Submmitted March 1997
Published in *Computer Vision and Image Understanding*, vol.73, no.1, pp.82-98, 1999,
Academic Press

The Visual Analysis of Human Movement: A Survey

D.M. Gavrila [1]

Image Understanding Systems
Daimler-Benz Research
Wilhelm Runge St. 11
89081 Ulm, Germany
gavrila@dbag.ulm.DaimlerBenz.com

Abstract

The ability to recognize humans and their activities by vision is key for a machine to interact intelligently and effortlessly with a human-inhabited environment. Because of many potentially important applications, "Looking at People" is currently one of the most active application domains in computer vision. This survey identifies a number of promising applications and provides an overview of recent developments in this domain. The scope of this survey is limited to work on whole-body or hand motion; it does not include work on human faces. The emphasis is on discussing the various methodologies; they are grouped in 2-D approaches with or without explicit shape models and 3-D approaches. Where appropriate, systems are reviewed. We conclude with some thoughts about future directions.

1 Introduction

A new application domain of computer vision has emerged over the past few years dealing with the analysis of images involving humans. This domain (sometimes called "Looking at People") covers, among others, face recognition, hand gesture recognition and whole-body tracking. The strong interest

[1] The majority of this work was done while at the Computer Vision Laboratory at the University of Maryland at College Park, U.S.A.; it was supported by the Advanced Research Projects Agency (ARPA Order No. C635) and the Office of Naval Research under Grant N00014-95-1-0521.

in this domain has been motivated by the desire for improved man-machine interaction for which there are many promising applications.

One of the general goals of artificial intelligence has been to design machines which act more intelligently or human-like. Natural language understanding, large knowledge bases and sophisticated reasoning have all made contributions towards reaching this goal, as embodied by the Turing test. Yet, they provide only a partial solution; for a machine to be truly intelligent and useful, it requires the ability to perceive the environment in which it is embedded. It needs to be able to extract information from its environment independently, rather than rely on information supplied to it externally by keyboard input (as in the original conception of the Turing test). Perhaps the most relevant information to be retrieved for interaction is where and who are the humans in the environment and what are their activities. Here, computer vision can play an important role. An added benefit of such a capability is that it makes communication with machines easier for humans, allowing input modalities such as gestures.

Traditionally, there has been keen interest in human movement from a wide variety of disciplines. In psychology, there have been the classic studies on human perception by Johansson [39]. His experiments with moving light displays (MLD) attached to body parts showed that human observers can almost instantly recognize biological motion patterns even when presented with only few of these moving dots. This raised the question whether recognition of moving parts could be achieved directly from motion, without structure recovery. In the hand gesture area, there have been many studies on how humans use and interpret gestures; see for example work by McNeill [52]. Quek [66] has put this in the context of vision-based human-computer interfaces.

In kinesiology (i.e. biomechanics) the goal has been to develop models of the human body that explain how it functions mechanically and how one might increase its movement efficiency. A typical procedure involves obtaining 3-D joint data, performing kinematic analysis, and computing the corresponding forces and torques for a movement of interest [12]. 3-D data is typically obtained in an intrusive manner, e.g. by placing markers on the human body.

In choreography, there has been long-term interest in devising high-level descriptions of human movement for the notation of dance, ballet and theatre. Some of the more popular notations have been the Labanotation, the Ekshol-Wachmann and the Effort-Shape notation. Across the variety of notation systems there has been little consensus, though, what these general-purpose descriptions should be. Badler and Smoliar [6] provide a good discussion of these issues.

Computer graphics has dealt with the synthesis of human movement. This has involved devising realistic models of human bodies for applications in crash simulations, workplace assessment and entertainment. Some of the issues have been how to specify spatial interactions and high-level tasks for the human models; see [5] [6] [50].

The recent interest in vision in the "Looking at People" domain is hardly surprising. From a technical point of view, this domain is rich and challenging because of the need to segment rapidly changing scenes in natural environments involving non-rigid motion and (self) occlusion. A number of potentially important applications exist; see the next Section. Additional momentum has been provided by recent technological advances, chief among them the introduction of real-time capture, transfer and processing of images on standard hardware systems (e.g. PCs). The extensive coverage in the vision literature is apparent from the many special workshops devoted to this topic: the "Looking at People" workshop in Chambery (1994), the "Motion of Non-Rigid and Articulated Objects" workshop in Austin (1994), and the two "Automatic Face and Gesture Recognition" workshops in Zürich (1995) and Killington (1996). Some of the material has now also reached the popular scientific press [63].

This papers surveys the work on visual analysis of gestures and whole-body movement. These are discussed together because of obvious similarities (i.e. both involve articulated structures). Section 2 discusses promising application scenarios of the "Looking at People" domain in some detail. Many criteria could be used to classify previous work, for example, the type of models used (e.g. stick figure-based, volumetric, statistical), the dimensionality of the tracking space (2-D vs. 3-D), sensor modality (e.g. visible light, infra-red, range), sensor multiplicity (monocular vs. stereo), sensor placement (centralized vs. distributed) and sensor mobility (stationary vs. moving). This survey is based on the first two criteria; it distinguishes

- 2-D approaches without explicit shape models (Section 3),
- 2-D approaches with explicit shape models (Section 4), and
- 3-D approaches (Section 5).

These classes do have some overlap. For example, some 2-D approaches use explicit shape models but also contain some elements of learning or self-adaptation. Nevertheless, this general classification provides a good framework for discussion throughout this survey.

Section 6 provides an overview of techniques for human action recognition; it takes a bottom-up view which assumes that all relevant features have been extracted from the images at this point, i.e. using one of the approaches of the last three Sections. A general discussion of past work is given in Section 7 together with some thoughts about future directions. The conclusions are listed in Section 8.

Face analysis (head pose estimation, face recognition, facial expressions, lip reading) is not covered by this survey, see for example [83]. Earlier reviews on non-rigid motion, motion-based recognition and gesture interpretation were given by Aggarwal *et al.* [1], Cedras and Shah [14] and Pavlovic, Sharma and Huang [61], respectively.

2 Applications

There are a number of promising applications in the "Looking at People" area in computer vision in addition to the general goal of designing a machine capable of interacting intelligently and effortlessly with a human-inhabited environment; for a summary see Table 1.

An important application domain is smart surveillance. Here "smart" describes a system that does more than motion detection, a straightforward task prone to false alarms (there might be animals wandering around, wind blowing, etc.). A first capability would be to sense if a human is indeed present. This might be followed by face recognition for the purpose of access control and person tracking across multiple cameras. In other applications, one needs to determine what a person in the scene is doing, rather than simply signaling human presence. In a parking lot setting, one might want to signal suspicious behavior such as wandering around and repeatedly looking into cars. Other surveillance settings involve supermarket or department stores, vending machines, ATMs and traffic. The benefits of such surveillance applications need in some cases to be balanced with possible drawbacks, e.g. regarding privacy.

Another application domain is virtual reality. In order to create a presence in a virtual space one needs to first recover the body pose in the physical space. Application areas lie in interactive virtual worlds, with the internet as a possible medium. The development of interactive spaces on the internet is still in its infancy; it is in the form of "chat rooms" where users navigate

general domain	**specific area**
virtual reality	- interactive virtual worlds - games - virtual studios - character animation - teleconferencing (e.g. film, advertising, home-use)
"smart" surveillance systems	- access control - parking lots - supermarkets, department stores - vending machines, ATMs - traffic
advanced user interfaces	- social interfaces - sign-language translation - gesture driven control - signaling in high-noise environments (airports, factories)
motion analysis	- content-based indexing of sports video footage - personalized training in golf, tennis, etc. - choreography of dance and ballet - clinical studies of orthopedic patients
model-based coding	- very low bit-rate video compression

Table 1: Applications of "Looking at People"

with icons in 2-D spaces while communicating by text. A more enriched form of interaction with other participants or objects will be possible by adding gestures, head pose and facial expressions as cues. Other applications in this domain are games, virtual studios, motion capture for character animation (synthetic actors) and teleconferencing.

In the user-interface application domain, vision is useful to complement speech recognition and natural language understanding for a natural and intelligent dialogue between human and machine. The contribution of vision to a speech-guided dialogue can be manifold. One can simply determine if a user is present to decide whether to initiate a dialogue or not. More detailed cues can be obtained by recognizing who the user is, observing facial expressions and gestures as the dialogue progresses and perhaps recalling some of the past interactions. It would certainly be useful to determine who is talking to whom in case of multiple participants. Vision can also provide speech recognition with a more accurate input in a noisy environment by focusing the attention to the spatial location of the user [80]. This is achieved either by a post-filtering step when using a phased array of microphones or, more actively, by directing a parabolic microphone to the intended source. Finally, vision can also prove helpful for phoneme disambiguation, i.e. lip reading.

An important application area in the user interface domain involves social interfaces. Social interfaces deal with computer-generated characters, with "human-like" behaviors, who attempt to interact with users in a more personable way [80]. Alternative application areas in the user interface domain are sign-language translation, gesture driven control of graphical objects or appliances, and signaling in high-noise environments such as factories or airports.

In the motion analysis domain, a possible application is content-based indexing of sports video footage; in a tennis context, one may want to query a large video archive with "give me all the cases where player X came to the net and volleyed". This would eliminate the need for a human to browse through a large data set. Other applications lie in personalized training systems for various sports; these systems would observe the skills of the pupils and make suggestions for improvement. Vision-based human motion analysis is also useful for choreography of dance and ballet, and furthermore, for clinical studies in orthopedy.

