UNIVERSITY OF AALBORG

March 1999

Summaries of 107 Computer Vision-Based Human Motion Capture Papers

Technical Report LIA 99-01
by:

Thomas B. Moeslund

Laboratory of Image Analysis
Institute of Electronic Systems
Fredrik Bajers Vej 7D
DK-9220 Aalborg East
Denmark

Abstract

This technical report contains summaries of 107 papers concerned with computer vision-based human motion capture. The report can be seen as an appendix to a survey, **Computer Vision-Based Human Motion Capture - A Survey**, which I wrote in parallel to this report. The survey gives a taxonomy for the papers described in this report. The summaries are presented in alphabetic order, with respect to the surname of the first Author. I have tried to give objective summaries of the different papers and only give subjective critique in an item, comments, reserved for this purpose. I have also tried to use the same amount of energy and space on the different summaries. But large variations between the length of the different summaries can be observed. This is mainly due to the length of the papers, but also due to my interest in the individual papers.

Preface

This technical report contains summaries of 107 papers concerned with computer vision-based human motion capture. The report can be seen as an appendix to a survey, **Computer Vision-Based Human Motion Capture - A Survey**, which I wrote in parallel to this report. This report has been written as a separate report due to its size alone. The writing of the two reports was carried out in the winter 98/99 while the reading of papers was done during the summer and fall of 1998. The work presented is the first step in my Ph.D.-study titled: "Multiple Cues for Model-Based Human Motion Capture". Whenever I use 'he' throughout the report is should be read as he/she.

Aalborg, Denmark, March 1999.

Thomas B. Moeslund (tbm@vision.auc.dk)

Contents

Chapter 1

Introduction

During the last two decades a growing interest within man-machine interfaces has been seen. This is mainly motivated by the fact that more and more information is processed and stored in computers which are sometimes hard to access. One of the problems is that the communication is done via interface devices, such as mouse and keyboard, which are not natural to humans. We are forced to communicate on the terms of the computer instead of natural human terms such as speech and gestures.

Since it is very hard for humans to learn the binary language of computers, the computers must learn to communicate on human terms. This means, among other things, that the computer should be equipped with the same senses as humans. Ultimately this is a humanoid, a robot designed as a human. As a first step it might not be necessary to replace the home PC with a robot - or perhaps it will! - but instead one can add a few devices to enhance its ability to interact in a human fashion. These devices should be: a camera to make it able to see, a microphone to make it able to hear, and a loudspeaker to make it able to speak. Other devices might be introduced in the future but the three mentioned are the most important ones. The speak-and hear-abilities are currently leaving the labs around the world and entering the consumer market, but the see-ability is still lacking. The main reason for this lack is the high complexity involved in the image processing task to interpret a scene.

One aspect of image processing in this context is to obtain the motion conducted by the human or a part of him. This process is known as human motion capture. Even though the term covers all aspects of human motion, it is usually only used in connection with the large scale body analysis in contrast to, e.g. the motion done by a hand. In other words, when people talk

about human motion capture they usually refer to the process of capturing the large scale body movements, which is the movements of the head, arms, torso, and legs. The motion of the toes, fingers, hands, and facial muscles are usually not considered when talking about human motion capture.

In this report I refer to large scale body movements when I use the term human motion capture. Formally I define it as:

> *Human motion capture is the process of capturing the large scale body movements at some resolution.*

I added "at some resolution" to point out that both precise tracking of a human and his limps, as well as overall tracking of a human and his whereabouts, is considered to fall within the above definition. Hence human motion capture is used both when the human is viewed as one single object and when he is viewed as a high degree of freedom skeleton structure with a number of joints.

I realize that this report does not contain every work published within the field of computer vision-based human motion capture. However, a large number of papers are presented and they represent the main references within this field of research. The field is rather new and therefore most of the publications are from the mid-and late 90s. The distribution of the publications with respect to their publication year is shown in figure 1.1.

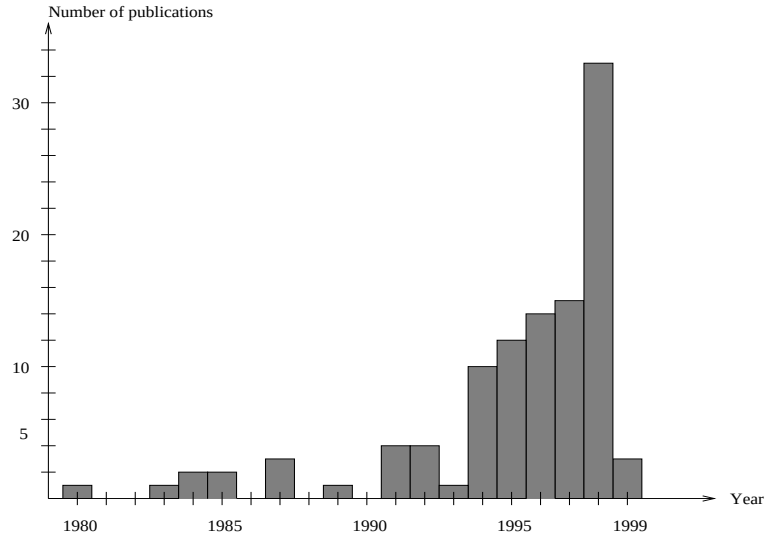

Figure 1.1: A histogram showing the number of publications per year since 1980.

Chapter 2

Summaries

This chapter contains summaries of all the papers reviewed for my survey **Computer Vision-Based Human Motion Capture - A Survey** [71]. The summaries are presented in alphabetic order, with respect to the first Author's last name. Each summary is presented using the structure shown below and two subsequent summaries are separated by an empty line.

Title: The title of the paper followed by its reference number
Author(s): The name(s) of the author(s)
Location: Where the authors are situated
Year: The publication year
Published: Where the paper has been published
Type: The type of publication
Key words: Key words for the paper
Summary: A summary of the paper
Comments: Any relevant comments I might have

The **Published** item is left empty when a paper hasn't been published. I have tried to give objective summaries of the different papers and only give subjective critique in the last item. When no relevant comments are present the last item is removed. I have also tried to use the same amount of energy and space one the different summaries. But, as can be seen in the following, large variations between the length of the different summaries can be observed. This is mainly due to the length of the papers, but also due to my interest in the individual papers.

Title: Human Motion Analysis: A Review [1]
Author(s): J.K. Aggarwal and Q. Cai
Location: Computer and Vision Research Center, University of Texas at Austin
Year: 1997
Published: Workshop on Motion of Non-Rigid and Articulated Objects, Puerto Rico, USA
Type: Paper (Review)
Key words: Overview and human motion
Summary: This is an overview paper dealing with human motion. They divide the topic into three main areas, body structure analysis, tracking and recognition. The first consists of work where the different body parts are located. This is again divided into model-based and non-model-based methods. The tracking area consists of work without using body parts and is divided into single camera systems and multiple camera systems. The last area is divided into state-space methods and template matching methods. All this is represented in a nice tree-structure. Most of the references are located in this structure which gives a very nice overview. In the rest of the paper they give examples from the different areas.
Comments: An extremely relevant paper which holds most relevant references up to 1997. Good taxonomy.

Title: Articulated and Elastic Non-rigid Motion: A Review [2]
Author(s): J.K. Aggarwal, Q. Cai, W. Liao and B. Sabata
Location: Computer and Vision Research Center, Department of Electrical and Computer Engineering, University of Texas, Austin
Year: 1994
Published: Workshop on Motion of Non-Rigid and Articulated Objects, pp. 2-14, Austin, Texas, USA
Type: Paper (Review)
Key words: Review, elastic non-rigid motion, articulated motion and taxonomy
Summary: This paper gives a review of different methods used in articulated and elastic non-rigid motion. First a good taxonomy of different motion types are given. Next the different approaches within articulated motion without a priori shape models and with a priori models are described. Then the elastic motion approaches are described in the two categories: without a shape model and with a shape model.
Comments: A very nice taxonomy of different motion types.

Title: Image sequence Analysis of Real World Human Motion [3]
Author(s): K. Akita
Location: Computer Systems Works, Mitsubishi Electric Corp., Kanagawa, Japan

Year: 1984
Published: Pattern Recognition Vol. 17, No. 1. pp. 73-83
Type: Paper
Key words: Human motion tracking, human model, cylinder model, stick-figure, key-frame, occlusion matrix and skeleton
Summary: This work is about human motion tracking using a cylinder model and key-frames. The application is to track a person during a gymnastic exercise which is known in advance. Therefore the different poses of the person are known beforehand and can be modeled using key-frames. A key-frame represents a state, with respect to occlusion, of the posture and is represented as a stick-figure with five segments and four joints. An occlusion matrix is formed which holds information about which segments are occluding which segments with respect to the key-frames. The system then knows what to expect regarding occlusion. Before the system is started different parameters are measured about the object. Then the outline and skeleton of the user are found using edge information. For each outline pixel a feature vector is created, called a window code. These feature vectors are tracked over time using a distance measure and thereby finding the trajectories of the different body parts. The skeleton is transformed into a cylinder model and it is calculated when a potential occlusion is present. In this case a different image is calculated in the occluded area and, together with the outline images, used to find the precise outline of the body parts in the occluded area.
Comments: It is not clear how the key-frames and occlusion matrix are used in practice. Also it is unclear how the difference image is used. I like the idea about the occlusion matrix.

Title: Three-Dimensional Analysis of Human Movement [4]
Editors: P. Allard*, I.A.F. Stokes** and J.P. Blanchi***
Location: * = Sainte-Justine Hostpital, Montreal, Quebec, Canada, ** = University of Vermont, Burlington, Vermont, USA, *** = Joseph Fourier University, Grenoble, France **Year:** 1995
Published:
Type: Book
Key words: Motion capture
Summary: The book is divided into three main parts: (i)Data capture and processing, (ii)Mechanical and neuromuscular modeling and (iii)implementation and scope of three-dimensional analysis. Only the first part is summarized below.
Part I: Chapter 1) An overview of different sensors for measuring parameters of different body segments and their pros and cons. These are: goniometer, electromagnetic and acoustic sensors, photogrammetric reconstruction and accelerometers. A good summary which includes six points that should be considered when choosing a technique. Chapter 2) Explains the principle

behind 3D reconstruction using photogrammetric (cameras) and markers. Touches upon the following subjects: the different coordinate systems, stereo, fixed and moving cameras, the errors involved and calibration. A illustrative error propagation graph is shown. Chapter 3) Kind of the same as chapter 2, but more focus on the technical side of cameras and recorders. Also some details of how to track the markers (image processing). Chapter 4) The same again but explained from an image processing/computer point of view with some focus on the sensor. The tasks are: coordinate enhancement (thresholding or cross-correlation), marker labeling, distortion correction and 3D-reconstruction. A discussion on passive versus active markers is presented. Finally an informative section about accuracy (measurements versus ground truth) and precision (the repeatability from frame to frame). Chapter 5) A signal processing chapter on how to avoid/reduce errors/noise in a system. The following topics are described: sampling frequency, filtering and model fitting. Chapter 6) This chapter is divided into two sections: computer graphics and representation of the human body. The former presents the SW and HW components used in computer graphics: raster display, vector image, hidden line removable, shadowing, shading, polygons, color, speed, memory requirements etc. The latter section describes how a human can be visualized and animated. Chapter 7) Explains how x-rays work and can be used to capture human motion. Also presents two different techniques for capturing 3D data: stereo (2 x-ray sources and one image plane) and biplanar (2 orthogonal x-ray sources and two image planes). Explains the need for calibration and how this can be performed using anatomic and metallic landmarks. Concludes with a survey of different applications.
Comments: So far I've only read part I, which is a collection of papers rather than one book. Therefore the same stuff is repeated again and again. It however still gives a good overview of the area of human motion capture up til 1992. Part II deals with mechanical modeling and might be interesting when working in more detail with the modeling of the human. Part III deals with applications of motion analysis and might have some good references.

Title: Human Motion Capture [5]
Author(s): B. Andersen, T. Dahl, M. Iversen, M. Pedersen and T. Søndergaard
Location: Laboratory of Image Analysis, Aalborg University, Denmark
Year: 1998
Published:
Type: Technical Report
Key words: Human tracking, chromatic colors, verging stereo, second order predictor, error function, avatar, known init-pose and real time
Summary: This work is about tracking the hands and head of a human and controlling an avatar using these data. The hands and head (face) are found using color thresholding in the chromatic rg-plane. The tracking

works with a normal background as long as the user, and the background, do not contain colors very similar to human skin. The tracking is carried out in 3D, meaning that the face and hands are found in two calibrated verging images, using Tsai's algorithm. To speed up the system, the images are down-sampled with a factor 1:8 when searching for regions of interest. Within each region of interest a region growing algorithm is used, on a 1:1 image resolution, to find the centroids. Using triangulation on the centroids of the hands and face yields a 3D estimate of both hands and the face. Due to the lack of framegrabbers which are able to grab two RGB signal, two different SGI Indy machines are used in a client-server architecture. Using a second order predictor, velocity and acceleration, the 3D positions of the hands and face are predicted. The predicted positions are compared with the measured positions in the current image using an error function. Optimizing this function yields continuous tracking of the three body parts. For the tracking to work the user must be in a special pose (no arm crossing) when the system is initialized. To control the avatar three assumptions are introduced. First the line spanned by the shoulders are assumed to be parallel to the image plane (or rather to the global coordinate system) and horizontal. Secondly the spine is assumed to be hanging down from the head, i.e. bending not allowed. Finally the elbows are also assumed to be hanging down all the time. Using these three assumptions the 3D position of the head and face can be used to control the avatar which consists of cylinders. After a moving average filtering of the positions of the head and hands, a user can smoothly control the avatar in real time (25Hz).
Comments: I supervised this project.

Title: Model-based Recognition of Human Posture Using Single Synthetic Images [6]
Author(s): C.I. Attwood, G.D. Sullivan and K.D. Baker
Location: Intelligent Systems Group, Dept. of Computer Science, University of Reading, UK
Year: 1989
Published: Fifth Alvey Vision Conf., University of Reading, UK
Type: Paper
Key words: Posture recognition, model based, cylinder model, synthetic input, constraints, collision detection, balance constraint, probability density function and pruning the search space
Summary: This work is about model-based recognition of three different body postures (standing, kneeling and sitting) using synthetic images. The idea is to have a system which as input takes the different joints and as output gives the posture of the observed (synthetic) human. A cylinder based model is introduced. It consists of 16 segments (where the length is known) and 14 joints. Since the 2D positions (in the image) of the joints are known together with the length of the individual segments, the 3D location

of a segment can be in one of two positions. This adds up to $2^{18} = 262.144$ different posture configurations. The search space is pruned using different constraints. First anatomical constraints are used to introduce limits on the different joint angles. Next a posture constraint is introduced where the assumption is that the body posture is close to one of the three above mentioned. In practise this is done by introducing probability density functions for the different postures and using them to measure the likelihood of the different postures. Next the different postures (which have survived the first two constraints) are subjected to a collision detection and a balance constraint. Finally the postures which only differ slightly in depth or have different extrem 'joints'(head, hands and feet) are excluded. The system is tested on three different sequences of synthetic images and the correct posture is found in every frame, but sometimes together with a few other postures.
Comments: Classical piece of work. No real-time and no real input data are used. I like the idea about using all kinds of different constraints to prune the search space.

Title: Real-Time 3-D Tracking of the Human Body [7]
Author(s): A. Azarbayejani, C.R. Wren, and A.P. Pentland
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1996
Published: In proceeding of IMAGE'COM 96, Bordeaux, France
Type: Paper
Key words: Pfinder, Spfinder and automatic camera calibration
Summary: This paper describes the Pfinder, the Spfinder and the different applications they have been used in. The Pfinder description is a shorter version of the one from [99]. The Spfinder is a 3D version of Pfinder. Two cameras are used and for each camera a Pfinder algorithm is used. The three blobs from each of the two cameras are used to do an automatic camera calibration and find the 3D position of the blobs (hands and face). In the end some applications which have used the P/SP-finder are presented.
Comments: [99] gives a better overview of the Pfinder algorithm.

