in Proc. IEEE Computer Vision and Pattern Recognition, San Francisco, 1996

3-D model-based tracking of humans in action: a multi-view approach

D.M. Gavrila and L.S. Davis
Computer Vision Laboratory, CfAR,
University of Maryland
College Park, MD 20742, U.S.A.
{gavrila,lsd}@cfar.umd.edu
http://www.umiacs.umd.edu/users/{gavrila,lsd}/

Abstract

We present a vision system for the 3-D model-based tracking of unconstrained human movement. Using image sequences acquired simultaneously from multiple views, we recover the 3-D body pose at each time instant without the use of markers. The pose-recovery problem is formulated as a search problem and entails finding the pose parameters of a graphical human model whose synthesized appearance is most similar to the actual appearance of the real human in the multi-view images. The models used for this purpose are acquired from the images. We use a decomposition approach and a best-first technique to search through the high dimensional pose parameter space. A robust variant of chamfer matching is used as a fast similarity measure between synthesized and real edge images.

We present initial tracking results from a large new Humans-In-Action (HIA) database containing more than 2500 frames in each of four orthogonal views. They contain subjects involved in a variety of activities, of various degrees of complexity, ranging from the more simple one-person hand waving to the challenging two-person close interaction in the Argentine Tango.

1 Introduction

The ability to recognize humans and their activities by vision is a key feature in the pursuit to design a machine capable of interacting intelligently and effortlessly with a human-inhabited environment. Besides this long-term goal, there are many applications possible in the more near term, e.g. in virtual reality, "smart" surveillance systems, motion analysis in sports, choreography of dance and ballet, sign language translation and gesture-driven user interfaces. In many of these applications a non-intrusive sensory method based on vision is preferable over a (in some cases not even feasible) method that relies on markers attached to the bodies of human subjects.

Our approach to looking at humans and recognizing their activities has two major components:

1. body pose recovery and tracking
2. recognition of movement patterns

We consider the case where we have multiple stationary (visible-light) cameras, previously calibrated, and we observe one or more humans performing some action from multiple viewpoints. The aim of the first component is to reconstruct from the sequence of multi-view frames the (approximate) 3-D body pose of the human(s) at each time instant; this serves as input to the movement recognition component. In an earlier paper [5] we considered movement recognition as a classification problem and we used a Dynamic Time Warping method to match a test sequence with several reference sequences representing prototypical activities. The features used for matching were various 3-D joint angles of the human body. In this paper, we deal only with the pose recovery and tracking component.

The outline of this paper is as follows. First, Section 2 provides a motivation for our choice of a 3-D recovery approach rather than a 2-D approach. In Section 3 we discuss 3-D human modeling issues and the (semi-automatic) model acquisition procedure used by our system. Section 4 deals with the pose recovery and tracking component. Included is a bootstrapping procedure to start the tracking or to re-initialize it if it fails. Section 5 presents new experimental results in which successful unconstrained whole-body movement is demonstrated on two subjects. These are initial results [1] derived from a large Humans-In-Action (HIA) database containing two subjects involved in a variety of activities, of various degree of complexity. We discuss our results and possible improvements in Section 6. Finally, Section 7 contains our conclusions.

2 2-D vs. 3-D

One may question whether it is desirable or feasible to try to recover 3-D body pose from 2-D image sequences for the purpose of recognizing human movement. An alternative approach is to work directly with 2-D features derived from the images, using some form of 2-D model [7] [11] or not [3] [16].

Recognition systems using 2-D model-free features have been able to claim early successes in matching human movement patterns. For constrained types of human movement (such as walking parallel to the image plane, involving periodic motion), many of these features have been successfully used for classification, as in [16]. This may indeed be the easiest and best solution for several applications. But we find it unlikely that reliable recognition of more unconstrained and complex human movement (e.g. humans wandering around, making different gestures while walking and

[1] The tracking results described in this paper are also available as video clips from our home pages.

turning) can be achieved using these types of features exclusively. With respect to using 2-D model-based features, we note that few systems actually derive the features they use for movement matching. Self-occlusion makes the 2-D tracking problem hard for arbitrary movements and thus existing systems assume some a-priori knowledge of the type of movement and/or the viewpoint under which it is observed. 2-D labeling and tracking under more general conditions is attempted by [11].

We therefore investigate in this paper the more general-purpose approach of recovering 3-D pose through time, in terms of 3-D joint angles defined with respect to a human-centered 3-D motion recovery from 2-D images is often an ill-posed problem. In the case of 3-D pose tracking, however, we can take advantage of the large available a-priori knowledge about the kinematic and shape properties of the human body to make the problem tractable. Tracking also is well supported by the use of a 3-D human model which can predict events such as (self) occlusion and (self) collision. Once 3-D tracking is successfully completed, we have the benefit of being able to use the 3-D joint angles as features for movement matching, which are viewpoint independent and directly linked to the body pose. Compared with 3-D joint coordinates, they are less sensitive to variations in the size of humans.