One final application domain is that of model-based image coding, with activity centered around the forthcoming MPEG-4 standard. In a video phone setting, one can track faces in images and code them in more detail than the background. More ambitiously, one might try to recover a 3-D head model initially and code only the pose and deformation parameters subsequently. It is unclear whether these applications will materialize; the 2-D head tracking application provides modest compression gains and is specific to scenes with human faces; the 3-D head (or body) tracking application has not been solved satisfactory yet. See Aizawa and Huang [2] for a good overview.

In all the applications discussed above, a non-intrusive sensory method based on vision is preferable over a (in some cases a not even feasible) method that relies on markers attached to the bodies of the human subjects or a method which is based on active sensing.

3 2-D approaches without explicit shape models

One general approach to the analysis of human movement has been to bypass a pose recovery step altogether and to describe human movement in terms of simple low-level, 2-D features from a region of interest. Polana and Nelson [65] refer to "getting your man without finding his body parts". Models for human action are then described in statistical terms derived from these low-level features, or by simple heuristics. The approach without explicit shape models has been especially popular for applications of hand pose estimation in sign language recognition and gesture-based dialogue management.

For applications involving the human hand, the region of interest is typically obtained by background image subtraction or skin color detection. This is followed by morphological operations to remove noise. The extracted features are based on hand shape, movement and/or location of the interest region. For shape, Freeman *et al.* [24] use x-y image moments and orientation histograms and Hunter *et al.* [38] use rotationally invariant Zernike moments. Others [16] [20] [77] [79] consider the motion trajectories of the hand centroids. Quek [66] has proposed using shape and motion features alternatively for the interpretation of hand-gestures. According to Quek, when the hand is in gross motion, the movements of the individual fingers are unimportant for gesture interpretation. On the other hand, gestures in which fingers move with respect to each other will be performed with little hand motion.

A similar technique to derive low-level features is to superimpose a grid on the interest region, after a possible normalization of its extent. In each tile of the grid a simple feature is computed, and these features are combined to form a $K \times K$ feature vector to describe the state of movement at time t. Polana and Nelson [65] use the sum of the normal flow (see Figure 1), Yamamoto et al. [86] use the number of foreground pixels and Takahashi et al. [78] define an average edge vector for each tile. Both Darell and Pentland [19] and Kjeldsen and Kender [44] use the image pixels directly as input. The work by Darell and Pentland [19] aims to build view models automatically by adding views to the model set whenever correlation with the existing views falls below a certain threshold.

For the above systems, action classification is based on hard-coded decision trees [16] [20] [79], nearest neighbor criteria [38] [65] or is based on general pattern matching techniques for time-varying data, as described in Section 6. Some additional constraints on actions can be imposed using a dialogue structure where the current state limits the possible actions that can be expected next.

Oren *et al.* [59] perform object detection in static images. They use (Haar) wavelet coefficients as low-level intensity features; these coefficients are obtained by applying a differential operator at various locations, scales and orientations on the image grid of interest. Many coefficients can be part of this representation. In a training stage, however, one selects a small subset of coefficients to represent a desired object, based on considerations regarding relative strength and positional spread over the images of the training set. Once it has been established which wavelet coefficients to use as features, a Support Vector Machine (SVM) classifier is applied to the training set. During the detection stage, one shifts windows of various sizes over the image, extracts the selected features and applies the SVM classifier to verify whether the desired object is present or not. Oren *et al.* apply this technique to detecting frontal and rear views of pedestrians; see Figure 2.

Another line of research involves statistical shape models to detect and track the contours of hands or persons. The work by Cootes *et al.* [18] uses "Active Shape Models" for this purpose; these are models derived from a training stage were example shapes are described in terms of known feature point locations. Cootes *et al.* perform principal component analysis on the feature locations to describe the example shapes using a reduced parameter set. With this compact representation one obtains, in addition to efficiency, some degree of generalization over the training set. This can be useful when tracking deformable shapes; using the new representation one allows, in essence, only those deformations which are consistent with the training set. Cootes *et al.* show some examples of tracking hands. The followed method also has some drawbacks. Features need to be present at all times (no occlusions). At initialization, a good initial estimate must be available for the method to converge properly. And finally, the chosen parametrization might include states which have implausible physical interpretations.

Baumberg and Hogg [8] apply Active Shape Models to the tracking of pedestrians. They use a somewhat different shape representation, based on B-splines; see Figure 3. By assuming a stationary camera, tracking is initialized on the foreground region; the latter is obtained by background subtraction. Spatio-temporal control is achieved using a Kalman filter formulation, similar to work by Blake *et al.* [9].

Recent work by Franke *et al.* [23] applies principal component analysis on a grid representation of pedestrians. The training set is obtained by blurring binary images which correspond to pedestrian silhouettes. Principal component analysis results, as before, in a compact representation of the training set in terms of various eigenvectors which span a linear sub-space. See Figure 4: the main variation is captured by the first few eigenvectors (corresponding to the largest eigenvalues), the 25th eigenvector already contains mostly noise. Pedestrian detection involves shifting windows of various sizes over the image, normalizing for gradient energy within the window, and determining the "distance" between the normalized (gradient) data enclosed by the window and the linear sub-space corresponding to the training set. One of the advantages of using grid representations (e.g. [23] [59]) is that dealing with partial occlusion is relatively straightforward.

General-purpose motion-based segmentation and tracking techniques have also been used for applications such as people tracking. Shio and Sklansky [75] aim to recover the average 2-D image velocity of pedestrians in a traffic setting. They obtain a motion field based on correlation techniques over successive frames. The motion field is smoothed both spatially and temporally to reduce the effects of non-rigid motion and measurement errors. A quantization of the field is then followed by an iterative merging step which results in regions with similar motion direction. Segen and Pingali [73] group partially-overlapping feature tracks over time in a real-time implementation. Heisele *et al.* [32] use groups of pixels as basic units for tracking. Pixels are grouped by clustering techniques in combined color (R,G,B) and spatial (x,y) dimensions; the motivation for this is that adding spatial information makes clustering more stable than using only color information. The obtained pixel groups are adapted iteratively from one image to the next image using a k-means clustering algorithm. Because of the fixed number of pixel groups and the enforced one-to-one correspondence over time, tracking these units is straightforward. Of course, there is no guarantee that units will remained locked onto the same physical entity during tracking, but initial results on tracking pedestrians appear promising; see Figure 5.

4 2-D approaches with explicit shape models

This section discusses work which uses explicit a priori knowledge of how the human body (or hand) appears in 2-D, taking essentially a model- and view-based approach to segment, track and label body parts. Since self-occlusion makes the problem quite hard for arbitrary movements, many systems assume a priori knowledge of the type of movement or the viewpoint under which it is observed. The human figure is typically segmented by background subtraction, assuming a slowly changing or stationary background and a fixed camera. The models used are usually stick figures, wrapped around with ribbons or "blobs". An example of a ribbon-based 2-D model is illustrated in Figure 6. The type of the model strongly influences what features are used for tracking; one can distinguish systems using edges or ribbons, "blobs" and points.

A number of researchers have analyzed scenes involving human gait parallel to the image plane. Geurtz [27] performs hierarchical and articulated curve fitting with 2-D ellipsoids. Niyogi and Adelson [56] [57] advocate segmentation over time because of robustness; their procedure involves finding human silhouettes with deformable contours in X-T space [56] or deformable surfaces in X-Y-T space [57]. See Figure 7. Guo *et al.* [30] propose obtaining a 2-D stick figure by obtaining the skeleton of the silhouette of the walking human and matching it to a model stick figure. They use a combination of link orientations and joint positions of the obtained stick figure as features for a subsequent action recognition step. Chang and Huang [15] detect ribbons corresponding to the arms and feet. Ju *et al.* [40] use a parametrized motion model to analyze gait constrained to a plane. The legs are modeled a set of connected planar patches.

An early attempt to segment and track body parts under more general conditions is made by Akita [3]. The assumption made is that the movement of the human is known a priori in the form of a set of representative stick figure poses or "key frames". These would be of help when the tracking of body parts fails. The foreground figure and its silhouette are easily obtained given the large dark-light differences. The recognition of body parts proceeds in the order: legs, head, arms and trunk following the assumption that legs are the most stable to detect and the trunk the least. Unfortunately, a number of unstated simplifications and procedures make evaluation of this approach difficult.

Without a priori knowledge of the type of movement being performed, Long and Yang [49] track the limbs of a human silhouette by tracking anti-parallel lines (apars). They develop methods to deal with the effects of occlusion, i.e. the appearance, disappearance, merging and splitting of apars. The work by Kurakake and Nevatia [47] is similar. Leung and Yang [48] report progress on the general problem of segmenting, tracking and labeling of body parts from a silhouette of the human. Their basic body model consists of five U-shaped ribbons and a body trunk, various joint and mid points, plus a number of structural constraints, such as support. In addition to the basic 2-D model, view-based knowledge is defined for a number of generic human postures (e.g. "side view kneeling model", "side horse motion"), to aid the interpretation process. The segmentation of the human silhouette is done by detecting moving edges. See Figure 8.

