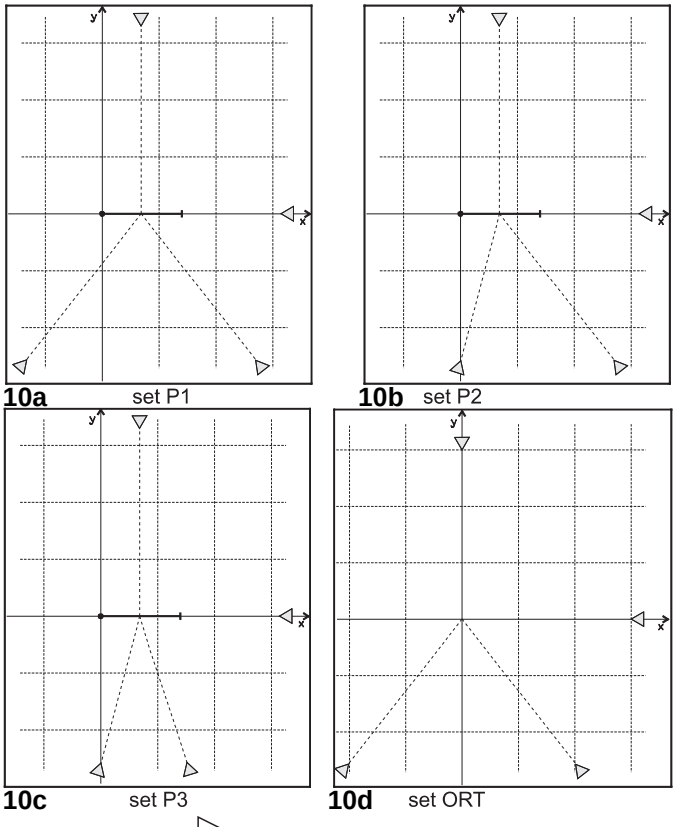
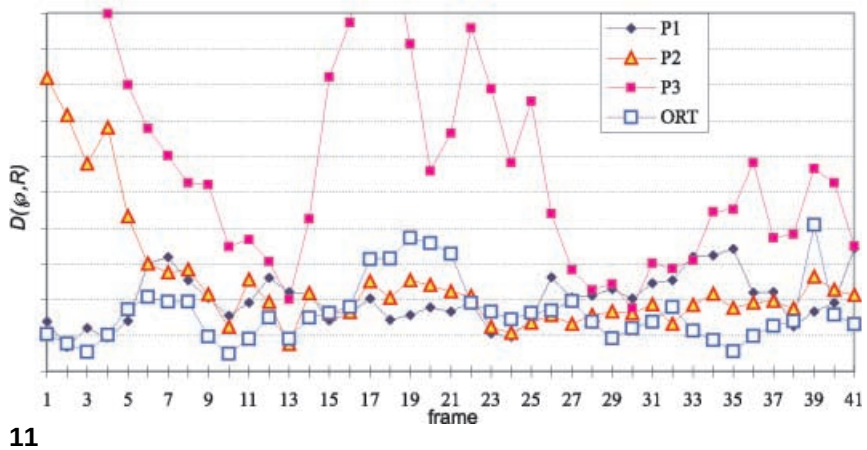
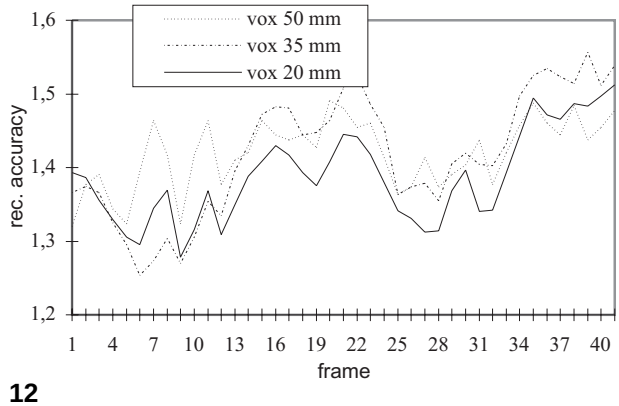


Fig. 9. The walking sequence
Fig. 10a–d. Four camera arrangements in the horizontal plane



11



12

Fig. 11. $D(\wp, R)$ versus the frame number for different arrangements
Fig. 12. Volume intersection precision for four cameras

the distance function $D(\wp, R)$ of the walking model obtained for each frame. The units of the scale are not given since the value of $D(\wp, R)$ cannot immediately be related to the error functions defined in the following paragraphs.

