A Survey of Computer Vision-Based Human Motion Capture

Thomas B. Moeslund and Erik Granum

Laboratory of Computer Vision and Media Technology, Aalborg University,
Niels Jernes Vej 14, 9220 Aalborg, Denmark
E-mail: tbm@cvmt.auc.dk, eg@cvmt.auc.dk

A comprehensive survey of computer vision-based human motion capture literature from the past two decades is presented. The focus is on a general overview based on a taxonomy of system functionalities, broken down into four processes: initialization, tracking, pose estimation, and recognition. Each process is discussed and divided into subprocesses and/or categories of methods to provide a reference to describe and compare the more than 130 publications covered by the survey. References are included throughout the paper to exemplify important issues and their relations to the various methods. A number of general assumptions used in this research field are identified and the character of these assumptions indicates that the research field is still in an early stage of development. To evaluate the state of the art, the major application areas are identified and performances are analyzed in light of the methods presented in the survey. Finally, suggestions for future research directions are offered.   © 2001 Academic Press

CONTENTS

1. INTRODUCTION

The analysis of human actions by a computer is gaining more and more interest. A significant part of this task is to register the motion, a process known as *human motion*

capture. Even though this term covers many aspects, it is mainly used in connection with capturing large scale body movements, which are the movements of the head, arms, torso, and legs. Formally we here define human motion capture as the process of capturing the large scale body movements of a subject at some resolution.

We included *at some resolution* to emphasize that tracking of a subject's limbs, as well as overall tracking of a subject, are considered to fall within the above definition. Hence, human motion capture is used both when the subject is viewed as a single object and when viewed as articulated motion of a high degree of freedom skeleton structure with a number of joints.

What is not covered by the above definition is small scale body movements such as facial expressions and hand gestures. A thorough review of hand gestures can be found in the survey by Pavlovic *et al.* [116].

1.1. Application Areas

The potential applications of human motion capture are the driving force of system development, and we consider the following three major application areas: surveillance, control, and analysis.

The *surveillance* area covers applications where one or more subjects are being tracked over time and possibly monitored for special actions. A classic example is the surveillance of a parking lot, where a system tracks subjects to evaluate whether they may be about to commit a crime, e.g., steal a car.

The *control* area relates to applications where the captured motion is used to provide controlling functionalities. It could be used as an interface to games, virtual environments, or animation or to control remotely located implements. For a comprehensive discussion of motion capture in the control application area, see [99].

The third application area is concerned with the detailed *analysis* of the captured motion data. This may be used in clinical studies of, e.g., diagnostics of orthopedic patients or to help athletes understand and improve their performance.

1.2. Alternative Technologies for Motion Capture

The systems used to capture human motion consist of subsystems for sensing and processing, respectively. The operational complexity of these subsystems is typically related, so that high complexity of one of them allows for a corresponding simplicity of the other. This trade-off between the complexities also relates to the use of active versus passive sensing. Active sensing operates by placing devices on the subject and in the surroundings which transmit or receive generated signals, respectively [99].

Active sensing allows for simpler processing and is widely used when the applications are situated in well-controlled environments. That is in particular the case for the third application area, analysis, and in some of the control applications.

Passive sensing is based on "natural" signal sources, e.g., visual light or other electromagnetic wavelengths, and requires no wearable devices. An exception is when markers are attached to the subject to ease the motion capture process. Markers are not as intrusive as the devices used in active sensing. Passive sensing is mainly used in surveillance and some control applications where mounting devices on the subject is not an option.

Computer vision with the passive sensing approach has challenged active sensing within all three application areas. Even though the use of markers may seem a good compromise

between passive and active sensing, it is still inconvenient for the subject (sometimes impossible) and computer vision allows in principle for touchfree and more discrete "pure" motion capture systems.

1.3. Content of This Paper

This paper is only concerned with computer vision-based approaches, i.e., passive sensing. It provides a compressive survey of publications in computer vision-based human motion capture from 1980 into the first half of 2000. The focus is on a general overview in relation to a functionally structured taxonomy rather than extended examples and summaries of individual papers.

Section 2 reviews briefly other surveys within the research field and the taxonomy of this survey is presented. It builds on the four primary functionalities of motion capture processing: initialization, tracking, pose estimation, and recognition. Approaches and techniques are presented in relation to these functionalities in Sections 3 to 6. The descriptions will focus on general principles and similarities among various systems and methods. Section 7 revisits the three application areas and discusses them in light of the survey and in the context of various performance parameters and examples of state-of-the-art systems. Furthermore, suggestions for future research directions are offered. Finally, Section 8 concludes the survey.

2. SURVEYS AND TAXONOMIES

Over the past two decades, the number of papers within the field of registering human body motion using computer vision has grown significantly. To structure an overview of the individual papers, their purpose, algorithms, etc. a taxonomy may be defined to arrange them into various groups having similar characteristics.

Various categories may be used for a taxonomy, e.g. (in random order): kinetic vs kinematic, model-based vs non-model-based, 2D approaches vs 3D approaches, sensor modality (visual light, infrared (IR) light, range data, etc.), number of sensors, mobile vs stationary sensors, tracking vs recognition, pose estimation vs tracking, pose estimation vs recognition, various applications, one person vs multiple persons, number of tracked limbs, distributed vs centralized processing, various motion-type assumptions (rigid, nonrigid, elastic), etc.

Which categories to use depends on the purpose of the survey, and the various published surveys have used different taxonomies.

2.1. Previous Surveys

Aggarwal *et al*. [2] give an overview of various methods used prior to 1995, in articulated and elastic nonrigid motion. After a good overview of various motion types the approaches within articulated motion with or without *a priori* shape models are described. Then the elastic motion approaches are described in two categories with and without a shape model.

Cedras and Shah [23] give an overview of methods within motion extraction prior to 1995, which are all classified as belonging to optical flow or motion correspondence. The human motion capture problem is described as action recognition, recognition of the individual body parts, and body configuration estimation.

An overview of the area of human motion estimation and recognition with special focus on optical flow techniques, prior to 1996, is given by Ju [72]. The overall taxonomy is motion estimation and motion recognition, both of which are divided into subclasses.
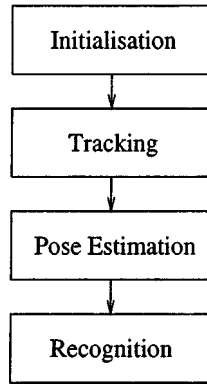
FIG. 1.   A general structure for systems analyzing human body motion.

In the survey by Aggarwal and Cai [1], which describes work prior to 1998, the same taxonomy as in Cedras and Shah [23] is applied even though they use different labels for the three classes. The classes are divided into subclasses yielding a rather comprehensive taxonomy.

A survey by Gavrila [44] describes work prior to 1998 and gives a good general introduction to the topic with a special focus on applications. The taxonomy covers 2D approaches with and without explicit shape models and 3D approaches. Across these three classes the approaches dealing with recognition are described.

2.2. A Taxonomy Based on Functionalities

The above taxonomies have different emphases depending on their purpose. We will focus on more general aspects such as the overall structure of a motion capture system and the various types of information being processed. The functional structure of a comprehensive motion capture system is shown in Fig. 1.

Before a system is ready to process data it needs to be *initialized*; e.g., an appropriate model of the subject must be established. Next the motion of the subject is *tracked*. This implies a way of segmenting the subject from the background and finding correspondences between segments in consecutive frames. The *pose* of the subject's body often needs to be *estimated* as this may be the output of the system, e.g., to control an avatar (the graphical representation of a human) in a virtual environment, or may be processed further by the recognition process. Some higher level knowledge, e.g., a human model, is typically used in *pose estimation*. The final process analyzes the pose or other parameters in order to *recognize* the actions performed by the subject.

A system need not include all four processes, especially since many of the systems described in this survey are research, where only a method within one of the processes is investigated. Still, all systems can be described within the structure.

More than 130 human motion capture papers published since 1980 are reviewed for this survey. They are all listed in the tables in the Appendix. Within the tables the papers are ordered first by the year of publication and second by the surname of the first author. Four columns allow the clarification of the contributions of the papers within the four processes of our taxonomy. The location of the reference number (in brackets) indicates the main topic of the work and an asterisk (∗) indicates that the paper also describes work at an interesting

level regarding this process. The tables show, among other things, that the majority of the work in human motion capture is carried out within tracking and pose estimation. This is reflected in the rest of the paper where these two processes receive more attention than the other two processes.

In earlier surveys detailed summaries of individual papers were used extensively to exemplify individual classes in the taxonomies. Generally we will not do this. First because we focus more on overall methods and general characteristics and second to avoid drowning the essence in implementation details. Therefore, only the ideas of individual papers are described, when relevant. Furthermore, when presenting an idea, concept, or method we will generally refer only to one paper where a good description is given. Readers interested in the contents of the papers can, beside earlier surveys, refer to [98] where detailed summaries of many of the papers used for this survey are presented. Actually, [98] could be considered an appendix to this paper.

2.3. Assumptions

As for computer vision papers in other fields, various assumptions on the conditions for motion capture are associated with the individual contributions. The actual assumptions made characterize the various systems and provide a useful reference for evaluation.

The typical assumptions may be divided into two classes: *movement assumptions* and *appearance assumptions*. The former concerns restrictions on the movements of the subject and/or the camera(s) involved. The latter concerns aspects of the environment and the subject. In Table 1 the relevant assumptions and their association with the two classes are listed.

The first three assumptions related to movements are very general and used in every system with a few exceptions; see, e.g., [19, 36, 105]. The fourth assumption is mainly used in human computer interaction (HCI) applications and simplifies the calculation of the overall body pose. The next assumption reduces the dimensionality of the problem from 3D to 2D and is often used in applications such as gait analysis. The sixth assumption concerns occlusion and simplifies the task of tracking the subject and limbs since the entire posture

TABLE 1

The Typical Assumptions Made by Motion Capture Systems Listed in Ranked Order According to Frequency

Assumptions related to movements	Assumptions related to appearance
	Environment
1. The subject remains inside the workspace	1. Constant lighting
2. None or constant camera motion	2. Static background
3. Only one person in the workspace at the time	3. Uniform background
4. The subject faces the camera at all time	4. Known camera parameters
5. Movements parallel to the camera-plane	5. Special hardware
6. No occlusion	
7. Slow and continuous movements	*Subject*
8. Only move one or a few limbs	1. Known start pose
9. The motion pattern of the subject is known	2. Known subject
10. Subject moves on a flat ground plane	3. Markers placed on the subject
	4. Special coloured clothes
	5. Tight-fitting clothes

of the subject is visible in every frame. The next assumption goes both for the movement of the camera (if it is allowed to move) and the subject. No sudden movements are allowed and the movements follow a simple and continuous trajectory. This assumption simplifies the calculation of the velocity of the subject and of the camera. The eighth assumption allows tracking to focus on only one or a few body parts. The next assumption is used to simplify the tracking and pose estimation problems by reducing the solution space. The final assumption allows a calculation of the distance between the camera and the subject using the camera geometry and the size of the subject.

The first environmental assumption in practice constrains the scene to be indoors. The next assumption requires a static background which makes it possible to segment the subject based on motion information. The third environmental assumption constrains the background further to have a uniform color and a simple thresholding may be used to segment the subject. The first two assumptions are used in many systems while the third assumption is used in approximately half of the systems. The fourth assumption concerns camera parameters which are necessary to know in order to obtain absolute measures in the registration. The last environmental assumption concerns the use of special hardware such as multiple cameras or an IR-camera.

The first subject assumption about the known start pose is introduced in many systems to simplify the initialization problem. The next assumption concerns prior knowledge of the subject, e.g., in terms of specific model parameters such as the subject's height, length, and width of limbs. The last three subject assumptions reduce the segmentation problems by making the subject's structure easier to detect.

The above assumptions are used to make the human motion capture problem tractable and they are applied in varying number and selection in all the reviewed papers. Which assumptions a particular system uses depends on its goals. Generally the complexity of a system is reflected in the number of assumptions introduced; i.e., the fewer the assumptions, the higher the complexity.

3. INITIALIZATION

As the first of the four major categories of our taxonomy we discuss initialization. Initialization covers the actions needed to ensure that a system commences its operation with a correct interpretation of the current scene. Sometimes the term initialization is also used for the preprocessing of data; see, e.g., Meyer *et al.* [97] or Rossi and Bozzoli [125]. We discuss preprocessing in Section 4 as part of the tracking procedure. Some of the initialization may be performed offline prior to the start of operation, while other parts preferably are included as the first phase of operation. Initialization may be simplified by relying on some of the assumptions discussed above. Initialization mainly concerns camera calibration, adaption to scene characteristics, and model initialization.

As for other computer vision systems the parameters of the camera often need to be known. These can be obtained through offline camera calibration, and for a stationary camera setup occasionally recalibration will suffice. If something in the setup regularly changes, a procedure for online calibration may be preferred as in the work by Azarbayejani *et al.* [7]. However, virtually all other systems are based on offline calibration.

Initialization to adapt the scene characteristics mainly relates to the appearance assumptions and the segmentation methods (described in Section 4) using them. In systems based on these assumptions a typical offline initialization is carried out to find the thresholds and

capture reference images which will be used during processing. In some systems initialized parameters are used in an adaptive procedure to calculate (and update) scene characteristics on the fly [53].