The techniques described in this paper lead to tracking on a fine scale, with the obtained joint angles being within a few degrees of their true values. Besides providing meaningful generic features for a movement matching component, such techniques are of independent interest for their use in virtual reality applications. In other applications, such as surveillance, continuous fine-scale 3-D tracking will not always be necessary, and can be combined with tracking on a more coarse level (for example, considering the human body as a single unit), changing the mode of operation from one to another depending on context.

3 3-D body modeling and model acquisition

3-D graphical models for the human body generally consist of two components: a representation for the skeletal structure (the "stick figure") and a representation for the flesh surrounding it. The stick figure is simply a collection of segments and joint angles with various degree of freedom at the articulation sites. The representation for the flesh can either be surface-based (using polygons, for example) or volumetric (using cylinders, for example). There is a trade-off between the accuracy of representation and the number of parameters used in the model. Many highly accurate surface models have been used in the field of graphics [1] to model the human body, often containing thousands of polygons obtained from actual body scans. In vision, where the inverse problem of recovering the 3-D model from the images is much harder and less accurate, the use of volumetric primitives has been preferred to "flesh out" the segments because of the lower number of model parameters involved.

For our purposes of tracking 3-D whole-body motion, we currently use a 22-DOF model (3 DOF for the positioning of the root of the articulated structure, 3 DOF for the torso and 4 DOF for each arm and each leg), without modeling the palm of the hand or the foot, and using a rigid head-torso approximation. See [1] for more sophisticated modeling. Regarding shape, we felt that simple cylindrical primitives (possibly with elliptic XY-cross-sections) [4] [8] [18] would not represent body parts such as the head and torso accurately enough. Therefore, we employ the class of *tapered super-quadrics* [12]; these include such diverse shapes as cylinders, spheres, ellipsoids and hyper-rectangles. So far, we have obtained satisfactory modeling results with these primitives alone (see experiments); a more general approach also allows deformations of the shape primitives [12] [14].

We derive the shape parameters from the projections of occluding contours in two orthogonal views, parallel to the zx- and zy-planes. This involves the human subject facing the camera frontally and sideways. We assume 2-D segmentation in the two orthogonal views; a way to obtain such a segmentation is proposed in [10]. Back-projecting the 2-D projected contours of a quadric gives the 3-D occluding contours, after which a coarse-to-fine search procedure is used over a reasonable range of parameter space to determine the best-fitting quadric. Fitting uses chamfer matching (see the next section) as a similarity measure between the fitted and back-projected occluding 3-D contours. Figure 3 shows frontal and side views of the recovered torso and head for two persons: DARIU and ELLEN. Figure 4 shows their complete recovered models in a graphics rendering. These models are used in the tracking experiments of Section 5.

4 Pose recovery and tracking

The general framework for our tracking component is inspired by the early work of O'Rourke and Badler [19] and is illustrated in Figure 1a. Four main components are involved: prediction, synthesis, image analysis and state estimation. The prediction component takes into account previous states up to time t to make a prediction for time $t+1$. It is deemed more stable to do the prediction at a high level (in state space) than at a low level (in image space), allowing an easier way to incorporate semantic knowledge into the tracking process. The synthesis component translates the prediction from the state level to the measurement (image) level, which allows the image analysis component to selectively focus on a subset of regions and look for a subset of features. Finally, the state-estimation component computes the new state using the segmented image.

The above framework is general and can also be applied to other model-based tracking problems. In the remainder of this section, we discuss how the components are implemented in our system for the case of tracking humans, and how this relates to ex-

isting work. In the first subsection we cover the pose estimation component, the second subsection briefly covers the other components.

4.1 Pose estimation

Our approach to pose recovery is based on a generate-and-test strategy. The problem is formulated as a search problem and entails finding the pose parameters of a graphical human model whose synthesized appearance is most similar to the actual appearance of the real human. (see Figure 1b). This approach has the advantage that the measure of similarity between synthesized appearance and actual appearance can now be based on whole contours and/or regions rather than on a few points. So far, existing systems which work on real images using this strategy have had limitations: Perales and Torres [15] describe a system which involves input from a human operator. Hogg [8] and Rohr [18] deal with the restricted movement of walking parallel to image plane, for which the search space is essentially one-dimensional. Downton and Drouet [4] attempt to track unconstrained upper-body motion, but must conclude that the tracking gets lost due to propagation of errors. Goncalves *et al.* [6] use a Kalman-filtering approach to track arm movement from single-view images where the shoulder remains fixed. Finally, work by Rehg and Kanade [17] is geared towards finger tracking. We aim to improve the previous approaches, where applicable, along the following lines.