Title: An Efficient Method for Contour Tracking using Active Shape Models [8]
Author(s): A.M. Baumberg and D.C. Hogg
Location: University of Leeds, UK
Year: 1994
Published: School of computer studies research report series, Report 94.11
Type: Paper
Key words: Active shape models, contour tracking, B-spline, Principal components, Kalman filter and Sobel-based edge
Summary: This work is about tracking the contour of a walking human

using active shape models. In previous work by the two authors a method for extraction the shape of a contour has been developed. A shape is represented by a closed uniform B-spline with a fixed number of control points equally spaced around the boundary of the object. Principal components are used to reduce the amount of information required to represent the state space (different model configurations/shapes). A framework for tracking the contour over time is presented based on a Sobel-based edge search from the estimated contour. A set of Kalman filters are used to predict and smooth the different parameters within the system. To improve the system global shape constraints and an iteration scheme are introduced. The system is tested on prerecorded image sequences and the result seems to be rather fair. It runs at 30Hz (using 80 contour points and three iterations per frame) on an R4000 Silicon Graphics Indigo.
Comments: Interesting tracking scheme, but since it relays on edge points it requires images with a rather good contrast between object and background.

Title: Staying Alive: A Virtual Reality Visualization Tool for Cancer Patients [9]
Author(s): D.A. Becker and A. Pentland
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1996
Published: Workshop on Entertainment and Alife/AI, Portland, Oregon, August
Type: Paper
Key words: Staying Alive, VR, gestures, T'ai Chi, Spfinder, STIVE and HMM
Summary: This work is a VR application, called Staying Alive, to help cancer patients. The idea is that by visualizing your immune system fighting of diseases, you will actually feel better. This system takes this idea to the extreme by making you able to navigate within a virtual blood stream and fighting the bad cells. To navigate different gestures are used (e.g. for up and down). The gestures are the ones used in T'ai Chi which makes the application more relaxing. The gestures are obtained using the STIVE system (Spfinder) together with HMM. The recognition rate of the T'ai Chi gestures are 90%.
Comments: An interesting application.

Title: Security Applications of Computer Motion Detection [10]
Author(s): A.P. Bernat*, J. Nelan*, S. Riter* and H. Frankel**
Location: *=Departments of Computer Science and Electrical Engineering, The University of Texas, El Paso. **=Office of Research and Development, U.S. Immigration and Naturalization Service, Washington D.C.
Year: 1987

Published: Applications of Artificial Intelligence V, Vol. 786
Type: Paper
Key words: Human motion detection, geopixels, mean, median, likelihood and synthetic images
Summary: This work is about detecting aliens at the US-Mexican border. The idea is to have a set of cameras to monitor the border. For each camera image motion is found. If a motion object occurs in a likely location in the image, has the correct size and has a realistic trajectory, it is considered to be a human and an alarm sounds. Three different motion detection methods are proposed and tested. Each operate in the following way. A measure for each geopixel (a rectangle of contiguous pixels) is calculated and compared to a measure for a geopixel (at that location) in a reference frame. The three methods/measures are: mean value, median value and likelihood value. They are each tested on synthetic images and the result is that the mean value has the best detection rate for a given false alarm rate.

Title: Lower Limb Kinematics of Human Walking with the Medial Axis Transformation [11]
Author(s): A.G. Bharatkumar*, K.E. Daigle**, M.G. Pandy**, Q. Cai* and J.K. Aggarwal*
Location: The University of Texas at Austin. *=Computer and Vision Research Center. **=Kinesiology Biomechanics Laboratory.
Year: 1994
Published: Workshop on Motion of Non-Rigid and Articulated Objects
Type: Paper
Key words: Gait recognition, stick-figure, Medial Axis Transformation, skeleton, cubic splines, correlation, markers and stereo
Summary: This work is about gait recognition using a stick figure. The idea is to match, by correlation, a trained model with a stick-figure representation of the lower limbs. An entire gait cycle is matched at the time. The model is obtained in the following way. Markers are attached to the left leg of a walking human and measured using a calibrated stereo setup. This is done for several different persons and an average trajectory for a gait cycle is obtained. Due to symmetry the trajectory of the other leg can be found. A stick-figure representation of the user in an image is found in the following way. First, the following constraints are put upon the test person in order to simplify the procedure. The subject wear dark tight-fitting clothes and walk against a light uniform background. Furthermore the subject walks perpendicular to the axis of the camera. The image is then thresholded and the Medial Axis Transformation is used to obtain a skeletonized image. The lines of the skeleton are fitted using first a polygonal approximation and then cubic splines. Two periodic signals (the model and the image) are scaled to the same size and sampled in a similar manner. They are then compared by a correlation method and the peak yields the recognized gait.

For this to work the subject has to walk with the same speed as the subjects used for training the system. In this setup all subjects have a correlation above 0.95, meaning that gait is recognized for all subjects. They state that a better method is needed to extract the stick-figure from the images.

Title: Computers Seeing Action [12]
Author(s): A.F. Bobick
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1996
Published: British Machine Vision Conference, Edinburgh, Scotland
Type: Paper
Key words: Motion understanding and overview
Summary: An earlier and shorter version of [13].

Title: Movements, Activity, and Action: The Role of Knowledge in the Perception of Motion [13]
Author(s): A.F. Bobick
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1997
Published: Royal Society Workshop on Knowledge-based Vision in Man and Machine, London, England
Type: Paper
Key words: Motion understanding, overview and taxonomy
Summary: This paper describes Mr. Bobick's view on motion understanding and gives examples from the Media lab. His idea is that motion can be divided into three different categories depending on the knowledge level involved. The three are: movements, activity and action. A movement is a predictable motion easily defined by a space-time trajectory in some space (e.g. recognizing a swinging baseball bat). An activity involves a sequence of movements (e.g. pitching a baseball). An action involves semantic knowledge related to the context of the motion (e.g. a runner being tagged). Examples of work within the first category are [19], [20] and [29]. Examples of the second category are different gesture work done at MIT. Finally he states that the mechanisms to manipulate time is very important to his taxonomy. Simple and linear for the first case, dynamic time warping for the second, and reasoning about temporal relationships for the last case.
Comments: Nice overview paper with a good motion distinction/taxonomy.

Title: An Appearance-based Representation of Action [14]
Author(s): A.F. Bobick and J.W. Davis
Location: M.I.T. Media Laboratory, USA
Year: 1996
Published: International Conference on Pattern Recognition
Type: Paper

Key words: View-based action recognition, motion-energy image, PCA, stick-figure and optical flow
Summary: This work is part of the work described in [29] but with another recognition method instead of the MHI. They do a simple recognition based on the MEI, which suggests which action and which viewing angle is present. The idea is now to find a stick-figure representation for the human and then match this against a trained model. The parameters of the sticks are compactly represented using PCA. They are found by hand in the training phase and as the center axis of optical flow patches in the recognition phase. The system is tested on three persons sitting and three persons doing aerobic moves. All are correctly classified.
Comments: This solution doesn't seem very noble and I guess that is the reason they abandoned it and starting using MHI.

Title: Toward Non-intrusive Motion Capture [15]
Author(s): A. Bottino, A. Laurentini and P. Zuccone
Location: Politecnico di Torino, Italy
Year: 1998
Published: Asian Conference on Computer Vision
Type: Paper
Key words: Shape-from-silhouettes, model based, VI algorithm, synthesized images and gait tracking
Summary: Human motion capture using the shape-from-silhouettes algorithm called VI. This technique is simple and non-intrusive but requires several cameras located around the subject. VI finds all boundary voxels of an object using camera projection. A boundary voxel is defined as a voxel where some, but not all, of its vertices belong to the object (determined by the 2D projections of the object on the different image planes). The voxel representation of the object is compared to a model of the object and through a gradient search in a 31-dimensional (the degrees of freedom in the model) space the posture of the model/object is found. All this takes place in a virtual world using synthesized imaged from 4-5 different cameras with different resolutions of the voxels. Finally they test the error rate of the recognition process.
Comments: Nice work but I think it will have difficulties on real world images.

Title: Learning and Recognizing Human Dynamics in Video Sequences [16]
Author(s): C. Bregler
Location: Computer Science Division, University of California, Berkeley
Year: 1997
Published: Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico
Type: Paper

Key words: Human dynamics, action recognition, EM and HMM
Summary: This paper describes a framework for learning and recognizing human movements, e.g. walking. The framework, which consists of four levels, is probabilistic and delays hard decision thresholds by using higher level statistical models and temporal context. At the first level each pixel is represented by optical flow and (optional) color (HSV values). In the next level blobs are defined to be areas having coherent motion, color and spatial "support-regions". The parameters behind this "layered motion" representation are estimated using the EM algorithm. Each blob is represented as a mixture of multivariate Gaussians. In the third level the blobs are grouped and represented as linear dynamical model using a Kalman filter. E.g., during a gait two different linear models are active. The first one is for the gait cycle and the second one when the leg has ground support. These models are called "movemes" in relationship to phonemes in speech. A complex gesture like gait (a "word" in speech terms) should be composed out of simple movemes. Since the goal is recognition a fourth level is introduced, HMM level. Here different HMM are set up for different types of actions. The framework is tested on three different gait categories: running, walking and skipping. The training data were tracked using MLD or were hand-labeled to get "ground-truth" information. A recognition rate between 86% and 93% is obtained. Finally they propose future work to include: texture, kinematic and 3D information.
Comments: I think he did one hell of a good job. He is using the newest techniques (EM, 'condensation') and some of the older but very powerful ones like optical flow, Gaussian modeling, Kalman-filters and HMM. I think it is a very noble idea to define "movemes" as some kind of generic types of movements. I, however, think he should have included a underlying explicit model of the human because it could have solved some of the occlusion problem which he must have. It would also help - is probable necessary? - when he wants to expand his algorithm to 3D.

Title: Tracking Human Motion Using Multiple Cameras [17]
Author(s): Q. Cai and J.K. Aggarwal
Location: Computer and Vision Research Center, University of Texas, Austin, USA
Year: 1996
Published: International Conference on Pattern Recognition
Type: Paper
Key words: Multiple human tracking, multiple cameras, human model, double-difference-method, Mahalanobis distance and window slicing
Summary: This paper is about tracking humans using multiple cameras and is based on previous work [18]. The human head is modeled by an ellipse having height/width ratio of 1 to 1.5. the trunk is modeled as a rectangle with different ratios depending on the viewing angle. A human

is segmented from the background using the double-difference-method described in [18]. Different non-background objects are found in the following way. The horizontal and vertical profiles of the binary image is found. Then the valleys of the smoothed profiles defines the boundaries of the rectangle boxes containing non-background objects. In each box a search for, first the head and then the trunk is carried out. When a match is found the medial axis of the trunk is found and different points on this axis is chosen. The location of these point yields a set of feature points as does the graylevel value at these positions. These feature points, together with the velocity, are compared to the feature vectors from the previous frames using the Mahalanobis distance. A method is presented for mapping the feature vector seen from one camera into a feature vector seen from another vector. The system is tested where two persons are being tracked.

Title: Tracking Human Motion in an Indoor Environment [18]
Author(s): Q. Cai, A. Mitiche and J.K. Aggarwal
Location: Computer and Vision Research Center, University of Texas
Year: 1995
Published: International Conference on Image Processing
Type: Paper
Key words: Tracking of humans, movable camera, double-difference-images and background segmentation
Summary: This work deals with tracking of humans using a movable camera. They divide the problem in to two situations, a viewing system with negligible motion and one with non-negligible motion. The idea is to reduce the latter to the former situation and then use the same technique. The technique for the simple situation goes like this. Find the background image and subtract it from the image and you have segmented the moving human(s). The background image is obtained using a double-difference-image technique and an examination of the result. For the situation with the non-negligible motion of the viewing system the idea is to find the motion of the viewing system and then correct the images according to this. Hereafter the same techniques as mentioned above can be used. The motion is found by matching line segments between successive frames and then translating the images accordingly. This operation will align all the images and the double-difference-image method can be used. When the human(s) have been segmented from the background they are tracked using the following three assumptions: 1) a person must move continuously in space, 2) a person does not undergo a sudden change in movement and 3) the coarse features such as hight/width ratio and texture does not change significant under indoor lighting. The system is tested in two ways. First to see that the motion of a rotating camera can be found and the images correctly compensated. Secondly a test is carried out where two persons are walking around in the lab and the system is able to track them.

Comments: The paper is too short.

Title: Recognition of Human Body Motion Using Phase Space Constraints [19]
Author(s): L. Campbell and A. Bobick
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1995
Published: International Conference on Computer Vision
Type: Paper
Key words: Human movement recognition, ballet dance steps, phase space representation and Euler angles
Summary: This work is about recognizing ballet dance steps using a phase space representation. They use a commercial system for getting 3D data from the markers on the user. A phase space is then generated where each axis correspond to each degree of freedom in the human articulating system (DOF). The system is trained on nine different ballet moves and the data is fitted in each 2D projection using cubic polynomials. A supervised learning paradigm is used to figure out which phase spaces should be used to recognize each ballet move. A measured point (set of joints) is said to belong to a certain ballet move if it is within a static threshold from the 2D projection of each of the chosen curves.
Comments: Nice work and a very well written paper.

Title: Using Phase Space Constraints to Represent Human Body Motion [20]
Author(s): L. Campbell and A. Bobick
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1995
Published: Workshop on Automatic Face- and Gesture-Recognition
Type: Paper
Key words: Human movement recognition, ballet dance steps, phase space representation and Euler angles
Summary: This is a shorter version of [19].

Title: Motion-Based Recognition: A Survey [21]
Author(s): C. Cedras and M. Shah
Location: University of Central Florida
Year: 1995
Published: Image and Vision Computing, volume 13, number 2
Type: Paper (survey)
Key words: Motion-based recognition, motion information and overview
Summary: From the moving light displays (MLD) the action or even the gender of a person can be recognized. But how is this done? Two main ideas are present: structure from motion and direct use of motion for recognition.

There are two main steps in motion-based recognition (MBR). First an appropriate representation of the object or motion must be found. Second, this representation must be matched against a model. An overview of the different motion extraction methods is shown. These are: trajectory-based features, optical flow and region-based features, which again is divided into several methods. But generally speaking all the methods can be divided into two overall methods: optical flow and motion correspondence. When a trajectory, from the motion correspondence, is found it is normally transformed into a single valued function in order to make the subsequent calculations easier. Relative motion may be useful when dealing with human motion. Motion events, which are defined as significant changes or discontinuities in motion, may also provide useful information in the recognition of motion. After the motion has been extracted from the images it can be used for recognition. Different recognition examples are given: cyclic motion detection, lip reading, gesture interpretation, motion verb recognition. The last part of the survey is the most interesting, it deals with human motion tracking and recognition. They start out by defining three ways of viewing this problem. The first one is to recognize the action performed by a human. The second is to be able to recognize the different body parts, this is called labeling. The last is called tracking of human motion and is defined as follows. Knowledge, e.g. a model, is used to guide the interpretation of an image sequence in order to analyze the motion between frames, to determine the most plausible configuration of the body. The human body is modeled either as a stick-figure model or as a volumetric model. The tradeoff is between speed in processing and details in the model. The motion must also be modeled, which usually is done by the use of joint angles. Some also use key frames, which is a sequence of stick-figure models, to model coarse movements of the body. Different examples of human motion recognition systems using models are presented. Finally the paper present a summary and some ideas of the trends of the future.
Comments: I find the paper very interesting and it gives a good overview over the motion-based recognition area, by presenting methods (divided into tree structures) and systems.