Wren *et al.* [84] take a region-based approach. Their real-time person finder system "Pfinder" models and tracks the human body using a set of "blobs"; each blob is described in statistical terms by a spatial (x, y) and color (Y, U, V) Gaussian distribution over the pixels it consists of (compare with the shape-color model used in [32]). The blobs typically correspond to the person's hands, head, feet, shirt and pants. A statistical model is also constructed for the background region; here each pixel is described by a Gaussian distribution in terms of color values. At initialization, the background model is used to identify a foreground region with pixel values other than expected given the background model. A model-building process follows where blobs are placed over the foreground region. This process is guided by a 2-D contour shape analysis that attempts to identify various body parts using heuristics. Tracking involves a loop of predicting the appearance of the person in the new image, determining for each pixel the likelihood that it is part of one of the blob models or background model, assigning it to one of the models, and updating the statistical models. See Figure 9.

Cai and Aggarwal [11] describe a system with a simplified head-trunk model to track humans across multiple cameras. In this work, tracking uses point features derived from the medial axis of the foreground region. Attributes used for tracking are position and velocity of the points, together with the average intensity of the local neighborhood of the points. The use of point features has the advantage that the features can be relatively easily brought into correspondence across multiple cameras, given constraints on epipolar geometry. It remains difficult, though, to robustly track points in long sequences when the points do not correspond to stable features on the human body.

Finally, Kahn and Swain [41] describe a system which uses multiple cues (intensity, edge, depth, motion) to detect people pointing laterally. Their system architecture is quite generic and could be described as being "object-oriented"; a number of generic objects are defined for a particular application (e.g. person, background, floor, lights) and visual routines are provided to detect these in the images. Once various object properties have been extracted from the image, the objects become "instantiated" and specialized visual routines apply afterwards.

5 3-D approaches

In this section we discuss work that aims to recover 3-D articulated pose over time, i.e. joint angles with respect to an object-centered [51] coordinate system. We will not consider intrusive techniques for motion capture, e.g. techniques which use markers or active sensing.

The general problem of 3-D motion recovery from 2-D images is quite difficult. In the case of 3-D human tracking, however, one 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 shape model which can predict events such as (self) occlusion and (self) collision.

A general framework for model-based tracking is illustrated in Figure 10, based on the early work of O'Rourke and Badler [60]. 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. This framework can be applied to any model-based tracking problem, whether involving a 2-D or 3-D tracking space. Many of the tracking system discussed in this section follows this general framework.

Once 3-D tracking is successfully implemented, one has the benefit of being able to use the 3-D joint angles as features for subsequent action matching; these have the advantage to be viewpoint independent and directly linked to the body pose. Compared to 3-D joint coordinates, joint angles are less sensitive to variations in the size of humans.

5.1 3-D body modeling

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. Relevant rotations are generally described by their three Euler angles [13] [76].

The representation for the flesh can either be surface-based (e.g. using polygons) or volumetric (e.g. using cylinders). 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 to model the human body [5], 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. After all, human models used for computer vision do not have to meet the standard of being highly realistic and natural looking as long as their shape approximates the real human shape well enough to support image segmentation.

An early example of human modeling is Badler's "Bubbleman" [60], where body parts consist of overlapping spheres. Another modeling choice has involved simple cylindrical primitives (possibly with elliptic XY-cross-sections) [22] [29] [36] [51] [71]. More accurate modeling of body parts is obtained using superquadrics [7]; these are generalizations of ellipsoids which have additional "squareness" parameters along each axis. They include such diverse shapes as cylinders, spheres, ellipsoids and hyper-rectangles. Superquadrics improve the modeling accuracy for body parts such as the head and torso and for regions close to articulation sites. Additional flexibility can be achieved by allowing global deformations (e.g. tapering, bending) and/or local deformations on the superquadrics [7] [26] [43] [53] [62]. Figure

11 shows an example of human modeling based on tapered superquadrics that was used for 3-D model-based tracking in [25] [26].

5.2 3-D pose recovery and tracking

We first discuss approaches which use articulated models to recover 3-D pose from a monocular image sequence. One possibility is to use a divide-and-conquer technique, where an articulated object is decomposed into a number of primitive (rigid or articulated) sub-parts; one solves for motion and depth of the sub-parts and verifies whether the parts satisfy the necessary constraints. Shakunaga [74] identifies such a set of primitive sub-parts for which he solves the pose recovery problem using the angles between projected line features.

To avoid unfavorable combinatorics at the verification step, it is beneficial to propagate constraints from part to part. The primitives of O'Rourke and Badler [60] are box-shaped regions which represent possible joint locations in 3-D. These regions are initially constrained by the measurement of joints in the images (essentially given to the system) and the orthography assumption. A constraint propagation procedure is then applied based on the known distances between connected joints. A further verification procedure involves an iterative search procedure, in which angular and collision constraints are verified using the 3-D model. Each step results in a refinement of the 3-D uncertainty regions of joints; the final regions can be used for prediction at the next time iteration.

Other work has used perspective projection models. The constraint propagation scheme of Chen and Lee [17] starts at the human head and continues via the torso to the limbs. An interpretation tree is built to account for the spatial ambiguity which arises from the fact that there are two possible poses of a link (of known length) in 3-D which result in the same 2-D projection. This interpretation tree is pruned later for physically implausible poses. Chen and Lee's assumption of six known feature points on the head to start the procedure and the overhead of the interpretation tree makes their approach somewhat unappealing for practical applications. Zhao [87] has a similar problem formulation but does not maintain the interpretation tree, considering instead only one pose at the time. He monitors when spatial ambiguities are encountered and disambiguates them by temporal coherence. Holt *et al.* [37] provide a constraint propagation scheme for human gait, where one joint remains at a fixed location. Motion constraints are also incorporated at the earliest stages. The core of their system involves solving a polynomial system of equations. Other approaches have imposed general constraints on the articulated motion, such as the "fixed-axis" [82] or "in-plane" [35] assumptions of rotations.

Hel-Or and Werman [33] describe a technique for articulated pose recovery based on the fusion of constraints and measurements using a Kalman filter framework. Kakadiaris and Metaxas [42] [43] use a physics-based approach where various forces act on the different parts to align them with the image data; constraint forces enforce point-to-point connectivity between the parts. They apply this approach to multi-camera tracking and, additionally, deal with the problem of active camera selection based on body-part visibility and motion observability.

Other approaches to 3-D articulated motion use parametrized models where the articulation constraints are encoded in the representation itself. This has the advantage that each representable state represents a physically valid pose (aside from joint-angle limitations and collisions); thus, the resulting approach takes advantage as much as possible of prior 3-D knowledge and relies as little as possible on error-prone 2-D image segmentation. On the downside, by considering the (coupled) parameters simultaneously, one needs to work in a high dimensional parameter space.

One approach using such parametrized models [21] [29] [69] [70] [81] [85] [87] updates pose by inverse kinematics, a common technique in robot control theory [76]. The state space maps onto image space by a non-linear measurement equation which takes into account the coordinate transformations at various articulation sites and the 3-D to 2-D projection. Inverse kinematics involves inverting this mapping to obtain changes in state parameters which minimize the residual between projected model and image features. The procedure involves a linearization of the measurement equation, as defined by the Jacobian matrix, and a gradient-based optimization scheme. The inverse kinematics approach can also be taken with multiple cameras when no feature correspondence between cameras is assumed. One simply concatenates the residual from the available camera views; see for example [70].

Another approach using parametrized models does not attempt to invert a non-linear measurement equation. Instead, it uses the measurement equation directly to synthesize the model and uses a fitting measure between synthesized and observed features for feedback; see [22] [26] [36] [46] [58] [64] [71]. Pose-recovery can then be formulated as a search problem which entails finding the pose parameters of a graphical human model whose synthesized appearance is most similar to the actual appearance of the real human. Because one need not to invert a measurement equation, one is quite flexible in choosing an appropriate evaluation measure between model and scene; typical measures are based on occluding contours or regions. No point correspondences between model and scene are required. To find a good fit, Ohya and Kishino [58] use a global search strategy based on genetic algorithms. Kuch and Huang [46] use a greedy search strategy based on perturbation of individual state parameters. Gavrila and Davis [26] use local search based on best-first search. The high-dimensional search space, which results from recovering whole-body pose, necessitates in the latter work a decomposition technique, in which pose-recovery is done successively for torso (without twist), arms and torso twist, and legs. Some of the combinatoric pose-recovery approaches have also been applied to the multi-camera case, in simulations [58] and with real data [26].

Comparing the above greedy gradient-based inverse kinematics approaches with the non-greedy combinatoric search approaches, one notes that the former have the advantage that they exploit gradient cues in the vicinity of a minimum and therefore are computationally more efficient, see for example [69]. On the other hand, concern is justified that a gradient-based scheme might get stuck in a local minimum (i.e. to converge to a sub-optimal or undesired solution) because the measurement equation is highly non-linear (composition of various non-linear rotation matrices and perspective mapping) and the sampling ratio at which one obtains image measurement is relatively low for fast movement such as locomotion and gesticulation. Furthermore, measurements are typically noisy and can be incorrect altogether, e.g. when corresponding the features with the wrong body parts. A non-greedy search method also promises to be more robust over time; if it fails to find a good solution at time t, there is still a chance that it may recover at time $t + 1$ if the search area is sufficiently wide. A combination of non-greedy search followed by a gradient-based technique is probably a good compromise between robustness and efficiency.