- Similarity measure

In our approach the similarity measure between model view and actual scene is based on arbitrary edge contours rather than on straight line approximations (as in [18], for example); we use a robust variant of *chamfer matching* [2]. The *directed* chamfer distance $DD(T,R)$ between a test point set T and a reference point set R is obtained by summing the distances between each point in set T to its nearest point in R

$$DD(T,R) = \sum_{t \in T} dd(t,R) = \sum_{t \in T} min_{r \in R} \parallel t - r \parallel \tag{1}$$

and its normalized version is

$$\overline{DD}(T,R) = DD(T,R)/|T| \tag{2}$$

$DD(T,R)$ can be efficiently obtained in a two-pass process by pre-computing the chamfer distance on a grid to the reference set. The resulting distance map is the so-called "chamfer image" (see Figures 6b and 6c). It would be efficient if we could use only $DD(M,S)$ during pose search (as done in [2]), where M and S are the projected model edges and scene edges, respectively. In that case, the scene chamfer image would have to be computed only once, followed by fast access for different model projections. However, using this measure alone has the disadvantage (which becomes apparent in experiments) that it does not contain information about how close the reference set is to the test set. For example, a single point can be really close to a large straight line, but we may not want to consider the two entities very similar. We therefore use the *undirected* normalized chamfer distance $\overline{D}(T,R)$

$$\overline{D}(T,R) = (\overline{DD}(T,R) + \overline{DD}(R,T))/2 \tag{3}$$

A further modification is to perform outlier rejection on the distribution $dd(t,R)$. Points t for which $dd(t,R) > \theta$ are rejected outright; the mean μ and standard deviation σ of the resulting distribution is used to reject points t for which $dd(t,R) > \mu + 2\sigma$.

We note that other measures could (and) have been used to evaluate a hypothesized model pose, which work directly on the scene image: correlation (see [6] and [17]) and average contrast value along the model edges (a measure commonly used in the snake literature). The reason we opted for preprocessing the scene image (i.e. applying an edge detector) and chamfer matching is that it provides a gradual measure of similarity between two contours while having a long-range effect in image space. It is gradual since it is based on distance contributions of many points along both model and scene contours; as two identically contours are moved apart in image space the average closest distance between points increases gradually. This effect is noticeable over a range up to threshold θ, in the absence of noise. The two factors, graduality and long-range, make (chamfer) distance mapping a suitable evaluation measure to guide a search process. Correlation and average contrast along a contour, on the other hand, typically provide strong peak responses but rapidly declining off-peak responses.

- Multiview approach

By using a multi-view approach we achieve tighter 3-D pose recovery and tracking of the human body than from using one view only; body poses and movements that are ambiguous from one view can be disambiguated from another view. We synthesize appearances of the human model for all the available views, and evaluate the appropriateness of a 3-D pose based on the similarity measures for the individual views (see Figure 1b).

- Search

Search techniques are used to prune the high dimensional pose parameter space (see also [13]). We currently use *best-first* search; we do this because a reasonable initial state can be provided by a prediction component during tracking or by a bootstrapping method at start-up. The use of a well-behaved similarity measure derived from multiple views, as discussed before, is likely to lead to a search landscape with fairly wide and pronounced maxima around the correct parameter values; this can be well detected by a local search technique such as best-first. Nevertheless, the fact remains that the search-space is very large and high-dimensional (22 dimensions per human, in our case); this makes "straight-on" search daunting. The proposed solution to this

is *search space decomposition*. Define the original N-dimensional search space Σ at time t as

$$\Sigma = \{\{p_1\} \times .. \times \{p_N\}\},$$
$$\{p_i\} = \{\hat{p_i} - \Delta_{1i}, .., \hat{p_i} + \Delta_{2i}\}, \ \ step\ \Delta_{3i} \quad (4)$$

where $\hat{\mathbf{P}} = (\hat{p}_1, .., \hat{p}_N)$ is the state prediction for time t. We define the decomposed search space Σ^* as

$$\Sigma^* = (\Sigma_1, \Sigma_2) \quad (5)$$
$$\Sigma_1 = \{\{p_{i_1}\} \times .. \times \{p_{i_M}\} \times \{\hat{p}_{i_{M+1}}\} \times .. \times \{\hat{p}_{i_N}\}\} \quad (6)$$
$$\Sigma_2 = \{\{\tilde{p}_{i_1}\} \times .. \times \{\tilde{p}_{i_M}\} \times \{p_{i_{M+1}}\} \times ... \times \{p_{i_N}\}\} \quad (7)$$

where $(\tilde{p}_{i_1}, .., \tilde{p}_{i_M})$ is derived from the best solution to searching Σ_1. The above search space decomposition can be applied recursively and can be represented by a tree in which non-leaf nodes represent search spaces to be further decomposed and leaf nodes are search spaces to be actually processed. The recursive scheme we propose for the pose recovery of K humans is illustrated in Figure 2. In order to search for the pose of the i-th human in the scene we synthesize humans 1, ..., $i-1$ with the best pose parameters found earlier, and synthesize humans $i+1$, ..., K with their predicted pose parameters. We search for the best torso/head configuration of the i-th human while keeping the limbs at their predicted values, etc.

