Non-Intrusive Silhouette Based Motion Capture

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

ABSTRACT

A growing number of promising applications requires recognizing posture and motion of humans. Conventional techniques require to attach foreign objects to the body, which in several applications is disturbing or impossible. In this paper we present a new non intrusive motion capture approach. It is able to reconstruct unconstrained human motion by means of silhouettes extracted from multiple-viewpoint images. Silhouettes are easy to obtain from intensity images, can directly provide a 3D reconstruction of the body and drive model-based motion capture. We also present results about the precision of the overall process, obtained in a virtual environment.

Keywords: Silhouettes, Volume intersection, Model based recognition, Motion Capture

1. INTRODUCTION

Capturing posture and motion of the human body is an important practical issue. Several applications already exist, and many others are foreseen. Among them: virtual reality (character animation, games, interactive virtual worlds), sport performance analysis and athletics training, clinical study of orthopedic patients, choreography, smart surveillance systems, gesture driven user interfaces, video annotation.

Several commercial MC systems exist based on optical or magnetic tracking of sensors attached to the body of the performer. However, in many application areas it is disturbing or impossible to attach foreign objects to the body of the subject. This calls for a new non intrusive technique.

A number of non intrusive MC techniques have been reported in the literature, different for the data used and for the approach to motion recovery. The reader is referred to [1], [6] and [11] for comprehensive references. However, as far as we know, these research did not result yet in practical devices.

The purpose of the authors is to develop an alternative approach sufficiently simple and robust to allow the implementation of practical equipment. The approach presented is based on multiple 2D silhouettes of the body extracted from 2D images.

The outline of our approach is 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*). 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.
- The above procedure is applied to each frame of the motion sequence. Exception made for the first time, at each frame the starting position of the model is that obtained for the previous frame. This reduces the computation and exploits continuity to avoid local minima.
- 3D poses are recovered as 3D rotations of the joints of the model. Once tracking is successfully completed the 3D joint angles can be used for reproducing motion or for template matching.

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 ([8], [15]). *R* is a bounding volume which more or less closely approximates *O*, depending on the viewpoints and the object itself.

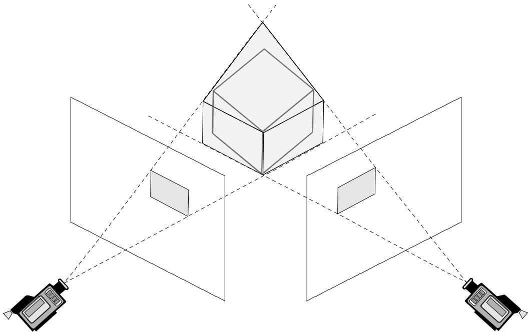

Fig. 1: the VI technique

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 correspondences between multiple images.

However, using VI for reconstructing the human body requires to face several difficulties. Bad placement and 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.

3. The Motion Capture System

First we describe the model of the human body used for pose recovery. 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 details together with the posture reconstruction algorithm.

The Model

Many approaches to motion recovery are model-based. The human body is represented with some kind of model whose 3-D posture and motion is matched with the physical data. Stick articulated models, as in [9], idealize the human skeleton. Ellipsoidal blobs [5], cylinders and generalized cylinders [11], deformed superquadrics [7], geons [3], parametric solids and finite elements [10], have been used to build models which mimic more or less closely the human body.

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

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

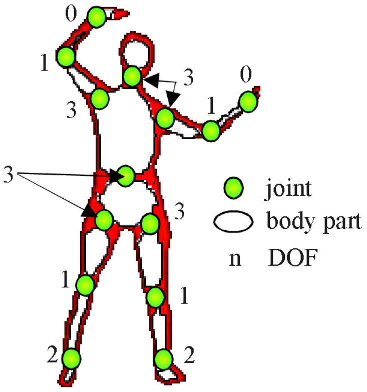

Fig. 2: the human body model

Body segments are organized in a tree whose root is located in the pelvis (see Fig. 2). 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 the physic 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.

The surface: The surface is defined through a triangular mesh consisting of more than 600 triangles depicted in Fig. 3. The complete set of shape parameters can be arranged to match the characteristics of the real performer. This surface representation has an high level of accuracy.

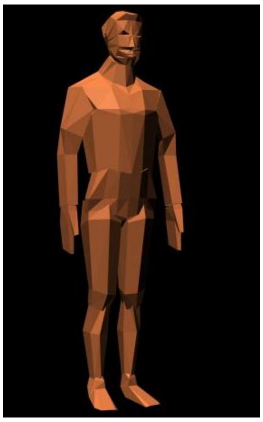

Fig. 3: model surface

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 [12] 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. 4), that is an indeformable structure containing a grid of squares of known position and dimension, whose centers are the reference points. The squares have different colors and are arranged on patterns that allows 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.