Model initialization is concerned with two things: the initial pose of a subject and the model representing the subject. Both are closely related to model-based pose estimation described in Section 5.

Rohr [123] uses a model-based approach to estimate the pose of a subject and he describes the overall problem as first finding the initial pose of the human and then incrementing it from frame to frame. This structure represents the design of many model-based pose estimation systems. In the majority of these systems the overall problem is reduced either by assuming that the subject's initial pose is known as a special start pose [27] or by having the operator of the system specify it [152]. Perales and Torres [117] take this idea to the extreme by having the operator specify the pose in every single frame. Zheng and Suezaki [155] use a similar approach but they only manually fit the pose at some key frames and have the system interpolate, using correlation, between frames.

Only a few systems actually have a special initialization phase where the start pose is found automatically [123]. In some systems the same algorithm is used for initialization and during tracking–pose estimation [108]. This indicates that no temporal information is used and nothing is learned by the system during processing. These systems are usually not considered initialization since they do not address the initialization problem in a general sense, but rather in a special situation constrained by the assumption of a known motion pattern.

The result of a model-based approach is usually dependent on how well the subject fits the human model in the system, i.e., the complexity of the model. Some systems use a general model which is an average of many individuals [9]. Others measure the current subject and generate a model based on these data. This may be done offline [49, 86] or online as by Wren *et al*. [142]. In the latter, analysis of the subject's initial pose is carried out to build an initial model which is refined as more information becomes available to the system. Generating a personalized model mainly relies on building up a 3D shape of the subject through multiple cameras, e.g., stereo reconstruction [102, 106], and then mapping texture onto the shape model. The personalized model may also be obtained by fitting a generic model to current data [57, 155].

In computer graphics the concept of using real images of humans to animate personalized human models is becoming more popular. A clear tendency toward the merging of computer vision and computer graphics is appearent [85] and personalized models are being incorporated into computer vision-based motion capture systems [55] to improve performance.

4. TRACKING

Tracking is a well-established research field which may be addressed from various viewpoints. In this context we define tracking as establishing coherent relations of the subject and/or limbs between frames. What is needed to achieve this depends on the context. Tracking may be seen as a separate process, as a means to prepare data for pose estimation, or as a means to prepare data for recognition.

If considered a separate process the subject is typically tracked as a single object (without any limbs) and no high-level knowledge is used. An example is the work by Tsukiyama

and Shirai [135]. They detect moving people in a hallway by first detecting moving objects, then finding the objects corresponding to humans, and finally tracking the moving humans over time.

If the tracking process prepares data for pose estimation its purpose is to extract specific image information, either low level, such as edges, or high level, such as hands and head. An example of low-level tracking is seen in the Walker system by Hogg [59]. Here edges are extracted from the image and matched against the edges of a human model to determine the pose of the model/human. An example of high-level tracking is seen in the Pfinder system by Wren *et al.* [142] where the human body is tracked in 2D using statistical models of the background and foreground to segment the image into blobs. Some of these blobs represent the hands and feet of the subject yielding some sort of limb representation.

If tracking prepares data for recognition, the task is usually to represent data in an appropriate manner. An example of this was published by Polana and Nelson [120] where flow information and down-sampling are used to represent image information in a compact manner which is processed by a classifier to recognize six different classes, e.g., walking and running.

Independent of the context of tracking, three common aspects can be identified. First, nearly every tracking algorithm within human motion capture starts with the figure–ground problem, i.e., segmenting the human figure from the rest of the image. Second, these segmented images are transformed into another representation to reduce the amount of information or to suit a particular algorithm. Third, how the subject should be tracked from frame to frame is defined.

4.1. Figure–Ground Segmentation

Figure–ground segmentation may be based on either temporal or spatial information. Table 2 shows how the data may be characterized in further detail.

4.1.1. Temporal data. The use of temporal data is mostly based on the assumption of a static background (and camera). In this case the differences between images from a sequence must originate from the movements of the subject. Two subclasses may be introduced: subtraction and flow.

Subtraction is widely used by simply subtracting the current image from the previous image in a pixel by pixel fashion, using either the intensity values [120] or the gradients [4]. An improved version is to use three consecutive images instead of two [78]. The result reflects movements (and noise) between the images unless the subject has the same intensity

TABLE 2
The Various Characteristics and Subclasses of
Figure–Ground Segmentation Approaches

Temporal data		Spatial data	
Subtraction	Flow	Threshold	Statistics
Two images	Points	Chroma-keying	Pixels
Three images	Features	Special clothes	Blobs
Background	Blobs	IR	Contours

or color as the background. The use of background subtraction is very popular. If the scene is static, a noise-free background image without any subject(s) may be recorded and used as a reference in a subtraction scheme [105]. A more advanced version is to update the background image during processing [53].

Flow is used here as a general term describing coherent motion of points or features between image-frames. An example of such a flow using points is described by Yamamoto and Koshikawa [152]. They find the motion parameters of a human body part by calculating the optical flow of several points within this part and compare them to the movements of a model of the part. Gu *et al*. [48] instead track edge features, defined by their length and contrast, in consecutive images using the optical flow constraint. In the work by Bregler [17] each pixel is represented by its optical flow which is grouped into blobs having coherent motion and represented by a mixture of multivariate Gaussians.

Figure-ground segmentation based on temporal data assumes the subject to be the only moving object in the scene (or rather in the region of interest). In many cases temporal data are a strong alternative to spatial data. Temporal data are usually simpler to extract and focus directly on the target of motion capture. A number of systems are based on mapping the motion data directly to human pose through an inverse kinematic framework; see, e.g., [18].

4.1.2. Spatial data. The use of spatial data falls into two distinct subclasses: *thresh-olding* and *statistical approaches*. The former is simple processing based on special environmental assumptions. The latter is a rather advanced class where some of the appearance assumptions exploited by the *subtraction* methods are relaxed.

If the subject's color or intensity appears different from the rest of the scene it may be segmented using simple thresholding. A good example is Chroma-keying where a person appears in front of a one-color, usually blue, screen wearing nonblue clothes. By thresholding, the person can easily be separated from the background [32, 65]. The opposite approach, where the subject wears one color, usually dark, clothes against a different background is also very popular [12]. A special version of this idea is to make the subject wear markers (passive or active) which are easily segmented by thresholding [21, 46]. A related approach is to use an IR camera. Thermal images can be obtained where the subject is easy to segment through thresholding as the only "hot" object in the scene [66].

The *statistical* approaches use the characteristics of individual pixels or groups of pixels to extract the figure from the background. The characteristics are typically colors and edges. Some approaches are inspired by the background subtraction methods described above. A sequence of background images of the scene is recorded and the mean and variance intensity or color of each pixel are calculated over time. In the current image each pixel is compared to the statistics of the background image and classified as belonging to the background or not [150]. This approach is becoming increasingly popular due to its robustness compared to the subtraction approaches. In the work by McKenna *et al*. [96] the approach is combined with the statistics of pixel gradients to remove the shadows cast by subjects. A more advanced version is used in the blob approaches, where the subject is modeled by a number of blobs with individual color and spatial statistics. Each pixel in the current image is then classified as belonging to one of the blobs according to its color and spatial properties [142].

Another statistical approach is to use static or dynamic contours, where dynamic is usually named the active contour, and static refers to the use of predefined static structures representing a part of the subject's outline. These structures consist of edge segments and other attributes. Long and Yang [92] uses *Logs*, which is an area defined by two parallel

edge segments and a number of attributes. These are found in the image and used to extract human body parts by comparing them with a human model consisting of Logs. The active contours are used in a similar manner except they have the ability to adjust their shape on the fly, hence active. Their deformation is controlled by external and internal energy functions. The former fits the curve to the image features, e.g., edges, while the latter adjusts the smoothness of the curve. This type of active contours is also called *snakes* and works relatively well when the changes in the structure of the object are unknown. If a shape model is used the active contour is known as a deformable template [13]. Active contours may be used to extract the entire outline of the subject [9] or to extract individual body parts [75]. A related approach is to extract the silhouette instead of the contour. Rigoll *et al*. [121] use a stochastic approach to silhouette extraction. They use pseudo-2D hidden Markov models (HMM) (nested one-dimensional HMMs), to extract the silhouette in a discrete cosine transformed (DCT) representation of the image.

The use of thresholding to process spatial data is strongly dependent on a number of appearance assumptions. This might seem unreasonable if the goal is an assumption-free vision system with human capabilities. However, there will always be a range of applications where the environment and the subject can be controlled, as underlined by the fact that one of the most widely used and robust figure–ground segmentation methods is Chroma-keying. In more unconstrained applications, the statistical methods are a far better choice due to their adaptability. The use of pixel statistics is a good concept, but region-based methods, using, e.g., blobs, tend to be more reliable. On the other hand it is more difficult to model larger entities of correspondingly higher complexity. The active contours aim directly at extracting the shape of the subject and can be very efficient. However, they require a good initial fit and have difficulties with complex articulated objects such as the human body. The best way to exploit them seems to require an active contour for each body part.

4.2. Representation

Segmented entities are described compactly by some convenient representation. There are in principle two types of representations: the *object-based representation* which is based on the figure–ground segmentation and the *image-based representation* which is derived directly from the image. In Table 3 the various types of representations are shown.

4.2.1. Object-based representation. The object-based representations rely mainly on the output from the figure–ground segmentation. Therefore some of the arguments and descriptions from Section 4.1 also apply here.

TABLE 3
The Various Types of Data
Representations

Object-based	Image-based
Point	Spatial
Box	Spatio-temporal
Silhouette	Edge
Blob	Features

The point representation is sufficient in systems using passive or active markers. The active markers yield a high contrast in the images and provide a robust representation [131]. If more than one camera is used in a marker-based system a 3D representation may be obtained [103].

The box representation is used in many systems. The philosophy is to represent the subject by a set of boundary boxes containing the pixels or regions found in the figure–ground segmentation process. Some systems track these boxes over time [105], while others use them as intermediate representations [142], which are processed (rerepresented) further in pose estimation.

The silhouette representation is popular due to its simplicity. It can be obtained using the thresholding or subtraction methods from figure–ground segmentation. It is used in both 2D and 3D. The 2D representation is usually straightforward [91], but can also be more complex as in the work by Baumberg and Hogg [9], were the silhouette (or rather active contour) is represented even further using closed uniform B-splines with a fixed number of control points equally spaced around the silhouette. This actually makes it a contour representation but equivalent to a silhouette. The 3D (volumetric) silhouette can be obtained using combined 2D silhouettes [15] or directly using stereo approaches [71]. The silhouette representation may, as was the case for the box representation, be tracked directly [53] or more likely processed (rerepresented) further in pose estimation [52].

The blob representation typically follows some of the figure–ground segmentation approaches described under flow and statistics. The subject is represented as a blob or a number of blobs each having some similar characteristics. The similarities can be coherent flow [73], similar colors [54], or both [17]. The main philosophy of grouping information according to similarities is inspired by research into the human visual system by the Gestalt school in the 1930s [82].

4.2.2. Image-based representation. This class of representations is based directly on the pixels of the image. The representations are either derived from the image independent of the presence of an object or possibly constrained to the interior of one of the representations described above. These images (or image parts) may be transformed into another space spanned by non-Cartesian basis functions, yielding a more compact representation of the data or image. Transformations used are, e.g., Fourier, principle component analysis (PCA), DCT, and wavelets.

A more advanced representation is obtained when including the temporal dimension in the representation [25]. This allows for motion-related features to be included in the description.

A third subclass of image-based representations is edges. They may be represented by points [45] or as line segments which are more robust to noise [123].

The last form of representation is features. Features are usually computed from one of the previously mentioned types of representation combined with additional information. Christensen and Corneliussen [27] use the length, area, and color to represent the individual body parts found by thresholding an image of a subject wearing special colored clothes.

4.3. Tracking over Time

Tracking over time is finding corresponding objects in consecutive frames where the objects may be any of the representations from Table 3. The difficulties of this task are related to the complexity of the scene and the complexity of the tracked objects. The

latter is again related both to the degrees of freedom of individual objects and to their representation. Tracking more points in an object is equivalent to tracking multiple objects simultaneously. The points or objects (hereafter objects) may split and merge into new objects due to occlusion or image noise, or the appearance of an object may change due to shadows and changes in the lighting.

The correspondence analysis is often supported by prediction. Based on previously detected objects and possibly high level knowledge the state of the objects (appearance, position, etc.) in the next frame is predicted and compared (using some metric) with the states of objects found in the actual image. Prediction introduces a region-of-interest in both image space and state space and hereby reduces the overall need for processing. The prediction of the various state parameters is based on a model of how they evolve over time. A model of velocity and acceleration [101] or more advanced models of movements such as walking [123] may be used. An alternative approach is to learn probabilistic motion models prior to operation [115]. A commonly used method for prediction is the Kalman filter, which is also capable of estimating the uncertainties of the prediction. These uncertainties may be used to determine the regions-of-interest.