There has also been work on using depth data for articulated pose recovery. Rather than requiring the typical point features, Azarbayejani and Pentland [4] "triangulate" using blob features [84]; a 3-D blob (shape, orientation) is recovered from a pair of corresponding 2-D blob features using non-linear estimation techniques. In other work, Pentland [62] fits deformable superquadrics to range data. A maximum-likelihood technique provides the initial part segmentation based on the object silhouette. The subsequent fitting procedure deformes superquadrics using modal dynamics.

Finally, work by Heap and Hogg [31] involves an example-based approach to articulated pose recovery. Their method involves a principal component analysis of 3-D positional (hand) data and allows shape deformations of a tracked object. This method was mentioned earlier in the 2-D context; see Section 3 [8].

5.3 Feature correspondence

A variety of features can be used to establish correspondence between model and image remains, from low-level to high-level. Using high-level features (e.g. joint locations) simplifies pose recovery but places a greater burden on segmentation. Approaches [17] [37] [60] [74] [87], which use joint locations as features and assume these are given make strong assumptions. In reality, the joints are hard to detect; no characteristic intensity distribution exists at their location; rather, joints are localized indirectly by segmenting the adjoining body parts. Moreover, relying exclusively on a few correspondences makes the resulting approach [21] [69] quite sensitive to occlusion.

This has lead many researchers to consider low- or intermediate-level features to establish correspondence between model and image. Some use occluding contours, where the evaluation measure for the model-to-image fit is based on image regions in the neighborhood of the projected model contours. Typical measures are correlation on a raw or smoothed LOG-filtered image [29] [70], perpendicular- [31] and chamfer-distance [26] (from projected model edges to image edges) and straight-line distance metrics [71]. Others have used evaluation measures derived from the regions corresponding to the projected body-parts, e.g. based on image intensities [46] [81] or optical flow [85]. A distinction between low and intermediate features can be made, as before, based on the segmentation effort involved to extract the features. Image intensities and optical flow can be considered low-level, and features derived by thresholding or perceptual grouping, intermediate-level.

What the best trade-off between segmentation effort and ease of pose recovery is difficult to determine. For example, a method which matches model and image edges based on a distance map approach (e.g. perpendicular or chamfer distance) has the advantage that the evaluation measure tends to be smooth in terms of the pose parameters; the measure is well suited to guide an iterative estimation process. A correlation measure on the unsegmented image, on the other hand, typically provides strong peak responses but rapidly declining off-peak responses. But then, no edge segmentation is needed for the latter. What might be worth considering is using intermediate-level features to provide a rough correspondence between model and image, and guiding the fine-tuning with low-level features.

5.4 Experimental results

This section reviews previous work on 3-D tracking in terms of experimental results on real data. Dorner [21] tracks articulated 3-D hand motion (palm motion and finger bending/unbending) with a single camera. Her system requires colored markers on the joints and cannot handle occlusions. Rehg and Kanade [69] do not require markers. Their "DigitEyes" system tracks an 8-DOF partial hand model (movement of palm in a 2-D plane and three fingers) with one camera and a full 27-DOF hand model with two cameras in real-time from the hand silhouette. Occlusion cannot be handled at this point. See Figure 12. A later version of the system [70] does tolerate partial occlusion; a successful tracking example is shown where one finger moves over the other finger, with the rest of the hand fixed. Heap and Hogg [31] show preliminary tracking results on hand model and hand pose recovery.

In terms of experimental results on whole (or upper body) movement using a single camera, Hogg [36] and Rohr [71] deal with the restricted movement of gait (parallel to image plane). The movement is essentially in 2-D with no significant torso-twist. Given that gait is modeled a priori, the resulting search space is one-dimensional. Downton and Drouet [22] attempt to track unconstrained upper-body motion but must conclude that tracking gets lost due to propagation of errors. Goncalves *et al.* [29] track one arm while keeping the shoulder fixed at a known position. Other results use multiple cameras. Kakadiaris and Metaxas [43] track one arm using three orthogonal cameras. See Figure 13. Azarbayejani and Pentland [4] obtain the 3-D locations of the face and hands by essentially triangulating on blobs representing the skin regions in the stereo views. Perales and Torres [64] describe a multi-view camera system for whole-body tracking which requires input from a human operator. Finally, Gavrila and Davis [25] [26] show initial results on whole-body tracking using four cameras placed in the corners of a room; see Figure 14.

In the above approaches working with real data it has often been difficult to quantify how good the 3-D pose recovery results are; typically, no ground truth has been established. This problem is alleviated somewhat in approaches which use multiple camera views; here one can at least visually verify the recovered pose along the depth dimension.

6 Action recognition

The prevalent view towards action recognition has been to consider it simply as a classification problem involving time-varying feature data; the feature data is derived from an earlier segmentation stage, using techniques of the last three sections. Recognition then consists of matching an unknown test sequence with a library of labeled sequences which represent the prototypical actions. A complementary problem is how to learn the reference sequences from training examples. Both learning and matching methods have to be able to deal with small spatial and time scale variations within similar classes of movement patterns.

Polana and Nelson [65] detect periodic activity such as persons walking lateral to the viewing direction using spatio-temporal templates. They argue that a template matching technique is effective here because a sufficiently strong normalization can be carried out on the region of interest with respect to spatial and time scale variations. For example, for the case of a stationary camera and a single object of interest, background subtraction and size normalization of the foreground region is sufficient to obtain spatial invariance, if perspective effects are small. Polana and Nelson also describe a technique to deal with the more complex case of a moving camera and/or multiple (overlapping) objects, based on detecting and tracking independently moving objects. Size changes of the object are handled by estimating the spatial scale parameters and compensating for them, assuming the objects have a fixed height throughout the sequence. Temporal scale variations are factored out by detecting the frequency of an activity. After these normalizations, spatio-temporal templates are constructed to denote one generic cycle of activity; a cycle is divided into a fixed number of subintervals for which motion features are computed. The features of a generic cycle are obtained by averaging corresponding motion features over multiple cycles. Temporal translation is handled in the matching stage in an exhaustive manner; the test template is matched with the reference template at all possible temporal translations. Matching uses a nearest centroid algorithm.

Rangarajan *et al.* [68] match motion trajectories of selected feature points on a human body (tracked manually). Their trajectories are described in terms of two one-dimensional signals, speed and direction. These one-dimensional signals are each converted into a two-dimensional representation, the scale-space, by computing the degree of zero-crossing of the original one-dimensional signal. The resulting representation has the advantage of being translation and rotation invariant. Using a Gaussian convoluted reference scale-image, one can account for a fixed amount of time-offset between reference and test trajectory.

Goddard [28] represents activities by scenarios: a sequence of events with enabling conditions, and time constraints between successive events. Each possible scenario is matched and given a measure of appropriateness, depending on the cumulative confidence in the scenario, the likelihood that its "next" event has occured, and the time-constraints. No learning takes place in the previous two methods.

Campbell and Bobick [13] use a phase-space representation in which the velocity dimensions are projected out, discarding the time component of the data altogether. This makes the learning and matching of patterns simpler and faster, at the potential cost of an increase in false positives.

Other general techniques for matching time-varying data have been used as well. Dynamic Time Warping (DTW) [55] is a well-known technique to match a test pattern with a reference pattern if their time scales are not perfectly aligned but when time ordering constraints do hold. If the sizes of the test pattern and reference pattern are N and M, an optimal match is found by dynamic programming in $O(N \times M^2)$ time (or in $O(N \times M)$ time, if one introduces local continuity constraints, see [55]). Because of conceptual simplicity and robust performance, Dynamic Time Warping was extensively used in the early days of speech recognition, and more recently in matching human movement patterns [10] [19] [25] [78].

More sophisticated matching of time-varing data is possible by employing Hidden Markov Models (HMMs) [67]. HMMs are non-deterministic state machines which, given an input, move from state to state according to various transition probabilities. In each state, HMMs generate output symbols probabilistically; these need to be related to image features in an application-dependent manner. The use of HMMs involves a training and a classification stage. The training stage consists of specifying the number of (hidden) states of a HMM and optimizing the corresponding state transition and output probabilities such that generated output symbols match the image features observed during examples of a particular motion class; a HMM is needed for each motion class. Matching involves the computation of the probability that a particular HMM could have generated the test symbol sequence which corresponds to the observed image features.

The ability to learn from training data and to develop internal representations under a sound mathematical framework make HMMs attractive when compared to DTW. Another advantage of HMMs are their ability to deal with unsegmented data, i.e. dealing with continuous data streams where the beginning of a desired data segment is unknown (DTW could be adapted to handle this as well; see Continuous Dynamic Time Warping [78]). Because of these benefits, HMMs are currently widespread in speech recognition and more recently in matching human movement patterns [77] [86]. A less investigated but equally interesting approach for matching time-varying data is given by Neural Networks (NN) [30] [72].

With all the emphasis on matching time-varying data, one should note that another aspect of human action recognition is static posture; sometimes it is not the actual movement that is of interest but the final pose (for example, pointing). Herman [34] describes a rule-based system to interpret body posture given a 2-D stick figure. Although the actual system is applied on a toy problem (in baseball), it does make the point of using qualitative pose measures together with other attributes such as facing direction and contact. It also emphasizes the importance of contextual information in action recognition.