Several arrangements resulted in similar accuracy, provided that no two optical axes were too close, as in case P3. We also found better accuracy when the optical axes converged at the gait-cycle center. Thus, we also considered the arrangement named ORT (Fig. 10d), which is equivalent to moving the cameras, the center of the model being always in the same position.

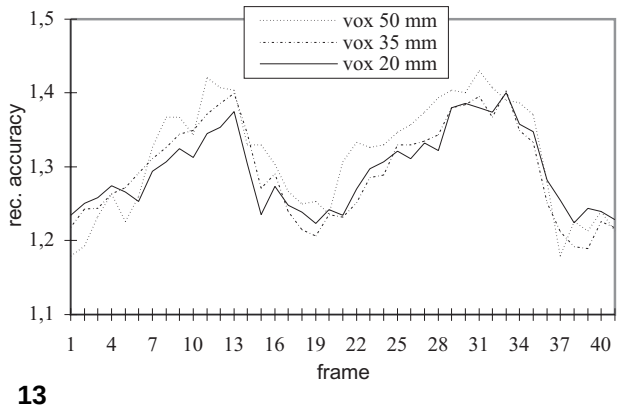
In conclusion, for the tests described in detail in the following subsections, we selected the arrangement labeled P1 in Fig. 10a for the four cameras in the horizontal plane.

4.1 Volume intersection precision

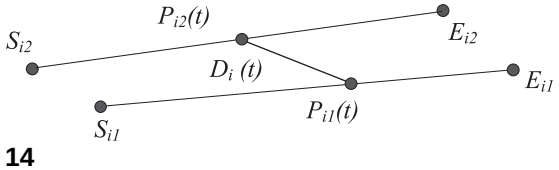
A common parameter used to define the precision of the reconstruction obtained by VI algorithms is the ratio between the reconstructed volume and the volume of the original object (Ahuja and Veenstra 1989; Noborio et al. 1988; Potemesil 1987). In order to minimize the consequences of the quantization, the reconstructed volume is computed with the formula:

$$\text{reconstructed volume} = \sum_{\text{inner voxels}} s^3 + \sum_{\text{boundary voxels}} \left(\frac{\text{no. of voxel vertices} \in R}{8} \right) \cdot s^3,$$

$s = \text{voxel size}$



13



14

Fig. 13. Volume intersection precision for five cameras
Fig. 14. Axes of the real and reconstructed segments

where the second term approximates the portion of volume of the boundary voxels belonging to R.
The reconstruction precision obtained with four and five cameras is plotted for all the frames of the gait cycle and for three resolutions in Figs. 12 and 13.
Even with five cameras, these data show a rather coarse reconstruction for many postures. Since decreasing the voxel size provides perceptible, but not substantial, improvements, little precision is essentially due to the few cameras in relation with the complex shape of the body. However, as we show in the next subsection, the posture of the dummy can be determined from these rather rough volumes with centimetric precision.