Title: Recognizing Motion Using Local Appearance [22]
Author(s): O. Chomat and J.L. Crowley
Location: INRIA, Grenoble, France
Year: 1998
Published: International Symposium on Intelligent Robotic Systems
Type: Paper
Key words: Human motion, Bayes rule, motion recognition, appearance-based, PCA and spatial-temporal
Summary: The paper presents a new way of using PCA. Instead of using only global temporal information from images, local spatial and temporal

information is used. The method is used in the action recognition area. The work is ongoing and therefore very few results are presented. The problem of recognizing the different actions in the feature space is not solved. They state that both structural and statistical methods can be used and they describe one statistical method in some detail: Bayes rule based on multidimensional histogram estimation.

Title: Tracking of Articulated Objects using Model-Based Computer Vision [23]
Author(s): C. Christensen and S. Corneliussen
Location: Laboratory of Image Analysis, Aalborg University, Denmark
Year: 1997
Published:
Type: Master's Thesis
Key words: articulated objects, model-based computer vision, human modeling, graph-match problems and human tracking
Summary: This work, which is based on their 9th semester project [24], deals with a model-based approach for tracking articulated objects. They define tracking in the following way: object recognition, pose estimation and motion registration. They only use one camera and they introduce the following limitations: only work with pose estimation and ignore the initial pose problem. They use the standard predict-match-update scheme where they assume that the articulated objects have been separated/segmented into their different parts. The matching is where they put most of there energy. They end up using a graph method for matching the input segments, e.g. body parts, with all the predicted (from the model) poses. They use a simple predictor to limit the search space. The features which are used in the matching process are simple features like number of pixels, center of mass, segment orientation etc. The model is visualized using OpenGl. The model used in the matching process is the same as the one used for visualization, making the image easier to compare to the visualized model. They test the system on synthetic images, with or without noise, and conclude that their general model-based method is promising.
Comments: Good general work on the matching process. The problem is the assumption about always having the articulated object in the image segmented into its respective parts, I guess this is where the real challenge is! They have some interesting stuff about: reconstruction versus recognition, tracking of humans, application examples, state of the art, predict-match-update and how to design and use a 3D model in model-based vision.

Title: Visualization of Human Motion using Model-based Vision [24]
Author(s): C. Christensen and S. Corneliussen
Location: Laboratory of Image Analysis, Aalborg University, Denmark
Year: 1996

Published:
Type: Technical Report
Key words: Pose estimation, predict-match-update, model-based, 2D, visualization, known init-pose and real time
Summary: This work is about a predict-match-update model-based approach for obtaining the pose of a human in motion. The system is only dealing with 2D motion, meaning that the user can only move his limbs in a plane parallel to the image plane. A dark background is used and the user is wearing tight fitting clothes where each limb is covered by a unique grey level. The images are down-sampled to speed up time. A labeling is carried out on a thresholded and edge filtered version of the images. Each limb in the image is located and represented using rectangles. The pose of the model is predicted (using velocity and acceleration) and information about the position and grey level value of each limb are calculated. A matching algorithm, based on input from the current image (rectangles and grey level values), the predicted model pose and general human constraints are used to figure out which predicted limbs match which found (in the image) limbs. The initial pose of the human must be known. When the correspondence has been established the parameters of the rectangles are used to update the pose of the model, which is visualized using a cylinder representation. The motion of the model is smoothed using a moving average filter. A number of tests have been made and show the system to be functional at 21Hz but a bit sensitive to noise.
Comments: I supervised this project.

Title: Cue Circles: Image Feature for Measuring 3-D Motion of Articulated Objects Using Sequential Image Pair [25]
Author(s): J.M. Chung and N. Ohnishi
Location: Bio-mimetic Control Research Center, Nagoya, Japan
Year: 1998
Published: International Conference on Automatic Face- and Gesture-Recognition
Type: Paper
Key words: Human body segmentation, body features, medial axis transformation and stereo
Summary: This work is about finding image features which represent a human. First the human is segmented from the background using edge image subtraction (X-OR). Then a medial axis transformation is used to find the medial axis of the torso, arms, legs and the head. Using the medial axis a number of circles are generated each with its center on the medial axis and within the contour of the human. Using a few rules the circles at the end of each segment - legs, arms and head - are found. These nine, or less, circles/features represent the shoulders, hands, hips, feet and head. The same procedure is carried out for a different camera and a correspondence

between the features in the two images are found using epipolar lines and the size of the different circles. The knees and elbows are not found but a suggestion is to use a kinematic model of a human. The system is tested and the results look nice.
Comments: Good ideas both for the segmentation and feature extraction. It is tested on very simple images without much joint bending and occlusion. I guess such phenomena would 'kill' the method!

Title: Real Time Tracking of a Human Arm [26]
Author(s): C.R. Corlin and J. Ellesgaard
Location: Laboratory of Image Analysis, Aalborg University, Denmark
Year: 1997
Published:
Type: Technical Report
Key words: Tracking a human arm, markers, stereo, error function, init-pose, avatar and Transom Jack.
Summary: This work is about controlling the arm of an avatar by tracking a human arm. The idea is to place markers on the arm of the user and find their 3D position using stereo. These 3D positions are then used to control the arm of an avatar. Three reflective ribbon markers are used on a user wearing dark clothes on a dark background. Using simple grey level thresholding the markers can be found. A marker is represented by its centroid and area. This process is done by two cameras in a calibrated, using Tsai' algorithm, parallel stereo setup. The features of the markers are matched to find the correspondence of markers in the two images. This is done using a local error function. Next the previous positions of the markers are used to track the markers from frame to frame. For this to work the user must be in a known pose when the system is initialized. The tracking of the 3D positions of the markers can now be used to control the pose of an avatar' arm. The system is tested using Transom Jack for visualization. It is able to run at 16Hz.
Comments: I supervised this project.

Title: Complex Object Tracking by Visual Servoing Based on 2D Image Motion [27]
Author(s): A. Cretual, F. Chaumette and P. Bouthemy
Location: IRISA/INRIA, France
Year: 1998
Published: International Conference on Pattern Recognition
Type: Paper
Key words: Servoing, tracking, movable camera, center of gravity, multi resolution, RMR algorithm and optical flow
Summary: This work is about visual servoing based on 2D image motion. The motion of an object is found and this information is used by a movable

camera to keep the object in the center at all time. The image motion is represented by its center of gravity (cog). The motion model used to approximate speed in the image is an affine model with six parameters. The motion parameters are computed using the multi-resolution robust estimation method (RMR) in an optical flow manner. Different image resolutions are used in each frame. Estimation at a finer resolution level is initialized by the value obtained at the preceding coarser one. In tracking mode the algorithm first find, using image differencing, a window wherein motion is present. This gives an area where the pixels are applied to the RMR algorithm. When the cog is found the control lows together with a Kalman filter are used to control the camera. The system has been tested on rigid as well as non-rigid objects. Both gave good results.
Comments: The approach will be sensitive to occlusion and multiple moving objects.

Title: A Novel Environment for Situated Vision and Behavior [28]
Author(s): T. Darrell, P. Maes, B. Blumberg and A.P. Pentland
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1994
Published: Workshop for Visual Behaviors at CVPR-94
Type: Paper
Key words: ALIVE, Looking at People, magic-mirror, Chroma-keying and figure-ground
Summary: This work describes the ALIVE (Artificial Life Interactive Video Environment) system developed at MIT. It is within the "Looking at People" domain. The idea is a "magic-mirror" paradigm, where a user's image is captured by a camera and shown (a mirrored version) on a large screen in a simulated graphical world. The good thing about this paradigm is that the user is situated in VR without wearing any special devices (gloves, HMD) which are cumbersome and can make the user feel "sea-sick". The user can be segmented using either single color background subtraction, Chroma-keying, or using a static background subtraction. The result of the figure-ground processing are binarized and the connected regions within a boundary box are found. Using knowledge about the user's anatomy and the fact that he will most likely face the screen (camera) due to the mirror effect, are used to locate the hands of the user. By tracking them over time a few gestures, e.g. waving and pointing, can be classified. The system includes two different sets of agents with whom the user can interact: a Puppet, and Hamster and a Predator. The puppet can follow the user around and imitated his actions, e.g. sitting down and jumping. The user can send it away by pointing and call it back by waving. The puppet has the ability to show emotions to convey some of its internal states. The Hamster can be feed by the user and chased by a Predator. Both will avoid objects in the scene. Both applications have been tried out by 500 people with great

success. The frame rate of the system is 10Hz.
Comments: Very nice system, but using Chroma-keying is kind of cheating :-)

Title: The Representation and Recognition of Action Using Temporal Templates [29]
Author(s): J.W. Davis and A. Bobick
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1997
Published: Conference on Computer Vision and Pattern Recognition
Type: Paper
Key words: View-based action recognition, temporal templates, Hu-moments, pooled independent Mahalanobis distance, motion-energy image, motion-history image, image differencing and KidsRoom
Summary: This work is about view-based action recognition using temporal templates. The idea is to obtain temporal templates of several different actions - aerobic moves - from all possible viewing angles. Each template is then represented using seven Hu-moments which are known to yield good shape descriptions. Several sequences of the same action is analyzed and a class is generated for each action. Each class is represented by its mean and covariance matrix. When a new action is carried out by a user the template is found, the seven Hu-moments calculated and it is classified to one of the classes using a pooled independent Mahalanobis distance. The temporal templates are defined in the following way. Each element in the template contains two types of information, or you can say that the template consists of two images. The first image is a motion-energy image (MEI). It is generated by ORing all motion images within a certain integration interval/time. A pixel in the MEI will be 0 if no motion has been present at this position for the entire integration cycle and 1 otherwise. The motion image is obtained using image differencing. The second image in the template is the motion-history image (MHI). It is basically the same as the MEI except that the intensity value is not binary. Instead it represent the motion history. The longer time it is since motion was present in a certain pixel, the darker this pixel will be. I.e., the pixels where motion was present in the last frame are white. To improve the result they use two cameras. They state that a fair recognition rate is obtained and the the system runs at 9Hz. Finally they talk a little about the KidsRoom where the technique has been applied.
Comments: Nice work. I like the idea about representing the motion using a simpler way than templates, namely Hu moments.

Title: SIDEshow: A Silhouette-based Interactive Dual-screen Environment [30]
Author(s): J.W. Davis and A. Bobick
Location: Vision and Modeling Group, Media Lab, MIT, USA

Year: 1998
Published:
Type: Technical Report No. 457
Key words: Silhouette extraction, IR image, morphology and virtual aerobic trainer
Summary: This work is about silhouette extraction in a dual-screen environment. The user is located between two video projection screens, a front and a rear screen. Behind the rear screen a set of infrared emitters are located, aiming at the projector screen (and through it). On top of the front screen a camera with an infrared-pass/visible-block filter is mounted. It first captures a background IR image and after that the silhouette of the user can be obtained by subtracting an IR image containing the user from the background image. The obtained silhouette is binarized and after a morphology operation the silhouette is smooth. Finally they mention that the system has been used in a project with a virtual aerobic trainer.
Comments: Very good idea to get stable silhouette images.

Title: Virtual PAT: A Virtual Personal Aerobics Trainer [31]
Author(s): J.W. Davis and A. Bobick
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1997
Published:
Type: Technical Report No. 436
Key words: Virtual aerobic trainer
Summary: This work is about a virtual aerobic trainer. The idea is that the user put on his own music, choose his own aerobic trainer (currently only one exist in the system) and choose the order of the exercises. The system then starts up a video session where the virtual trainer follows the music by finding the beat in the music and use this information as a synchronization signal to the video signal, where the same exercise is being looped over and over again until the next exercise begins. Two cameras pick up the user and act accordingly. When a user enters the workspace the system starts up and when the user leaves the system shuts down. The vision module also figures out if the user is doing the exercises correct and gives appropriate comments. How the vision system works is not explained, instead a reference is given.
Comments: A nice application but not many technical information in the paper.

Title: Visual Surveillance of Human Activity [32]
Author(s): L. Davis, S. Fejes, D. Harwood, Y. Yacoob, I. Hariatoglu and M.J. Black
Location: University of Maryland
Year: 1998
Published: Asian Conference on Computer Vision, Mumbai, India

Type: Paper
Key words: Overview, directional motion, tracking people/body parts and activity recognition
Summary: This paper gives an overview of what goes on in the computer vision lab. of the university of Maryland within the area of ground-based mobile surveillance system. The ideal system should, during motion, be able to detect objects and identify them as humans, animals and vehicles. When one or more humans are detected it must be determined what kind of activities they are involved in. The overall system is divided into several subsystems which are described individually. The first project deals with detecting independent motion in a long image sequence where the camera is moving. Their approach is based on a projection of optical flow fields hereby simplifying the problem. Secondly their W^4 system is described. It is a real time system for tracking people and their body parts. It uses difference images, connected component analysis and morphological analysis to detect the people. It also tries to predict where the tracked objects will be in the next frame. The body parts are tracked using dynamic template matching. For more details see [44]. Next they talk about activity recognition. They present two different systems where the first one is a model-based approach and the second is a parameterized optical flow approach.
Comments: Good overview of the activities in their lab, but very few details of course.

Title: Real Time Human Action Recognition for Virtual Environment [33]
Author(s): E. Luc
Location: Computer Graphics Lab, Swiss Federal Institute of Technology, Switzerland
Year: 1996
Published: Computer Science Postgraduate Course in Virtual Reality
Type: Paper
Key words: Flock of Birds, action recognition, motion capture, action primitives and agent response
Summary: This work is about how to recognize human actions. The input is a motion capture system based on Flock of Birds sensors. The idea is to divide each action into a number of action primitives. A primitive can be e.g. the upper body moving down or vertical down movement. The primitives are divided into gestures (dynamic) and postures (static). The recognition is carried out in a hierarchical manner. First the center of mass is used to limit the search space (number of matches). Then the end effectors are used and finally the skeleton joints are used. At each step an Euclidean distance measure is used to match the input against a trained posture prototype. A test with 16 gestures and 15 postures yields a recognition rate of approximate 90%. The rest of the paper deals with how an agent can respond to the avatar.

Comments: Interesting work in the action recognition and agent response areas.

Title: Computer vision for computer games [34]
Author(s): W.T. Freeman*, K. Tanaka**, J. Ohta** and K. Kyuma**
Location: *=MERL, a Mitsubishi Electric Research Lab., Cambridge, USA **=Mitsubishi Electric, Advanced Technology R&D Center, Amagasaki City, Japan
Year: 1995
Published: Workshop on Automatic Face- and Gesture-Recognition
Type: Paper
Key words: Computer game, special CCD chip, on-board processing, moments and orientational histograms
Summary: This work is about simple image routines for computer game. A special CCD chip with on-board processing has been constructed. It can, in real time, calculate horizontal, vertical and diagonal projections together with x and y derivatives. This can be down in real time. The chip is connected to a 16 bit 1MHz microprocessor making the solution very fast and cheap. The hardware has been used in two different applications. First the projections have been used to calculate the moments of a hand and later a person (with the background subtracted). The hand moments are used to control a graphic representation of a car while the moments of the person gives the input to a skateboarding game. The derivatives are used to calculate orientation histograms of a person where the background has been subtracted. The derivatives at each pixel are used to calculate the orientation. These orientations are represented in a histogram which are matched, using Euclidean distances, against trained histograms of different body postures. The correct posture is found and can then be used as input to e.g. a flying game.
Comments: The idea about a cheap chip with on-board processing is good but the applications are a bit naive.