We have found in practice that it is more stable to include the torso-twist parameter in the arms (or legs) search space, instead of in the torso/head search space. This is because the observed contours of the torso alone are not very sensitive to twist. Given that we keep the root of the articulated figure fixed at the torso center, the dimensionalities of the search spaces we actually search are 5, 9, and 8, respectively.

- Initialization

Our bootstrapping procedure for starting the tracking currently handles the case where moving objects (i.e. humans) do not overlap and are positioned against a stationary background. The procedure starts with background subtraction, followed by a thresholding operation to determine the region of interest; see Figure 5. This operation can be quite noisy, as shown in the figure. The aim is to determine from this binary image the major axis of the region of interest; in practice this is the axis of the prevalent torso-head configuration. Together with the major axis of another view, this allows the determination of the major 3-D axis of the torso. Additional constraints regarding the position of the head along the axis (currently, implemented as a simple histogram technique) allow a fairly precise estimation of all torso parameters, with the exception of the torso twist which is searched for, together with the arms/legs parameters, in a coarse to fine fashion.

The determination of the major axis can be achieved robustly by iteratively applying a principal component analysis (PCA) [9] on data points sampled from the region of interest. This process results in the removal of the data points corresponding to the hands if they are located lateral to the torso, and also of other types of noise. In Figure 5 the successive approximations to the major axis are shown by straight lines in increasingly light colors.

4.2 The other components

Our prediction component works in batch mode and uses a constant acceleration model for the pose parameters. In other words, a second degree polynomial is fitted at times $t, ..., t-T+1$, and its extrapolated value at time $t+1$ is used for prediction. The synthesis component uses a standard graphics renderer to give the model projections for the various camera views. Finally, the image analysis component applies an edge detector to the real images, performs linking, and groups the edges into constant curvature segments. These segments are each considered as a unit and either accepted or rejected into the filtered scene edge map, a decision which is based on their directed chamfer distances to the projected model edges; see Figure 6. This process facilitates the removal of unwanted contours which could disturb the scene chamfer image (in Figure 6, for example, background edges around the head area in the original edge image are absent in the filtered edge image).

5 Experiments

We compiled a large data base containing multi-view images of human subjects involved in a variety of activities. These activities are of various degrees of complexity, ranging from single-person hand waving to the challenging two-person close interaction of the Argentine Tango. The data was taken from four (near-) orthogonal views (FRONT, RIGHT, BACK and LEFT) with the cameras placed wide apart in the corners of a room for maximum coverage; see Figure 7. The background is fairly complex; many regions contain bar-like structures and some regions are highly textured (observe the two VCR racks in lower-right image of Figure 7). The subjects wear tight-fitting clothes. Their sleeves are of contrasting colors, simplifying edge detection somewhat in cases where one body part occludes another.

Because of disk space and speed limitations, the more than one hour's worth of image data was first stored on (SVHS) video tape. A subset of this data was digitized (properly aligned by its time code (TC)) and makes up the HIA database, which currently contains more than 2500 frames in each of the four views.

The cameras were calibrated using an iterative, non-linear least squares method developed by Szeliski and Kang [20] and kindly made available to us. Figure 7 illustrates the outcome; the epipolar lines shown in the RIGHT, BACK and LEFT views correspond to the selected points in the FRONT view. One can see that corresponding points lie very close to or on top of the epipolar lines. Observe how all the epipolar lines emanate from one single point in the BACK view: the FRONT camera center lies within its view.

Our system is implemented under A.V.S. (Advanced Visualization System). Following its data flow network model, it consists of independently running modules, receiving and passing data through their interconnections. The implemented A.V.S. net-

work bears a close resemblance to Figure 1(b). The parameter space was bounded in each angular dimension by $\pm$ 15 degrees, and in each spatial dimension by $\pm$ 10 cm around the predicted parameter values. The discretization was 5 degrees and 5 cm, respectively. We kept these values constant during tracking.

Figure 9 (a)-(c) illustrates tracking for persons DARIU and ELLEN. The movement performed can be described as raising the arms sideways to a 90 degree extension, followed by rotating both elbows forward. Moderate opposite torso movement takes place for balancing as arms are moved forward and backwards. The current recovered 3-D pose is illustrated by the projection of the model in the four views, shown in white. The displayed model projections include for visual purposes the edges at the intersections of body parts; these were not included in the chamfer matching process. It can be seen that tracking is quite successful, with a good fit for the recovered 3-D pose of the model for the four views. Figure 8 shows some of the recovered pose parameters for the DARIU sequence.

6 Discussion

As we process more sequences of our HIA database our aim is to be able to process the more complex sequences, involving fast-varying poses, multiple bodies and close interaction (see for example Figure 9(d)). We consider several improvements to our system. On the image processing level, we are interested in a tighter coupling between prediction and segmentation. Currently, the image processing component applies a general- purpose edge-detector and uses prediction only for filtering purposes. We are interested in more actively using the prediction information through the use of deformable templates. On the algorithmic level, we are interested in methods of further constraining the search space, based on either image flow or stereo correspondence.