4.2 Posture determination in the gait cycle

In this subsection, we present the experimental results concerning the precision of the postures of the model determined in the gait cycle.
For each segment i, $i = 1 \ldots 14$, we define a reconstruction error d_i as the average distance between a point $P_{i1}(t)$ of the axis of the real segment and the corresponding point $P_{i2}(t)$ of the axis of the reconstructed segment (Fig. 14):

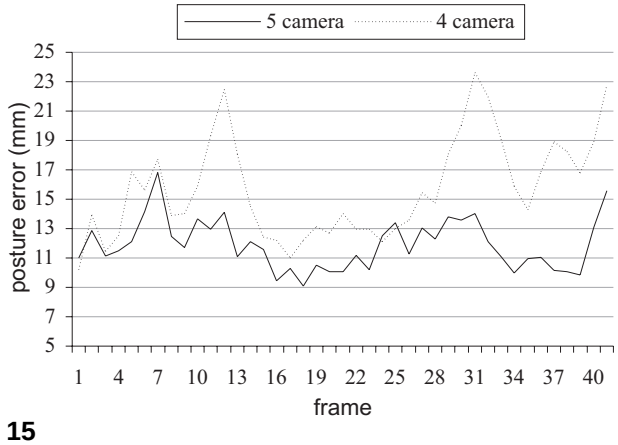
$$d_i = \frac{1}{T}\int_0^T D_i(t)dt = \int_0^1 \|P_{i1}(t),\, P_{i2}(t)\|dt$$

$$= \int_0^1 \|S_{i1} + (E_{i1} - S_{i1})t,\, S_{i2} + (E_{i2} - S_{i2})t\|dt.$$

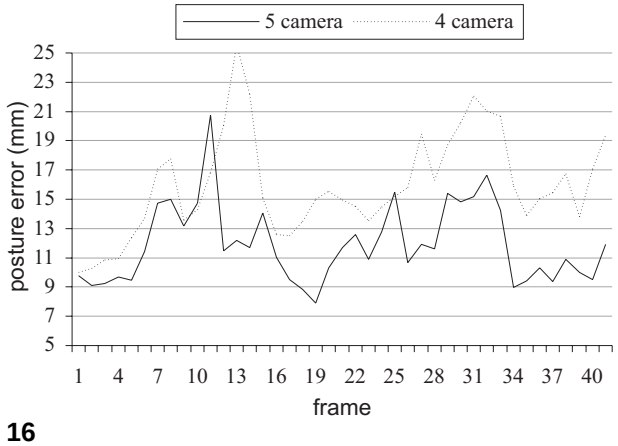
To summarize the overall difference between the true posture and the posture determined with our algorithm for the entire dummy, we compute an average distance weighted with the length of the segments for each frame as follows:

$$\text{Posture error} = \sum_{i=1}^{14} L_i \cdot d_i \Big/ \sum_{i=1}^{14} L_i \ ,$$

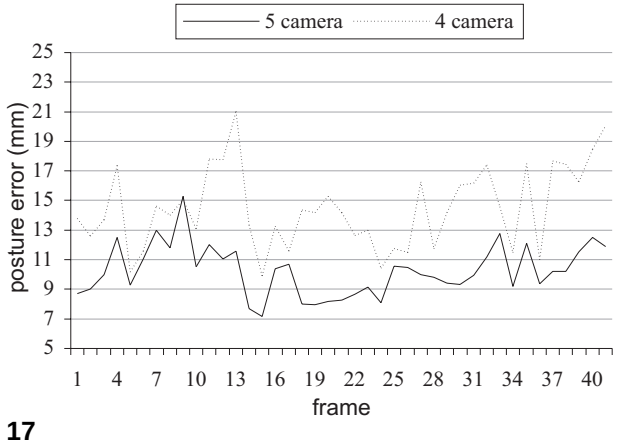
where L_i is the length of each segment. The average posture errors for each frame of the gait cycle, are reported in Figs. 15–17 for decreasing voxel size. In each diagram we plot the errors for both the four- and five-camera arrangements.
A substantial error reduction is provided by the fifth camera, with a vertical axis, which strongly reduces phantoms and bulges. It can also be observed that the larger posture errors are in the frames where the arms are close to the trunk and the legs are close to each other. Images of the reconstructed sequence can be seen in Fig. 18 in which the reconstructed posture (depicted in red) and the corresponding posture of the walking dummy (depicted in blue) are overlaid.
The results obtained are summarized in Table 1, where we report the posture errors averaged over all the frames of the gait cycle. We stress the results obtained with 5 cameras and 50-mm voxels: an average error of 10 mm and a maximum error of less than 16 mm. In Table 2, we briefly report the average reconstruction error per body part when five cameras and voxels of 50 mm were used. A unoptimized version of the system runs on a Pentium 500 between 7 s/frame with 50-mm voxels and 55 s/frame with