The Kalman filter is unfortunately restricted to situations where the probability distribution of the state-parameters is unimodal. In the presence of occlusion, cluttered background resembling the tracked objects, and complex dynamics, the distribution is likely to be multimodal. Alternative tracking algorithms have therefore been developed capable of tracking multiple hypotheses, i.e., support multimodal distributions. Most recognized is perhaps the Condensation algorithm [64]. It is based on sampling the posterior distribution estimated in the previous frame and propagating these samples to form the posterior for the current frame. The method has shown to be a powerful alternative to the Kalman filter [112, 130]. However, since it is nonparametric it requires a relatively large number of samples to ensure a fair maximum likelihood estimate of the current state. In high-dimensional problems a more efficient method might be necessary. In the work by Cham and Rehg [24] only the peaks of the posterior distribution are sampled and propagated to the next frame, resulting in relatively few samples. Furthermore, a distribution is formed by piecewise Gaussians where the means are given by the propagated (predicted) samples and the covariances by the uncertainties of the predictions. Hence the distribution is parametric and allows for a direct maximum likelihood estimate.

Another tracking aspect arises when multiple cameras are used. Some systems have access to more cameras than required and therefore need to decide which camera(s) or image(s) to use at each time instant. Various measurements are introduced to quantify the individual viewpoints in terms of ambiguity [112], occlusion [75], and in general reliability of data [67, 137].

5. POSE ESTIMATION

Pose estimation is the process of identifying how a human body and/or individual limbs are configured in a given scene. Pose estimation can be a postprocessing step in a tracking algorithm or it can be an active part of the tracking process. Various levels of accuracy may be required in pose estimation. At one extreme coarse estimations are carried out yielding only information about the subject's hands and head (or even simpler, the body's center of mass) which could be used in a surveillance system or in an HCI system. At the other extreme the precise pose in terms of positions, orientation, width, etc. of each limb

is estimated. The latter allows for, e.g., direct copy-cat interfaces to virtual environments or as input data to medical studies. Due to the complexity involved in this type of pose estimation, only one subject or a few body parts are considered.

A common aspect of pose estimation is the use of a human model. Usually a geometric model of the human body is applied, but other models, such as motion models, may also be applied. The model may be used for various purposes and at various levels. The general concept of using a human model is to exploit the fact that the system is dedicated to analyze a human and therefore may incorporate knowledge about humans into its processing.

Due to the extensive use of human models it is a good cue for a taxonomy. In former surveys the dimensionality of the applied (geometric) model is used to distinguish between various model-based approaches. This provides a good overview of 2D and 3D approaches, as seen in the surveys by Aggarwal and Cai [1] and Gavrila [44], respectively. However, in some cases it is not clear whether a system is a 2D or a 3D approach, e.g., when 2D data are combined into 3D data using triangulation. Instead, we suggest investigating how various systems apply a human model. Pose estimation is therefore divided into three classes. The first class, *model-free* covers approaches where no *a priori* model is used. The other two classes of pose estimation methods, *indirect model use* and *direct model use*, are characterized by having a human model beforehand. In the indirect case the model is used as a reference to constrain and guide the interpretation of measured data. In the direct case the model is maintained and updated by the data, and hence at any time it includes an estimate of the pose.

5.1. Model-Free

In this class of methods no *a priori* model is used. The methods do, however, build some sort of model to represent the pose. These pose representations are points, simple shapes, and stick-figures. The two first are similar to the object-based representations described in the previous section, while the stick-figure representation is more advanced.

The pose of the subject may be represented by a set of points and is widely used when markers are attached to the subject. Without the markers, the hands and the head may be estimated from the image represented by just three points [142]. This is a very compact representation which suffices for a variety of applications. These three points are usually found using color segmentation [101] or blob segmentations [142].

A subject may be represented by simple boundary boxes [32]. The box representation is, however, mainly used as an intermediate representation during processing and not as the final representation. Instead, shapes which are more human-like, such as ellipses [105], may be used as a final representation.

The stick-figure representation includes structure information resembling the human skeleton. It is a popular representation in systems where the gait of a subject is studied. The stick-figure is obtained in various ways, e.g., directly using a medical axis transformation [12] or a distance transformation [66]. A distance transformation is slightly more advanced allowing for suppression of noninteresting parts, e.g., arms when searching for the torso. Both methods give a direct but noisy estimate of the human skeleton.

A rather different approach to pose estimation without the use of an *a priori* model is to learn a mapping directly from image features to pose data. These types of systems rely on extensive training using ground truth data obtained with commercial motion capture systems. Below two examples are given.

Rosales and Sclaroff [124] use 3D ground truth data of a subject's joints to train their system. For a number of viewpoints the 3D data are mapped into 2D joint positions. For the same viewpoints a cylinder model is used to synthesize a silhouette and its Hu-moments are calculated. The 2D joint positions are for all training sets clustered into a number of Gaussians using the expectation maximization (EM) algorithm [39]. For each cluster a neural network is trained which maps Hu-moments into 2D joint positions. During processing the Hu-moments are calculated from an image silhouette and feed to each neural network, resulting in a pose candidate for each cluster. The candidates are each synthesized and their Hu-moments are compared to those found from the image in order to evaluate the different hypotheses. This comparison is similar to the use of silhouettes described in Section 5.3.

In the work by Brand [16], the paths through the 3D state space obtained through training are modeled using an HMM where the states are linear paths modeled by multivariate Gaussians. As in [124] the moments (central) are found by synthesizing various poses. The moments are associated to the HMM; i.e., a new HMM is obtained with moments as input and pose parameters as hidden states. Altogether a sequence of moments is mapped to the most likely sequence of 3D poses. Obviously this approach is either offline or includes a significant time-lag, but it has the ability to resolve ambiguities using hindsight and foresight.

5.2. Indirect Model Use

The methods in this class use an *a priori* model when estimating the pose of a subject. They use the model as a reference or look-up table from which relevant information may be extracted to guide the interpretation of measured data.

Various types of models and various levels of detail are used. The level of details ranges from just the height of the subject to both structure and dynamic information about subjects. The estimated pose of systems in this class is generally not very detailed. Typically positions of the head, hands, and feet or a rough description of the entire human body is used to represent the pose.

A simple human model is the aspect ratios between the various limbs [20, 53] which may be used to guide the pose estimation. In the work by Leung and Yang [90] the outline of the subject is estimated as edge regions described using 2D ribbons which are U-shaped edge segments. A 2D ribbon model guides the labeling of the image data by searching for structures similar to the model. To select among the various labeled image parts they apply more high-level knowledge such as expected motion. Njåstad *et al*. [108] also assume a known motion pattern which is used to define where to search for the individual body parts with respect to the torso. A related concept is to have key frames beforehand as used by Attwood *et al*. [6] and Akita [3], who exploits them to predict occlusion in the next frame, which again guides the processing. Haritaoglu *et al*. [52] went a step further by first calculating which overall pose is present (standing, craw-bend, laying down, or sitting) and then using this information to find the individual parts. Wren and Pentland [145] generalized the concept of knowing the motion beforehand by suggesting that various behavior models may be used to improve the pose estimation. An example of this was also published by Iwai *et al*. [65] where six different motion models are used.

On the borderline to direct model use O'Rourke and Badler [114] describe a rather detailed human model. The model is used to ensure that the predicted human pose is realistic. The

same approach is used in Hunter *et al.* [62] where the estimated pose is projected onto a feasible manifold in the solution space, ensuring a realistic pose. Ioffe and Forsyth [63] use the probability of various model configurations to guide the estimation of the pose of one or more people. In the work by Wren and Pentland [144] model information is used to estimate the pose of the elbows after the head and hands have been located.

5.3. Direct Model Use

By direct model use we mean that an *a priori* human model is used as the model representing the observed subject. It is then continuously updated by the observations. Hence, this model provides any desired information including pose at any time. Approximately 40% of the surveyed papers use a model in this manner and they are all listed in Table 4.

The models used in this section are generally very detailed. They explicitly exist within the computer program and are used intensively in the processing phase. An important benefit gained by introducing a human model is the ability to handle occlusion and the ease by which various kinematic constraints may be incorporated into a system.

A human model is represented by a number of joints and the sticks (bones) connecting them. The sticks and the "flesh" surrounding them may be represented in various ways depending on the level of detail needed by a system. In Table 4 the type of model used by each system, the number of segments in the model, and the parts of the subject being estimated are listed in columns 3, 4, and 5, respectively. The more complex the model, the better results may be expected but on the expense of more processing and training.

The concrete representation of the human model is a state space where each axis represents a degree of freedom of a joint in the model. One pose of the subject may be expressed as one point in the state space as oppose to many points in the 2D image space. The problem is how to use this state space representation and, hence, how to relate image data to pose data. The general approach is known as *analysis-by-synthesis* and is used in a *predict–match–update* fashion. The philosophy is to predict the pose of the model corresponding to the next image. The predicted model is then synthesized to a certain abstraction level for comparison with the image data.

When comparing the real and synthesized data a similarity measure is used to evaluate how alike the image and the synthesized model are. Typically this is done for a number of predicted model poses until the correct (best) pose is found and used to update the model in the system.

Clearly the state space, even with a coarse resolution, describes a very large number of possible model poses unreasonable to synthesis for matching. Therefore constraints are introduced to prune the state space. An obvious and substantial reduction of the range of each parameter and its derivatives in the state space is achieved by introducing the kinematic constraints of the human motor system, e.g., the bending of the elbow is between 0 and 145°. This may be used directly to partition the state space into legal and illegal regions, as in [100], or the constraints may be defined as forces acting on an unconstrained state phase, as in [144]. The fact that two human body parts cannot pass through each other also introduces constraints. Another approach to reduce the number of possible model poses is to assume a known motion pattern—especially cyclic motion such as walking and running. In the work by Rohr [123] gait parallel to the image plane is considered. Using a cyclic motion model of gait all pose parameters are estimated by just one parameter, which specifies the current phase of the cycle. This is the most efficient pruning encountered in the surveyed

TABLE 4
The Papers where a Human Model Is Used Directly

Year	First author	Model type	Parts	Object	Ab. level	Dimension
1983	Hogg [58]	Cylinders	14	Body	Edges	$2\frac{1}{2}$
1984	Hogg [59]	Cylinders	14	Body	Edges	$2\frac{1}{2}$
1985	Lee [83]	Stick-Figure	17	Body	Joints	3
1989	Attwood [6]	Cylinders	16	Body	Joints	3
1991	Wang [140]	Cylinders	2	Leg	Motion	3
1991	Yamamoto [152]	CAD Model	4	U Body	Motion	$2\frac{1}{2}$
1992	Lee [84]	Stick-Figure	17	Body	Joints	3
1992	Luo [94]	Stick-Figure	6	Body	Silhouettes	3
1992	Wang [141]	Cylinders	2	Leg	Motion	3
1993	Kameda [79]	Patches	17	Body	Silhouettes	$2\frac{1}{2}$
1994	Guo [50]	Stick-Figure	10	Body	Sticks	2
1995	Goncalves [47]	R-circular cones	2	Arm	Edges	3
1995	Kameda [80]	Patches	15	Body	Silhouettes	$2\frac{1}{2}$
1996	Gavrila [45]	Super-Quadrics	10	Body	Edges	3
1996	Ju [73]	Planar Patches	2	Leg	Motion	2
1996	Kakadiaris [77]	D Silhouettes	2	Arm	Silhouettes	3
1996	Kameda [78]	Patches	9	U Body	Silhouettes	$2\frac{1}{2}$
1997	Christensen [27]	Cylinders	10	Body	Sticks	2
1997	Lerasle [86]	CAD Model	2	Leg	Texture	3
1997	Meyer [97]	Boxes	6	Body	Contours	$2\frac{1}{2}$
1997	Rohr [123]	Cylinders	14	Body	Edges	$2\frac{1}{2}$
1997	Wachter [138]	R-Elliptical Cones	10	Body	Edges	3
1998	Bregler [18]	Ellipsoids	10	Body	Motion	3
1998	Fua [42]	Ellipsoids	2	Arm	Silhouettes	3
1998	Jojic [71]	D Super-Quadrics	(5)6	(U) Body	Contours	3
1998	Kakadiaris [75]	D Silhouettes	4	Arms	Silhouettes	3
1998	Li [91]	Rectangles	14	Body	Silhouettes	2
1998	Munkelt [103]	Modified Cylinders	10	Body	Joints	3
1998	Silaghi [131]	Stick-Figure	16	Body	Sticks	3
1998	Wren [145]	Stick-Figure	5	U Body	Blobs	3
1998	Yamamoto [151]	CAD Model	9	Body	Motion	3
1998	Yaniz [154]	Stick-Figure	16	Body	Sticks	3
1999	Cham [24]	Scaled Prismatics	10	Body	Texture	2
1999	Delamarre [37]	Cones and Spheres	15	Body	Silhouettes	3
1999	Iwai [65]	Stick-Figure	6	U Body	Silhouettes	3
1999	Lerasle [87]	CAD Model	2	Leg	Texture	3
1999	Njåstad [108]	Cylinders	10	Body	Contours	$2\frac{1}{2}$
1999	Ong [112]	Stick-Figure	4	Arms	Contour	3
1999	Pavlović [115]	Scaled Prismatics	6	Body	Texture	2
1999	Plänkers [119]	Ellipsoids	6	Arm	Depth	3
1999	Wachter [139]	R-Elliptical Cones	10	Body	Edges	3
1999	Wren [146]	Stick-Figure	5	U Body	Blobs	3
2000	Hu [60]	Rectangles	10	Body	Silhouette	2
2000	Moeslund [100]	Cylinders	2	Arm	Silhouette	3
2000	Moeslund [101]	Cylinders	2	Arm	Silhouette	3
2000	Okada [111]	Boxes & Cylinders	4	U Body	Motion/Depth	3
2000	Rosales [124]	Cylinders	10	Body	Silhouette	2
2000	Sidenbladh [130]	Cylinders	10	Body	Texture	3
2000	Wren [144]	Stick-Figure	5	U Body	Blobs	3
2000	Yamamoto [153]	CAD Model	11	Body	Motion	3

Note. The first and second column states the publication year and first author. The third shows how the human is modeled. The fourth shows the number of parts in the model. The fifth shows which part of the subject is analyzed. The sixth shows which abstraction level the system works at and the last shows the dimensionality of the pose estimated data. D, deformable; R, right; U, upper.

literature. Ong and Gong [112] map training data into the state space and use PCA to find a linear subspace where the training data can be compactly represented without losing too much information. Pavlović *et al.* [115] take this idea a step farther by learning the possible or rather likely trajectories in the state space from training data; i.e., dynamic models are learned. Yet another approach is to rerepresent the state space more efficiently (without losing any information as in PCA). In the work by Moeslund and Granum [100] the state space representation of an arm is given in just two parameters instead of the usual four (two for the shoulder and two for the elbow) using geometric reasoning.