7 Conclusions

We have presented a new vision system for the 3-D model-based tracking of unconstrained human movement from multiple views. A large Humans In Action database has been compiled for which initial tracking results were shown. We draw the following two conclusions from these initial experimental results. First, our calibration and human modeling procedures support a (perhaps surprisingly) good 3-D localization of the model such that its projection matches the all-around camera views. This is good news for the feasibility of *any* multi-view 3-D model-based tracking method, not just ours. Second, the proposed pose recovery and tracking method based on, among others, the chamfer distance as similarity measure, is indeed able to maintain a good fit over time. This is encouraging as we turn to the more complex sequences.

8 Acknowledgements

We would like to thank Ellen Koopmans, P.J. Narayanan and Pete Rander for their help in acquiring the Humans-In-Action database at CMU's 3-D Studio. This work was supported by the Advanced Research Projects Agency (Order No. C635) and by the Office of Naval Research (Grant N00014-95-1-0521).

References

[1] N.I. Badler, C.B. Phillips, and B.L. Webber, "Simulating Humans," Oxford University Press, Oxford, UK, 1993.

[2] H.G. Barrow *et al.,* "Parametric Correspondence and Chamfer Matching: Two New Techniques For Image Matching," *Proc. IJCAI,* vol.2, pp.659-663, 1977.

[3] T. Darrell and A. Pentland, "Space-Time Gestures," *Looking at people, Proc. IJCAI,* Chambery, France, 1993.

[4] A.C. Downton and H. Drouet, "Model-Based Image Analysis for Unconstrained Upper-Body Motion," *Proc. Int. IEE Conf. on Image Processing and its Applications,* pp. 274-277, 1992.

[5] D. M. Gavrila and L.S. Davis, "Towards 3-D Model-based Tracking and Recognition of Human Movement," *Proc. Int. Work. on Face and Gesture Recognition*, Zurich, Switzerland, 1995.

[6] L. Goncalves *et al.,* "Monocular Tracking of the Human Arm in 3D," *Proc. ICCV,* pp.764-770, 1995.

[7] Y. Guo, G. Xu and S. Tsuji, "Understanding Human Motion Patterns," *Proc. ICPR,* 1994.

[8] D. Hogg, "Model Based Vision: A Program to See a Walking Person," *Image and Vision Computing,* vol.1, nr.1, pp.5-20, 1983.

[9] I.T. Jolliffe, *Principal Component Analysis*, Springer Verlag, New York, 1986.

[10] I. Kakadiaris and D. Metaxas, "3D Human Body Model Acquisition from Multiple Views," *Proc. ICCV*, 1995.

[11] M.K. Leung and Y.H. Yang, "First Sight: A Human Body Outline Labeling System," *IEEE Trans. on PAMI,* vol.17, no.4, pp.359-377, 1995.

[12] D. Metaxas and D. Terzopoulos, "Shape and Nonrigid Motion Estimation through Physics-Based Synthesis," *IEEE Trans. on PAMI,* vol.15, no.6, pp.580-591, 1993.

[13] J. Ohya and F. Kishino, "Human Posture Estimation from Multiple Images Using Genetic Algorithm," *Proc. ICPR*, 1994.

[14] A. Pentland, "Automatic Extraction of Deformable Models," *Int. J. Computer Vision,* vol.4, pp.107-126, 1990.

[15] F.J. Perales and J. Torres, "A System for Human Motion Matching between Synthetic and Real Images," *IEEE Work. on Motion of Non-Rig. and Art. Objects,* Austin, TX, 1994.

[16] R. Polana and R. Nelson, "Low Level Recognition of Human Motion," *IEEE Workshop on Motion of Non-Rigid and Articulated Objects,* Austin, TX, 1994.

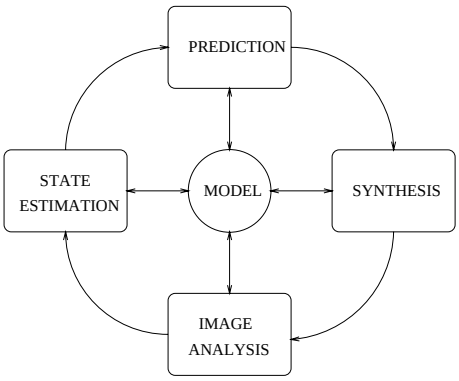
[17] J. Rehg and T. Kanade, "Model-Based Tracking of Self-Occluding Articulated Objects," *Proc. ICCV*, pp.612-617, 1995.

[18] K. Rohr, "Towards Model-Based Recognition of Human Movements in Image Sequences," *CVGIP: Image Understanding,* Vol.59, No.1, pp.94-115, 1994.