15



16



17

18

Fig. 15. The average posture error for four and five cameras. The voxel size is 20 mm
Fig. 16. The average posture error for four and five cameras. The voxel size is 35 mm
Fig. 17. The average posture error for four and five cameras. The voxel size is 50 mm
Fig. 18. The reconstructed posture (*red*) and the original posture (*blue*)

Table 1. Posture errors averaged over all frames

	Five cameras 50-mm Voxels	Five cameras 35-mm Voxels	Five cameras 20-mm Voxels	Four cameras 50-mm Voxels	Four cameras 35-mm Voxels	Four cameras 20-mm Voxels
Mean	10.259	11.910	11.900	14.492	15.944	15.717
Standard deviation	1.721	2.673	1.715	2.747	3.498	3.447
Variance	2.965	7.147	2.942	7.550	12.236	11.883
Minimum	7.160	7.896	9.070	9.921	9.982	10.184
Maximum	15.300	20.761	16.812	21.023	25.567	23.637

Table 2. The average reconstruction error per body part with five cameras and voxels of 50 mm

Body part	Mean	Body part	Mean
Neck	11.27	Breast	7.86
Right shoulder	8.09	Right pelvis	8.70
Left shoulder	8.62	Left pelvis	8.44
Right upper arm	9.16	Right thigh	8.12
Left upper arm	9.43	Left thigh	8.61
Right forearm	12.91	Right shank	10.22
Left forearm	13.70	Left shank	10.42

20 mm voxels. In fact, the computational speed was not the primary concern of this work.
Finally, it must be underlined that the posture precision is practically unaffected by the voxel size.

4.3 Experimenting with other kinds of motion

In this section, we briefly report the results of the tests on two other kinds of motion: a run through the working space and a sort of gymnastic exercise in which the dummy move its limbs and bends down. The run is 32 frames long, while the exercise sequence is 40 frames long. Some images extracted from the two sequences are shown in Figs. 19 and 20. The two sequences have been reconstructed with the same camera arrangement used for the walk sequence and three voxel resolutions (50 mm, 35 mm, and 20 mm). The average posture errors for each frame of the sequences, are given in Figs. 21 and 22. Table 3 summarizes the results obtained.
The increase of the average error is mainly due to the fact that the arms are very close to the trunk or all the limbs are close together in many frames (as at the end of the gymnastic sequence). The results obtained are still satisfactory (the best average error was 12 mm for the run sequence, and 11 mm was the best for the gymnastic sequence).

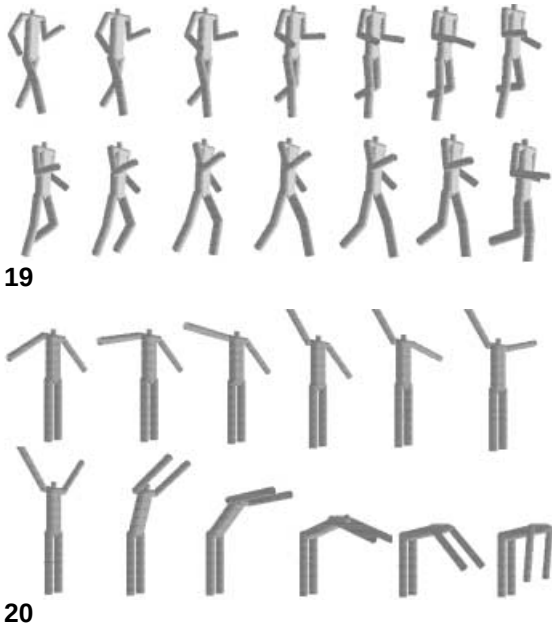
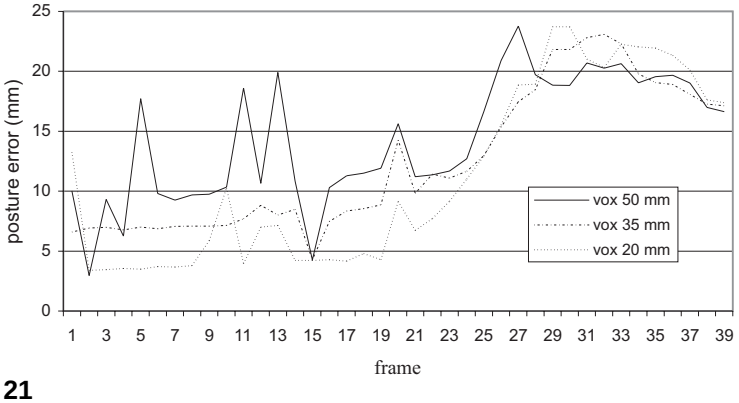