The systems also differ in the way they compare the synthetic data and the image data. Even after the constraints have been applied a brute force comparison is seldom realistic. Therefore, the matching problem is formulated as a function which is optimized. Due to the high dimensionality of the problem an analytic solution is not possible and a numeric iterative approach is generally used [139]. An alternative approach is to predict just one state and let the difference between the synthesized data and the measurements be used as an error signal to correct the state of the model [47, 146]. An excellent framework for this type of approach is the Kalman filter. Yet another approach is to predict the most likely states using a multiple hypothesis framework (see Section 4.3) [24].

If proper constraints are introduced, the state space is reduced significantly. Combining this with an effective match scheme the analysis-by-synthesis approach can be very successful. The papers in Table 4 all use the analysis-by-synthesis approach and they have produced some of the best results within human pose estimation. To give a better understanding of how the analysis-by-synthesis approach is used the various abstraction levels which may be used in a system are described in the following.

5.3.1. Abstraction levels. The various abstraction levels used for comparing image data and synthesized data are edges, silhouettes, contours, sticks, joints, blobs, depth, texture, and motion. In Table 4 (column 6) each system and its level of abstraction can be seen. Below the various levels are exemplified using a number of references.

The *edges* of the model and the subject in the image are relatively easy to find, especially when the appearance-related assumptions are used. Edges are therefore a good representation for the matching process. One of the first analysis-by-synthesis publications was by Hogg [58] in 1983. He used image subtraction to obtain a boundary box of a human. In this box he found edges and compared them with edges projected from a human model. He successfully showed the effect of the analysis-by-synthesis approach. Rohr [123] used the same idea, but in a more sophisticated system where he used the Kalman filter, edge segments, and a motion model tuned to walking to obtain a more robust result. Gavrila and Davis [45] worked with the same problem but without assuming a known motion model. They instead utilized four camera views and tight-fitting colored clothes to obtain good results. They compare image and model edges using a robust variant of chamfer matching. In the work by Wachter and Nagel [139] both edges and regional information is applied in the matching. They also work with unconstrained motion patterns, but only using monocular vision. Another interesting issue is that they only use edge segments which are expected to be visible in the image.

A *silhouette* and its *contours* are, like edges, relatively easy to extract from both the model and the image. A silhouette has the advantage over edges of being less sensitive to noise, since it is a region-based data type. On the other hand, fine details may be lost in the extraction of the silhouette. In the work by Kameda *et al.* [79] the silhouette extracted

from the image is matched against the silhouette of the model and the similarity measure is simply given as the area difference. They use a local match strategy, i.e., one limb at the time. Hu *et al.* [60] also use a local match strategy but they consider both positive and negative matching errors and introduce, weighting of the two in their similarity measure. Furthermore, they apply multiscale morphologic matching to obtain better results. The idea is that the main structural information is reserved in large morphologic scale. So both the input and model silhouettes are dilated in large scale, removing a significant amount of noise and thereby improving the result of the matching.

If the contour, rather than the silhouette, is used the fitting between the image and model data is usually done via active contours, i.e., the forces which should be applied to the model (a deformable rather than a geometric model) to make it fit the image contours are calculated [71, 75]. In the work by Meyer *et al.* [97] a geometric model is used. The parts of the subject are found from optical flow and represented by their contours. These are obtained by active rays which are similar to active contours, except the problem is reduced from 2D to 1D. The contours are then compared to predicted model contours. In some systems a stick-figure model is matched against an image silhouette. This may, for example, be seen in the work by Luo *et al.* [94] where the silhouette of the subject from two different views are compared with a synthesized stick-figure model using kinematic match criteria.

The human is represented by its joints or stick-figure since it reflects anatomic features of the human. Both may be easily obtained from the model since it is basically defined in these terms. However, it may be hard to obtain them directly from the image and usually various assumptions are introduced to simplify matters. In the work by Lee and Chen [83] the positions of the joints in the image and the 3D length of each segment are known beforehand. Given the 3D position of the neck a partial tree is build. At each node one of two solutions for the next joint is possible, since the joint's projection in the image and the 3D length are known. A path through the tree is equal to one body pose. The tree is pruned using kinematic constraints and the assumption that the subject is walking. In [84] they improve their system by introducing a smooth motion constraint. In the work by Attwood [6] a similar approach is taken, except that he uses three static postures (standing, kneeling, and sitting) instead of a walking assumption to prune the solution space. In Munkelt *et al.* [103] the 3D positions of the joints are estimated using markers and stereo. These are compared with 3D model joints using a graph-based scheme to find the correct pose of the subject.

A stick-figure representation is closely related to a joint representation and also reflects anatomic features which make it tractable. In the work by Guo *et al.* [50] a model stick-figure is compared to the image-skeleton found from the silhouette. To reduce the complexity of the matching, a potential field is introduced. It transforms the problem into one of finding a stick-figure with minimal energy in a potential field. Prediction and angle constraints of the individual joints are introduced to simplify the matching process further.

In the work by Wren and Pentland [144–146] blobs are used as the abstraction level. The Pfinder algorithm [142] (see Section 4) is run on two images from two different cameras producing 3D blobs for the hands and the head. Together with a dynamic model, kinematic constraints, and a Kalman filter the blobs provide enough information to estimate the pose parameters of the upper body. Another interesting aspect of this work is, as mentioned earlier, the inclusion of behavior models into the control loop. These models resemble the effect of an active complex controller (nervous system) as the kinematic and dynamic models resemble the motor system. Due to the difficulties in explicitly modeling the nervous system they apply learned probabilistic models, each corresponding to a prototypical behavior. They

choose a model according to current observations and thereby allow multiple hypotheses as in the Condensation algorithm.

In the work by Plänkers *et al*. [119] depth data of a subject's right arm is estimated using three cameras. The subject wears tight fitting textured clothes to simplify the correspondence problem. The depth data are compared to model ellipsoids in an optimization framework.

In the work by Lerasle *et al*. [86, 87] a texture representation is used; i.e., the image data are used without much preprocessing. Offline they generate a model of a subject's leg, which is texture mapped using the real image data where the subject wears a pair of textured tights. During processing they compare the texture in the image to the texture on the model using correlation. This is done for several cameras, and by merging the results the 3D motion of the leg is estimated. In the work by Sidenbladh *et al*. [130] a texture model of each limb is generated offline. They use a commercial motion capture system to obtain ground truth 3D pose data. Based on these data and a camera model they derive a mapping between an image pixel and its position on a cylinder—modeling a limb. In this way they automatically build a texture model of each cylinder/limb. Using PCA the models are compactly represented in a linear subspace. During processing they synthesize various poses, i.e., texture maps, and compare them to the image data to evaluate the similarity. The similarities of the synthesized poses constitute the posterior distribution in the Condensation algorithm where either a smooth motion model or a gait model is used to propagate the distribution over time.

The previous abstraction levels have been used in a spatial context where the structure of the subject is synthesised and compared with image measurements in each frame. Obviously the temporal context is also considered, but it is mainly to constrain the search space and thereby making the approach suitable for real-time implementation. The systems described later all use the temporal context more intensively by estimating the motion between consecutive frames.

Instead of first finding the pose of the subject in several individual frames and then using this information to calculate the motion, one may measure the motion of the subject between images and set up an inverse kinematic framework which makes it possible to calculate the corresponding motion in the 3D model. That is, the model and therefore the pose are updated based on the motion in the images. This was first done by Yamamoto and Koshikawa [152] in 1991. They measured optical flow within various body parts and used that through a Jacobian matrix to update the model. Ju *et al*. [73] use two planar patches to model a leg—a cardboard model. The motion of each patch is defined by eight parameters. For each frame the eight parameters are estimated by applying the optical flow constraint on all pixels in the predicted patches. The distance between the corners of the predicted patches are constrained to reduce the complexity of the estimation. Bregler and Malik [18] extended the concept by introducing a twist motion model and exponential maps which simplify the relation between image motion and model motion. Their novel formulation also has the advantage of being open to both single and multiple views. In addition to lab video sequences they also tested their system on a number of the famous Muybridge images recorded in 1884.

5.3.2. Dimensionality of pose estimation approaches.   The distinction between 2D and 3D systems concerns the dimensionality of the pose estimation approaches based on a direct use of a human model. Classification of the relevant systems is provided in the last column in Table 4 where the dimensionality is indicated as 3D, 2D, and $2\frac{1}{2}$D according to the following definitions.

3D refers to estimation of 3D movements. 2D refers to estimating either motion carried out in 2D, e.g., swinging an arm fronto-parallel to the camera, or to the motion observed directly in the image, which per definition is 2D.

$2\frac{1}{2}$D refers to either estimating 3D pose data based on 2D processing or testing a 3D pose estimating framework on pseudo-3D data. The former occurs, for example, when a subject is walking fronto-parallel to the camera. Estimating the pose of the arm and leg closest to the camera is a 2D problem but due to symmetry and *a priori* knowledge the pose of the opposite arm and leg may be estimated producing 3D pose data, or rather $2\frac{1}{2}$D. The latter refers to situations where a 3D model is used but for some reason the test data are only $2\frac{1}{2}$D, e.g., when the subject is walking fronto-parallel to the camera. This makes it hard to judge the success of a 3D pose estimating system and therefore these systems are also referred to as being $2\frac{1}{2}$D.

5.3.3. Evaluating pose estimation.   Another aspect arising when discussing systems estimating 3D (and pseudo-3D) poses is how to evaluate their results. In the case of 2D pose data a straightforward comparison with the image data may be used in most cases, and pose estimation tailored to a recognition task should of course be tested in this context; see Section 6.

The problem of evaluating the estimated poses is that usually no ground truth is available. Therefore alternative test methods are used which may be divided into quantitative tests and qualitative tests.

Quantitative tests rely on estimating the ground truth in some way and comparing it to the estimated data. One way is to move the subject and/or limbs along a well-defined path where the ground truth is known, e.g., a rectangular pattern on a table [47] or a circular groove engraved in a glass plate [75]. Another approach is to use a static object which is measurable, e.g., a doll [71], and yet another is to use synthetic data [77] or hand-segmented data [18].

Qualitative tests are widely used and rely on visual inspection. The most straightforward way is to project the estimated 3D pose into the image and inspect, visually, how alike the two are. The projection may be on top of the subject [45] or next to him or her [80]. Another form of visual inspection is to apply the estimated motion to a virtual character and see (from various viewpoints) if the movements seem human-like [119].

In the future when the motion capture devices based on active sensing, see Section 1.2, become cheaper, less noisy, and easier to use, this might be a good way to obtain ground truth.

6. RECOGNITION

The recognition aspect of human motion capture can be seen as a kind of post processing. It is relevant to include since the recognition guides the development of many motion capture systems as it is their final or long term goal. The recognition is usually carried out by classifying the captured motion as one of several types of actions. The actions are normally simple, such as walking and running, but more advanced actions such as different ballet dance steps have also been studied.

Traditionally two different paradigms exist: recognition by reconstruction and direct recognition. The former is based on the concept of first reconstructing the scene and then recognizing it, while the latter recognizes directly on the low-level data, e.g., motion, without

much preprocessing. Which one is preferable is hard to say and both paradigms may be supported by studies into the human visual system. Johansson [69] showed in his moving lights displays (MLD) experiments that the actions of a human may be recognized solely on motion (of the lights). This totally agrees with the idea of recognition directly through motion. Later Sumi [133] tried to redo Johanson's experiments but turned the data upside down. This resulted in a very poor recognition rate suggesting that a variant model of some kind is used when recognizing motion.[1] This could, of course, be a motion model but it could also be a geometric model which agrees with the former paradigm. Another fact which supports the latter is that humans may recognize different postures from a single frame, i.e., without any motion cues.