Finally, work by Kollnig *et al.* [45] goes beyond the narrow interpretation of action recognition as a classification problem. They investigate ways of describing scene motion in terms of natural language ("motion verbs"); this is achieved within a logic-based framework. Their particular application is vehicle motion in traffic scenes. See also work by Mohnhaupt and Neumann [54].

7 Discussion

Table 2 lists the previous work on the analysis of human movement, which was discussed in this survey. Whether to pursue a 2-D or a 3-D approach is largely application-dependent. A 2-D approach is effective for applications where precise pose recovery is not needed or possible due to low image resolution (e.g. tracking pedestrians in a surveillance setting). A 2-D approach also represents the easiest and best solution for applications with a single human involving constrained movement and single viewpoint (e.g. recognizing gait lateral to the camera, recognizing a vocabulary of distinct hand gestures made facing the camera).

A 3-D approach makes more sense for applications in indoor environments where one desires a high level of discrimination between various unconstrained and complex (multiple) human movements (e.g. humans wandering around, making different gestures while walking and turning, social interactions such as shaking hands and dancing). It is unlikely that this

2-D approaches without explicit shape models	**2-D approaches with explicit shape models**	**3-D approaches**
Baumberg and Hogg [8]	Akita [3]	Azarbayejani and Pentland [4]
Bobick and Wilson [10]	Cai and Aggarwal [11]	Campbell and Bobick [13]
Charayaphan and Marble [16]	Chang and Huang [15]	Chen and Lee [17]
Cootes *et al.* [18]	Geurtz [27]	Dorner [21]
Darell and Pentland [19]	Goddard [28]	Downton and Drouet [22]
Davis and Shah [20]	Guo *et al.* [30]	Gavrila and Davis [25] [26]
Franke *et al.* [23]	Herman [34]	Goncalves *et al.* [29]
Freeman *et al.* [24]	Ju *et al.* [40]	Heap and Hogg [31]
Heisele *et al.* [32]	Kurakake and Nevatia [47]	Hel-Or and Werman [33]
Hunter *et al.* [38]	Leung and Yang [48]	Hoffman and Flinchbaugh [35]
Johansson [39]	Long and Yang [49]	Hogg [36]
Kjeldsen and Kender [44]	Niyogi and Adelson [56] [57]	Holt *et al.* [37]
Oren *et al.* [59]	Wren *et al.* [84]	Kahn and Swain [41]
Polana and Nelson [65]		Kakadiaris and Metaxas [42] [43]
Quek [66]		Kuch and Huang [46]
Rangarajan *et al.* [68]		Ohya and Kishino [58]
Segen and Pingali [73]		O'Rourke and Badler [60]
Shio and Sklansky [75]		Pentland [62]
Starner and Pentland [77]		Perales and Torres [64]
Takahashi *et al.* [78]		Rehg and Kanade [69] [70]
Tamura and Kawasaki [79]		Rohr [71]
Turk [80]		Shakunaga [74]
Yamato *et al.* [86]		Wang *et al.* [81]
		Webb and Aggarwal [82]
		Yamamoto and Koshikawa [85]
		Zhao [87]

Table 2: A selection of previous work on the visual analysis of human movement

can be achieved by a purely 2-D approach; a 3-D approach leads to a more accurate, compact representation of physical space which allows a better prediction and handling of occlusion and collision. It leads to meaningful features for action recognition, which are directly linked to body pose. Furthermore, 3-D recovery is often required for virtual reality applications.

2-D approaches have shown some early successes in the analysis of human movement. In some cases these successes were obtained relatively easily; for example, some work on motion-based recognition involved classification of a few, well separable, motion classes for which a multitude of features and classification methods could have been applied to obtain good results. In other cases, the application involved seemingly complex activities [65] [77] with no straightforward recognition solution. A main design choice for 2-D systems has been whether to use prior explicit models or to take a learning approach. It has been especially important for systems without explicit shape models to be able to accurately determine the foreground region in the image. Techniques based on background subtraction, color spotting, obstacle detection and independent motion detection have all been employed to provide this initial segmentation. Another issue for these systems has been the proper normalization of the features extracted from this foreground region, with respect to both the spatial and time dimension. Examples have included the use of scaled image grids and detection of periodicity. One of the challenges of 2-D systems on the topic of pose recovery is to show that they scale up to unconstrained movement.

It is fair to say that the results of vision-based 3-D tracking are still limited at this point. Few examples of 3-D pose recovery on real data exist in the literature and most of these introduce simplifications (e.g. constrained movement, segmentation) or limitations (e.g. processing speed) that still require improvement with respect to robustness. Robust 3-D tracking results have been particularly scarce for approaches using only one camera. The benefit of using multiple cameras to achieve tighter 3-D pose recovery has been quite evident [26] [43] [69]; body poses and movements that are ambiguous from one view (by occlusion or depth) can be disambiguated from another view. The added calibration effort has been worthwhile.

There are a number of challenges that need to be resolved before vision-based 3-D tracking systems can be deployed widely.

- The model acquisition issue. Almost all previous work assumes that the 3-D model is fully specified a priori and only addresses the pose recovery problem. In practice, the 3-D model is parametrized by various shape parameters that need to be estimated from the images. Some work has dealt with this issue by decoupling model acquisition and pose recovery, i.e. requiring a separate initialization stage where either known poses [26] or known movements [42] simplify the acquisition of the shape parameters. Although work in [42] represents a step forward on this matter, no approach has been provided that can recover both shape and pose parameters from uncontrolled movement, e.g. the case of a person walking into a room and moving freely around.

- The occlusion issue. Most systems cannot handle significant (self) occlusion and do not provide criteria when to stop and restart tracking of body parts. There is no notion of pose ambiguity either.

- The modeling issue. Human models for vision have been adequately parametrized with respect to shape and articulation, but few have incorporated constraints such as joint angle limits and collision, and even less have considered dynamical properties such as balance. In contrast to graphics applications, they have made little or no use of color and texture cues to capture appearance. Lacking entirely is the ability to deal with loose-fitting clothes. Finally, there is also a need to model the objects the human interacts with.

- Using ground truth. A quantitative comparison between estimated and true pose is very important to evaluate and compare systems. For simulations to be realistic, they have to include modeling, calibration and segmentation errors. Even better is obtaining ground truth on real data using markers and active sensing.

- Using 3-D data. Few systems (e.g. [62]) have used range data so far, given sensor-related drawbacks (e.g. high cost, low resolution, limited measuring range, safety concerns). Also, relatively few systems (e.g. [4] [41]) have obtained 3-D data by passive sensing techniques (i.e. triangulation) without relying on markers. Combining the use of 3-D data with some of the monocular techniques described in the previous Sections is likely to alleviate a number of problems related to figure-background separation, model acquisition and model fitting.

For both 2-D and 3-D approaches, the issue of tracking versus initialization remains open. Most work only deals with incremental pose estimation and does not provide ways for bootstrapping, either initially or when tracking gets lost. But it is the availability of an easy initialization procedure, which can be started up from a wide range of situations, what makes a system robust enough to be deployed in real world settings (e.g. [84]).

Another desirable extension to past work is the ability to detect and track multiple humans in the scene (one might even try crowds). Naive techniques which rely on background subtraction to obtain a segmented human figure, will no longer be feasible here. Stronger models might be necessary to handle occlusion and the correspondence problem between features and body-parts.

Action recognition is also an area which could welcome further attention. Particularly interesting is the question of whether a set of generic human actions can be defined which can be applied to a variety of applications. These generic actions might include those given in Table 3; a distinction is made between stand-alone actions and interactions with objects or other people. If indeed such a useful set of generic actions can be defined, would it be possible to identify corresponding features and matching methods which are, to a large degree, application-independent?

The classification of various actions also facilitates the introduction of a symbolic component on top of the image processing in order to reason about the scene. A variety of logic-based approaches come to mind for implementing this (e.g. conventional first-order logic, fuzzy logic, temporal logic). The connection from the sensory to the symbolic level can be provided by action recognizers such as those described in Section 6. The connection in the opposite direction, from symbolic to sensory level, also seems very useful; this would allow controlling what vision tasks are to be executed. For example in some person-tracking application, one might want to alternate the tracking mode from a fine-scale (with each body part tracked) to a coarse scale (with human body considered as a whole), depending on context.

Finally, it will be important of test the robustness of any of the resulting systems on large amounts of data, many different users, and in a variety of environments.

8 Conclusions

The visual analysis of human movement has become a major application area in computer vision. This development has been driven by the many interesting applications that lie ahead in this area and the recent technological advances involving the real-time capture, transfer and processing of images on widely available low-cost hardware platforms (e.g. PCs).

A number of promising application scenarios were discussed: virtual real-

Stand-alone Actions	**Interactions with objects**	**Interactions with people**
walking running jumping turning around bending over looking around squatting falling sitting (down) standing (up) climbing pointing waving clapping	grasping, carrying, putting down examining transferring (from one hand to another) throwing dropping pushing hitting shaking drinking, eating writing, typing	shaking hands embracing, kissing pushing hitting

Table 3: A sample of action verbs

ity, surveillance systems, advanced user interfaces and motion analysis. The scope of this survey was limited to the analysis of human gesture and whole-body movement; three main approaches were discussed: 2-D approaches without explicit shape models, 2-D approaches with explicit shape models and 3-D approaches. It was argued that which of the above approaches to pursue depends on the application; some general guidelines were given. Action recognition was considered in the context of matching time-varying feature data.