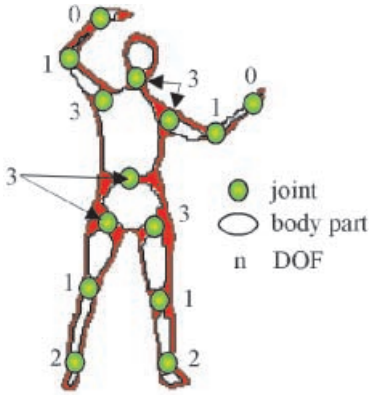
Fig. 19. Images of the running dummy
Fig. 20. Images of the gymnastic dummy

4.4 Experimenting with a more complex dummy

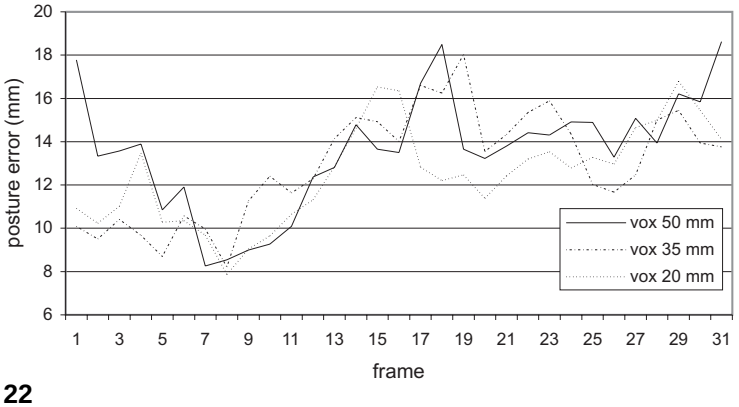
In order to verify the results obtained, we applied our technique to a more realistic dummy and a different definition of posture error. This model has 15 segments that are connected by spherical joints. The segments are organized in a tree whose root is located in the pelvis (Fig. 23), where the numbers represent the DOFs of the various segments. The total number of DOFs of the model, including the (x, y, z) position of the radix of the tree, is 32 (trunk and pelvis are both connected to the root and have 3 DOFs each).



21



23



22

24

Fig. 21. The average posture error for the gymnastic sequence
Fig. 22. The average posture error for the run sequence
Fig. 23. The human body model
Fig. 24. Model surface

Table 3. Posture errors in millimeters averaged over all frames, standard deviation, variance, maximal and minimal values of the distributions of Figs. 21 and 22

	Run with 50-mm Voxels	Run with 35-mm Voxels	Run with 20-mm Voxels	Gymnastics with 50-mm Voxels	Gymnastics with 35-mm Voxels	Gymnastics with 20-mm Voxels
Mean	13.578	12.953	12.511	14.310	12.425	11.280
Standard deviation	2.682	2.529	2.270	5.219	5.857	7.433
Variance	7.193	6.400	5.157	27.244	34.308	55.253
Minimum	8.252	8.211	7.870	2.955	4.280	3.401
Maximum	18.613	18.027	16.790	23.774	23.083	23.718

The surface is defined by means of a triangular mesh consisting of more than 600 triangles depicted in Fig. 24.