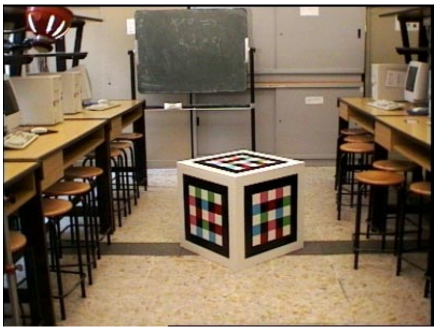

Fig. 4: image of the calibration object

Our approach to silhouette extraction is mainly based on the ideas presented in [13], [14]. 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 is 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 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 pre-processing 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:

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

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 α_{tol} < 1. For each component of pixel $\boldsymbol{p}$ we evaluate the inequality:

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

If the inequality is false for both U and V component or false for Y and at least one of the UV component, 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 to the background the pixel statistic is updated using a simple adaptative filter:

$$\boldsymbol{\mu}_t = \alpha \cdot \boldsymbol{p} + (1-\alpha) \cdot \boldsymbol{\mu}_{t-1}$$

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 then the expected value. In this case, we consider a second threshold for Y, given by:

$$T_s(x,y) = \alpha_{shadow} \cdot T_Y(x,y) \tag{2}$$

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

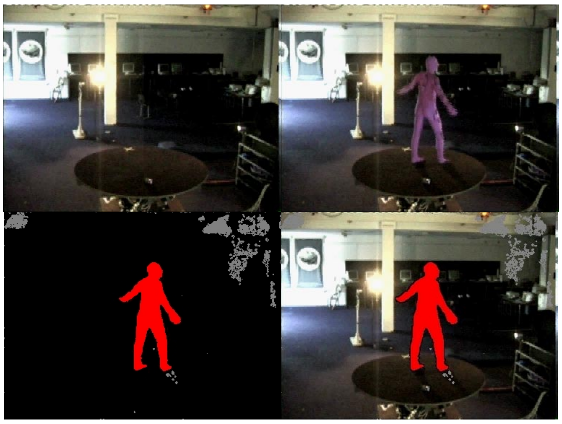

Fig. 5 (a-d): the scene model, a frame of the sequence and the extracted silhouette

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

In Fig. 5a is depicted the scene model built for one of the cameras used into the test sequence, while Fig. 5b is shown one of the frames of the sequence, in which a person moves into the active area. Fig. 5c-d show the result of the silhouette extraction process.

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 dummy. 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 $\mathrm{C_i}$ to the closest segment of the model.

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

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

The contribution to $D(\wp,R)$ of each segment depends on the number of voxels assigned to the segment and on the dimension of the correspondent 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.

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

In order to reduce the number of computations required, each segment is approximated by an oriented bounding ellipsoid (OBE, see Fig. 6). 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.

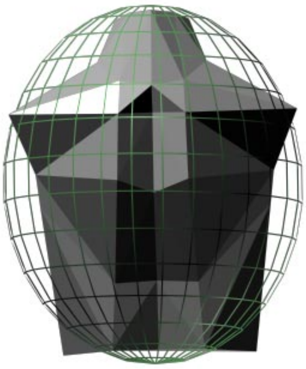

Fig. 6: oriented bounding ellipsoid (OBE) of the trunk

To recover the motion of the dummy, the above procedure is applied to each frame of the motion sequence. Exception made for the first time, the starting position of the model is that obtained for the previous frame. Since each time the dummy is close to its final position, 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.

5. EXPERIMENTAL RESULTS

The experimental work has been divided in 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.

Accuracy in a Virtual Environment

In order to cover many significant postures and typical movements, we evaluated the reconstruction accuracy for several different image sequences (see Fig. 7), that is:

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

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 camera located in an horizontal plane, one meter and a half above the floor and the fifth shooting the dummy from above). The active area is 4x4 meters wide.

The model used to create the motion sequences is 1.80 m high. The results obtained are summarized from Table 1 to Table 3, where we report the posture errors averaged over all the frames of the sequences. Diagram 1 to Diagram 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.

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

Recovering model postures from real image sequences

The video sequences we have used in our tests are courtesy of MOTEK Motion Technology of Amsterdam (The Netherlands). We 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 the starting video frames containing the flash.