Although one appreciates from this survey the large amount of work that already has been done in this area, many issues are still open, e.g. regarding image segmentation, use of models, tracking versus initialization, multiple persons, occlusion and computational cost. One of the challenges for 2-D systems is to show that the approaches scale up to allow pose recovery for large set of movements from different viewpoints. 3-D systems still have to resolve issues dealing with model acquisition, detail of modeling and obtaining ground truth. Scenes, such as Figure 15, are far too complex currently. An interesting question is whether a set of generic human actions can be defined which are useful across applications and if so, what the features of interest would be. Added functionality and performance is likely to be gained by adding a symbolic component on top of the image processing to reason about the scene and control image tasks. Work on different sensor modalities (range, infrared, sound) will furthermore lead to systems with combined strengths.

By addressing the above issues, vision systems will have improved capabilities to successfully deal with complex human movement. This might transform the "Looking at People" domain into the "Understanding People" domain.

9 Acknowledgements

The author would like to thank the past authors who have contributed figures to this survey. Also, the support of Larry Davis (University of Maryland) and Franz May (Daimler-Benz, Ulm) is gratefully acknowledged.

References

[1] J. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Articulated and elastic non-rigid motion: A review. In *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 2–14, Austin, 1994.

[2] K. Aizawa and T. Huang. Model-based image coding: Advanced video coding techniques for very low bit-rate applications. *Proceedings of IEEE*, 83(2):259–271, 1995.

[3] K. Akita. Image sequence analysis of real world human motion. *Pattern Recognition*, 17(1):73–83, 1984.

[4] A. Azarbayejani and A. Pentland. Real-time self-calibrating stereo person tracking using 3-D shape estimation from blob features. In *Proc. of International Conference on Pattern Recognition*, Vienna, 1996.

[5] N. Badler, C. Phillips, and B. Webber. *Simulating Humans*. Oxford University Press, Oxford, 1993.

[6] N. Badler and S. Smoliar. Digital representations of human movement. *ACM Computing Surveys*, 11(1):19–38, 1979.

[7] A. Barr. Global and local deformations of solid primitives. *Computer Graphics*, 18(3):21–30, 1984.

[8] A. Baumberg and D. Hogg. An efficient method for contour tracking using active shape models. In *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 194–199, Austin, 1994.

[9] A. Blake, R. Curwen, and A. Zisserman. A framework for spatiotemporal control in the tracking of visual contours. *International Journal of Computer Vision*, 11(2):127–145, 1993.

[10] A. Bobick and A. Wilson. A state-based technique for the summarization and recognition of gesture. In *Proc. of International Conference on Computer Vision*, pages 382–388, Cambridge, 1995.

[11] Q. Cai and J. Aggarwal. Tracking human motion using multiple cameras. In *Proc. of International Conference on Pattern Recognition*, pages 68–72, Vienna, 1996.

[12] T. Calvert and A. Chapman. Analysis and synthesis of human movement. In T. Young, editor, *Handbook of Pattern Recognition and Image Processing: Computer Vision*, pages 432–474. Academic Press, 1994.

[13] L. Campbell and A. Bobick. Recognition of human body motion using phase space constraints. In *Proc. of International Conference on Computer Vision*, pages 624–630, Cambridge, 1995.

[14] C. Cedras and M. Shah. Motion-based recognition, a survey. *Image and Vision Computing*, 13(2):129–154, 1995.

[15] I.-C. Chang and C.-L. Huang. Ribbon-based motion analysis of human body movements. In *Proc. of International Conference on Pattern Recognition*, pages 436–440, Vienna, 1996.

[16] C. Charayaphan and A. Marble. Image processing system for interpreting motion in American Sign Language. *Journal of Biomedical Engineering*, 14(15):419–425, 1992.

[17] Z. Chen and H. Lee. Knowledge-guided visual perception of 3-D human gait from a single image sequence. *IEEE Transactions on Systems, Man and Cybernetics*, 22(2):336–342, 1992.

[18] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models - their training and applications. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[19] T. Darrell and A. Pentland. Space-time gestures. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 335–340, New York, 1993.

[20] J. Davis and M. Shah. Gesture recognition. Technical Report CS-TR-93-11, University of Central Florida, 1993.

[21] B. Dorner. Hand shape identification and tracking for sign language interpretation. In *Looking at people, International Joint Conference on Artificial Intelligence*, Chambery, 1993.

[22] A. Downton and H. Drouet. Model-based image analysis for unconstrained human upper-body motion. In *IEE International Conference on Image Processing and its Applications*, pages 274–277, 1992.

[23] U. Franke, D. Gavrila, S. Görzig, F. Lindner, F. Pätzhold, and C. Wöhler. Autonomous driving approaches downtown. *submitted to IEEE Expert (Special Issue on Vision-based Driving Assistance in Vehicles of the Future)*, 1997.

[24] W. Freeman, K. Tanaka, J. Ohta, and K. Kyuma. Computer vision for computer games. In *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 100–105, Killington, 1996.

[25] D. Gavrila. *Vision-based 3-D tracking of humans in action.* PhD thesis, Department of Computer Science, University of Maryland, 1996.

[26] D. Gavrila and L. Davis. 3-D model-based tracking of humans in action: a multi-view approach. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–80, San Francisco, 1996.

[27] A. Geurtz. *Model-based Shape Estimation.* PhD thesis, Department of Electrical Engineering, Polytechnic Institute of Lausanne, 1993.

[28] N. Goddard. Incremental model-based discrimination of articulated movement direct from motion features. In *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 89–94, Austin, 1994.

[29] L. Goncalves, E. Di Benardo, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3-D. In *Proc. of International Conference on Computer Vision*, pages 764–770, Cambridge, 1995.

[30] Y. Guo, G. Xu, and S. Tsuji. Understanding human motion patterns. In *Proc. of International Conference on Pattern Recognition*, pages 325–329 (B), 1994.

[31] T. Heap and D. Hogg. Towards 3-D hand tracking using a deformable model. In *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 140–145, Killington, 1996.

[32] B. Heisele, U. Kressel, and W. Ritter. Tracking non-rigid, moving objects based on color cluster flow. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 257–260, San Juan, 1997.

[33] Y. Hel-Or and M. Werman. Constraint fusion for recognition and localization of articulated objects. *International Journal of Computer Vision*, 19(1):5–28, 1996.

[34] M. Herman. *Understanding body postures of human stick figure*. PhD thesis, Department of Computer Science, University of Maryland, 1979.

[35] D. Hoffman and B. Flinchbaugh. The interpretation of biological motion. *Biological Cybernetics*, 42:195–204, 1982.

[36] D. Hogg. Model based vision: A program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[37] R. Holt, A. Netravali, T. Huang, and R. Qian. Determining articulated motion from Perspective views: A decomposition approach. In *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 126–137, Austin, 1994.

[38] E. Hunter, J. Schlenzig, and R. Jain. Posture estimation in reduced-model gesture input systems. In *Proc. of International Workshop on Automatic Face and Gesture Recognition*, pages 290–295, Zurich, 1995.

[39] G. Johansson. Visual perception of biological motion and a model for its analysis. Perception and Psychophysics, 14(2):201–211, 1973.

[40] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parametrized model of articulated image motion. In *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 38–44, Killington, 1996.

[41] R. Kahn, M. Swain, P. Prokopowicz, and J. Firby. Gesture recognition using the perseus architecture. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 734–741, San Francisco, 1996.

[42] I. Kakadiaris and D. Metaxas. 3-D human body model acquisition from multiple views. In *Proc. of International Conference on Computer Vision*, pages 618–623, Cambridge, 1995.

[43] I. Kakadiaris and D. Metaxas. Model-based estimation of 3-D human motion with occlusion based on active multi-viewpoint selection. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 81–87, San Francisco, 1996.

[44] R. Kjeldsen and J. Kender. Toward the use of gesture in traditional user interfaces. In *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 151–156, Killington, 1996.

[45] H. Kollnig, H.-H. Nagel, and M. Otte. Association of motion verbs with vehicle movements extracted from dense optical flow fields. In *Proc. of European Conference on Computer Vision*, pages 338–347, 1994.

[46] J. Kuch and T. Huang. Vision-based hand modeling and tracking for virtual teleconferencing and telecollaboration. In *Proc. of International Conference on Computer Vision*, pages 666–671, Cambridge, 1995.

[47] S. Kurakake and R. Nevatia. Description and tracking of moving articulated objects. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 491–495, The Hague, 1992.

[48] M. Leung and Y. Yang. First Sight: A human body outline labeling system. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 17(4):359–377, 1995.

[49] W. Long and Y. Yang. Log-tracker, an attribute-based approach to tracking human body motion. *International Journal of Pattern Recognition and Artificial Intelligence*, 5(3):439–458, 1991.

[50] N. Magnenat-Thalmann and D. Thalmann. Human modeling and animation. In *Computer Animation*, pages 129–149. Springer-Verlag, 1990.

[51] D. Marr and H. Nishihara. Representation and recognition of the spatial organization of three dimensional shapes. *Proc. of Royal Society London B*, 200:269–294, 1978.

[52] D. McNeill. *Hand and Mind - What Gestures Reveal about Thought.* The University of Chicago Press, Chicago and London, 1992.

[53] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 15(6):580–591, 1993.

[54] M. Mohnhaupt and B. Neumann. On the use of motion concepts for top-down control in traffic scenes. In *Proc. of European Conference on Computer Vision*, pages 598–600, Antibes, 1990.