Another error function can be defined with the polygonal surface. In this case, the posture error is computed as the average of the distance between corresponding vertices of the reference model and the reconstructed model. We evaluated the posture error for several image sequences, obtained as before with simple motion algorithms (Fig. 25):

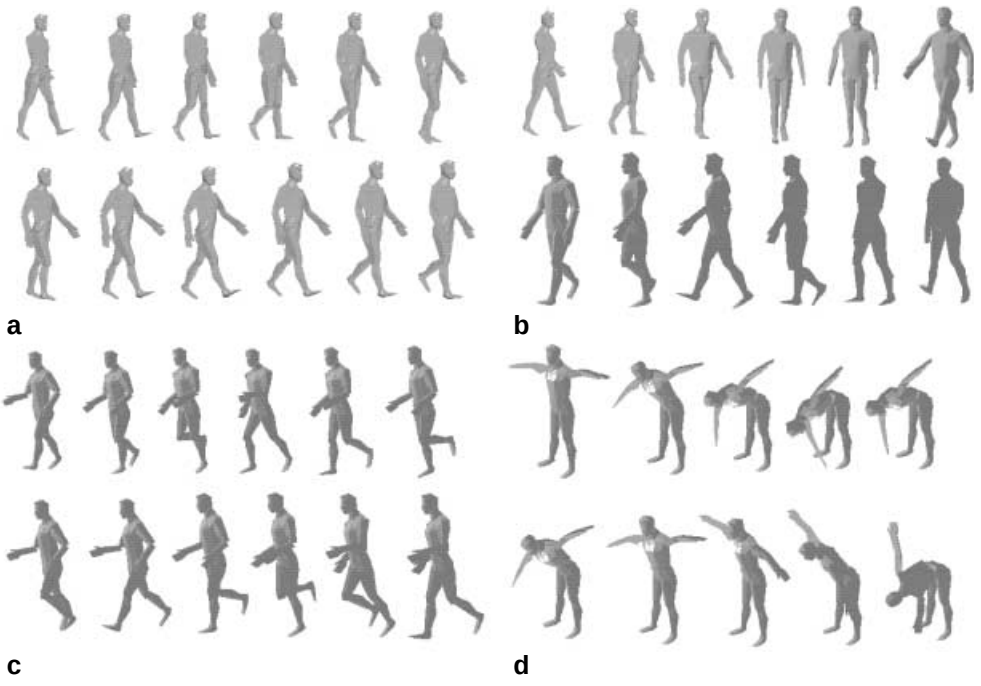


Fig. 24a. Linear walk; **b** circular walk; **c** run; and **d** gymnastic sequence

Table 4. Summary of the results for the linear walk sequence

Voxel	Mean error	Maximum error	Minimum error	Standard deviation
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 5. Summary of the results for the circular walk sequence

Voxel	Mean error	Maximum error	Minimum error	Standard deviation
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 6. Summary of the results for the run sequence

Voxel	Mean error	Maximum error	Minimum error	Standard deviation
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 7. Summary of the results for the gymnastic sequence

Voxel	Mean error	Maximum error	Minimum error	Standard deviation
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

- 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)

The volume has been reconstructed with three voxel sizes (45 mm, 35 mm, and 25 mm). Five cameras have been used for all the tests, and their orientation is the same as the arrangement named P1 in Fig. 10. The active area is 4 m $\times$ 4 m. The model used to create the motion sequences is 1.80 m high. The results obtained are summarized in Tables 4 to 7, where we

report the posture errors averaged over all the frames of the sequences. The best average error obtained for the various sequences is between 16 mm and 21 mm; that is, almost 1% of the body size. The best reconstruction for all the sequences is the one using voxels of 35 mm.