The actor performed freely in the work area since we wanted to test our approach for real unconstrained motion. Even if it was not possible to put a camera over the head of the performer, the reconstructed sequence looks satisfactory when seen at real frame rate (25 frame/s). Fig. 9 shows the composition of real and virtual model (Fig. 8 contains the same frames with only the real performer).

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 and its mean value for the first 55 frames is 8.9 mm.

6. CONCLUSIONS AND FUTURE WORKS

We have demonstrated an approach that is able to reconstruct unconstrained human motion in realistic situations without requiring 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 silhouette the performer can be reconstructed with a technique known as *Volume Intersection*. The 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 centimeters (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 also for real images. 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 highly artificial condition) with that of other motion capture techniques. However, this does not appear an easy task. A 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 whit real images, but usually no precision data are available. Clearly, the reason is that this would require to know the true posture. The only attempt to perform precise error analysis known to the authors is described in [2]. However, the measurements reported only refer to the position of a hand and are not easily comparable with our results.

In order to improve our technique, we are considering several issues.

In our approach we only addressed the problem of pose recovery: as a matter of fact, we assumed that the 3D model was fully specified a priori. We are currently developing an automatic measurement process to calibrate the model by means of an initialization stage which exploits both known poses and known movements.

Dynamic filtering (such as Kalman filtering) can remove noisy component of the recovered sequence and could be used to boost the reconstruction process. Finally, we will attempt to extend our technique beyond stationary cameras, which could be of paramount importance for many applications.

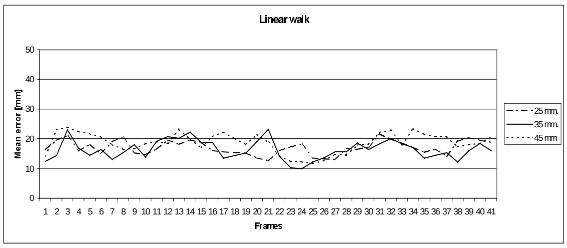

Diagram 1: avg. posture error in mm for linear walk sequence

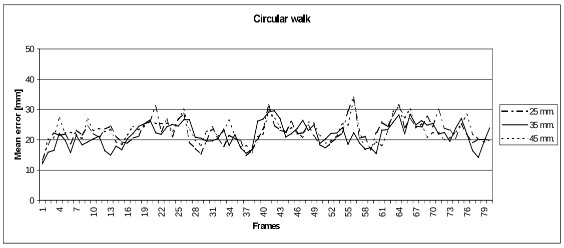

Diagram 2: avg. posture error in mm for circular walk sequence

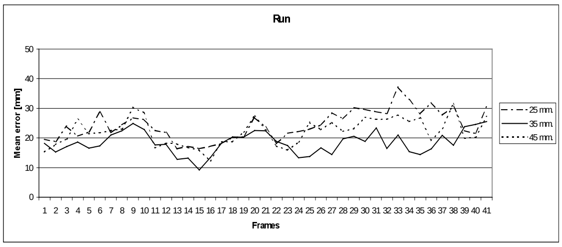

Diagram 3: avg. posture error in mm for run sequence

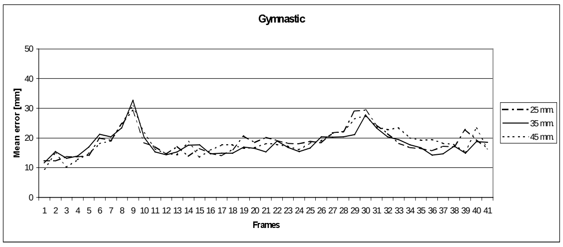

Diagram 4: avg. posture error in mm for gymnastic sequence

Voxel	Mean Err.	Max Err.	Min. Err.	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 1: summary results for linear walk sequence

Voxel	Mean Err.	Max Err.	Min. Err.	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 2: summary results for circular walk sequence

Voxel	Mean Err.	Max Err.	Min. Err.	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 3: summary results for run sequence

Voxel	Mean Err.	Max Err.	Min. Err.	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

Table 4: summary results for gymnastic sequence

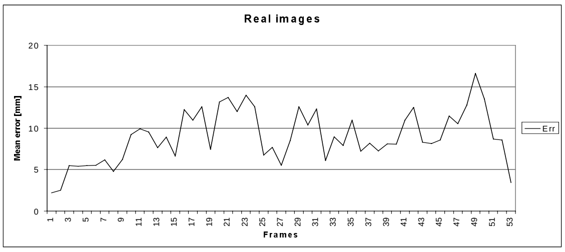

Diagram 5: avg. posture difference for reconstruction of real image sequences using 30 and 50 mm voxels