[55] C. Myers, L. Rabinier, and A. Rosenberg. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on ASSP*, 28(6):623–635, 1980.

[56] S. Niyogi and E. Adelson. Analyzing and recognizing walking figures in XYT. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 469–474, 1994.

[57] S. Niyogi and E. Adelson. Analyzing gait with spatiotemporal surfaces. In *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 64–69, Austin, 1994.

[58] J. Ohya and F. Kishino. Human posture estimation from multiple images using genetic algorithm. In *Proc. of International Conference on Pattern Recognition*, pages 750–753 (A), 1994.

[59] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 193–199, San Juan, 1997.

[60] J. O'Rourke and N. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 2(6):522–536, 1980.

[61] V. Pavlovic, R. Sharma, and T. Huang. Visual interpretation of hand gestures for human-computer interaction: A review. Technical Report UIUC-BI-AI-RCV-95-10, University of Central Florida, 1995.

[62] A. Pentland. Automatic extraction of deformable models. *International Journal of Computer Vision*, 4:107–126, 1990.

[63] A. Pentland. Smart rooms. *Scientific American*, 274(4):54–62, 1996.

[64] F. Perales and J. Torres. A system for human motion matching between synthetic and real images based on a biomechanic graphical model. In *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 83–88, Austin, 1994.

[65] R. Polana and R. Nelson. Low level recognition of human motion. In *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–82, Austin, 1994.

[66] F. Quek. Eyes in the interface. *Image and Vision Computing*, 13(6):511–525, 1995.

[67] L. Rabinier. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings IEEE*, 77(2):257–285, 1989.

[68] K. Rangarajan, W. Allen, and M. Shah. Matching motion trajectories using scale space. *Pattern Recognition*, 26(4):595–610, 1993.

[69] J. Rehg and T. Kanade. Visual tracking of high DOF articulated structures: an application to human hand tracking. In *Proc. of European Conference on Computer Vision*, pages 35–46, Stockholm, 1994.

[70] J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. of International Conference on Computer Vision*, pages 612–617, Cambridge, 1995.

[71] K. Rohr. Towards model-based recognition of human movements in image sequences. *Computer Vision, Graphics and Image Processing: Image Understanding*, 59(1):94–115, 1994.

[72] M. Rosenblum, Y. Yacoob, and L. Davis. Human emotion recognition from motion using a Radial Basis Function network architecture. In *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 43–49, Austin, 1994.

[73] J. Segen and S. Pingali. A camera-based system for tracking people in real-time. In *Proc. of International Conference on Pattern Recognition*, pages 63–67, Vienna, 1996.

[74] T. Shakunaga. Pose estimation of jointed structures. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 566–572, 1991.

[75] A. Shio and J. Sklansky. Segmentation of people in motion. In *IEEE Workshop on Visual Motion*, pages 325–332, 1991.

[76] M. Spong and M. Vidyasagar. *Robot Dynamics and Control.* John Wiley and Sons, Inc., 1989.

[77] T. Starner and A. Pentland. Real-time American Sign Language recognition from video using hidden markov models. In *International Symposium on Computer Vision*, pages 265–270, Coral Gables, 1995.

[78] K. Takahashi, S. Seki, H. Kojima, and R. Oka. Recognition of dexterous manipulations from time-varying images. In *Proc. of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 23–28, Austin, 1994.

[79] S. Tamura and S. Kawasaki. Recognition of sign language motion images. *Pattern Recognition*, 21(4):343–353, 1988.

[80] M. Turk. Visual interaction with lifelike characters. In *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 368–373, Killington, 1996.

[81] J. Wang, G. Lorette, and P. Bouthemy. Analysis of human motion: A model-based approach. In *7th Scandinavian Conference on Image Analysis*, Aalborg, 1991.

[82] J. Webb and J. Aggarwal. Structure from motion of rigid and jointed objects. *Artificial Intelligence*, 19(1):107–130, 1982.

[83] C. Wilson, C. Barnes, R. Chellappa, and S. Sirohey. Face recognition technology for law enforcement applications. Technical Report 5465, NIST, 1994.

[84] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 19(7):780–785, 1997.

[85] M. Yamamoto and K. Koshikawa. Human motion analysis based on a robot arm model. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 664–665, Maui, 1991.

[86] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using Hidden Markov Model. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–385, 1992.

[87] J. Zhao. *Moving Posture Reconstruction from Perspective Projections of Jointed Figure Motion*. PhD thesis, University of Pennsylvania, 1993.

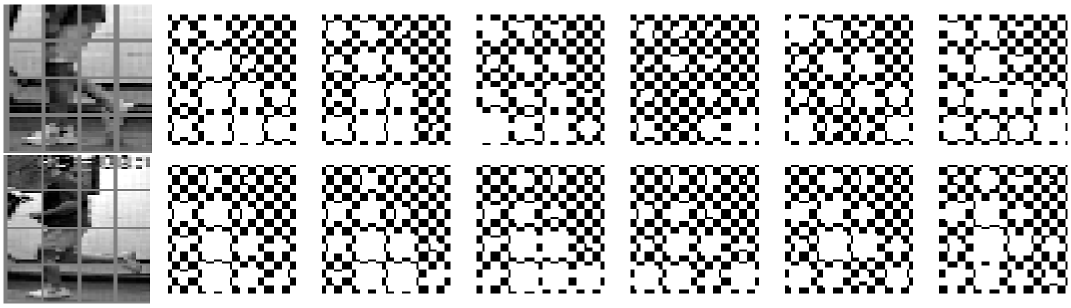
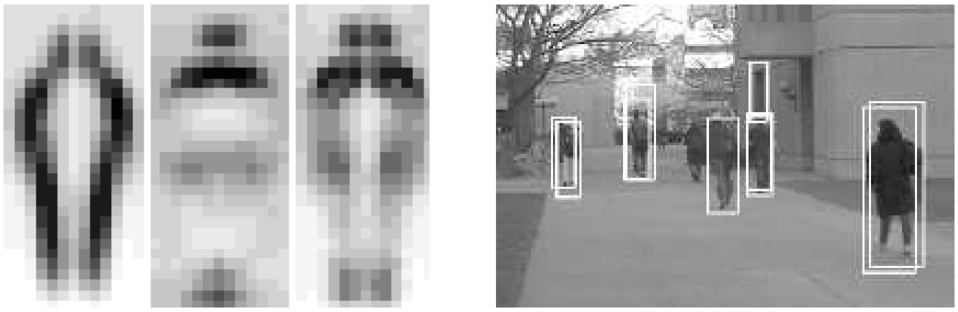


Figure 1: Detection of periodic activity using low-level motion features (from Polana and Nelson [65], © 1994 IEEE).



(a) (b)

Figure 2: Detecting frontal and rear views of pedestrians; (a) the features: vertical, horizontal and corner wavelet coefficients (b) the detection results using the SVM classifier (from Oren *et al.* [59], © 1997 IEEE).

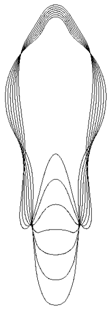


Figure 3: Principal component analysis on a data set of pedestrians represented by B-splines; shown is the shape variation along the principal component. (from Baumberg and Hogg [8], © 1994 IEEE).

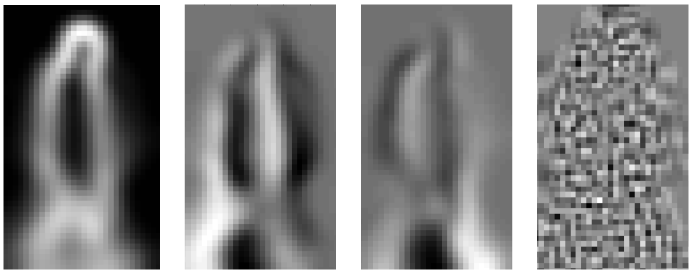


Figure 4: Principal component analysis on a data set of pedestrians represented by images of size 30 by 50 pixels; shown are eigenvectors 0 (mean), 1, 2, and 25, in order of decreasing eigenvalues (from Franke *et al.* [23]).

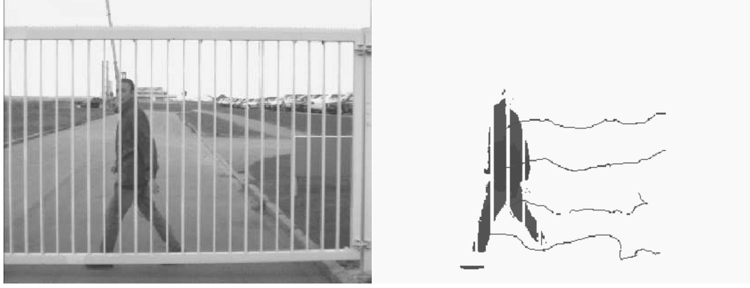


Figure 5: Tracking pedestrians with the Color Cluster Flow (from Heisele, Kressel and Ritter [32], © 1997 IEEE).

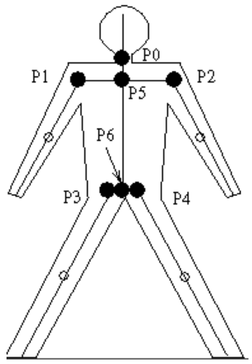


Figure 6: A 2-D stick-figure model fleshed out with ribbons (from Leung and Yang [48], © 1995 IEEE).