5 Concluding remarks and future work

We have shown and experimented with a technique for determining posture and motion of the human body in a virtual environment. This technique, based on multiple silhouettes obtained with a set of stationary cameras, is nonintrusive, which could be important or even necessary in many current and future applications.

The results obtained suggest the practical feasibility of the proposed approach. The experiments in a virtual world show that, using silhouettes only, model-based motion can be captured with a reasonable number of cameras, and this can be done with a precision that is apparently sufficient for many practical applications. More precisely, we have found that, although the volumes determined by four or five silhouettes are usually rather coarse, a sequence of postures with an average error of roughly 1% of the dimension of the model can be determined by fitting a model to a sequence of such volumes. Another interesting result is that the precision is relatively unaffected by reconstructing the 3D volumes at low resolution. This reduces the amount of computation required, and could be important in cases where a wide area is observed.

It could be interesting to compare the precision of the reconstruction we have found (even though in highly artificial conditions) with that of other motion capture techniques. However, this does not appear to be an easy task. One reason is that no comparable data are available as far as we know. Optical markers are tracked with millimetric precision, and similar data are claimed for magnetic tracking. However, this precision only refers to some points (at most some dozens, but usually many fewer) lying more or less close to the body, and the actual posture of the body must be worked out with nontrivial computations. The only attempt to analyze errors precisely in computer-vision-based, motion-capture studies known to the authors is described by Azarbayejani and Pentland (1996). However, their measurements only refer to the position of a hand moving along a straight trajectory of known dimensions, and such measurements cannot easily be compared with our results.

In any case, much more work, both in the virtual and real worlds, is needed for transforming the idea into an effective practical technique. Clearly, to apply our algorithms to real-world images, we must overcome a number of difficulties. One is the effect of clothing. Another important point is the size of the model. Measuring the subject could be disturbing or impossible. A possible solution we are exploring is a self-adjusting, realistic model exploiting both known poses and known movements [examples can be found in works by Kakadiaris and Metaxas (1998) and Gavrila and Davis (1996)]. Finally, real-world precision can be improved through dynamic filtering, and the silhouette data could be integrated with other image clues like optical flow and texture information.

References

1. Aggarwal JK, Cai Q (1999) Human motion analysis: a review. Comput Vision Image Understanding 73:428–440
2. Ahuja DV, Veenstra J (1989) Generating octrees from object silhouettes in orthographic views. IEEE Trans Patt Anal Machine Intell 11:137–149
3. Azarbayejani A, Pentland A (1996) Real-time self-calibrating stereo person tracking using 3D shape estimation from blob features. Patt Recognition 3:627–632
4. Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev 94:115–147
5. Blunno I, Falcione A (1996) Realistic animation of articulated models. Internal Report DAUIN-CGVG-011-96 Politecnico di Torino, Turin, Italy
6. Bregler C, Malik J (1998) Tracking people with twists and exponential maps. Proceedings of CVPR ’96 IEEE Comput Soc Conference, pp 8–15
7. Chian, Aggarwal JK (1989) Model reconstruction and shape recognition from occluding contours. IEEE Trans Patt Anal Machine Intell 11:372-389
8. Fantin S, Dias M (1994) 3D virtual mannequin database. ENEA, Document E65334-ENEA-WP2-03.00
9. Gavrila DM (1999) The visual analysis of human movement: a survey. Comput Vision Image Understanding 73:82–98
10. Gavrila DM, Davis LS (1996) 3D model-based tracking and recognition of human movement: a multi-view approach. Proceedings of CVPR ’96, IEEE Comput Soc Conference, pp 73–80
11. Kakadiaris I, Metaxas D (1998) Three-dimensional human body acquisition from multiple views. Int J Comput Vision 30:191–218