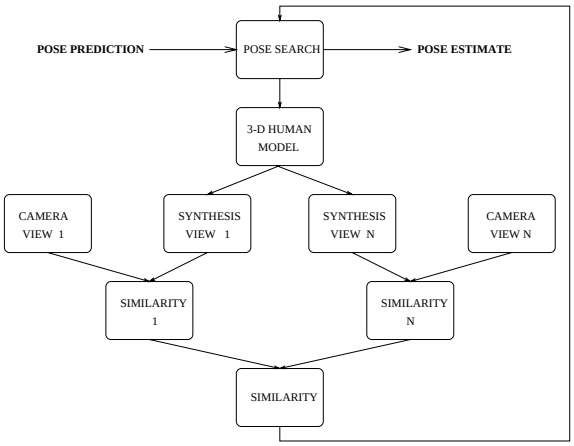
[19] J. O'Rourke and N.I. Badler, "Model-based image analysis of human motion using constraint propagation," *IEEE Trans. on PAMI,* vol.2, pp.522-536, 1980.

[20] R. Szeliski and S.B. Kang, "Recovering 3D Shape and Motion from Image Streams Using Nonlinear Least Squares," *J. Vis. Comm. and Im. Rep.*, vol.5, no.1, pp.10-28, 1994.

[21] J. Yamato, J. Ohya and K. Ishii, "Recognizing Human Action in Time-Sequential Images using Hidden Markov Model," *Proc. IEEE CVPR,* pp 379-385, 1992.



(a) tracking cycle



(b) pose-search cycle

Figure 1: Tracking and pose-search cycle

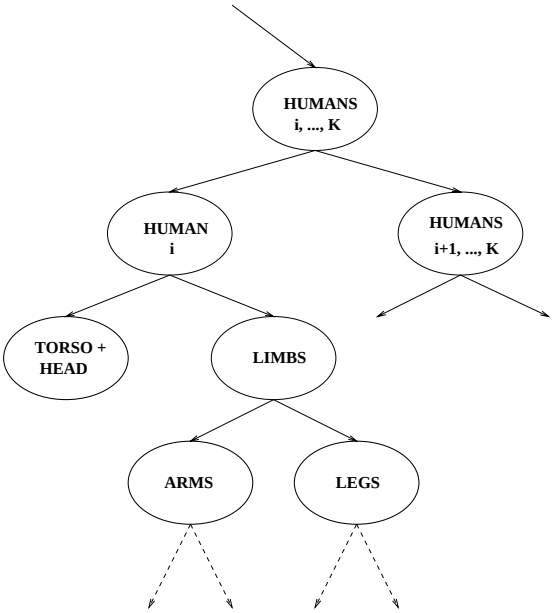


Figure 2: A decomposition of the pose-search space

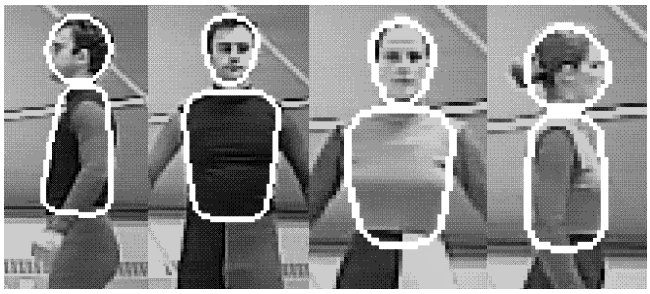


Figure 3: Frontal and side views of the recovered torso and head for the DARIU and ELLEN model

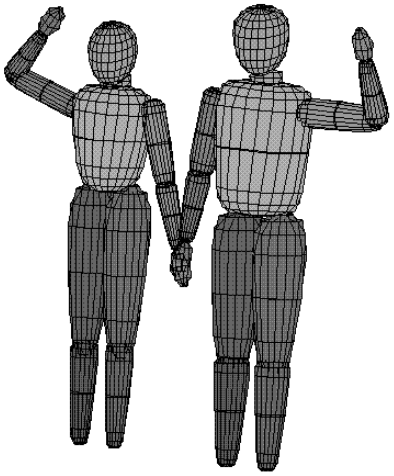


Figure 4: The recovered 3-D models ELLEN and DARIU say "**hi!**"

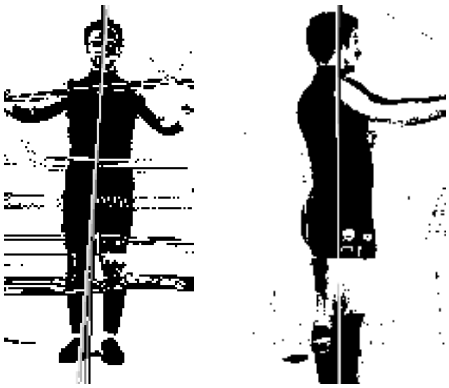
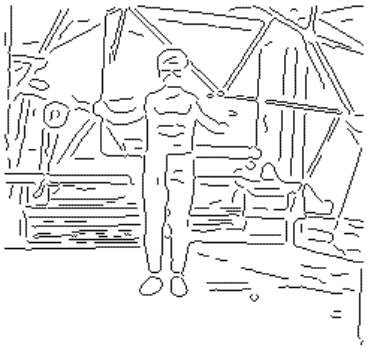


Figure 5: Robust major axis estimation using iterative PCA (cameras FRONT and RIGHT). Successive approximations to the major axis are shown in lighter colors.