The reconstruction paradigm is most relevant for motion capture due to its strong relation to pose estimation, but both paradigms can be seen in the literature reviewed for this survey. Common for the reviewed recognition systems is that if pose estimation is used, it is simply a tool for generating higher level data. The various systems may be taxonomized with respect to the two paradigms, but some of the systems do not follow one of the paradigms in a strict sense, e.g., when recognition is based directly on image data (not motion). Instead a structure of first representing and then classifying the data is used.

A more relevant distinction is to look at whether the recognition is static or dynamic, i.e., whether the recognition is based on one or more frames.

6.1. Static Recognition

Static recognition is concerned with spatial data, one frame at a time. The approaches usually compare prestored information with the current image. The information may be templates [132], transformed templates [113], normalized silhouettes [52], or postures [22]. The goal of static recognition is mainly to recognize various postures, e.g., pointing [74], standing and sitting [6], or specially defined postures used in interfaces [4, 41].

In the interactive karaoke system build by Sul *et al.* [132] the postures of the subject are used to trigger and control the system. The postures are recognized by comparing the image data to different prerecorded templates. Templates are also used in the work by Oren *et al.* [113]. Offline they segment pedestrians in a number of images and generate a common template based on Haar wavelets. In run-time the template is compared to various parts of an image to find pedestrians. In the work by Freeman *et al.* [41] a computer chip capable of doing on-board image calculations is presented. The chip is used to calculate the orientation histogram of an image in real time. These histograms are matched against prerecorded histograms of various human postures and the current pose may be found. In the work by Jojic *et al.* [70] a dense depth map of the scene wherein a subject is pointing is used. After a depth-background subtraction the data are classified into points belonging to the arm and points belonging to the rest of the body. The index finger and top of the head are found as the two extreme points in the two classes and the line through them defines the pointing direction.

---

[1] In a panel session at a CVPR'2000 postconference workshop "Human Modeling, Analysis and Synthesis," Pietro Perona performed a similar experiment. He showed a MLD sequence which everybody in the audience recognized as human motion. He then showed another MLD sequence which the majority of the audience believed originated from a huge spider (without considering the inherent difficulties involved in actually estimating the motion of a spider's limbs). In fact, it was the exact same MLD sequence, but visualised from above!

6.2. *Dynamic Recognition*

These approaches use temporal characteristics in the recognition task. Relatively simple activities, such as walking, are typically used as test scenarios. The systems may use low-level or high-level data.

Low-level recognition is typically based on spatio-temporal data without much processing. The data are spatio-temporal templates [25] and motion templates [120]. The goal is usually to recognize whether a human is walking in the scene or not [54]. More high level methods are usually based on pose estimated data. Such methods vary from correlation [12] and silhouette matching [32] to HMMs [17] and neural networks [54]. The objective is to recognize actions such as walking [123], carrying objects [51], removing and placing objects [96], pointing and waving [32], gestures for control [4], standing vs walking [53], walking vs jogging [115], walking vs running [43], and classifying various aerobic exercises [122] or ballet dance steps [21].

Chomat and Crowley [25] generate motion templates by using a set of temporal-spatial filters computed by PCA. A Bayes classifier is used to perform action selection. In the work by Polana and Nelson [120] motion templates are also used but in a different way. Six subsequent motion images are computed. Each motion image is rerepresented by a subsampling where each new pixel contains the number of motion pixels from the original motion image. By representing the six subsampled images as a vector it can be classified using standard techniques. Niyogi and Adelson [107] also use temporal templates, but in a very special way. They generate an XT-slice by concatenating one of the lowest rows from each image in a sequence. If a walking figure is present an ankle profile of the figure can be seen in the XT-slice. By comparing this to prerecorded templates a walking figure can be recognized, and using *k*-nearest neighbors with Euclidean distance measures various persons may be recognized by their walking pattern. The idea of using only a small part of the body to recognize the walking person is also used in the work by Heisele and Wohler [54]. They segment the area containing the legs over time and process it by a neural network. In this way they can recognize whether a pedestrian is present or not. Davis and Bobick [33, 35] also use temporal templates which are created based on motion. They use the information about where and how much motion has been present in a sequence of frames. Both a motion-energy image and a motion-history image are used. By representing the templates by its seven Hu-moments [61] a Mehalanobis distance can be used to classify the action of the subject by comparing it to the Hu-moments of prerecorded actions.

More high-level dynamic recognition is usually based on pose estimated data for the different limbs. In the work by Ju *et al.* [73] the recognition of movements is based on the motion parameters of the individual body parts (legs). The problem is viewed as matching the curves of the motion parameters against a set of known curves. Yacoob and Black [149] use the results from [73] and build a classifier on top. After translating and scaling the motion parameters a PCA is used to obtain a more compact and discriminative representation. Four different activities are recognized (four variants of walking) by comparing the PCA-transformed data to offline generated (and PCA-transformed) data sets. In the work by Bharatkumar *et al.* [12] a stick-figure representation of the legs of a walking human is matched against a kinematic human walking model. This is done by correlating the two data sets with each other. Fujiyoshi and Lipton [43] also use a stick-figure representation of the legs of a person walking. They transform the time signal of the various parameters into the frequency domain, enabling them to separate walking from running. In the work

by Bregler [17] the idea of representing motion data by *movemes* (similar to phonemes in speech recognition) is suggested. This makes it possible to compose a complex activity (word) out of simple movemes. An HMM is used to classify three different gait categories: running, walking, and skipping. This type of high level symbolic representation is also used in the work by Wren *et al.* [143]. They automatically build a behavior alphabet (a behavior is similar to a moveme) and model each behavior using an HMM. The alphabet is used to classify different types of actions in a simple virtual reality game and to distinguish between the playing style of different subjects.

7. DISCUSSION

It is difficult to compare and grade the different systems and methods described in the four previous sections. The systems are based on different test data and different assumptions. However, some general comments can be made concerning the application areas in relation to some performance parameters.

7.1. Performance Characterization

In the Introduction, three main application areas were identified. In Table 5 these are related to three concepts which may be considered the main performance parameters in any motion capture system: robustness, accuracy, and speed. The symbols in the table indicate the performance requirements for each of the applications.

The *robustness* of a system is here related to the various assumptions shown in Table 1. The fewer assumptions a system imposes on its operational conditions, the more robust it is considered to be. Surveillance systems are aiming at very robust performance since they will often be working continuously at remote and uncontrolled locations, e.g., a parking lot. They should not be sensitive to changes in lighting, weather, number of people, clothes, etc. Furthermore, they are required to work autonomously and for long periods of time. Some control applications, e.g., gestures for signaling, are for the same reasons also subjected to this high level of robustness. Other systems in the control applications, e.g., direct avatar control, do not necessarily need this level of robustness, since they may operate in well-defined environments and for shorter periods of time where a number of assumptions conveniently may be applied. In the analysis applications the situation is similar. It is often possible to use a highly controlled setup and therefore the robustness is not the most important issue.

The *accuracy* of a system refers to how close the captured motion corresponds to the actual motion performed by the subject. Generally speaking the accuracy is directly proportional to the size of the human in the image. In surveillance systems the accuracy is rarely a key

TABLE 5
The Three Application Areas and Their Requirements of
the Three Main Performance Parameters

	Surveillance	Control	Analysis
Robustness	++	+/−	−
Accuracy	−	+	++
Speed	+	++	−

issue. It may not matter whether a surveyed subject is recognized to be walking around a car in a 3- or 4-m radius. The important thing is that the behavior is recognized. For the control purpose the situation is different. In many applications a one-to-one mapping between the movement carried out by the subject and the action he or she controls is critical, e.g., in tele-surgery and other futuristic telepresence applications. In the last area the need for accuracy is even more pronounced, since the applications rely directly on the pose estimation output.

The processing *speed* of a system is usually divided into real-time and offline processing. In this context the definition of real time is not clear even though many researchers use it in their systems. One definition is that each frame is processed before a new frame is recorded. Another definition relates to the motion which is being captured. A simpler way to view speed is to divide it into online and offline processing. In that sense surveillance systems require high speed since the images need to be processed before the car is stolen! In the control applications an even higher requirement for speed is needed. Actually the second definition of real time would apply in this case. In the last application area the processing may be performed offline, and if so no special requirements for high speed are needed.

7.2. State of the Art

In this section we will discuss state of the art within the three application areas. Furthermore, we will relate the three areas to the major issues which have emerged from the survey.

Surveillance applications are mainly carried out in uncontrolled environments. Therefore the figure–ground segmentation relies mostly on motion data, since these are less dependent on various assumptions such as a known subject, known lighting, and different markers. For the same reason, object-based representation (Table 3) is the natural way of representing the images at a higher level. Surveillance applications are generally more focused on tracking than on any of the three other processes in Fig. 1. Therefore, the surveillance applications mainly use one of the two simple forms of pose estimation where no model or only an indirect use of a model is used. Due to the nature of this application area the dynamic recognition approach is widely used.

An example of state of the art is the W4-system by Haritaoglu *et al.* [53] where the aim is to survey and recognize interactions between people and people or objects in an outdoor setting.[2] They detect and track multiple people and their body parts. The system works with monocular gray-scale images and infrared images. It uses a standard predict–match–update scheme, where it matches predicted objects or persons with measured objects or persons (in the image). The objects are obtained by detecting movements using an adaptive background subtraction, yielding a motion boundary box. The position (and motion) parameters of a person are estimated in two steps, first median matching for a coarse matching and then silhouette correlation between two consecutive frames for the fine matching. The individual body parts, head, torso, hands, legs, and feet, are found using a cardboard model of a walking human as reference. Online, the system is able to track multiple people and their limbs and cope with occlusions. Furthermore, it can detect and track objects carried and exchanged by people [51].

---

[2] The W4-system has also been used in an indoor setting [110] where it was extended with Kalman filters and kinematic constraints from [144].

The ultimate surveillance system tracks multiple humans and all their (inter)actions in real time. Haritaoglu *et al.* [53] may be heading in the right direction, but improved performance is required for handling a dynamic background and nontrivial movement patterns. Furthermore, a more intelligent handling of multiple objects and their occlusion is needed. Perhaps a further step could be to use 3D data in a direct model-based pose estimation scheme.

Some control applications are concerned with the recognition of gestures. In this case the methods used are generally similar to the ones used in the surveillance applications. However, if the application is more in the form of direct animation, e.g., avatar control, different methods are used. This type of application is carried out in an indoor setting where a number of assumptions may be introduced, e.g., known subject, known background, and known start pose. Then the appearance-based figure–ground segmentation methods are applied. To obtain good accuracy, direct use of a human model is usually used.

As an example of state of the art we consider the work by Wren *et al.* [143]. First of all they use the Pfinder algorithm [142] as the underlying tracking methods. It is a probabilistic method which segments the subject into a number of blobs and tracks those over time. This method has proven to be fast, robust, and able to directly estimate the positions of the head and hands, which are of great importance in control applications. They apply two Pfinder algorithms to obtain 3D estimates of the hands and head. Using a human model and kinematic constraints they estimate the 3D pose of the upper body. In the framework of a Kalman filter the model is predicted into the next frame to support the blob segmentation and tracking. The innovation of the Kalman filter is used to learn the various motion patterns (behaviors) of the subject. These can then be incorporated into the filter to improve the state estimates and predictions, i.e., a better pose estimation result.

The last application area is concerned with analysis of the captured motion and is typically used for clinical studies. These applications are carried out in well-controlled environments meaning that a number of assumptions may be introduced. In commercial systems markers are used which allow point representation of the data. A model of the human is necessary for interpreting the data. Usually it is not used directly in the pose estimation, but rather indirectly. The use of markers yields stable tracking, but the obtained points are not placed directly on the skeleton (for obvious reasons). Therefore an offset distance is introduced between the markers and the physical skeleton. This is, besides initialization and calibration, where the main problems are in state of the art commercial systems. In future systems one goal is to move away from the marker approach and aim at the more pure computer vision solution without the use of markers [38]. This will make systems more flexible and less cumbersome. The solution may be based on detailed human models used directly in the processing.

An example of direct use of a model is the work by Gavrila and Davis [45]. They use a model-based approach to track a subject in 3D. A recognition cycle goes as follows. Based on the current and previous states, the allowed intervals for each body parameter (e.g., joint angles) are predicted. For each combination of the 22 body parameters the human model is synthesized from the cameras' point of view. They compare edges between the synthesized model and the images and thereby (re)formulate the problem as a search problem—how to compare two edge images (a real image with a synthesized image). The search problem is solved using a robust variant of chamfer matching. When they find the best fit (highest similarity measure) the model is updated using these parameters. They use four synchronized sequences from four different cameras and run the algorithm for each

view. In order to obtain stable edges they wear tight-fitting colored clothes. The high number of joints in their relatively detailed model, the four cameras, and relatively few assumptions make it a rather complex system which, to some extent, is able to estimate the pose of an entire subject.