Title: Real-Time Human Motion Analysis by Image Skeletonization [35]
Author(s): H. Fujiyoshi and A.J. Lipton
Location: Carnegie Mellon University
Year: 1998
Published: Workshop on Applications of Computer Vision
Type: Paper
Key words: Cyclic motion, walking and running recognition, background subtraction, dilation, erosion, "star" skeletonization, Fourier domain and autocorrelation
Summary: In this paper they work with recognition of cyclic motion, walking and running. First they detect a moving human by adaptive background subtraction. After a binarization the detected object is dilated twice fol-

lowed by an erosion to filter the segmented object. Next a border extraction is performed and this data is represented as a distance map with respect to the center of mass. The distance data are filtered either in the Fourier domain or in the spatial domain. A "star" skeletonization is carried out by finding extreme points in the distance representation. The angle between the vertical line passing through the center of mass and the lower left most point is found. This represents the movement of the feet. A similar angle is found for the head as the top most extreme point. The angular data are filtered by a autocorrelation and then transformed to the Fourier domain. The system is tested on 20 image sequences containing a walking human and 20 of a running human. The average walking frequency is 1.75Hz and 2.875Hz for running. A recognition rate of 97.5% is archived.

Title: The Visual Analysis of Human Movement: A Survey [36]
Author(s): D.M. Gavrila
Location: Image Understanding Systems, Daimler-Benz Research, Ulm, Germany
Year: 1999
Published: Computer Vision and Image Understanding, Vol 73, no. 1
Type: Paper (survey)
Key words: Survey, applications, taxonomy, recognition and future research
Summary: This paper is a survey describing state of the art work within the "Looking at People"-domain. Especially whole-body and hand motion. First a number of applications are presented which are divided into the following categories: virtual reality, "smart" surveillance systems, advanced user interfaces, motion analysis and model-based coding. Next the previous work is classified into the following three sections: 2D approaches without explicit shape models, 2D approaches with explicit shape models, and 3D approaches. The first class of approaches is characterized as bypassing the pose recovery step and describe human movement in terms of simple 2D low-level features from a region of interest. Or as Polana and Nelson describe it: "getting your man without finding his body parts". The next class of approaches are characterized as using a priori knowledge of how the human body (hand) appears in 2D. The last class of approaches involved 3D model informations. It is again divided into: 3D body modeling, 3D pose recovery and tracking, feature correspondence and experimental results. On top of the three different classes of approaches is a recognition layer which is described independent of the three classes. A discussion on the above subjects are given. It is argued that the choice between 2D and 3D is largely application-dependent. A number of issues for future research are presented: model acquisition, occlusion, model constraints, the clothes problem, using ground truth, using 3D data, initialization, initialization versus tracking, and multiple humans.

Comments: A very good survey. Good reference list, good application section and good taxonomy. I, however, think that he divide the 3D part in a strange way.

Title: 3-D Model-Based Tracking of Humans in Action: A Multi-View Approach [37]
Author(s): D.M. Gavrila and L.S. Davis
Location: University of Maryland
Year: 1996
Published: Conference on Computer Vision and Pattern Recognition, San Francisco
Type: Paper
Key words: Model-based vision, multiple view, predict-match-update and chamfer images/matching
Summary: They use a model-based approach to track humans in 3D. A recognition cycle goes like this: predict, based on the current and previous states, the intervals in where each body parameter (e.g. joint angles) lie. For each combination of these 22 parameters synthesize the projection of the human model from the cameras point of view. Calculate a similarity measure between each synthesized (view) image and the real image obtained by the camera. They compare edges in the images and hereby (re)formulate the problem as a search problem of how to compare two edge images - a real image and a synthesize image. The search is carried out in a 22-D space, but using search space decomposition the search is reduced to three searches in the following dimensions: 5, 9 and 8. 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 a sequence of multi-view (4) frames and run the algorithm for each view. In order to obtain better edges they wear tight-fitting colored clothes. They conclude that their system, even though it is slow, is capable of tracking multiple humans.
Comments: Very nice system. They have some good comments on the topic of why to use model based approaches. Once the model pose which best fit the data has been found, all the parameters of the model is given in a viewpoint independent manner. Such an approach is more robust since it is based on whole contours rather than on a few points - joint angle recognition. A more robust prediction can be done at pose level instead of pixel level - low level features and edges - and it is also easier to include semantics and high level a priori knowledge, e.g. kinematic, at this high level.

Title: Reach Out and Touch Space (Motion Learning) [38]
Author(s): L. Goncalves, E.D. Bernardo and P. Perona
Location: California Institute of Technology, USA
Year: 1998

Published: International Conference on Automatic Face- and Gesture-Recognition
Type: Paper
Key words: Motion models, human motion, markers and Polynomial basis functions
Summary: This work is about a method for learning models of human motion. A user is equipped with a number of ping-pong balls (14) which are tracked over time. Polynomial basis functions are used to model the motion. The idea is to use the obtained motion pattern in a human motion tracking setup. Two experiments are conducted and the results were that users find model-generated synthetic motions very realistic and have a hard time discriminating between real data and synthetic data.

Title: Monocular Tracking of the Human Arm in 3D [39]
Author(s): L. Goncalves*, E.D. Bernardo*@, E. Ursella@ and P. Perona*@
Location: *=California Institute of Technology, Pasadena, CA, USA @=Universita di Padova, Italy
Year: 1995
Published: International Conference on Computer Vision
Type: Paper
Key words: Analysis-by-synthesis, arm tracking, model-based, truncated right-circular cones, extended Kalman filter and ground truth
Summary: This paper deals with tracking of a human arm. They use an analysis-by-synthesis approach where the arm is modeled as two truncated right-circular cones. The shoulder is assumed to be fixed and the shoulder and elbow joints are represented as spherical joints. The hand is assumed to be stretched in the direction of the lower arm at all time. The user in the image is blurred and thresholded. Due to a dark background the edges around his arm are very clear. They are compared with a projected version of the model of the arm. An error measure, representing the differences between the arm in the image and the projected model arm, is calculated and used to update the pose of the model. This is done through an extended Kalman filter. The system is tested where the user moves his arm in a rectangular pattern. Using ground truth information the average error is calculated to be maximum 5cm in all three directions. This correspond to a relative depth error of less than 7.5%.
Comments: A well-written paper. A very clear example of the analysis-by-synthesis approach. The paper has a good section on different applications. It also has an interesting analysis on the depth sensitivity with respect to a given image coordinate error.

Title: MDL-Based Spatiotemporal Segmentation from Motion in a Long Image Sequence [40]
Author(s): H. Gu, Y. Shirai and M. Asada

Location: Osaka University, Suita Osaka, Japan
Year: 1994
Published: Conference on Computer Vision and Pattern Recognition
Type: Paper
Key words: Tracking human motion, long image sequence, spatially stable edge segment (SSES), motion models and minimum description length (MDL)
Summary: This work is about tracking motion in a long images sequence. The idea is to represent an image by spatial stable edge segments (SSES) and then track these edges by different motion models. A SSES is an edge with an associated number representing its strength which is calculated as the product of its length and greylevel contrast. Each SSES is being tracked using the optical flow constraint between consecutive images and is classified to follow one of the following motion patterns: stationarity, translation and rotation, using minimum description length (MDL). SSESes are merged and split according to the MDL. In order to describe motion in a long sequence more complex motion models are defined. These eight new models include the three mentioned above but with constant velocity and/or acceleration in all combinations. This of course means that more then two consecutive images must be considered in order to obtain the correct motion model. A framework is set up where the individual body parts (or SSESes) can change motion model during a sequence. The system is tested on a walking person moving parallel to the image plane and his head, torso, left arm and left foot is correctly tracked.
Comments: Very hard to understand the MDL since they mostly refer to another work. They don't deal with occlusion where, I think, the method will fail. Moreover, the motion models are only in 2D.

Title: A 3D Reconstruction System for Human Body Modeling [41]
Author(s): J. Gu*, T. Chang@, I. Mak*, S. Gopalsamy*@, H.C. Shen* and M.M.F. Yuen**
Location: Location: *=Department of Computer Science, Hong Kong University of Science and Technology. **=Department of Mechanical Engineering, Hong Kong University of Science and Technology. @=NCST, Bombay, India.
Year: 1998
Published: Lecture Notes in Artificial Intelligence 1537. Modeling and Motion Capture Techniques for Virtual Environments
Type: Paper
Key words: Human model, framework, multiple cameras, contours, deformable superquadrics, manufacturing mannequins and parametric surface representation
Summary: This paper is about a system for generating a model of a human. The idea is to mount 13 cameras on the poles of a hexagonal framework and place the user, whose body is to be modeled in this framework. The

cameras are calibrated and a multiplex switch is used to feed the individual camera signals into the same computer. This capturing process needs 15 seconds to complete its task. In this period the user has to stand absolutely still. When the images are captured the edges are found and combined into contours which are segmented into various body parts. A human model consisting of deformable superquadrics is introduced. The image data is fitted to the model in two steps. First a rough match using the contours of the human is used. Next the model is finetuned using information from 3D data obtained through a stereo algorithm by projecting a structured light pattern onto the user. The estimated human model pose is compared to ground truth data and very good results are achieved. They want to use their system for manufacturing mannequins. Therefore a parametric surface representation is needed. To obtain this they use key feature points on the human body which are located using special markers attached to the user. At, or with respect to, these key feature points it is calculated how a plane intersects the superquadrics models, and parametric curves are found. Finally the paper states how long time the different tasks require and how much user interaction is needed.
Comments: I don't understand why the image capturing process needs 15 seconds. Not many details are given on the generation and matching of the model.

Title: Tracking Human Body Motion Based on a Stick Figure Model [42]
Author(s): Y. Guo, G. Xu, and S. Tsuji
Location: Department of Systems Engineering, Osaka University, Toyonaka, Japan
Year: 1994
Published: Journal of Visual Communication and Image Representation, Vol. 5, No. 1, March, pp. 1-9
Type: Paper
Key words: Human motion, model-based, stick-figure, skeleton, silhouette, morphology, potential field, energy function and analysis-by-synthesis
Summary: This work is about finding human body motion based on a stick figure. The idea is to find the skeleton on a human and then match this up against a stick-figure model with 6 joints and 10 sticks. The movements are all done in parallel to the image plane, making it a 2D problem. The silhouette of a human is found by subtracting each frame from a static background. Noise is removed using morphological operations. The skeleton is found using a method proposed by Montanari. It is matched against the stick-figure using an energy function which are minimized. To reduce the complexity of the problem a potential field is introduced. It transforms the problem into finding a stick-figure with the minimal energy in the potential field. The problem is further reduced by the use of prediction and angle constraints of the individual joints. The system is tested on a person walk-

ing on a treadmill and a person running in an outdoor scene. The system has problems handling occlusion but beside that it performs well. It runs on prerecorded image sequences at a frame rate worse than 1Hz, i.e. off-line processing.
Comments: Good work in the analysis-by-synthesis field. Of course its only in 2D.

Title: Ghost: A Human Body Part Labeling System Using Silhouettes [43]
Author(s): I. Haritaoglu, D. Harwood and L.S. Davis
Location: Computer Vision Laboratory, University of Maryland, USA
Year: 1998
Published: International Conference on Pattern Recognition
Type: Paper
Key words: Body part labeling, silhouette, histograms, convex hull and relative path distances
Summary: This work deals with a system for body part labeling using monochromatic images. They use the silhouette of a human found by the W^4 system from their lab, see [44]. The idea is to identify the main features of a human body: head, hands, feet, torso, elbows, knees, shoulders, armpits, hip and upper back. Based on these features an estimate of the different body parts and their positions can be found. The part are found in the following way. They calculate the horizontal and vertical histograms of the silhouette of a person in four main postures, standing, crawling-bending, laying down and sitting. For each posture several different images are obtained and an average for each two histograms is calculated. This is done for three different camera views, front, right and left. When a new image is captured the silhouette is found and the two histograms calculated. These histograms are compared to the off-line-obtained average histograms and the most likely pose and action is found as the one with the highest match. Having this pose it can be predicted which of the parts can be found and which will be occluded. A convex hull algorithm - Graham scan - is used to find the convex and concave hull vertices which are candidates for the parts. The topology of the estimated body posture and relative path distances are used to figure out which vertices correspond to which parts.
Comments: I don't understand how they come from the vertices to the (body)parts.

Title: W^4: Who? When? Where? What? - A Real Time System for Detecting and Tracking People [44]
Author(s): I. Haritaoglu, D. Harwood and L.S. Davis
Location: Computer Vision Laboratory, University of Maryland
Year: 1998
Published: International Conference on Automatic Face- and Gesture-Recognition

Type: Paper
Key words: Tracking of people, predict-match-update, part tracking and dynamic template
Summary: This paper is about an early version of the W^4 system which is a surveillance system aiming at recognition interactions between people and people/objects. In this version they work with detecting and tracking of multiple people and their parts. The system works with monocular gray scale images and infrared images (since it should work in the dark). They use a standard predict-match-update scheme, where they match predicted data against image data to find the correspondence between a predicted object/person and a measured (in the image) person/object. The image data is obtained by detecting movements using an adaptive background subtraction, yielding a motion boundary box. Noise is removed from the motion boundary box by thresholding and morphology. The position and motion parameters of a person are estimated based on, first median matching for a coarse matching and then silhouette correlation between two consecutive frames, for the fine matching. This works fine until either a split or a merge situation happens. A split situation is when a motion blob splits into more blobs and a merge situation is when more motion blobs merge into one blob. In the split situations, which is not following a merge situation, W^4 follows the new objects for a while and if they stay separated it concludes that it actually is two different objects and they are tracked individually. The merge situation, which happens frequently when people occlude each other, is solved in the following way. When two or more objects merge into one they are tracked as one. When they emerge again it can be hard to tell which object correspond to which of the objects prior to the merging. Therefore a dynamic template - called temporal texture template - is constructed in each frame where it can be done and a correlation can solve the problem. If that is not enough the average intensities in the different body parts can be used. They also want to track the individually body parts: head, torso, hands, legs and feet. For that reason they use a cardboard model. Based on the motion boundary box and the defined scaling of the different parts of the cardboard model an initial guess (part boundary boxes) can be stated about the vertical position of the different parts in the image. Using the width of the part in each boundary box and calculating the principal axis of each part a good estimate of the position of the parts can be estimated. Finally the parts are refined using temporal textured templates, as mentioned earlier. The system runs at 20Hz on a 200MHz on a dual processor Pentium PC. In the future they want to remove the assumption about persons being in an upright position and they also want to use stereo.
Comments: Nice work on the visual surveillance problem. God idea to use templates to refine the positions of the parts.

Title: Motion-Based Recognition of Pedestrians [45]

Author(s): B. Heisele and C. Wohler
Location: Daimler-Benz, Research and Technology, Image Understanding, Germany.
Year: 1998
Published: International Conference on Pattern Recognition
Type: Paper
Key words: Recognition of pedestrians, color clustering, frequency domain, time delay neural network and spatio-temporal receptive fields
Summary: This work is about recognition of pedestrians from a single moving color camera. The idea is to first divide the image into clusters having the same color/position. These clusters are tracked over time and each represented by a rectangle whose dimensions are found using PCA. Since both legs usually are segmented into the same cluster and since only motion parallel to the image plane is considered, only the width of a cluster is considered. The width of the leg-cluster roughly follows a sinus-wave over time. Therefore the width signals are transformed to the frequency domain where only the two first lines of the spectrum (1.56Hz and 3.125Hz) are used. A complete quadratic polynomial classifier is used to find pedestrians. The recognition rate is not very impressive and therefore another classifier is introduced. A time delay neural network (TDNN) is used. The input is a sequence of image regions cropped out according to the location of the clusters found by the first classifier. The regions are scaled to a fixed size to make them consistent for the neural network. In the network the concept of spatio-temporal receptive fields is introduced. This takes into account that the input is three-dimensional and therefore each neuron is not dependent on the entire number of neuron in the layer below, but only on the neurons present in its receptive field. The system is tested and robust results are obtained, but not in real time.
Comments: The details about the clustering algorithm and the neural network is not presented.