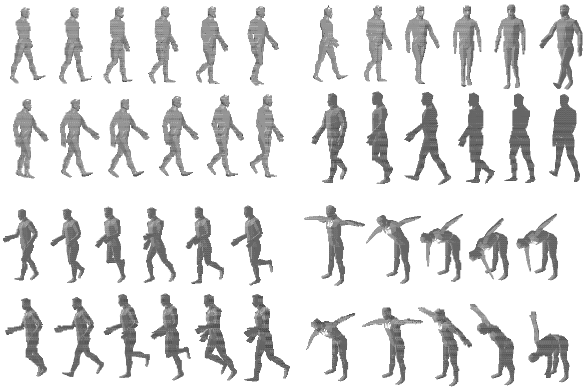

Fig. 7 (a-d): linear walk, circular walk, run and gymnastic sequence

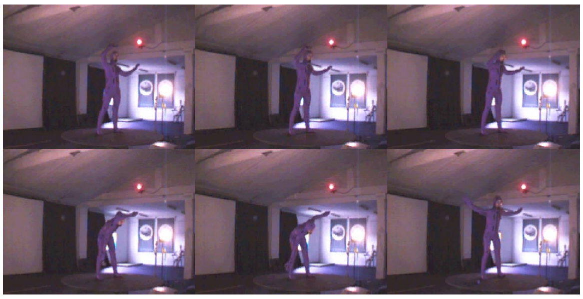

Fig. 8: outtakes from camera 2

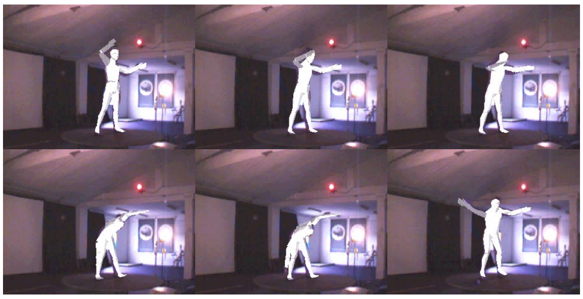

Fig. 9: reconstructed postures

7. References

[1] Aggarwal JK, Cai Q, Human motion analysis: a review, CVGIP: Image Understanding, 73(3), 1999, pp. 428-440

[2] Azarbayejani A, Pentland A, Real-time self-calibrating stereo person tracking using 3D shape estimation from blob features, Proc. of International Conf. on PR, Vienna, 1996.

[3] Biederman, Recognition-by-components: A theory of human image understanding, Psychological Rev. 94, 1987, pp.115-147

[4] Bottino A, Laurentini A, Zuccone P, Toward Non-intrusive Motion Capture, Proc. of third Asian Conf. on Computer Vision, Hong Kong, China, (1), 1998, pp. 417-423

[5] Bregler C, Malik J, Tracking People with twists and exponential maps, Proc. IEEE Conf. on CVPR, 1998, pp. 8-15

[6] Gavrila DM, The Visual Analysis of Human Movement: A Survey, Computer Vision and Image Understanding 73(1), 1999, pp. 82-98

[7] Gavrila DM, Davis LS, 3D Model-based Tracking and Recognition of Human Movement: a Multi-view Approach, Proc. IEEE CS Conf. On CVPR, San Francisco, CA, 1996, pp. 73-80

[8] Laurentini A, How far 3D shapes can be understood from 2D silhouettes, IEEE transactions on PAMI, 17(2), 1995, pp. 188-195

[9] Leung, Yang Y, First sight: a human body outline labeling system, IEEE Trans. on PAMI 17, 1995, pp. 359-377

[10] Pentland, Sclaroff S, Closed-form solutions for physically based shape modeling and recognition, IEEE Trans. on PAMI 13, 1991, 715-729

[11] Rohr K, Toward model-based recognition of human movements in image sequences, CVGIP: Image Understanding 59, 1994, pp. 94-115

[12] Tsai, A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. IEEE Journal of Robotics and Automation RA 3, 1987, pp. 323-344

[13] Wren C, Pentland A, Dynamic models of human motion, Proc. of third IEEE International conf. on Automatic Face and Gesture Recognition, 1998, Nara, Japan, pp. 22-27

[14] Yamada M, Kazuyuki E, Ohya J, A new robust real-time method for extracting human silhouettes from color images, 1998, pp. 528-53

[15] Zheng J, Acquiring 3D models from sequences of contours, IEEE transactions on PAMI 16(2), 1994, pp. 163-177