Future work in marker-free systems includes improved initialization to obtain a good model and the pose of the current subject fast and reliably. Although the analysis-by-synthesis approach seems to be the right one, it is still rather slow and computationally demanding. Methods to prune the state space and faster optimization schemes are required. Generally we may expect to see workable systems in the analysis area before we see them in the control area because the requirements for speed are relaxed in the analysis applications.

7.3. Future Directions

Although assumptions might be acceptable (e.g., Chroma-keying) for some applications, it is evident from the number of assumptions applied in the papers reviewed for this survey that the research field is still in a phase of development. Perhaps inspiration may be found in related research fields, e.g., speech recognition. First of all a tremendous amount of time is spent on recording and labeling training data in this field. These data are of a general nature, i.e., they suit a number of speech recognition tasks, and are represented in a well-defined modeling language which are the atoms (e.g., phonemes) of the spoken language. One reason for not spending the same amount of time on the training phase in computer vision might be the lack of a general underlying modeling language, i.e., how to map the images into symbols. An alphabet consisting of motion-entities would make computer vision-based human motion capture much easier, since it will transform the pose estimation problem into a recognition problem, i.e., recognize a sequence of symbols. This has already happened in Bregler [17] where the letters are called movemes, and Wren and Pentland [143] where the letters are called behaviors. Even though their alphabets are rather limited it is still a step in a very interesting direction. Furthermore, by having such an alphabet a vocabulary may be introduced to constrain the task at hand, as is the case in speech recognition.

Besides the current lack of a general alphabet, another reason for not using extensive training data is the amount of time required to actually capture and label human motion data. One solution is to use commercial motion capture systems (e.g., magnetic sensors) [16] which, when calibrated, easily produce thousands of labeled data sets. Another solution is to apply computer graphics to synthesize the appearance of a human model from various viewpoints, as in [124].

Another aspect of speech recognition, which is actually being seen more often in computer vision, is the use of probabilistic models for aspects other than recognition, e.g., modeling the position of the head using a Gaussian density. Some of these models are learned automatically using unsupervised methods, such as the EM algorithm. The entire tracking framework is also widely based on probabilistic methods such as the Kalman filter and the Condensation algorithm. Also, HMMs [121] and neural networks [124] have found their way into tracking and pose estimation. In future systems more of this may be expected due to the methods' ability to handle uncertainties and to suppress noise.

Even though interesting results such as [16] have recently arisen from methods not using a human model, the direct use of a human model seems to be the preferred trend. From Table 4 it can be seen that the choice of model type differs while silhouettes seem to be the preferred abstraction levels.

The use of silhouettes is motivated by the presence of simple algorithms for their estimation. They are easier to estimate than joints and a stick-figure, and their region-based nature makes them more robust to noise than local information such as edges. Furthermore, useful silhouettes might be extracted from relatively low resolution images. Due to the global nature of silhouettes details are likely to be missing. This results in additional complexity when trying to estimate a 3D pose from 2D silhouettes. Future work should therefore consider combining a silhouette representation with data capable of representing the interior of the silhouette, i.e., its relation to the human skeleton structure. An example is to combine silhouettes with the positions of the hands and head as seen in [100] and [112].

The use of motion as an abstraction level in pose estimation is also rather popular due to its inherent relation to the application. The motion in the images may be linked directly to the motion of the various limbs. Furthermore, many image points might be used to estimate the motion parameters. We expect the use of motion as a cue to be used more extensively in the future. However, to achieve success a number of issues still need to be addressed. First, the methods are based on incremental updates which rely on local (both spatial and temporal) smoothness. Therefore they often rely on a number of assumptions such as no occlusion and the subject being the only moving object in the image. Moreover, due to the incremental update, the initial pose is required and the systems have no way to recover after a total loses of track, lacking a mechanism for globally searching the entire image. Another problem is the risk of accumulating errors due to the incremental procedure. One solution is to use key frames as suggested in [153]. Given the initial and final pose parameters, both forward and backward iteration may ensure a consistent pose sequence. Alternatively, one might combine a motion-based method with a method based on spatial data. For example, it could be interesting to see image measurements and a human model linked by both the motion framework of Bregler and Malik [18] and the edge comparisons of Gavrila and Davis [45].

In addition to the problems related to incremental updates, another issue also has to be considered. Many movements become ambiguous when projected into the image plane; e.g., rotation about an axis parallel to the image plane will produce the same optical flow field as a translation in a certain direction will. Furthermore, movements along the optical axis are difficult to register robustly. To solve these problems multiple cameras are required or multiple data types as in the work by Okada *et al.* [111]. They combine motion data and depth data to resolve the ambiguities, thereby making the pose estimation more robust. In [109] a more general discussion on combining motion and depth data is given.

Generally speaking it seems to be a good and necessary approach to combine various data types to broaden invariance and robustness to all possible situations. Another promising approach is to use future measurements when processing the current data, i.e., allow a lag in the output. This helps to resolve ambiguities [16, 127].

8. CONCLUSION

Human motion capture goes back to at least the 1870s when Marey [95] and Muybridge [104] started their work. But recently, new technologies have made the motion capture problem popular as more convenient and affordable devices such as cameras, magnetic trackers, and computer power have become available.

Advances in active sensors, e.g., magnetic trackers, are making them cheaper, smaller, more precise, and generally easier to use. They will, however, still be cumbersome and limited in their use due to the need for special hardware. Therefore, computer vision could provide an attractive touch-free alternative.

The solutions developed to date are all based on a number of assumptions to make the problem tractable. This, together with the relatively simple methods being used, can be seen as an indication of the current state of the field—as being in its early development. The latest systems, however, use more advanced methods based on comprehensive probabilistic models and advanced training. Nevertheless, some assumptions are still required and we are far from a general solution to the human motion capture problem. Some of the general key issues needing to be addressed are initialization, recover from failure, and robustness.

Many systems are based on knowing the initial state of their system and/or a well-defined model fitted (offline) to the current subject. In a real life scenario we may expect a system to run on its own, i.e., adapt to the current situation. This might seem a minor problem, but what if none of the current research directions result in a system capable of autonomy? Should two parallel direction be followed: one for initialization and one for processing or should we aim at a common solution?

Related to this is the problem of how to recover from failure. A number of systems are based on incremental updates or searching around a predicted value. Many of these fail due to occlusion, bad predictions, and a change in the framerate/camera focus/image resolution and are not able to recover. This is an important problem since real life applications are likely to challenge a system by new situations not included in the design and training and hereby making it fail from time to time.

The robustness relates to the number of assumptions applied in systems, but also to the fact that most systems are tested on less than 1000 frames. How can one justify evaluating the robustness of a system within such a short lifespan? Long test sets available to everybody need to be generated (as in the face recognition community) to evaluate the robustness of individual systems and compare various systems.

For future systems to be more successful and less dependent of various assumptions new methods and a combination of current methods should be developed, i.e., the combination of various image cues, such as motion and silhouettes, and more extensive and adaptive use of human models. Furthermore, new sensors or combinations of sensors might also be an interesting path into the future.

The rapid developments in computer graphics may benefit human motion capture. Until recently the computer graphics field has been mostly interested in visual realism[3] and personalized human models, while the motion capture community has been more interested in spatial accuracy of the human models. We expect that the commercial interest in both fields will accelerate the development in human modeling and make the two fields approach and benefit from each other.

The applications of human motion capture are numerous and it is expected that we will see a continuous growth in the resources devoted to this topic and hence that interesting new results in spite of everything will appear in the not too distant future.

---

[3] A more comprehensive discussion on human animation can be seen found in [8].

APPENDIX A: THE DIFFERENT PUBLICATIONS' RELATION TO THE TAXONOMY

Year	First author	Initialization	Tracking	Pose estimation	Recognition
1980	O'Rourke			[114]	
1983	Hogg		*	[58]	
1984	Akita		*	[3]	
1984	Hogg		*	[59]	
1985	Lee			[83]	
1985	Tsukiyama		[135]		
1987	Bernat		[11]		
1987	Leung		*	[88]	
1987	Leung		[89]		
1989	Attwood			[6]	
1991	Long		*	[92]	
1991	Shio		[129]	*	
1991	Wang		*	[140]	
1991	Yamamoto			[152]	
1992	Kepple			[81]	
1992	Lee			[84]	
1992	Luo		*	[94]	
1992	Wang		*	[141]	
1993	Kameda		*	[79]	
1994	Baumberg	*	[9]		
1994	Bharatkumar		[12]	*	*
1994	Darrell		[32]	*	*
1994	Gu		[48]		
1994	Guo		*	[50]	
1994	Niyogi		[107]	*	*
1994	Perales	*		[117]	
1994	Polana		*		[120]
1994	Rossi	*	[125]		
1994	Schneider		[126]		
1995	Cai		[20]		
1995	Campbell			*	[21]
1995	Campbell			*	[22]
1995	Freeman		[41]		*
1995	Goncalves		*	[47]	
1995	Kakadiaris	[76]		*	
1995	Kameda		*	[80]	
1995	Leung		*	[90]	
1995	Tesei		[134]	*	
1996	Azarbayejani	*	[7]	*	
1996	Becker		*		[10]
1996	Bobick		[14]		*
1996	Cai		[19]	*	

Year	First author	Initialization	Tracking	Pose estimation	Recognition
1996	Gavrila		*	[45]	
1996	Ju		*	[73]	*
1996	Kahn		[74]		*
1996	Kakadiaris		*	[77]	
1996	Kameda		*	[78]	
1996	Luc				[93]
1996	Moezzi	[102]			
1996	Turk		[136]	*	*
1997	Bregler		*		[17]
1997	Christensen		*	[26]	
1997	Christensen		*	[27]	
1997	Davis		*		[33]
1997	Hunter		*	[62]	
1997	Iwasawa		[66]	*	
1997	Lerasle	*		[86]	
1997	Meyer	*	*	[97]	
1997	Oren		*		[113]
1997	Rohr	*	*	[123]	*
1997	Wachter		*	[138]	
1997	Wren	*	[142]	*	
1998	Bottino		*	[15]	
1998	Bregler	*	*	[18]	
1998	Chomat		*		[25]
1998	Chung		*	[28]	
1998	Corlin		[29]	*	
1998	Cretual		[30]		
1998	Davis		[34]		
1998	Davis		[35]		*
1998	Davis		[36]	*	*
1998	Fua		*	[42]	
1998	Fujiyoshi		[43]	*	*
1998	Goncalves			*	[46]
1998	Gu	[49]		*	
1998	Haritaoglu		[52]	*	*
1998	Haritaoglu		[53]	*	*
1998	Heisele		[54]		*
1998	Isard		[64]		
1998	Jojic		*	[71]	
1998	Kakadiaris	*	*	[75]	
1998	Li		*	[91]	
1998	Munkelt		*	[103]	
1998	Nakazawa		[105]		
1998	Narayanan	[106]			
1998	Nordlund		[109]		

Year	First author	Initialization	Tracking	Pose estimation	Recognition
1998	Pinhanez		[118]		
1998	Silaghi			[131]	
1998	Sul		*		[132]
1998	Utsumi		[137]	*	
1998	Wren		*	[145]	
1998	Wren		[147]		*
1998	Yacoob		*	*	[149]
1998	Yamada		[150]		
1998	Yamamoto			*	[151]
1998	Yaniz		*	[154]	
1998	Zheng	*		[155]	
1999	Amat		[4]	*	*
1999	Andersen		[5]	*	
1999	Brand		*	[16]	
1999	Cham		[24]	*	
1999	Cutler		[31]	*	
1999	Delamarre		*	[37]	
1999	Douros	[40]			
1999	Haritaoglu		*		[51]
1999	Hilton	[55]	*	*	
1999	Hilton	[56]		*	
1999	Ioffe		*	[63]	*
1999	Iwai		*	[65]	
1999	Iwasawa	*	[67]	*	
1999	Lerasle	*		[87]	
1999	Njåstad	*	*	[108]	
1999	Ohya		[110]	*	
1999	Ong		*	[112]	
1999	Pavlović		*	[115]	*
1999	Plänkers		*	[119]	
1999	Rittscher		*		[122]
1999	Segawa		[128]	*	
1999	Wachter		*	[139]	
1999	Wren		*	[146]	
2000	Hilton	[57]		*	
2000	Hu		*	[60]	
2000	Iwasawa	*	[68]		
2000	Jojic		[70]	*	*
2000	McKenna		[96]		*
2000	Moeslund	*	*	[100]	
2000	Moeslund		*	[101]	
2000	Okada	*	*	[111]	
2000	Rigoll	*	[121]		
2000	Rosales		*	[124]	

Year	First author	Initialization	Tracking	Pose estimation	Recognition
2000	Segawa		*	[127]	
2000	Sidenbladh	*	*	[130]	
2000	Wren		*	*	[143]
2000	Wren		*	[144]	
2000	Wu		*	[148]	
2000	Yamamoto		*	[153]	
$\sum$	Total = 136	8	48	64	16

ACKNOWLEDGMENTS

REFERENCES

1. J. K. Aggarwal and Q. Cai, Human motion analysis: a review, *Comput. Vision Image Understanding* **73**(3), 1999.

2. J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata, Articulated and elastic non-rigid motion: a review, in *Workshop on Motion of Non-Rigid and Articulated Objects, Austin, TX, 1994*, pp. 2–14.