Title: Popup People: Capturing human models to populate virtual worlds [46]
Author(s): A. Hilton and T. Gentils
Location: Center for vision, speech and signal processing, University of Surrey
Year: 1998
Published: Siggraph
Type: Paper
Key words: Animation, human model, silhouette and texture mapping
Summary: The paper is about how to incorporate individuality into human animation for VR-application. The idea is to use a generic human model, based on VRML-2 H-Anim, and then fit it to an individual person and map texture on top. Then you have a VMRL model of an individual

human with correct texture - cloth, face, hair etc - which can be used in VR animations. The match between the generic model and the human is found based on silhouette feature point matching. The results are some scaling factors (between the model and the data) which are applied to the generic model. Four orthogonal views are used and these result in a 3D reconstruction of each slice of the human, see figure 8, in the paper. Finally the 3D model is texture mapped using data from the four orthogonal images. Some nice results are presented where they generate textured models of three individually persons and animate them (running).
Comments: Its not perfect but it looks good. Of course the texture is static meaning that the clothes do not obey the lows of physics. They don't mention anything about the framerate, but I guess its slow due to the complex model.

Title: Model-Based Vision: A Program to See a Walking Person [47]
Author(s): D. Hogg
Location: University of Sussex, UK
Year: 1983
Published: Image and Vision Computing, Vol. 1, No. 1, February
Type: Paper
Key words: Model-based, tracking a walking person, plausibility value, image subtraction, constraints and "Walker"
Summary: This work is about a model-based approach for tracking a walking person. A model of a human body is used. It consists of 14 cylinders arranged in a hierarchy. The idea is to compare edge projections of the model with the edges in an image. The best fit defines the configuration of the walking human. The comparison is based on a plausibility value which is calculated in the following manner. For each model line all image edges which are within a certain range - distance and orientation - are summed and that sum is divided by the length of the model line. To avoid searching the entire search space constraints are introduced: positions, movements and postures. To narrow the search further the area in the image where the person is located is obtained using image subtraction. The idea are implemented in a system named "Walker", which is successfully tested on a single image sequence.
Comments: A classical piece of work in the analysis-by-synthesis field.

Title: Interpreting Images of a Known Moving Object [48]
Author(s): D.C. Hogg
Location: University of Sussex, UK
Year: 1984
Published:
Type: PhD-thesis
Key words: Model-based, tracking a walking person, plausibility value,

image subtraction, constraints and "Walker"
Summary: Basically the same as [47].

Title: Estimation of Articulated Motion Using Kinematically Constrained Mixture Densities [49]
Author(s): E.A. Hunter, P.H. Kelly and R.C. Jain
Location: Visual Computing Laboratory, UCSD, USA
Year: 1997
Published: Workshop on Motion of Non-Rigid and Articulated Objects, Puerto Rico
Type: Paper
Key words: Human motion estimation, 3D model, EM-algorithm, eigenvalues, manifold, kinematic, "feasible"-space and Newton-Raphson
Summary: This work is about a model-based approach to human motion estimation using a constrained EM algorithm. Images are segmented into moving objects and background objects using a simple statistical background model. A five-part (arms and torso) 3D model with 14 DOF are introduced and projected into the image plane. The segmented image and the projected image are compared and the expected foreground pixel ownership is computer, the E-step of an EM algorithm. Next the spatial distribution of the foreground pixels are estimated using an iterative maximum likelihood method, the M-step of the EM algorithm. Using the eigenvalues of the distributions the orientation of the human limbs are estimated. To use the kinematic constraints which are present in an articulated object a "feasible"-space is introduced. It consists of a manifold within the model parameter space. Each EM iterate is projected onto this subspace by a Newton-Raphson procedure. In this way the estimated solution of the model parameters are restricted to belonging to the set of points on the manifold, or in other words, the solution will be feasible/realistic. The system is tested on non-occluded data, since it can't handle any occlusions. With a framerate of 1Hz on a standard SGI Indy machine, the system is capable of tracking a human moving his arms. The accuracy of the estimation is somewhat dependent on the true posture.

Title: Real-Time Estimation of Human Body Posture from Monocular Thermal Images [50]
Author(s): S. Iwasawa*@, K. Ebihara*, J. Ohya* and S. Morishima@
Location: *=ATR Media Integration and Communications Research Laboratories, Japan. @=Faculty of Engineering, Seikei University, Japan
Year: 1997
Published: Conference on Computer Vision and Pattern Recognition
Type: Paper
Key words: Human body posture, thermal images, infrared camera, center of gravity, distance transformation and genetic algorithm (GA)

Summary: This work deals with estimation of human body posture from thermal images. An infrared camera is used to obtain thermal images of a human. Using a threshold the human can be segmented from the background. The system needs to be initialized. This is done by placing the human in a predefined position and extracting the different features. Then the center of gravity is found using a distance transformation, which can be weighted and thereby reducing the influence of the arms and legs. Using another distance transformation the principal axis of inertia of the upper body is found. Using different rules and distance transformations significant points - top of head, tip of feet and tip of hands - are found. The major joints, elbows and knees, are found using a learning procedure based on a genetic algorithm (GA). Tests show that the GA can find the joints rather accurately. As an evaluation an interactive system is build. The users posture is reproduced by a Kabuki character. The system runs at 20Hz.
Comments: Nice work and good method for finding the joints. I also like the way they use the distance transformation. I guess that the system only can work when the user is facing the camera rather directly. Also it can not deal with 3D movements. It is not clear how the initialization is done and how it is used.

Title: 3-D Reconstruction of Multipart Self-Occluding Objects [51]
Author(s): N. Jojic*, J. Gu**, H.C. Shen** and T. Huang*
Location: *= Beckman Institute, University of Illinois, USA, **= Department of computer Science, Hong Kong University of Science and Technology, Hong Kong
Year: 1998
Published: Asian Conference on Computer Vision
Type: Paper
Key words: Deformable superquadric models, multiple cues, contours and stereo
Summary: They work with reconstruction of multipart objects using superquadrics as a model of each part. The models are governed by two cues: occluding contours and structured light based stereo, where the latter is guided by the former. First the models are placed by hand in a position close to the true position. Then the forces, applied to make the models fit the image contours, are calculated using a modified chamfer matching. Next this result is used to guide the stereo calculation of the forces which will make the 3D data match the model data. The two forces are added yielding the total force which is used to calculate the final deformation and position of the models. Two experiments are performed. First on a human upper body modeled by five deformable superquadrics and then on a doll using six deformable superquadrics. For the second experiment the result is compared with the result obtained using only stereo and a better result is gained. The advantage of the approach is that no special camera configur-

ation is needed.
Comments: Very interesting work even though their result section is lacking. It is only done on still images and has to be manually initialized. It is not clear how self-occlusions are solved, except perhaps for multiple views. Also the part about the forces and the chamfer image is a bit hard to understand, but they have references.

Title: Human Motion Estimation and Recognition [52]
Author(s): S. Ju
Location: University of Toronto
Year: 1996
Published:
Type: Review - Depth Oral Report
Key words: Review, human motion estimation and recognition and optical flow
Summary: This is a review of work done within the area of human motion estimation and recognition with special focus on optical flow techniques. Both coarse human motion as well as hand and face motion are described. After an introduction the report is divided into the following sections: human motion estimation, segmentation, human motion recognition and future directions. The first is divided into these subsections: robust multiple motion estimation, model-based human motion tracking, tracking based on deformable models and tracking without a prior shape model. The first is basically optical flow in several variants. The next three are what the names imply. The segmentation section is divided into a part about segmenting moving people from the background and segmenting an image into piecewise-smooth surfaces. In the human motion recognition section the following topics are discussed: direct recognition methods, phase space methods, HMM methods and PCA methods. In the last chapter several general questions are asked and different answers given.
Comments: Too much focus on optical flow for my taste. It lacks some illustrations and a better taxonomy. I find it a bit messy and the division into the different sections is not clear. It is not a general review of this area but still it has a lot of relevant references.

Title: Cardboard People: A parameterized Model of Articulated Image Motion [53]
Author(s): S.X. Ju*, M.J. Black** and Y. Yacoob***
Location: * = University of Toronto, ** = Xerox Palo Alto Research Center, *** = University of Maryland
Year: 1996
Published: International Conference on Automatic Face- and Gesture-Recognition
Type: Paper

Key words: Model-based tracking of humans, image motion, optical flow and motion recognition
Summary: The paper describes a system for tracking the different limbs of a human and propose a classification method for different activities. The tracking, which is based on previous work by Black and Yacoob on motion of a human head, is done in the following way. They have a model of a human which consists of 10 rigid planar patches, the cardboard person model. The motion of each patch is defined by 8 parameters. For each frame the system tries to calculate these parameters, using the optical flow constraint equation. This is done by defining an energy function which is minimized in a standard way. The function is calculated by taking the intensity of each pixel in a patch from the last frame and then see which eight parameters best explains the position of each pixel in the current frame (using the optical flow constraint). The function does this for each patch making it a global optimization problem. The function is expanded by a term describing the distance between a predicted corner point in one patch and a predicted corner point in the patch next to it. These two points must be close to each other in real life and therefore the two model points must also be close to each other. By extending the function in this way, the model is taken into consideration. To make the approach work the four corner points of each patch is manually defined in the first frame. They test their approach using a person walking on a treadmill where they track the left leg and it seems to work. Next they propose a method for recognizing movements. It is a kind of DTW where the patterns are transformed using PCA to capture the dominant curve components. They end up with a discussion where they state two problems for their system: clothing and self occlusion.

Title: Vision-Based Animation of Digital Humans [54]
Author(s): I. Kakadiaris* and D. Metaxas**
Location: *=University of Houston, TX, USA. **=University of Pennsylvania, USA
Year: 1998
Published: Computer Animation
Type: Paper
Key words: Pose estimation, model-based, multiple cameras, Kalman filter, initialization, visibility criterion and observability criterion.
Summary: This work is about a model-based approach to human motion capture using multiple cameras. The idea is to start out by generating a model which fit the current user. This model is then predicted, using a Kalman filter, and the forces which should be applied to the model to make it match the image are calculated. The forces are then used to update the pose of the model. Since the model is fitted to the current user in the initial phase, the system is able to animate customized virtual humans. The system consists of three orthogonal cameras, but only one is used at the time.

To figure out which view to use a visibility criterion and an observability criterion are defined. The first is a ratio between the visible area with or without occlusion included. The last criterion concerns the observability of motion, how good a change can be observed. In the matching phase the silhouette of the human in the chosen view is fitted by deformable models. Next the correspondence between nodes on the predicted model and on the fitted model is found, and represented by a force which is used to update the human pose model. The system is tested on the arms of a user and the result looks good. The system is also used in a project about rapid prototyping of rehabilitation aids customized for a special physically challenged user.
Comments: Most of the work is not described in detail and instead references are given.

Title: 3D Human Body Model Acquisition from Multiple Views [55]
Author(s): I.A. Kakadiaris and D. Metaxas
Location: Grasp lab., Dep. of Computer and Information Science, University of Pennsylvania, Philadelphia
Year: 1995
Published: International Conference on Computer Vision, Boston, Massachusetts, USA
Type: Paper
Key words: Human body part identification, deformable silhouettes and model builder
Summary: A paper about a human body part identification system. They use deformable silhouettes to do the job. The system is kind of an initialization system where the user needs to do a preset sequence of movements. First you tread the human as one object. Then, as it deforms by moving, the object is split up into new objects and in the end you should have each limb represented by an object. How the current image data are matched against the found model/object data is not explained in detail, but they say that the physics-based shape and motion estimation framework is used and gives a reference. Next they describe how they obtain 3D data by combining data from three orthogonal views. First they build a 3D model of the human standing and then it is incrementally refined as the human starts to move. To get a continuously surface of the model a thin-plate deformation energy-function is imposed during the fitting process. They have some experiments where the human moves and the model seems to be able to coupe with it. It should be noticed that the human is wearing tight fitting uniform white cloth, the background is uniform black and the movements are very stiff, slow and simple.
Comments: I don't like the way they wrote the paper, too much focus on the algorithms and the implementation, and too little on the principles behind. The idea about gradually refining a model is good.

Title: Model-Based Estimation of 3D Human Motion with Occlusion Based on Active Multi-Viewpoint Selection [56]
Author(s): I.A. Kakadiaris and D. Metaxas
Location: Department of Computer and Information Science, University of Pennsylvania, USA
Year: 1996
Published: Conference on Computer Vision and Pattern Recognition
Type: Paper
Key words: Model-based, three camera, deformable model, occlusion and Kalman filter
Summary: This work is an extension of the work done in [55]. It deals with model-based estimation of a human body. It is about how to choose the camera view(s) which makes the estimation easier and how to update the human model. The model is the one developed in [55] and is a deformable model based on silhouettes. They use three orthogonal cameras to obtain the model. The idea is to find the forces applied to the deformable model using the images and then use this information to update the model. First the best camera view of a given non-occluded body part is found using two criteria. The first use the model to check the visibility of the part and the second use a Kalman filter to check the observability of its motion. The calculation of the force is done using a theorem from projective geometry that relates points on the occluding contour of an object to points on the surface of the object itself. Some experiments are carried out both on synthetic and real images. Both use an arm and uniform background. The result is mapped to a computer graphic model and the results look nice.
Comments: Very hard to read due to lots of math and the fact that it is based on earlier work. I don't quite understand how the image is used (which pixels/points).

Title: A Human Motion Estimation Method Using 3-Successive Video Frames [57]
Author(s): Y. Kameda and M. Minoh
Location: IMEL, Faculty of Engineering, Kyoto University, Japan
Year: 1996
Published: International Conference on Virtual Systems and Multimedia
Type: Paper
Key words: Estimate pose of a human, model-based approach and double-difference image
Summary: This works is about using double-difference images to estimate the pose of a human in a model-based approach. A double-difference image is calculated in the following way. Calculate a difference image from t-1 and t, and binarize it. Do the same for images at t and t+1. AND these two images and you have a double difference image. Noise will be present in an image like this and therefore some filtering is done. Each four by four pixels

is grouped into one new square pixel which is said to be in motion if more than half of its pixels are motion pixels. After an outlier removing process the image has been filtered and the resolution reduced to save computational cost. The human model is the same as they use in the two other papers [58] and [59]. They, however, only seem to be using the upper half of the model (upper body). They assume camera calibration and that the initial pose of the human is known in advance. If no motion is present in a double difference image then they assume that no motion has been present and the model is updated, without any calculations, using the parameters from the last frame. When motion is present they take one model body part at the time and project it to the image plane and match it against the motion in the double difference image. The best match yields the pose parameters. This is done for all body parts starting at the root, the pelvis. They test the system and get a frame rate for the image processing at approximately 4.4Hz. With the matching process included it drops to 1.1Hz.
Comments: I see a potential problem with the matching scheme since they are matching a solid object (model) against a potential non-solid object (the double difference image). I also see a potential problem in their local matching scheme. Perhaps a global matching would be better. As for earlier work done in their lab they don't explain anything about the model they use and as for the rest of their work it must be sensitive to the size/shape and clothes of the user using it. They don't have any figures showing what the model will look like for a given input image. It seems to be a too unprecise method for finding precise pose parameters.

Title: Three Dimensional Pose Estimation of an Articulated Object from its Silhouette Image [58]
Author(s): Y. Kameda, M. Minoh and K. Ikeda
Location: Faculty of Engineering, Kyoto University, Japan
Year: 1993
Published: Asian Conference on Computer Vision
Type: Paper
Key words: 3D pose, silhouette image, model-based approach, analysis-by-synthesis and XOR operation
Summary: This paper is about estimating the 3D pose of a human from one silhouette image. They use a model-based approach in a analysis-by-synthesis manner. The idea is to have a rather good model of a human. The silhouette is projected for a number of different angles and matched, using a XOR operation, against the silhouette of the human. The best match corresponds to the pose of the human. They use a local match strategy, i.e. one limb at the time. When a limb is found it's silhouette is removed from the matching process.
Comments: Nice work but must be very sensible to which human is using the system, to the silhouette segmentation, to scaling and to occlusions.

Perhaps a global matching strategy would be better. I like the idea about using the XOR-operation as a metric for the matching process.