(a) (b)

Figure 7: (a) One image of a sequence with walking people (b) various slices in the XYT volume reveal characteristic patterns (from Niyogi and Adelson [57], © 1994 IEEE).

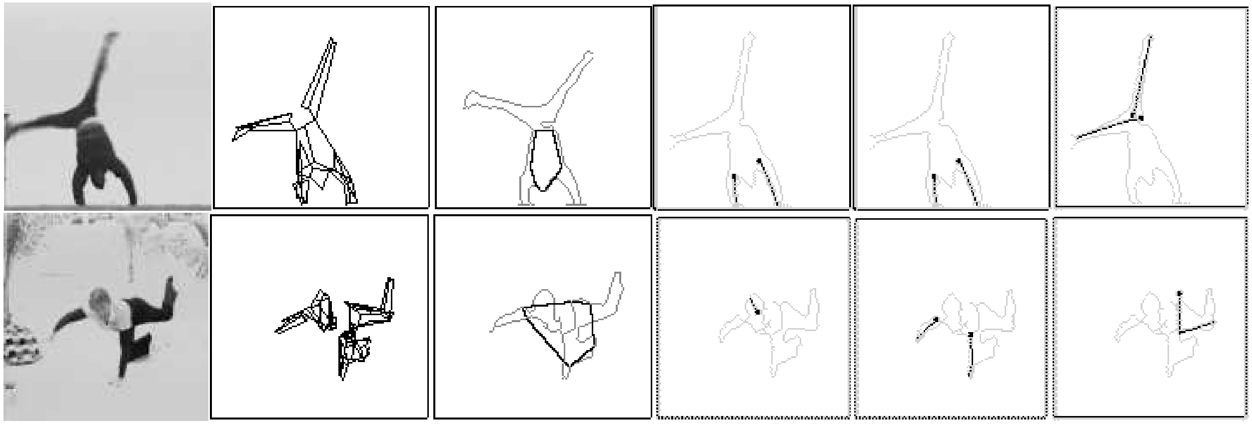


Figure 8: Original images, ribbon detection and body part labeling using the First Sight system (from Leung and Yang [48], © 1995 IEEE).

Figure 9: Detecting and tracking human "blobs" with the Pfinder system, (work by Wren *et al.* [84]).

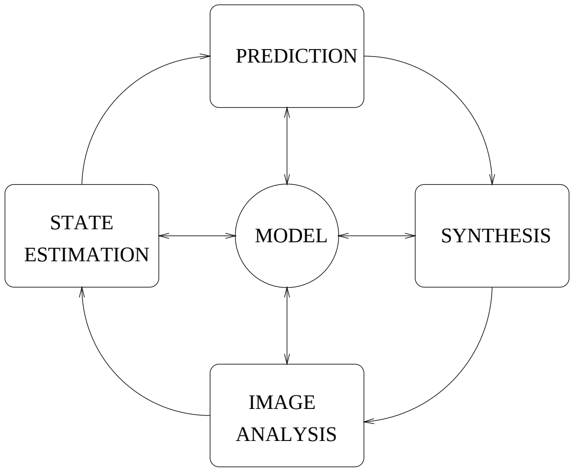


Figure 10: Model-based tracking (adapted from O'Rourke and Badler [60]).

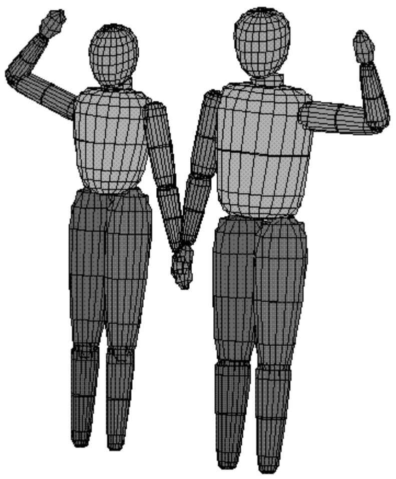
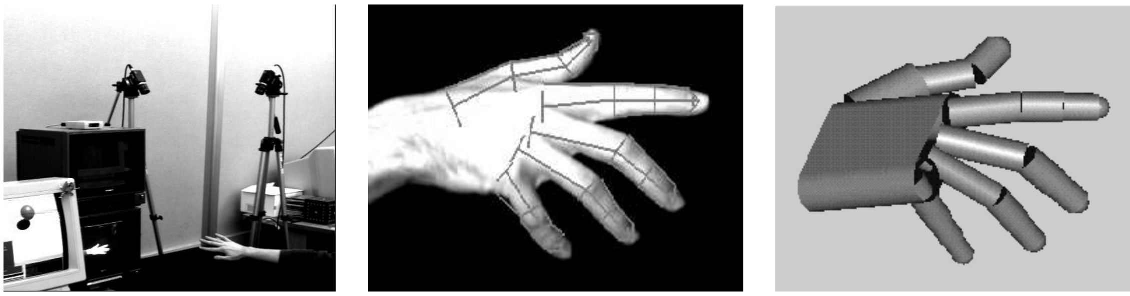


Figure 11: 3-D human models "ELLEN" and "DARIU" using tapered superquadrics (from Gavrila and Davis [26], © 1995 IEEE).



(a) (b) (c)

Figure 12: Hand tracking with the DigitEyes system (a) multi-camera setup (b) motion estimate superimposed on one of the two camera views (c) corresponding pose of 3-D hand model (from Rehg and Kanade [69], © 1994 Springer-Verlag).

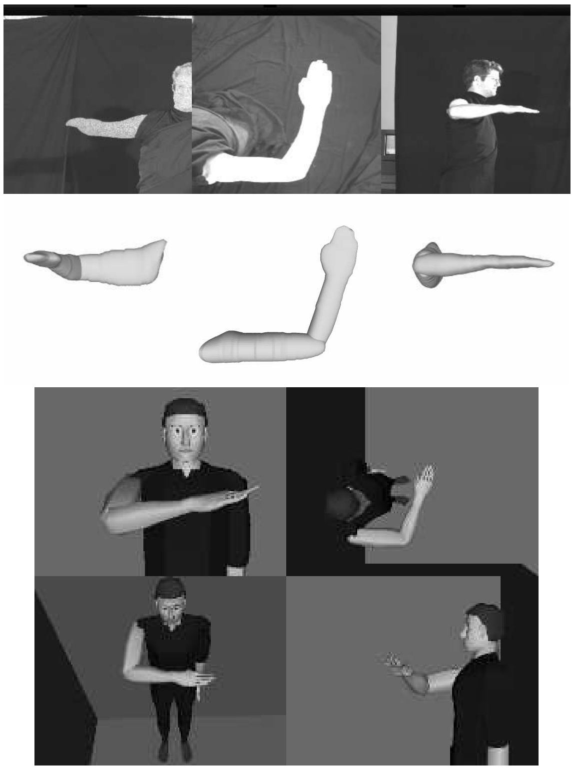


Figure 13: Multi-camera arm tracking: original images, recovered arm model and application to a whole-body graphical model (from Kakadiaris and Metaxas [43], © 1996 IEEE).

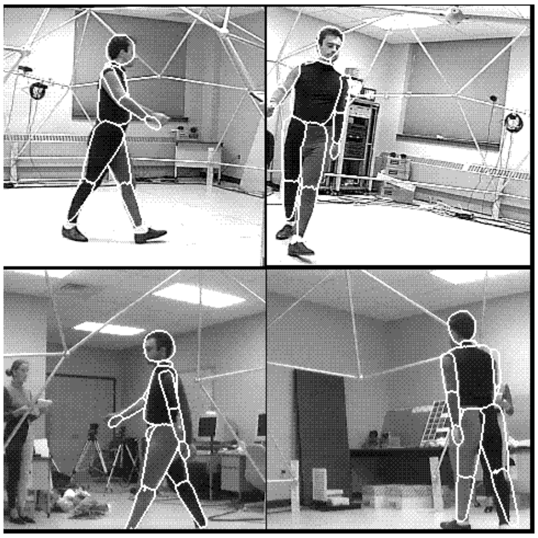


Figure 14: Multi-camera whole-body tracking; the current pose of the 3-D model is superimposed onto the four camera views (from Gavrila [25]).

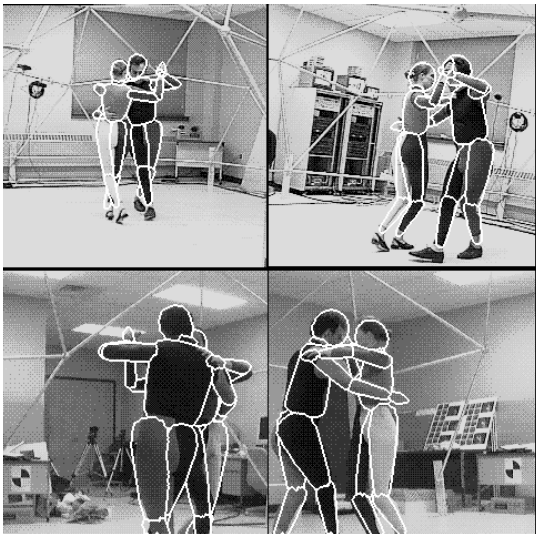


Figure 15: Will the Argentine Tango be danced in virtual reality? (from Gavrila and Davis [26], © 1996 IEEE).