3. K. Akita, Image sequence analysis of real world human motion, *Pattern Recognition* **17**, 1984, 73–83.

4. J. Amat, M. Casals, and M. Frigola, Stereoscopic system for human body tracking in natural scenes, in *International Workshop on Modeling People at ICCV'99, Corfu, Greece, September 1999*.

5. B. Andersen, T. Dahl, M. Iversen, M. Pedersen, and T. Søndergaard, *Human Motion Capture*, Technical report, Laboratory of Image Analysis, Aalborg University, Denmark, January 1999.

6. C. I. Attwood, G. D. Sullivan, and K. D. Baker, Model-based recognition of human posture using single synthetic images, in *Fifth Alvey Vision Conference, University of Reading, UK, 1989*.

7. A. Azarbayejani, C. R. Wren, and A. P. Pentland, Real-time 3-D tracking of the human body, in *IMAGE'COM 96, Bordeaux, France, May 1996*.

8. N. Badler, Virtual humans for animation, ergonomics, and simulation, in *Workshop on Motion of Non-Rigid and Articulated Objects, Puerto Rico, 1997*.

9. A. M. Baumberg and D. C. Hogg, An efficient method for contour tracking using active shape models, in *Workshop on Motion of Non-Rigid and Articulated Objects, Austin, TX, 1994*, pp. 2–14.

10. D. A. Becker and A. Pentland, Staying alive: a virtual reality visualization tool for cancer patients, in *AAAI'96 Workshop on Entertainment and Alife/AI, Portland, OR, August 1996*.

11. A. P. Bernat, J. Nelan, S. Riter, and H. Frankel, Security applications of computer motion detection, *Appl. Artificial Intell.* **V**, 1987, 786.

12. A. G. Bharatkumar, K. E. Daigle, M. G. Pandy, Q. Cai, and J. K. Aggarwal, Lower limb kinematics of human walking with the medial axis transformation, in *Workshop on Motion of Non-Rigid and Articulated Objects, Austin, TX, 1994*.

13. A. Blake and M. Isard, *Active Contours*, Springer-Verlag, Berlin/New York, 1998.

14. A. F. Bobick and J. W. Davis, An appearance-based representation of action, in *International Conference on Pattern Recognition, 1996*.

15. A. Bottino, A. Laurentini, and P. Zuccone, Toward non-intrusive motion capture, in *Asian Conference on Computer Vision, 1998*.

16. M. Brand, Shadow puppetry, in *International Conference on Computer Vision, Corfu, Greece, September 1999*.

17. C. Bregler, Learning and recognizing human dynamics in video sequences, in *Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, 1997*.

18. C. Bregler and J. Malik, Tracking people with twists and exponential maps, in *International Conference on Computer Vision and Pattern Recognition, 1998*.

19. Q. Cai and J. K. Aggarwal, Tracking human motion using multiple cameras, in *International Conference on Pattern Recognition, 1996*.

20. Q. Cai, A. Mitiche, and J. K. Aggarwal. Tracking human motion in an indoor environment, in *International Conference on Image Processing, 1995*.

21. L. Campbell and A. Bobick, Recognition of human body motion using phase space constraints, in *International Conference on Computer Vision, Cambridge, MA, 1995*.

22. L. Campbell and A. Bobick, Using phase space constraints to represent human body motion, in *International Workshop on Automatic Face- and Gesture-Recognition, Zurich, Switzerland, 1995*.

23. C. Cedras and M. Shah, Motion-based recognition: a survey, *Image Vision Comput* **13**(2), 1995, 129–155.

24. T. J. Cham and J. M. Rehg, Multiple hypothesis approach to figure tracking, in *Conference on Computer Vision and Pattern Recognition, Fort Collins, CO, June 23–25 1999*.

25. O. Chomat and J. L. Crowley, Recognizing motion using local appearance, in *International Symposium on Intelligent Robotic Systems, University of Edinburgh, 1998*.

26. C. Christensen and S. Corneliussen, *Tracking of Articulated Objects Using Model-Based Computer Vision*, Technical report, Laboratory of Image Analysis, Aalborg University, Denmark, June 1997.

27. C. Christensen and S. Corneliussen, *Visualization of Human Motion Using Model-based Vision*, Technical report, Laboratory of Image Analysis, Aalborg University, Denmark, January 1997.

28. J. M. Chung and N. Ohnishi, Cue circles: image feature for measuring 3-D motion of articulated objects using sequential image pair, in *International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998*.

29. C. R. Corlin and J. Ellesgaard, *Real Time Tracking of a Human Arm*, Technical report, Laboratory of Image Analysis, Aalborg University, Denmark, January 1998.

30. A. Cretual, F. Chaumette, and P. Bouthemy, Complex object tracking by visual serving based on 2D image motion, in *International Conference on Pattern Recognition, 1998*.

31. R. Cutler and L. Davis, Real-time periodic motion detection, analysis, and applications, in *Conference on Computer Vision and Pattern Recognition, Fort Collins, CO, June 23–25, 1999*.

32. T. Darrell, P. Maes, B. Blumberg, and A. P. Pentland, A novel environment for situated vision and behavior, in *Workshop for Visual Behaviors at CVPR-94, 1994*.

33. J. W. Davis and A. Bobick, The representation and recognition of action using temporal templates, in *Conference on Computer Vision and Pattern Recognition, 1997*.

34. J. W. Davis and A. Bobick. *SIDEshow: A Silhouette-based Interactive Dual-screen Environment*, Technical Report 457, MIT Media Lab, 1998.

35. J. W. Davis and A. Bobick, Virtual PAT: A virtual personal aerobics trainer, in *Workshop on Perceptual User Interfaces, San Francisco, November 1998*.

36. L. Davis, S. Fejes, D. Harwood, Y. Yacoob, I. Hariatoglu, and M. J. Black, Visual surveillance of human activity, in *Asian Conference on Computer Vision, Mumbai, India, 1998*.

37. Q. Delamarre and O. Faugeras, 3D articulated models and multi-view tracking with silhouettes in *International Conference on Computer Vision, Corfu, Greece, September 1999*.

38. B. Delaney, On the trail of the shadow women: the mystery of motion capture, *Comput. Graphics Appl.* **18**(5), 1998, 14–19.

39. A. Dempster, N. Laird, and D. Rubin, Maximum likelihood estimation from incomplete data via the EM algorithm, *J. Roy. Stat. Soc. (B)* **39**(1) 1977, 1–38.

40. I. Douros, L. Dekker, and B. F. Buxton, An improved algorithm for reconstruction of the surface of the human body from 3D scanner data using local B-spline patches, in *International Workshop on Modeling People at ICCV'99, Corfu, Greece, September 1999*.

41. W. T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma, Computer vision for computer games, in *International Workshop on Automatic Face- and Gesture-Recognition, Zurich, Switzerland, 1995*.

42. P. Fua, A. Gruen, R. Plänkers, N. D'Apuzzo, and D. Thalmann, Human body modeling and motion analysis from video sequences, in *International Symposium on Real Time Imaging and Dynamic Analysis, Hakodate, Japan, June 1998*.

43. H. Fujiyoshi and A. J. Lipton, Real-time human motion analysis by image skeletonization, in *Workshop on Applications of Computer Vision, 1998*.

44. D. M. Gavrila, The visual analysis of human movement: a survey, *Comput. Vision Image Understanding* **73**(1), 1999, 82–98.

45. D. M. Gavrila and L. S. Davis, 3-D model-based tracking of humans in action: a multi-view approach, in *Conference on Computer Vision and Pattern Recognition, San Francisco, CA, 1996*.

46. L. Goncalves, E. D. Bernardo, and P. Perona, Reach out and touch space (motion learning), in *International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998*.

47. L. Goncalves, E. D. Bernardo, E. Ursella, and P. Perona, Monocular tracking of the human arm in 3D, in *International Conference on Computer Vision, Cambridge, MA, 1995*.

48. H. Gu, Y. Shirai, and M. Asada, MDL-based spatiotemporal segmentation from motion in a long image sequence, in *Conference on Computer Vision and Pattern Recognition, 1994*.

49. J. Gu, T. Chang, I. Mak, S. Gopalsamy, H. C. Shen, and M. M. F. Yuen, A 3D reconstruction system for human body modeling, in *Modeling and Motion Capture Techniques for Virtual Environments, Lecture Notes in Artificial Intelligence, Vol. 1537*, Springer-Verlag, Berlin/New York, 1998.

50. Y. Guo, G. Xu, and S. Tsuji, Tracking human body motion based on a stick figure model, *J. Visual Comm. Image Representation* **5**, 1994, 1–9.

51. I. Haritaoglu, R. Cutler, D. Harwood, and L. S. Davis, Backpack: Detection of people carrying objects using silhouettes, in *International Conference on Computer Vision, Corfu, Greece*, September 1999.

52. I. Haritaoglu, D. Harwood, and L. S. Davis, Ghost: A human body part labeling system using silhouettes, in *International Conference on Pattern Recognition*, 1998.

53. I. Haritaoglu, D. Harwood, and L. S. Davis, W^4: Who? When? Where? What? - A real time system for detecting and tracking people, in *International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998*.

54. B. Heisele and C. Wohler, Motion-based recognition of pedestrians, in *International Conference on Pattern Recognition, 1998*.

55. A. Hilton, Towards model-based capture of a persons shape, appearance and motion, in *International Workshop on Modeling People at ICCV'99, Corfu, Greece, September 1999*.

56. A. Hilton, D. Beresford, T. Gentils, R. Smith, and W. Sun, Virtual people: Capturing human models to populate virtual worlds, in *International Conference on Computer Animation*, pp. 174–185, May 1999.

57. A. Hilton, D. Beresford, T. Gentils, R. Smith, and W. Sun. Whole-body modelling of people from multi-view images to populate virtual worlds, *The Visual Computer* **16**(7), 2000, 411–436.

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

59. D. C. Hogg, *Interpreting Images of a Known Moving Object*, Ph.D. thesis, University of Sussex, UK, 1984.

60. C. Hu, Q. Tu, Y. Li, and S. Ma, Extraction of parametric human model for posture recognition using genetic algorithm, in *The fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000*.

61. M. Hu, Visual pattern recognition by moment invariants, *IRE Trans. Inform. Theory* **8**(2), 1962, 179–187.

62. E. A. Hunter, P. H. Kelly, and R. C. Jain, Estimation of articulated motion using kinematically constrained mixture densities, in *Workshop on Motion of Non-Rigid and Articulated Objects, Puerto Rico*, 1997.

63. S. Ioffe and D. Forsyth, Finding people by sampling, in *International Conference on Computer Vision, Corfu, Greece, September 1999*.

64. M. Isard and A. Blake, CONDENSATION-conditional density propagation for visual tracking. *Int. J. Comput. Vision*, pp. 5–28, 1998.

65. Y. Iwai, K. Ogaki, and M. Yachida, Posture estimation using structure and motion models, in *Int. Conf. Comput. Vision, Corfu, Greece, September 1999*.

66. S. Iwasawa, K. Ebihara, J. Ohya, and S. Morishima, Real-time estimation of human body posture from monocular thermal images, in *Conference on Computer Vision and Pattern Recognition, 1997*.

67. S. Iwasawa, J. Ohya, K. Takahashi, T. Sakaguchi, S. Kawato, K. Ebihara, and S. Morishima, Real-time estimation of human body posture from trinocular images, in *International Workshop on Modeling People at ICCV'99, Corfu, Greece, September 1999*.

68. S. Iwassawa, J. Ohya, K. Takahashi, T. Sakaguchi, K. Ebihara, and S. Morishima, Human body postures from trinocular camera images, in *The Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000*.

69. G. Johnsson, Visual motion perception, in *Scientific American*, June 1975, 76–88.

70. N. Jojic, B. Brumitt, B. Meyers, S. Harris, and T. Huang, Detection and estimation of pointing gestures in dense disparity maps in *The Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000*.

71. N. Jojic, J. Gu, H. C. Shen, and T. Huang, 3-D Reconstruction of multipart self-occluding objects, in *Asian Conference on Computer Vision, 1998*.

72. S. Ju, *Human Motion Estimation and Recognition (Depth Oral Report)*, Technical report, University of Toronto, 1996.

73. S. X. Ju, M. J. Black, and Y. Yacoob, Cardboard people: A parameterized model of articulated image motion, in *International Conference on Automatic Face and Gesture Recognition, Killington, VT, 1996*.

74. R. E. Kahn and M. J. Swain, *Gesture Recognition Using the Perseus Architecture*, Technical Report TR-96-04, Department of Computer Science, University of Chicago, 1996.

75. I. Kakadiaris and D. Metaxas, Vision-based animation of digital humans, in *Conference on Computer Animation, 1998*, pp. 144–152.

76. I. A. Kakadiaris and D. Metaxas, 3D human body model acquisition from multiple views, in *International Conference on Computer Vision, Cambridge, MA, June 20–23, 1995*, pp. 618–623.

77. I. A. Kakadiaris and D. Metaxas, Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection, in *Conference on Computer Vision and Pattern Recognition, 1996*.

78. Y. Kameda and M. Minoh, A human motion estimation method using 3-successive video frames, in *International Conference on Virtual Systems and Multimedia, 1996*.