Title: Three Dimensional Motion Estimation of a Human Body Using a Difference Image Sequence [59]
Author(s): Y. Kameda, M. Minoh and K. Ikeda
Location: Faculty of Engineering, Kyoto University
Year: 1995
Published: Asian Conference on Computer Vision
Type: Paper
Key words: Model-based, pose estimation and image differencing
Summary: This paper deals with a model-based approach to human body pose estimation using image differencing. They have, what seems to be, a rather advanced human model. The idea is to find motion in an image and then divide it into three regions, generated moving regions, continuous moving regions and vanishing moving regions. These regions are compared with projected model body parts. This is done for different values of the model/pose parameters and the configuration which give the best overlap is the correct pose. When a body part is occluded a predictor based on inertia is used to update the pose parameters. The system is tested on a set of images and the result is shown.
Comments: Interesting work. It is not absolutely clear to me how they decide if a body part is occluded or not. Also I have some problems believing in the way they define the three types of motion. It will only work if the background is dark, I think. In the test it looks like the subject is moving parallel to the image-plane. The surface of the model seems to be rather complex, but they just state that it has 15 nodes. The approach must somehow be dependent on the size, shape and clothes of the user.

Title: MOVE3D - Software for Analyzing Human Motion [60]
Author(s): T.M. Kepple
Location: Department of Rehabilitation Medicine, National Institutes of Health, Maryland, USA
Year: 1992
Published: In Proc. of Johns Hopkins National Search for Computing Applications to Assist Persons with Disabilities, Laurel, Maryland
Type: Paper
Key words: Software package, post-processing of human motion data and visualization
Summary: This is about a software package which can be used to calculate and visualize human body motion. The software is used to study motion disorders for people with different syndromes, e.g. Rheumatoid and Post-Polio. The input to the system is three-dimensional coordinates obtained from fixed landmarks on the human body. The system is for post-processing

and an entire different system is needed to obtain the three-dimensional coordinates.

Title: Determination of 3D Human Body Posture from a single View [61]
Author(s): H.J. Lee and Z. Chen
Location: The Institute of Computer Engineering, National Chiao-Tung University, Hsinchu, Taiwan, Republic of China
Year: 1985
Published: Computer Vision, Graphics, and Image processing 30, 148-168
Type: Paper
Key words: Human posture, human model, stick-figure, partial binary tree, constraints and synthesized data
Summary: This work is about finding the posture of a human from a single view. They use a human model, stick-figure, with 14 joints and 17 segments. They assume an image where all joints of the human are segmented. Several feature points in the face are used to determine the 3D position of the neck. By assuming to know the 3D length of each segment, the 3D position of an end joint of a segment can be in one of two positions when the start joint's 3D position is known. A partial binary tree is build. The top level is the neck and the bottom level is the left ankle. At each node in the tree one of two solutions for the next joint is possible. Say that the left upper arm is found to be in a certain 3D position, then the lower arm must be in one of two positions, since the end joint's image projection and 3D length are known. A path through the tree is equal to one body posture. With 13 joints/levels the total number of possible body postures adds up to 8196. By using angle, distance and collision constraints this number can be reduced to below 200 configurations. By assuming the human is walking several additional motion constraints are added and the correct posture can be found. The system is tested on synthesized data generated by a walking-person simulator, and the correct post is found in both image sequences.
Comments: There is a nice table of human joint limitations and also a nice model. They also have good constraints on the joints. Also a nice presentation of different 3D methods. A mayor drawback is of cause that all joints need to be segmented beforehand.

Title: Knowledge-Guided Visual Perception of 3-D Human Gait from a Single Image Sequence [62]
Author(s): H.J. Lee and Z. Chen
Location: The Institute of Computer Engineering, National Chiao-Tung University, Hsinchu, Taiwan, Republic of China
Year: 1992
Published: Transactions on Systems, Man, and Cybernetics, Vol. 22, No. 2.
Type: Paper

Key words: Human posture, human model, stick-figure, partial binary tree, constraints, synthesized data, temporal considerations, graph search method and A*
Summary: A new step along the work descriped in [61]. New is that temporal considerations in taking into account to obtain smooth motion in the model. A graph search method, A*, is used to find a unique solution. New experiments show that the results are improved.
Comments: Seems ok, but still the problem about the joints being given beforehand.

Title: Human Body Limbs Tracking by Multi-ocular Vision [63]
Author(s): F. Lerasle, G. Rives and M. Dhome
Location: Blaise-Pascal University, France
Year: 1997
Published: Scandinavian Conference on Image Analysis
Type: Paper
Key words: Model-based, leg tracking, CAD-model, textual, correlation, Kalman and Levenberg-Marquardt algorithm
Summary: In this paper they use a two-camera model-based approach to track a leg of a cycling person. The model is a 3D textual CAD model. The 3D model is provided by a RMI scan which consists of 34 cross-sections of the leg. The texture originates form a pair of tights which the cyclist is wearing and is constructed by finding image features in different camera views. A correlation method is used for finding these features. During the design of the model, 'good' feature areas are stored. These textured areas are sought in the images by a correlation method. In order to reduce the search-space a Kalman filter is used to predict where the 'good' feature points will be in the next frames. When several feature points have been matched in two different camera images, a Levenberg-Marquardt algorithm is used to search for an optimal solution to the problem of finding the 3D pose of the leg. Finally they test the system and conclude that they obtain satisfactory results at video rate.
Comments: The work is nice but requires a lot of off-line processing before you can use the system. The part about how to go from the matching result and to the 3D pose localization is not clear to me.

Title: A Region Based Approach for Human Body Motion Analysis [64]
Author(s): M.K. Leung and Y.H. Yang
Location: Department of Computational Science, University of Saskatchewan, Canada
Year: 1987
Published: Pattern Recognition. Vol. 20, No 3, pp. 321-339
Type: Paper
Key words: Human body labeling, antiparallel, apars and human model

Summary: This paper is about human body labeling. The human body analysis problem is divided into three steps: the segmentation process, the labeling process and the motion type identification. The first step is discussed in [65] while this paper deals with the second step. The work is a bottom-up approach where the starting point is the edges of the human, segmented by the method described in [65], which are processed until they are represented by a few straight and antiparallel line pairs. Then a simple 2D human model is introduced consisting of six apars (a special case of ribbons). The model is used to label the different parts (line pairs) of the image. The low level edge data are processed by a line approximation algorithm which outputs orientated line segments. Two lines being antiparallel are expressed as an apar. Using different criteria irrelevant apars are eliminated and the remaining ones are concatenated into a few large apars. Now the apar model comes into play. Its six parts have the following width ratios 1:2:2:4 for the arms, legs, head and trunk. These ratios are compared with the ratios of the remaining apars in order to label the data. A likelihood of each found apar being one of the body parts is also included into the labeling process. These two cues are combined in an global iterative maximization process to do a correct labeling. The system is tested on three images sequence and has a success rate of approximately 60%.
Comments: I like the segmentation method but I'm not sure the approach is very robust in the labeling phase.

Title: Human Body Motion Segmentation in a Complex Scene [65]
Author(s): M.K. Leung and Y.H. Yang
Location: Department of Computational Science, University of Saskatchewan, Canada
Year: 1987
Published: Pattern Recognition. Vol. 20, No 1, pp. 55-64
Type: Paper
Key words: Human body segmentation, image subtraction, histogram and voting
Summary: This paper is about segmenting the human body in a complex scene. The idea is to use image subtraction in combination with edges to segment movements, and thereby the user, in an image sequence. Two images are subtracted and the values are plotted in a histogram. This reveals the static objects and the dynamic objects, and an adaptive threshold can be found. The edges in the difference image are found and compared with the edges in the current image. Votes are assigned to the different edges according to the information they carry. By doing a simple counting of votes the edges which corresponds to movements can be found and the none-moving edges eliminated. The scheme is tested on a sequence involving a human in motion. The result is that most of the human is segmented together with a few background regions.

Comments: Interesting idea to combine the edges from the difference image with the ones from the current image.

Title: First Sight: A Human Body Outline Labeling System [66]
Author(s): M.K. Leung and Y.H. Yang
Location: Department of Computational Science, University of Saskatchewan, Canada
Year: 1995
Published: Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, No. 4
Type: Paper
Key words: Human body labeling, outline subtraction, coincidence edges, 2D ribbons, model patterns, constraints and gymnastic image
Summary: This work describes a system for labeling the human body. The system consists of four parts: subtraction of the outline of the moving human body, description of the regions in the outline image, tracking of the regions over time and during an interpretation phase to generate a stick-figure of the moving human. The outline of the human body is found using coincidence edges, which are edges common to an edge image and a difference image, see [65] for a detailed description. The regions are described using 2D ribbons which are U-shaped edge segments. A 2D human model is used which consists of five U-shaped ribbons, a torso with associated spines and a number of points/joints. In the labeling/interpretation phase a number of constraints are used to prune the solution space. Structural constraints, such as length and relations between different joints, are introduced and combined with shape constraints. A number of model patterns are defined and used to interpret the data further. These are support postures (which parts of the body holds the weight of the human), side view kneeling and side horse motion. The last two models are introduced because the system is tested on gymnastic image sequences. Three image sequences are used each having large body movements and the last two include significant occlusion. The error rates are 15.5%, 15% and 21.5% respectively.
Comments: Nice work in the lines of earlier work from their lab. The outline segmentation algorithm and the support posture ideas looks interesting.

Title: Human Posture Recognition Using Multi-Scale Morphological Method and Kalman Motion Estimation [67]
Author(s): Y. Li, S. Ma and H. Lu
Location: National Lab. of Pattern Recognition, Automation Institute, Chinese Academy of Sciences
Year: 1998
Published: International Conference on Pattern Recognition
Type: Paper
Key words: Estimation of posture parameters, human model, silhouettes,

Kalman, connected tree, multiscale morphologic matching, similarity measure and metropolis algorithm
Summary: The paper is about a system for estimating the posture parameters of a moving human body. They use a monocular model based silhouette approach to solve the problem. The human model consists of 14 round-cornered rectangles which are contained in a connected tree (CT) describing the structural knowledge and constraints. Different CT's are defined for different camera view points. A Kalman filter is used to predict the position of the different body parts and a matching is carried out between the image silhouette and the model silhouette. The similarity measure which is used, is defined based on the area difference between the two silhouettes. To get a more robust solution a multiscale morphologic matching is used. 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 lot of noise and thereby improving the result of the matching. After the main parts have been found the matching process moves down in scale and find the small parts. A metropolis algorithm is used to find a solution to the matching problem. Finally a small test is made which shows input images, silhouette images and the identified model-pose.
Comments: The paper is too short - only 3 pages - and not many details are presented. It is unclear how the CT tree is designed and how many views are used. Also it is not evident how they obtain the silhouettes and how they define the constraints and the length and width of the model parts.

Title: Log-Tracker: An Attribute-Based Approach to Tracking Human Body Motion [68]
Author(s): W. Long and Y.H. Yang
Location: Department of Computational Science, University of Saskatchewan, Canada
Year: 1991
Published: International Journal of Pattern Recognition and Artificial Intelligence. Vol. 5, No. 3, pp. 439-458
Type: Paper
Key words: Human body tracking, human model, labeling, logs and attributes
Summary: This work is about tracking the human body in motion. The idea is to track "logs" from frame to frame. A log is an area defined by two parallel lines and a number of attributes. The methods described in [65] and [64] are used to segment the edges of the moving human and find the edges which are parallel. In each area surrounded by the parallel lines the following attributes are found. Log-length which is the average length of the two sides of the log, log-location which is the center of the log, log-area which is the area of the log, log-color which is the average intensity of the non-background pixels in the log and log-orientation which is the

orientation of the log. The logs are tracked in the image sequence using a normal split/merge, called Forks, criterion. During a period of 10 frames the different logs are linked together to larger units/logs, which move in a similar manner. The five best linked structures are chosen as body parts and labeled, using a simple human model, as the head, arms and legs. The system runs at 60Hz/half frame at a resolution of 128 times 128. The system is tested on two sequences where some problems are present, especially due to occlusion.

Title: An Automatic Rotoscopy System for Human Motion based on a Biomechanic Graphical Model [69]
Author(s): Y. Luo*, F. J. Perales* and J.J. Villanueva**
Location: *=Department of Mathematics and Computer Science, University of Balearic Islands, Spain. **=Department of Computer Science, Autonomic University of Barcelona, Spain
Year: 1992
Published: Computers & Graphics, Vol. 16, No. 4
Type: Paper
Key words: Model-based, analysis-by-synthesis, stick-figure model, match criteria and data visualization
Summary: This work is about human motion capture using a model-based approach. A classic analysis-by-synthesis work. They use two cameras. A person is segmented from the background in both images using uniform white background. A human stick-figure model is projected and matched against the segmented body. This is done for both images. The following five match criteria are used. 1) Both projected stick-figure models must be within the segmented human body. 2) The distance from the projected stick model and to either of the two edges of the segmented body part should be the same. 3) The position of a joint should be physical possible with respect to a real human. 4) The position of a joint can not move more then a certain distance between two frames. 5) Only one joint can be found at each 3D position. The projection which best fits these criteria is the correct match. The sequence of estimated pose parameters is smoothed and interpolated using B-splines. They test the system on a jumping and a walking human. They have developed a tool which can be used to visualize the data and superimpose them on top of the input image sequence.
Comments: They don't state the processing time, but I have a feeling its off-line. Some interesting references to MLD and kinematics modeling.

Title: Model Based Extraction of Articulated Objects in Image Sequences [70]
Author(s): D. Meyer, J. Denzler and H. Niemann
Location: Universitaet Erlangen-Nurnberg, Germany
Year: 1997

Published: Fourth int. conf. on Image Processing
Type: paper
Key words: motion tracking, gait analysis, displacement vectors (optical flow), monotony operators, contour, active rays, analysis-by-synthesis, human model and Kalman filter
Summary: The work is about motion tracking for gait analysis. The method goes as follows. The two first images in the sequence is used to initiate the system by finding a seed point. This is done by calculating the displacement vectors, using optical flow, between the two images. The method used is called monotony operators. The vectors are grouped by direction and the different body parts are found based on these groups. Next the contour of the different parts are segmented by a method called active rays using a seed point from the initialization phase. The contour segmentation is similar to active contours, except that the problem is reduced from 2D to 1D. A 3D model of the human is used in an analysis-by-synthesis scheme. The model consists of 6 boxes, head, torso, arms and legs. The best match between the contour and a projected version of the model is found and the parameters of the human can be obtained. This information is also used to predict a seed point for the contour search in the next image. Finally a Kalman filter is used to smooth the estimated parameters. Test show that the system can track the head, torso and legs of a walking person. The segmentation can be carried out in real time, but the parameter estimation takes 60 seconds for on image!
Comments: Neither the analysis-by-synthesis and the contour tracking methods are described very well, but instead references are used. Also it is not clear how the model looks like and how the different parts actually are found in the initialization process and in the tracking mode. I don't expect the approach to be very precise, mostly due to the very coarse modeling.

Title: Reality Modeling and Visualization from Multiple Video Sequences [72]
Author(s): S. Moezzi, A. Katkere, D.Y. Kuramura and R. Jain
Location: University of California, San Diego, USA
Year: 1996
Published: Computer Graphics and Applications
Type: Paper
Key words: Visualization of real video sequences, Immersive Video, background subtraction, 3D model, hypermosaicing, multiple projective texture mapping, and model tessellation and coloring
Summary: This work deals with visualization of real video sequences from an arbitrary viewpoint, which is called Immersive Video. The idea is to have a 3D model of an environment and several calibrated cameras which are observing it. Then, by background subtraction, dynamic objects can be segmented by the cameras and a 3D model of the entire (static and dynamic)

environment is obtained. The idea is to visualize this 3D model from any viewpoints by moving a virtual camera around in the constructed 3D model. Three different methods for constructing the virtual view are presented: hypermosaicing, multiple projective texture mapping, and model tessellation and coloring. The system is implemented and tested in different settings. The result looks good.
Comments: Nice work on immersive video.