(a) Scene edge image (after preprocessing)

(b) filtered edge image (model prediction in grey, accepted edges in black)

(c) chamfer image

Figure 6: Image processing

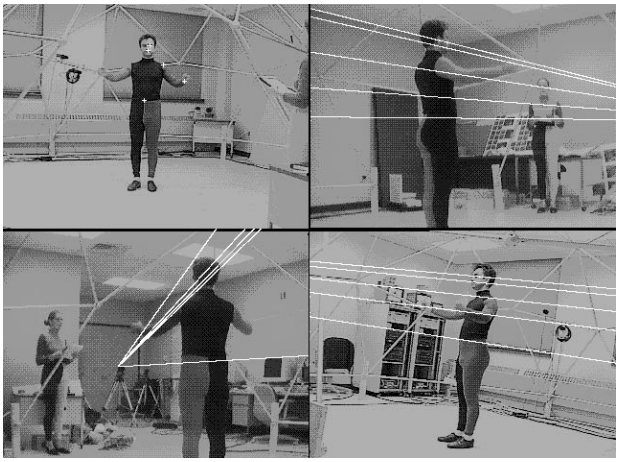


Figure 7: Epipolar geometry of cameras FRONT (upper-left), RIGHT (upper-right), BACK (lower-left) and LEFT (lower-right): epipolar lines are shown corresponding to the selected points from the view of camera FRONT

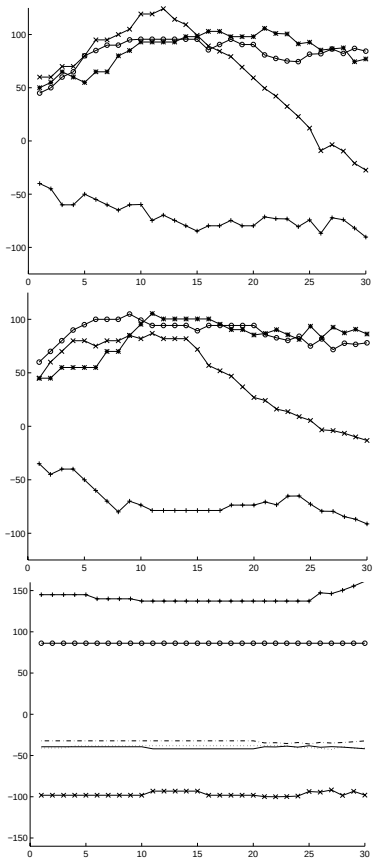
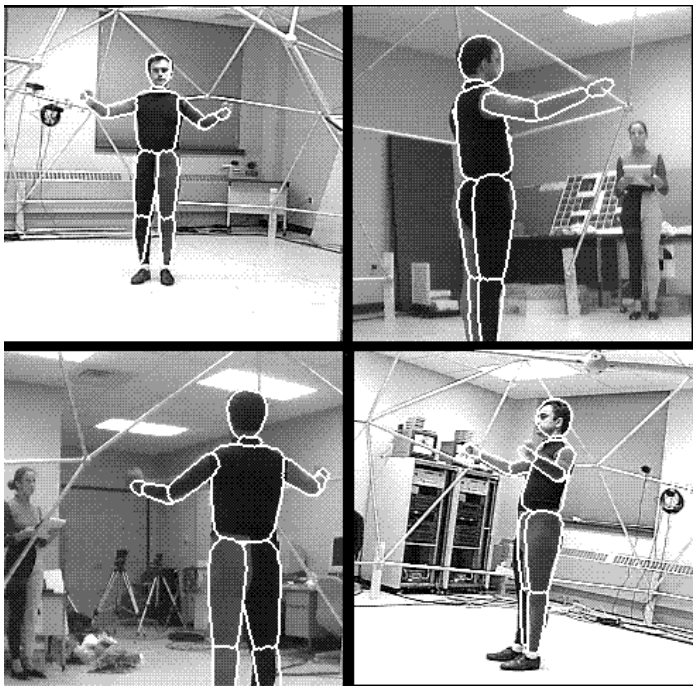
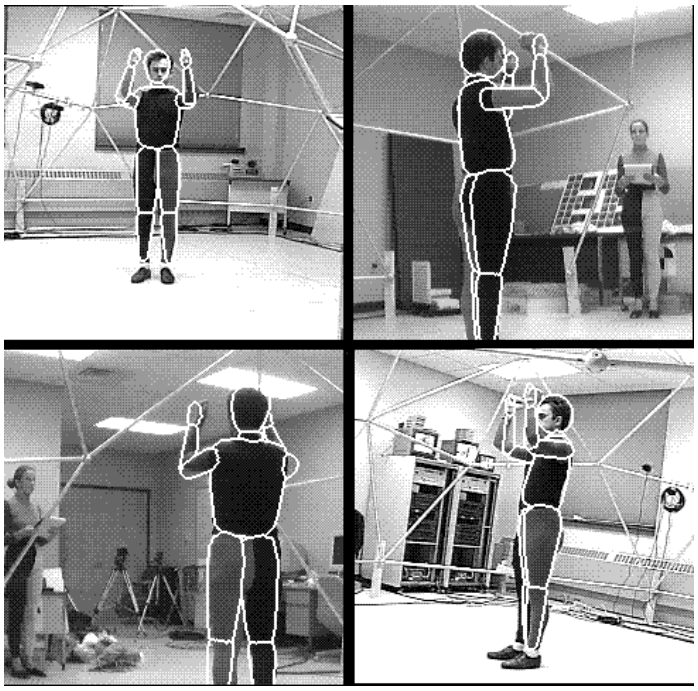