79. Y. Kameda, M. Minoh, and K. Ikeda, Three dimensional pose estimation of an articulated object from its silhouette image, in *Asian Conference on Computer Vision, 1993*.

80. Y. Kameda, M. Minoh, and K. Ikeda, Three dimensional motion estimation of a human body using a difference image sequence, in *Asian Conference on Computer Vision, 1995*.

81. T. M. Kepple, MOVE3D-Software for analyzing human motion, in *Proc. of Johns Hopkins National Search for Computing Applications to Assist Persons with Disabilities, Laurel, MD, February 1992*.

82. K. Koffka, *Principle of Gestalt Psychology*, Harcourt Brace, New York, 1935.

83. H. J. Lee and Z. Chen, Determination of 3D human body posture from a single view, *Comp. Vision, Graphics, Image Process*. **30** 1985, 148–168.

84. H. J. Lee and Z. Chen, Knowledge-guided visual perception of 3-D human gait from a single image sequence, *Trans. Systems, Man, Cybernetics* **22**(2), 1992, 336–342.

85. J. Lengyel, The convergence of graphics and vision, *Computer* **31**(7), 1998, 46–53.

86. F. Lerasle, G. Rives, and M. Dhome, Human body limbs tracking by multi-ocular vision, in *Scandinavian Conference on Image Analysis, Lappeenranta, Finland, 1997*.

87. F. Lerasle, G. Rives, and M. Dhome, Tracking of human limbs by multiocular vision, *Comp. Vision Image Understanding* **75**, 1999, 229–246.

88. M. K. Leung and Y. H. Yang, A region based approach for human body motion analysis, *Pattern Recognition*, **20**, 1987, 321–339.

89. M. K. Leung and Y. H. Yang, Human body motion segmentation in a complex scene, *Pattern Recognition*, **20**, 1987, 55–64.

90. M. K. Leung and Y. H. Yang. First sight: A human body outline labeling system, *Trans. Pattern Anal. Mach. Intelligence* **17**(4), 1995, 359–377.

91. Y. Li, S. Ma, and H. Lu, Human posture recognition using multi-scale morphological method and Kalman motion estimation, in *International Conference on Pattern Recognition, 1998*.

92. W. Long and Y. H. Yang, Log-Tracker: an attribute-based approach to tracking human body motion, *Pattern Recognition Artificial Intelligence* **5**, 1991, 439–458.

93. E. Luc, Real time human action recognition for virtual environments, in *Computer Science Postgraduate Course*, Computer Graphics Lab, Swiss Federal Institute of Technology, Lausanne, Switzerland, September 1996.

94. Y. Luo, F. J. Perales, and J. J. Villanueva, An automatic rotoscopy system for human motion based on a biomechanic graphical model, *Comput. Graphics* **16**(4), 1992, 355–362.

95. E. J. Marey, *Animal Mechanism: A Treatise on Terrestrial and Aerial Locomotion*, Appleton, New York, Appleton, 1873. [Republished as Vol. XI of the International Scientific Series.]

96. S. J. McKenna, S. Jabri, Z. Duric, and H. Wechsler, Tracking interacting people, in *The Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000*.

97. D. Meyer, J. Denzler, and H. Niemann, Model based extraction of articulated objects in image sequences, in *Fourth International Conference on Image Processing, 1997*.

98. T. B. Moeslund, *Summaries of 107 Computer Vision-Based Human Motion Capture Papers*, Technical report, Laboratory of Image Analysis, Aalborg University, Denmark, 1999.

99. T. B. Moeslund, Interacting with a virtual world through motion capture, in *Interaction in Virtual Inhabited 3D Worlds* (L. Qvortrup, Ed.), chap. 11, Springer-Verlag, Berlin/New York, 2000.

100. T. B. Moeslund and E. Granum, 3D human pose estimation using 2D-data and an alternative phase space representation, in *Workshop on Human Modeling, Analysis and Synthesis at CVPR, Hilton Head Island, SC, June 2000*.

101. T. B. Moeslund and E. Granum, Multiple cues used in model-based human motion capture, in *The Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000*.

102. S. Moezzi, A. Katkere, D. Y. Kuramura, and R. Jain, Reality modeling and visualization from multiple video sequences, *Comp. Graphics Appl.* **16**(6), 1996, 58–62.

103. O. Munkelt, C. Ridder, D. Hansel, and W. Hafner, A model driven 3D image interpretation system applied to person detection in video images, in *International Conference on Pattern Recognition, 1998*.

104. E. Muybridge, Animal locomotion, reprinted in *Animal in Motion* (L. S. Brown, Ed.), Dover, New York, 1957.

105. A. Nakazawa, H. Kato, and S. Inokuchi, Human tracking using distributed video systems, in *International Conference on Pattern Recognition, 1998*.

106. P. J. Narayanan, P. W. Rander, and T. Kanade, Constructing virtual worlds using dense stereo, in *International Conference on Computer Vision, Bombay, India, January 1998*.

107. S. A. Niyogi and E. H. Adelson, Analyzing and recognizing walking figures in XYT, in *Conference on Computer Vision and Pattern Recognition, 1994*.

108. J. Njåstad, S. Grinaker, and G. A. Storhaug, Estimating parameters in a $2\frac{1}{2}$ D human model, in *11th Scandinavian Conference on Image Analysis, Kangerlussuaq, Greenland, 1999*.

109. P. Nordlund, *Figure-Ground Segmentation Using Multiple Cues*, Ph.D. thesis, Kungi Tekniska Hogskolan, Sweden, 1998.

110. J. Ohya, J. Kurumisawa, R. Nakatsu, K. Ebihara, S. Iwasawa, D. Harwood, and T. Horprasert, Virtual metamorphosis, *MultiMedia* **6**(2), 1999, 29–39.

111. R. Okada, Y. Shirai, and J. Miura, Tracking a person with 3-D motion by integrating optical flow and depth, in *The Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000*.

112. E. J. Ong and S. Gong, Tracking hybrid 2D-3D human models from multiple views, in *International Workshop on Modeling People at ICCV'99, Corfu, Greece, September 1999*.

113. M. Oren, C. Papageorigiou, P. Sinha, E. Osuna, and T. Poggio, Pedestrian detection using wavelet templates, in *Conference on Computer Vision and Pattern Recognition, 1997*.

114. J. O'Rourke and N. I. Badler, Model-based image analysis of human motion using constraint propagation, *Trans. Pattern Anal. Mach. Intelligence* **2**(6), 1980, 522–536.

115. V. Pavlović, J. M. Rehg, T. J. Cham, and K. P. Murphy, A dynamic Bayesian network approach to figure tracking using learned dynamic models, in *International Conference on Computer Vision, Corfu, Greece, September 1999*.

116. V. I. Pavlovic, R. Sharma, and T. S. Huang, Visual interpretation of hand gestures for human-computer interaction: a review, *Trans. Pattern Anal. Mach. Intelligence* **19**(7), 1997, 677–695.

117. F. J. Perales and J. Torres, A system for human motion matching between synthetic and real images based on a biomechanic graphical model, in *Workshop on Motion of Non-Rigid and Articulated Objects Austin, TX, 1994*, pp. 83–88.

118. C. Pinhanez and A. Bobick, Using computer vision to control a reactive computer graphics character in a theater play, in *International Conference on Vision Systems, 1998*.

119. R. Plänkers, P. Fua, and N. D'Apuzzo, Automated body modeling from video sequences, in *International Workshop on Modeling People at ICCV'99, Corfu, Greece, 1999*.

120. R. Polana and R. Nelson, Low level recognition of human motion, in *Workshop on Motion of Non-Rigid and Articulated Objects, Austin, TX, October 1994*.

121. G. Rigoll, S. Eickeler, and S. Müller, Person tracking in real world scenarios using statiscal methods, in *The Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000*.

122. J. Rittscher and A. Blake, Classification of human body motion, in *International Conference on Computer Vision, Corfu, Greece, September 1999*.

123. K. Rohr, *Human Movement Analysis Based on Explicit Motion Models*, chap. 8, pp. 171–198, Kluwer Academic, Dordrecht/Boston, 1997.

124. R. Rosales and S. Sclaroff, Learning and synthesizing human body motion and posture, in *The Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000*.

125. M. Rossi and A. Bozzoli, *Tracking and Counting Moving People*, Technical Report 9404-03, IRST, Trento, Italy, April 1994.

126. M. Schneider and M. Bekker, *Tracking of Human Motion*, Master's thesis, LIFIA, Grenoble, France and LIA, AAU, Denmark, 1994.

127. H. Segawa, H. Shioya, N. Hiraki, and T. Totsuka, Constraint-conscious smoothing framework for the recovery of 3D articulated motion from image sequences, *in The Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000*.

128. H. Segawa and T. Totsuka, Torque-based recursive filtering approach to the recovery of 3D articulated motion from image sequences, in *International Conference on Computer Vision, Corfu, Greece, September 1999*.

129. A. Shio and J. Sklansky, Segmentation of people in motion, in *Workshop on Visual Motion*, October 1991, pp. 325–332.

130. H. Sidenbladh, F. De la Torre, and M. J. Black, A framework for modeling the appearance of 3D articulated figures, in *The Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000*.

131. M. C. Silaghi, R. Plänkers, R. Boulic, P. Fua, and D. Thalmann, Local and global skeleton fitting techniques for optical motion capture, in *International Workshop on Modelling and Motion Capture Techniques for Virtual Environments, Geneva, Switzerland, November 1998*.

132. C. Sul, K. Lee, and K. Wohn. Virtual stage: a location-based karaoke system, *Multimedia* **5**(2), 1998, 42–52.

133. S. Sumi, Upside-down presentation of the Johnsson moving light-spot pattern, *Perception* **13**, 1984, 283–286.

134. A. Tesei, G. L. Foresti, and C. S. Regazzoni, Human body modeling for people localization and tracking from real image sequences, in *Image Processing and Its Applications, July 1995*.

135. T. Tsukiyama and Y. Shirai, Detection of the movements of persons from a sparse sequence of TV images, *Pattern Recognition* **18**, 1985, 207–213.

136. M. Turk, Visual Interaction with lifelike characters, in *International Conference on Automatic Face and Gesture Recognition, Killington, VT, 1996*.

137. A. Utsumi, H. Mori, J. Ohya, and M. Yachida, Multiple-view-based tracking of multiple humans, in *International Conference on Pattern Recognition, 1998*.

138. S. Wachter and H.-H. Nagel, Tracking of persons in monocular image sequences, in *Workshop on Motion of Non-Rigid and Articulated Objects, Puerto Rico, 1997*.

139. S. Wachter and H.-H. Nagel, Tracking persons in monocular image sequences, *Comp. Vision Image Understanding* **74**, 1999, 174–192.

140. J. Wang, G. Lorette, and P. Bouthemy, Analysis of human motion: a model-based approach, in *Scandinavian Conference on Image Analysis, 1991.*

141. J. Wang, G. Lorette, and P. Bouthemy, Human motion analysis with detection of sub-part deformations, *SPIE-Biomedical Image Processing and Three-Dimensional Microscopy*, **1660**, 1992, 329–335.

142. C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, Pfinder: real-time tracking of the human body, *Trans. Pattern Anal. Mach. Intelligence* **19**(7), 1997, 780–785.

143. C. R. Wren, B. P. Clarkson, and A. P. Pentland, Understanding purposeful human motion, in *The Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000.*

144. C. R. Wren and A. P. Pentland, Dynaman: recursive modeling of human motion, *Image Vision Comp.*, in press.

145. C. R. Wren and A. P. Pentland, Dynamic models of human motion, in *International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998.*

146. C. R. Wren and A. P. Pentland, Understanding purposeful human motion, in *International Workshop on Modeling People at ICCV'99, Corfu, Greece, September 1999.*

147. C. R. Wren *et al.*, Perceptive spaces for performance and entertainment, in *ATR Workshop on Virtual Communication Environments: Bridges over Art/Kansei and VR Technologies, Kyoto, Japan, April 13, 1998.*

148. A. Wu, M. Shah, and N. Lobo, A virtual 3D blackboard: 3D finger tracking using a single camera, in *The Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000.*

149. Y. Yacoob and M. J. Black, Parameterized modeling and recognition of activities, in *International Conference on Computer Vision, Bombay, India, 1998.*

150. M. Yamada, K. Ebihara, and J. Ohya, A new robust real-time method for extracting human silhouettes from color images, in *International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998.*

151. M. Yamamoto, T. Kondo, T. Yamagiwa, and K. Yamanaka, Skill recognition, in *International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998.*

152. M. Yamamoto and K. Koshikawa, Human motion analysis based on a robot arm model, in *Conference on Computer Vision and Pattern Recognition, 1991.*

153. M. Yamamoto, Y. Ohta, T. Yamagiwa, and K. Yagishita, Human action tracking guided by key-frames, in *The Fourth International Conference on Automatic Face and Gesture Recognition, Grenoble, France, March 2000.*

154. C. Yaniz, J. Rocha, and F. Perales, 3D region graph for reconstruction of human motion, in *Workshop on Perception of Human Motion at ECCV, 1998.*

155. J. Y. Zheng and S. Suezaki, A model based approach in extracting and generating human motion, in *International Conference on Pattern Recognition, 1998.*