12. Laurentini A (1994) The visual hull concept for silhouette-based image understanding. IEEE Trans Patt Anal Machine Intell 16:150–162
13. Laurentini A (1995) How far 3D shapes can be understood from 2D silhouettes. IEEE Trans Patt Anal Machine Intell 17:188–195
14. Laurentini A (1997) How many 2D silhouettes does it take to reconstruct a 3D object? Comput Vision Image Understanding 67:81–87
15. Lee H, Chen Z (1985) Determination of 3D human body posture from a single view. Comput Vision, Graph Image Processing 30:148–168
16. Leung MK, Yang YH (1995) First sight: a human body outline labeling system. IEEE Trans Patt Anal Machine Intell 17:359–377
17. Marr D, Nashihara HK (1978) Representation and recognition of the spatial organization of three-dimensional shapes. Proc R Soc London B 200:269–294
18. Martin WN, Aggarwal JK (1983) Volumetric description of objects from multiple views. IEEE Trans Patt Anal Machine Intell 5:150–158
19. Mohan R, Nevatia R (1989) Using perceptual organization to extract 3D structures. IEEE Trans Patt Anal Machine Intell 11:1121–1139
20. Noborio H, Fukuda S, Arimoto S (1988) Construction of the octree approximating three-dimensional objects by using multiple views. IEEE Trans Patt Anal Machine Intell 10:769–782
21. Pentland A, Horowitz B (1991) Recovery of nonrigid motion and structure. IEEE Trans Patt Anal Machine Intell 13:730–742
22. Pentland A, Sclaroff S (1991) Closed-form solutions for physically based shape modeling and recognition. IEEE Trans Patt Anal Machine Intell 13:715–729
23. Potemesil M (1987) Generating octree models of 3D objects from their silhouettes in a sequence of images. Comput Vision Graph Image Processing 40:1–29
24. Rashid R (1980) Toward a system for the interpretation of moving light displays. IEEE Trans Patt Anal Machine Intell 2:574–581
25. Rohr K (1994) Toward model-based recognition of human movements in image sequences. CVGIP: Image Understanding 59:94–115
26. Sabel JC (1996) Optical 3D motion measurement. Proceedings of IMTC-96. Quality measurements: the indispensable bridge between theory and reality. IEEE 1:367–370
27. Solina F, Bajcsy R (1990) Recovery of parametric models from range images: the case for superquadrics with global deformation. IEEE Trans Patt Anal Machine Intell 12:131–147
28. Srinivasan P, Ping L, Hackwood S (1990) Computational geometric methods in volumetric intersection for 3D reconstruction. Patt Recog 23:843–857
29. Tsai RY (1987) A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. IEEE J Robotics Automation 3:323–344
30. Wachter S, Nagel HH (1999) Tracking persons in monocular image sequences. Comput Vision Image Understanding 74:174–192
31. Webb J, Aggarwal JK (1982) Structure from motion of rigid and jointed objects. Artif Intell 19:107–130
32. Wells M, Tutt M (1998) Practice and application of motion capture. IEE Colloquium on Computer Vision for Virtual Human Modelling Savoy Place, London (Ref. No. 1998/433), pp 6/1–6/3
33. Wren C, Azarbayejani A, Darrell T, Pentland A (1997) Pfinder: real-time tracking of the human body. IEEE Trans Patt Anal Machine Intell 19:780–785
34. Zheng J (1994) Acquiring 3D models from sequences of contours. IEEE Trans Patt Anal Machine Intell 16:163–177

ANDREA BOTTINO was born in Torino, Italy, in 1971. He received his Master's Degree in Computer Science Engineering and his PhD from Politecnico di Torino in 1995 and 2000. He is currently a Teaching Assistant in Computer Science at the Dipartimento di Automatica e Informatica. His research interests include computer graphics, computer vision, motion capture systems, and object-oriented technology.

ALDO LAURENTINI was born in Genova and received the degree of Ingegneria Elettronica from the Politecnico di Milano in 1963. In 1965 he joined the Politecnico di Torino, where he is now a Professor of Computer Science at the Dipartimento di Automatica ed Informatica. He is a member of the IEEE and ACM, and author of more than fifty scientific papers. His current research interests include computer vision, computer graphics, and computational geometry.