Title: A Real Time Anatomical Converter for Human Motion Capture [73]
Author(s): T. Molet, R. Boulic and D. Thalmann
Location: LIG - Computer Graphics Lab, Swiss Federal Institute of Technology, Lausanne, Switzerland
Year: 1996
Published: EUROGRAPHICS Int. Workshop on Computer Animation and Simulation, Poitier, France
Type: Paper
Key words: Motion capture and Flock of Birds
Summary: This work is about how to use 'A flock of Birds' sensor to capture human motion. They place one sensor at each limb of the user. A skeleton of a virtual human is build and fitted to the input data. First the virtual skeleton is calibrated by scaling an average human model to the performer. The height is directly measured by the head sensor while the other limb parameters are measured by hand. Next the sensors are calibrated by finding the relationship between the sensor attached to a performer's limb and the corresponding proximal joint of the virtual model. Only one sensor is calibrated with respect to position, the spine sensor, while the rest only uses rotation information. The performer stands in a predefined posture and the transformations are found. The system runs at 8Hz with twelve active sensors on a Onyx processor. It is demonstrated with some examples: soccer and tennis.
Comments: Includes references on this kind of motion capture.

Title: A Model Driven 3D Image Interpretation System Applied to Person Detection in Video Images [74]
Author(s): O. Munkelt, C. Ridder, D. Hansel and W. Hafner
Location: FORWISS, Munchen, Germany
Year: 1998
Published: International Conference on Pattern Recognition
Type: Paper
Key words: Human pose detection, stereo, graph matching and human model
Summary: They present a system for detecting the pose of a human. Markers are placed on the joints of the human and these are found in 3D using stereo. By using a graph-based method these markers are matched

against a 3D model of the human and the pose is found.
Comments: Too short to get a good feeling of their work. They mention something about using colors to segment the head, but I don't see where this comes in.

Title: Human Tracking Using Distributed Video Systems [75]
Author(s): A. Nakazawa, H. Kato and S. Inokuchi
Location: Department of Systems and Human Science, Osaka University
Year: 1998
Published: International Conference on Pattern Recognition
Type: Paper
Key words: Human tracking, distributed system, background subtraction and human ellipse model
Summary: This work is about a distributed system for tracking a human. The idea is to place different cameras in different locations in a building. Each camera is connected to a vision system which tracks the human when ever it can and sends out the position of the human to the other vision systems via a network. Since only positions are send the network will not be a bottleneck. The different vision systems know where it is and where the other systems are. Therefore it can 'track' the human even though it can't see him. When he is about to enter its workspace it can start tracking in this region of the image. When it knows that he will not enter the workspace right now it goes into idle mode, meaning that it can allocate its resources to other tasks. The tracking of the human goes like this. The input image is subtracted from a background image and binarized. The human is modeled by an ellipse. Using its position in the current frame different positions of the human is predicted. The predicted ellipses are matched against the binary image and the location with the largest overlap gives the position of the human. The system is tested in their lab using three machines and vision systems. It runs successfully at 15Hz.
Comments: Nice and simple. I like the idea of only calculating when necessary. I don't think they change/use the size and rotation of the ellipse which could give problems due to the 3D position of the human with respect to the camera. But in one figure the ellipse is rotated! They mention something about expanding regions, which I don't understand. Perhaps its something to do with 'old' motion. Meaning motion from the previous image. Perhaps they should do real correlation instead of their current matching procedure.

Title: Constructing Virtual Worlds Using Dense Stereo [76]
Author(s): P.J. Narayanan*, P.W. Rander** and T. Kanade**
Location: *=Center for Artificial Intelligence & Robotics, Bangalore, India. **=Robotics Institute, CMU, USA
Year: 1998

Published: International Conference on Computer Vision
Type: Paper
Key words: Virtualized reality, 51 cameras, multibaseline stereo, volumetric model, texture mapped and Z-keying
Summary: The paper describes a system to synthesis objects into a virtual world. They call the process virtualized reality which differ from virtual reality by being based on rendering of real images instead of virtual objects. The process is based on real images obtained from 51 cameras placed in a geodesic dome, named the 3D Dome. From all these cameras intensity images are, together with depth maps from multibaseline stereo, obtained. This information is combined to form a visible surface model (VSM), which can be visualized from different viewpoints. A virtual camera defines which viewpoint should be synthesized and then the nearest real image (depth map and intensity image) is used. Holes can appear in the synthesized image due to occlusion and therefore two neighbour images are used to fill out these holes. Besides the VSM representation they also build a complete surface model (CSM), which is a volumetric model represented by texture mapped triangles which can be handled by standard rendering tools, e.g. Open Inventor. So far, in the paper, they have only dealt with static objects. In the last part they work on a dynamic scene, 11 frames of a baseball stroke, where also a virtual object, a ball, is introduced into the scene. They tread this sequence as a sequence of static frames and obtain results with some blurring around the person as a result of different intensity levels in the different cameras. Also some motion problems are observed. Finally they talk about separating the object from the background so it can be inserted into another stage/background. They suggest to use 3D information, Z-keying.
Comments: A well written paper where they seem to be very open and honest about their methods and results. It should be noticed that the video signals from the 51 cameras are digitized off-line making the system incapable of real time performance.

Title: Analyzing and Recognizing Walking Figures in XYT [77]
Author(s): S.A. Niyogi and E.H. Adelson
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1994
Published: Conference on Computer Vision and Pattern Recognition
Type: Paper
Key words: Individual gait recognition, change detection, snakes and stick-figure model
Summary: They work with individual gait recognition in the following way. They have a video sequence of one person walking frontoparallel relative to the camera. A change detection algorithm is applied to detect moving objects. Each object is analyzed in the lower XT-plans using a correlation mask searching for braided walking patterns (two twisted lines).

Next, several snakes are used to find the contour of the human body, using the correlation result as an initial guess. A skeleton of the human is found and after another snake operation and a line fitting a simple stick model of the human is obtained. It consists of four angles: between the torso and left and right leg, and between each upper and lower leg. A recognition system is build using a k-nearest neighbours classifier. The result on 24 different image sequences of five persons was a recognition rate of 79%.
Comments: Nice idea to use XT slices. Too few tests to be able to conclude anything. The approach is very sensible to occlusion.

Title: Figure-Ground Segmentation Using Multiple Cues [78]
Author(s): P. Nordlund
Location: CVAP, Kungl Tekniska Hogskolan, Sweden
Year: 1998
Published:
Type: Ph.D.-thesis
Key words: Figure-ground segmentation, tracking and multiple cues
Summary: The thesis is divided into two main parts: a general description of his approach together with a state of the art section, and six papers where all the details are. This review only concerns the first part. He deals with the figure-ground segmentation problem by using multiple cues and applying a system's approach. The former is to make it more human like and thereby more robust, while the latter is in order to obtain real time performance. His main approach is the following. For each image calculate the disparity map and a horizontal flow map. Make a 3D histogram of these data having the depth at one axis and flow at the other (see page 10 and 11 in the thesis for good illustrations). In the histogram some peaks will appear: one for the background, with zero motion, and one for each person having either different motion direction or different depth. By taking the pixels which correspond to the peaks and back projecting them into the original image, blobs representing the different persons will appear (see page 12 in the thesis). Beside the main idea he also investigate the role of attention, the importance of 3D cues, cue integration and reliability measures. He talks about other features than motion and depth which he have worked with: corners, color and texture. He, however, doesn't explain how he used them and instead refers to the papers.

Title: Pedestrian Detection Using Wavelet Templates [79]
Author(s): M. Oren, C. Papageorigiou, P. Sinha, E. Osuna and T. Poggio
Location: CBCL and AI Lab, MIT, USA
Year: 1997
Published: Conference on Computer Vision and Pattern Recognition
Type: Paper
Key words: Detection pedestrians, wavelets, Haar wavelet, bootstrapping,

template matching and structural minimization risk
Summary: The work is about detection pedestrians in images using wavelets. They start out by the Haar wavelet-representation and then adjust it to get a more dense representation, i.e. a better spatial resolution. The coefficients of the Haar wavelet-representation yields a template. Then they take a lot of images containing a pedestrian and transform each image to the dimension of 128x64 such that the people are centered and approximately the same size. An average template is computed. Using multiple scaling and shifting of the template, pedestrians of different size and location can be detected in new images. A bootstrapping technique is used to train the system and after that two different classification methods are used. First a normal template matching and secondly a support vector machine (SVM) which uses structural minimization risk. Both methods are tried on approximately 1100 examples and the results are 52.7% and 69.7%. Since they use difficult outdoor scenes they think the detection rate is high.
Comments: Good paper with very little math. The wavelet stuff is not clear to me except that a wavelet method is like Fourier or PCA where you use a new basis set to represent your image data.

Title: Model-Based Image Analysis of Human Motion Using Constraint Propagation [80]
Author(s): J. O'Rourke and N.I. Badler
Location: Department of Computer Science, University of Pennsylvania, USA
Year: 1980
Published: Transactions on Pattern Analysis and Machine Intelligence, Vol. 2, No. 6
Type: Paper
Key words: Human motion tracking, model-based, constraints, propagation and prediction
Summary: This work is about model-based human motion analysis in a predict-match-update scheme. The idea is to use constraints which are propagated through the system. They have five modules in their system. An Image analysis module which, based on input from the Simulation module, extract features from the image. This information is send to the Parser module which apply a temporal-filter to these positions. Next a Predictor module predicts the position of the parameters in the next frame. These results are send to the Simulation module which place a human model, consisting of 25 segments and 600 overlapping spheres, in the nearest legal position with respect to the predicted parameters. The output from the Simulation module is regions where the Image module should look for features: feet, hands and head. Distance constraints are used to propagate through the system. Other constraints are figure/ground distinction, occlusion, collision detection and joint angle limitation. The system is tested on

synthetic images where the hands, feet and head are easy to segment due to different gray-level values. The system is tested on three different image sequences.
Comments: Very novel pioneering work in this field. Even though the input is simple and the framerate is slow (5Hz), it is still interesting. I like the idea of using several different constraints which together reduces the search space. The propagation theory is very hard to follow. They 'just' use their human model to predict and not in the matching process.

Title: Interactive Video Environments and Wearable Computers [81]
Author(s): A. Pentland
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1995
Published: Workshop on Automatic Face-and Gesture-Recognition
Type: Paper (Invited speaker)
Key words: Overview at MIT, looking at people domain, interactive video environment and wearable computers
Summary: The paper is about the current activity at the media lab at MIT. It is an overview explaining what is going on at MIT, which MIT-work will be presented at the workshop and finally a view into the future. He states two directions at MIT within the looking at people domain. The interactive video environment and wearable computers. The first is interactive spaces where people, using cameras and microphones, can control different stuff. The latter aims at understanding the user's situation and help him work using wearable cameras, microphones and wireless technology. He talks about the goals of the media lab and the-boy-in-the-closet-story. In the future the focus will be on attempting to understand what a person is thinking, doing, and intending. The general approach for doing so, is to model a human as a number of internal mental states.
Comments: This is the first time I hear about the looking-at-people paradigm.

Title: Machine Understanding of Human Action [82]
Author(s): A. Pentland
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1995
Published: Int'l Forum on Frontier of Telecommunication Technology, Tokyo, Japan
Type: Paper (Overview)
Key words: Overview paper, IVE spaces, Pfinder, face recognition, facial expression, vision-driven audio and action recognition
Summary: This is an overview paper from MIT's media lab. It starts out with "the boy in the closet" story and describe the different research directions that they are currently following. Focus is put on the IVE spaces

where cameras and microphones are used by a computer to observe the user and immerse the user into a virtual world/environment. The other elements of the paper are the Pfinder, face recognition, facial expression recognition, vision-driven audio and action recognition.
Comments: A good overview paper.

Title: Smart Rooms, Smart Clothes [83]
Author(s): A. Pentland
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1998
Published: International Conference on Pattern Recognition
Type: Paper (Overview)
Key words: Overview, smart rooms, smart clothes, perceptual intelligence
Summary: This paper gives an overview of the work done by Pentland's students. He defines two areas of interest (like in [81]): smart rooms and smart clothes. The latter is a general version of the idea of wearable computers presented in [81]. The idea is to give clothes perceptual intelligence. The former contains work on: person tracking, gesture recognition, face tracking, face recognition, face expression recognition, audio interpretation and voice recognition.
Comments: A very good overview of the work in his group.

Title: Using Computer Vision to Control a Reactive Computer Graphics Character in a Theater Play [84]
Author(s): C. Pinhanez and A. Bobick
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1998
Published: International Conference on Vision Systems
Type: Paper
Key words: Two-character theater, 'It'/'I', autonomous computerized character, script, silhouette images, stereo setup, depth subtraction and recognition
Summary: This work is about a two-character theater play called 'It'/'I'. 'I' is a human performer whose actions are recognized by a computer vision system. This, together with a script, controls a autonomous computerized character, called 'It'. 'It' controls sound and stage lights, and "appears" using computer graphics. The computer vision system is based on silhouette images which are obtained using a (three camera) stereo setup. Instead of doing image subtraction depth subtraction is used. The idea is to have a depth map of the static background. This is then used to segment pixels as belonging to the background or to a new object. The good thing about this approach is that it is insensitive to change in the illumination. The computer vision system is used to track the performer and then to recognize his actions. This recognition is done with respect to the current scene in the

play. Hereby assigning different meaning to the different actions depending on the context. The recognition is done using temporal templates [29]. The scripting, known as interval scripting, is based on a paradigm suggested by James F. Allen. The system has been tested six times for a total of 500 people in sections of 30 minutes.

Title: Low Level Recognition of Human Motion [85]
Author(s): R. Polana and R. Nelson
Location: University of Rochester
Year: 1994
Published: Workshop on Motion of Non-Rigid and Articulated Objects
Type: Paper
Key words: Appearance-based, activity recognition and periodic motion
Summary: Describes a system for low level recognition of periodic human actions. They use a bottom up approach where they don't need to find the separate body parts, but instead they look at the entire low level information, body(motion). First they use prediction to find the motion pixels corresponding to one actor in each frame, normalize for scale and then crop out a box containing the actor. Next the background (which will be moving with the same velocity as the object, but in the opposite direction) can be found by computing the flow field between successive frames and it is eliminated (motion=0). Each image now contains a normalized and centered version of the actor performing the activity. Motion images are calculated, using subtraction, based on these images and each motion image is divided into 4x4 features where each feature holds a number corresponding to the amount of motion within this image area. A new feature vector is generated by concatenating six of these motion images, yielding 4x4x6=96 elements per feature vector. The idea is that this feature vector should correspond to one or close to one cycle of the activity. A classifier is build/trained on six different classes, like walking and running, and a 100% correct classification is reported! The stuff about the classifier and the frequency/periodicity detection is described in another paper. They conclude that the system is robust to varying illumination and contrast because they only use motion which is invariant to these.
Comments: Good work using only low level motion information. Also a neat way to get rid of the background. However, the paper lacks a great deal of implementation details which makes it hard for me to judge how sensitive it is etc.