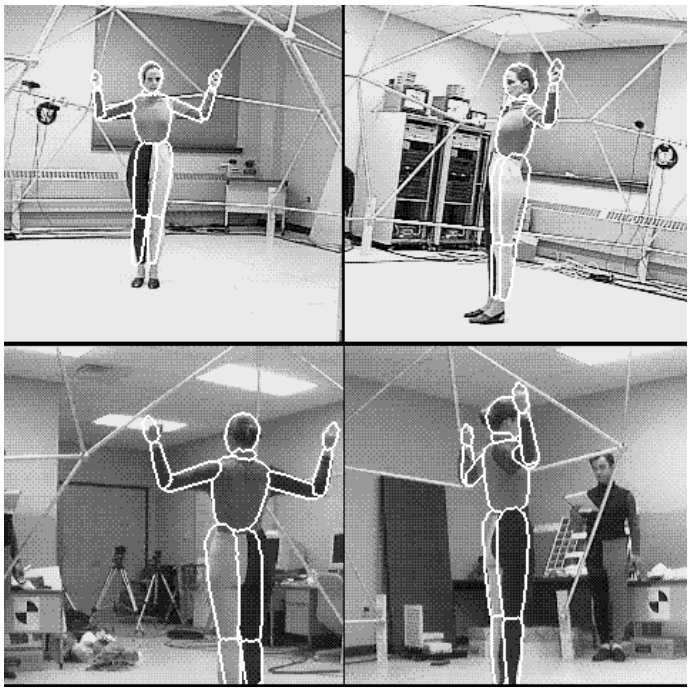
Figure 8: Recovered 3-D pose parameters vs. frame number, **D-TwoElbRot**; **(top)** and **(middle)**: LEFT and RIGHT ARM, abduction- (x), elevation- (o), twist- (+) and extension-angle (*) **(bottom)**: TORSO, abduction- (x), elevation- (o), twist-angle (+) and x- (dot), y- (dashdot) and z-coordinate (solid)



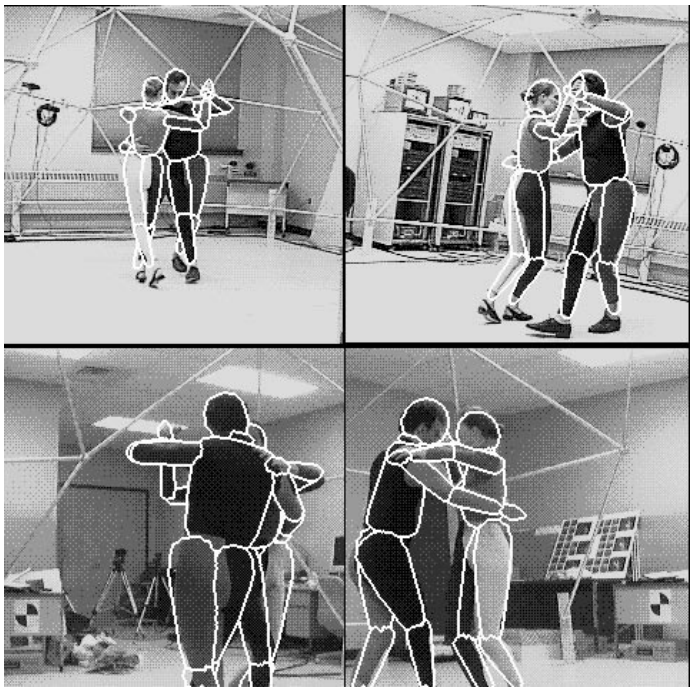
(a) D-TwoElbowRot, $t = 0$



(b) D-TwoElbowRot, $t = 25$



(c) E-TwoElbowRot, $t = 10$



(d) DE-Tango $t = -1$

Figure 9: (a)-(b) Tracking sequence D-TwoElbowRot ($t = 0, 25$), (c) Tracking sequence E-TwoElbowRot ($t = 10$), (d) Manual 3-D positioning for DE-Tango; cameras FRONT, RIGHT, BACK and LEFT.