Title: Human Movement Analysis Based on Explicit Motion Model [86]
Author(s): K. Rohr
Location: Arbeitsbereich Kognitive Systeme, Fachbereich Informatik, Universitat Hamburg
Year: 1997

Published: Chapter 8 of the book "Motion-based Recognition" edited by M. Shah and J. Rain
Type: Book chapter
Key words: Motion models, human cylinder model, recognition cyclic movements, walking, kinematic model, change detection algorithm, analysis-by-synthesis, Kalman filter and line matching
Summary: In this text Rohr describes his work on human movement recognition using motion models and an explicit human model. He sets up a framework for recognition cyclic movements which can be described by motion models. First he has a short, but very informative, state of the art section with a taxonomy. He then describes his framework and finally some experimental results. The type of movement which is used to explain the framework is walking, with the assumptions of movements parallel to the image plan at constant velocity. His approach has two steps: an initialization phase and a processing phase. The model of the human, which is inspired by Hogg, is a volume model consisting of 14 cylinders - head, torso and three for each arm and leg - connected by joins. The motion model is based on a kinematic model, as oppose to dynamic models. The model is a set of curves of the joints at the shoulder, elbow, hip, knee and the vertical displacement of the whole body. Each of these curves are periodic and obtained from a medical study into human motion. The initialization phase is described in the following. First a change detection algorithm is applied to the image in order to segment a moving person. The motion blob is represented by a boundary box. From the hight of this box, a camera calibration and an assumption of the persons height, the distance to the person from the camera can be estimated. Next an edge detection and later a line detection is performed on the image inside the boundary box. This is matched up against a contour representation of the human model, making it an analysis-by-synthesis approach, in the following way. For each visible model contour (straight line) a search window is computed. Within this search window the image lines are match against the model line and a similarity measure is calculated. This is done for all model contour lines of one pose and a total similarity measure is calculated. This is again done for all poses and the pose with the highest similarity measure determines the correct pose of the person. This procedure (all of the above) is done on 10-15 frames in order to get a stable amount of values for the next phase. In the processing phase a Kalman filter, initiated in the initialization phase, is applied to estimate the true pose and predict the pose in the next frame and thereby reducing the searchspace. He tests the system and obtains good results both for walking and for cycling. For future work he mentions the following. Deal with the fact that the underlying motion model is based on an average person instead of the test person. Incorporate nonconstant velocity. Compare motion velocity field with image velocity fields. Using an aspect graph for matching.

Comments: It is not clear how the measurements of the pose is obtained in the processing phase. It could either be as in the initialization phase but only for a small interval of the pose parameter(s) or he could use the best matched image line within the predicted search space. I strongly believe it is the former since he writes that he fits the model as a whole instead of as individual segments (page 14). I would argue that he only uses a $2\frac{1}{2}$ model since he has the assumption that the person is walking parallel to the image plan. He describes the Kalman filter very clearly, which is hard to do using only a few pages.

Title: Tracking and Counting Moving People [87]
Author(s): M. Rossi and A. Bozzoli
Location: IRST, Trento, Italy
Year: 1994
Published:
Type: IRST Technical Report #9404-03
Key words: Tracking and counting people moving, temporal change detection, histograms, correlation method and agglomerative algorithm
Summary: This work is about tracking and counting people moving e.g. through an exit. The idea is to divide the camera's field of view into three areas: an alerting area, a tracking area and a counting area. In the first area moving objects are detected using temporal change detection and histograms. Different templates are generated for each human, i.e. more that one template is associated to one human. In the tracking phase a correlation method is used to track the templates. The method is called three-step search and the idea is to do the correlation at nine points. Eight new correlations are carried out around the one of the nine which yields the highest correlation value and so on. In the counting phase the different templates from one human are grouped together based on a hierarchical procedure using an agglomerative algorithm. Whenever a cluster of templates belonging to one human cross the counting line the counting number is increased by one and the templates discharged. The system is tried out in a railway station environment and about 90% is correctly counted.
Comments: Nice way of doing the correlation but it must be very sensitive. I don't quit follow their explanation of the alerting phase. They don't mention how the system works when movements are carried out in both directions.

Title: Tracking of Human Motion [88]
Author(s): M. Schneider and M. Bekker
Location: LIFIA, Grenoble, France and LIA, AAU, Denmark
Year: 1994
Published:
Type: Master Thesis

Key words: Tracking of human motion, motion detection, image subtraction, correlation, Kalman filter and Lattice filter
Summary: The work deals with tracking of human motion in a lab environment. The tracking system consists of two modules, a motion detection module and a tracking module. The former is implemented as image subtraction and the latter using a correlation mask. Three different techniques are implemented and tested, sum of absolute differences, normalized cross correlation and the zero mean normalized cross correlation. No significant differences were found between the three techniques. Two prediction filters are implemented, the Kalman filter and the Lattice filter. The latter was found to converge faster, but also to be more sensitive to outliers.
Comments: Solid work. Even though they don't have a fancy system they are very thorough in their description of the different methods. Especially about the filters/predictors and the correlation methods. They also have a nice section about human mobility.

Title: Segmentation of People in Motion [89]
Author(s): A. Shio* and J. Sklansky**
Location: *=NTT Human Interface Laboratories, Nippon Telegraph and Telephone Corp., Yokosuka, Japan **=Department of Electrical Engineering, University of California, Irvine, USA
Year: 1991
Published: Workshop on Visual Motion, p. 325-332
Type: Paper
Key words: Segmenting people, background subtraction, quasi cross correlation and merge/split
Summary: This work is about segmenting people in motion. The idea is first to estimate the motion within an image and then to segment all people in the image. The motion estimation is done like this. The background is subtracted and the moving objects segmented. A motion field is found using quasi cross correlation. This mostly yields movements around the edges. A propagation technique is used to expand the motion into the different objects. Next the motion fields are temporal smoothed, which average out the movements originating from the different body parts. In the people-detection-phase each connected motion region is divided into subregions. These subregions are merged to form the regions of a person based on a blob model of a person in motion. The system is tested on an outdoor scene monitoring a walking pedestrian. The resolution is 256x240 at 15Hz and the results are almost correct.

Title: Virtual Stage: A Location-Based Karaoke System [90]
Author(s): C. Sul, K. Lee and K. Wohn
Location: Computer Science Department, Korea Advanced Institute of Science and Technology (KAIST), Korea

Year: 1998
Published: Multimedia, Vol 5, no 2
Type: Paper
Key words: Interactive virtual environment, scripting and chromakeying
Summary: This work is about a interactive virtual environment. They have build up a interactive Karaoke system. The idea is to have a script which consists of a song, a lyric and a virtual band. The script is processed and the image of the user is immersived into the virtual world. A tool for constructing the script has been developed making it much easier. Different actions can be associated with different virtual objects. E.g. the camera, which shows the audience the scene, can be panned, moved etc. Another action could be for one of the agents, band members and dancers, to follow or avoid the avatar of the user. The user is segmented from the background using chromakeying (using a blue background like in the weather forecast) and then represented as a bitmap which can be placed in VR. The posture of the user is recognized by comparing it to different template e.g. bowing and the VR reacts to these gestures. E.g. the song/script is not started before the user bows. The band members mostly perform key-animations which change according to cues in the music stated in the script. The dancer will try to copy the posture of the user and react to collision events. The system has been successfully demonstrated to the public and a prototype has been set up in a theme-park. In future versions a more sophisticated method for participant sensing will be implemented.

Title: Human Body Modeling for People Localization and Tracking from Real Image Sequences [91]
Author(s): A. Tesei, G.L. Foresti and C.S. Regazzoni
Location: University of Genoa, Italy
Year: 1995
Published: Image Processing and Its Applications
Type: Paper
Key words: Tracking a moving human, statistical morphological operators, heuristic algorithm, simple human model and Kalman filter
Summary: This work is about tracking a moving human. First the human is segmented from the background using statistical morphological operators. Next a minimum rectangle which bounds the human is found. Then a set of five points - head, neck, torso and feet - are found using a heuristic algorithm, which relay on a simple human model, or rather the different relations between the different body parts. The parameters are tracked and filtered using a Kalman filter. They are converted into 3D coordinates using the fact that the human is walking on a plane and the assumption that the person is walking upright. The system is tested and the result is that 88% of the points are tracked correct.
Comments: Extremly short paper.

Title: Detection of the Movements of Persons from a sparse sequence of TV Images [92]
Author(s): T. Tsukiyama and Y. Shirai
Location: Electrotechnical Laboratory, Ibaraki, Japan
Year: 1985
Published: Pattern Recognition Vol. 18. Nos 3/4. pp. 207-213
Type: Paper
Key words: Detecting people, tracking and feature points
Summary: This work is about detection of moving people in a hallway. The idea is to have a camera observe a hallway and track the people who walk in it. The method goes like this. First segment moving objects, then detect persons and finally track people over time. For the segmentation part the following is done. For each pixel the mean and variance is calculated using the eight nearest neighbour pixels. Each pixel is then classified belonging to either the floor (white) or to an object using thresholds. Next the object pixels are grouped into regions and mapped to a region map where the floor-plane is transformed to the X-Y-plane. In this region map each region blob is characterized by several feature points, e.g. the toe and the width. Based on these points the people are found. If the object blob is too small it is removed. If its too large it is split into a new blob. This continues until only one person, per blob, is found. By using a distance measure the correspondence between different people is found. They test the system at 1.5Hz on special hardware.

Title: Visual Interaction with Lifelike Characters [93]
Author(s): M. Turk
Location: Vision Technology Group, Microsoft Research, Redmond, WA, USA
Year: 1996
Published: International Conference on Face and Gestures, Killington, VT, USA
Type: Paper
Key words: Multi modalities, agent, vision modules, Peedy the Parrot, segmentation, draping, silhouette, eigeneyes and gesture recognition
Summary: This paper describe the vision part of a large interface project at Microsoft. The project is called Persona and deals with multi modalities (image and sound) in an agent-based system. The idea is to develop an alternative to the desktop environment. The user can speak to an intelligent agent, Peedy the Parrot, who speaks back. To make the interface more useful different vision modules are implemented: user segmentation by color and motion for both background and foreground, draping which is a representation of the head and shoulder silhouette, head tracking by histograms or eigenfaces, head counting by 'humps' detection in the head and shoulder silhouette, moving lips, pose recognition using the head and

shoulder silhouette, gesture recognition using the recognized poses over time, person identification using color moments and gaze determination using 'eigeneyes'. The idea is only to use the different vision module when necessary.
Comments: I like the idea about having different vision modules which are only invoked when necessary. The draping idea also seems very interesting.

Title: Multiple-View-Based Tracking of Multiple Humans [94]
Author(s): A. Utsumi*, H. Mori**, J. Ohya* and M. Yachida**
Location: *=ATR Media Integration & Communications Research Laboratories, Kyoto, Japan **=Department of Systems and Human Science, Graduate School of Engineering Science, Osaka University
Year: 1998
Published: International Conference on Pattern Recognition
Type: Paper
Key words: Tracking multiple humans, multiple cameras, center of gravity, distance transformation and elliptic pillar model
Summary: This work is about tracking multiple humans using multiple cameras. Three steps are carried out in their algorithm. First the position of the humans are found. Then their normal axes are found and finally it is determined how each person is oriented by choosing between the two results per human which are possible according to the normal axis. When calculating any of the three thing the cameras with the best view of the scene are used, together with the previous obtained results. Before finding the positions of the humans, they are each represented by one point, the center of gravity (cog). Cog is found by calculating a distance transformation (DT) to each object segmented from the background. Each cog is associated with an uncertainty. Combining different cogs and their uncertainty from different camera views (only if they can observe the human) yields a good estimate of the cog. Next the normal axis of each human is found using the assumption that a human can be modeled as an elliptic pillar with constant dimensions. Based on this result it is calculated which camera view can get the best view of the face of the human. The head of the human is then found in this view and it is calculated whether the human is facing the camera, light skin pixels, or has his back to the camera, dark hair pixels.
Comments: Good idea to base a calculation on the result of the previous one and to use the camera which will give the best result.

Title: Tracking of Persons in Monocular Image Sequences [95]
Author(s): S. Wachter and H.H. Nagel
Location: Fraunhofer-Institut fur Informations- und Datenverarbeitung (IITB), Karlsruhe, Germany
Year: 1997
Published: Workshop on Motion of Non-Rigid and Articulated Objects, Puerto Rico

Type: Paper
Key words: Human motion capture, human modeling, Right-elliptical cones, edges, regions, synthesis-by-analysis approach, maximum-a-posteriori (MAP), Newton-Raphson, extended Kalman filter and walking person
Summary: This work is about a model based synthesis-by-analysis approach to human motion capture. Right-elliptical cones are used to model the human in 3D. The edges, and their orientation, of the human are found in each image and also in each projected model image. Only edges which are located in "reliable" regions, and not in potential overlapping areas, are used. They are compared using a maximum-a-posteriori (MAP) method, which are minimized using the Newton-Raphson method. To improve the result the MAP is extended to include the current and predicted state of the system found by an extended Kalman filter. The MAP is also extended to include regional information in a kind of template matching manner, where the template consists of the grey level values measured at the projected surface of the 3D person model. The system is tested on a walking person both in a simple indoor environment and in an outdoor environment. The tracking went well except for the left arm which is not being tracked correctly. The system uses 5-10s for each half frame on a SUN Ultra-Sparc-1 with 167MHz.
Comments: Very interesting system in the spirit of Hogg. There is a nice review in the beginning of the paper. Good idea to combine edge information with regional information, but it is way too slow.

Title: Analysis of Human Motion: A Model-Based Approach [96]
Author(s): *J. Wang, **G. Lorette and *P. Bouthemy
Location: *=Universitaire de Beaulieu. **=Universite de Rennes. France
Year: 1991
Published: Scandinavian Conference on Image Analysis
Type: Paper
Key words: Model-based, cycling tracking, 'optical flow' and criteria function
Summary: This work is about tracking one leg of a person cycling ('on the spot'). They define a cylinder model of the leg and present a set of equations which can be used to define the angular velocities for the two body parts with respect to each other and the image coordinate system. They also define equations which maps from 3D model velocity to velocity in the image plane. Using the motion constraint equation it is possible to set up a criteria function which can be minimized and the 3D model parameters obtained. The system is initialized by hand where the different parts are identified in the image and matched against the 3D model. The tracking procedure goes like this. At time t the 3D model is projected to the image plane, called reference image, and compared to the image from time (t+1) using the criteria function. In this way the image velocity, and thereby the

3D velocity, can be obtained. To loosen up on the brightness constraint a predict-compensation procedure is used to update the reference image making it less sensitive to noise than optical flow. The system is tested and they claim to have good results.
Comments: It is not very easy to understand the method and I'm not absolutely certain that I understand it correctly. It is not mentioned how pixels representing the leg are segmented but some kind of markers can be seen in a figure showing the user so perhaps that is the way they do it. I'm not sure why they use a projected version of the model as reference image instead of the last image.

Title: Human Motion Analysis with Detection of Sub-Part Deformations [97]
Author(s): *J. Wang, **G. Lorette and *P. Bouthemy
Location: *=Universitaire de Beaulieu. **=Universite de Rennes. France
Year: 1992
Published: SPIE Vol. 1660 Biomedical Image Processing and Three-Dimensional Microscopy
Type: Paper
Key words: Model-based, cycling tracking, 'optical flow' and criteria function
Summary: Basically the same as [96].

Title: Pfinder: Real-Time Tracking of the Human Body [98]
Author(s): C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1996
Published: International Conference on Automatic Face- and Gesture-Recognition
Type: Paper
Key words: Pfinder, Gaussian distribution, color and spatial informations, blob tracking and Human body tracking
Comments: Very similar to [99].

Title: Pfinder: Real-Time Tracking of the Human Body [99]
Author(s): C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1997
Published: Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 7
Type: Paper
Key words: Pfinder, Gaussian distribution, color and spatial informations, blob tracking and Human body tracking
Summary: This paper describes the Pfinder system from MIT. It is a

system for tracking a human body in 2D. The overall idea is to model the scene as a static background and a dynamic foreground. Each pixel in the background is modeled using a Gaussian distribution and the foreground is modeled as a number of blobs each having similar color and spatial properties modeled by a Gaussian distribution. When a new image enters the system the Mahalanobis distance from this pixel's feature vector, color and spatial informations, to each blob in the previous image is calculated and a support map is generated. Each pixel in the support map is either a foreground pixel or a background pixel. The foreground pixels are "grown" until only one connected blob is present. Then the individual motion blobs are "grown" until they fill out the entire foreground blob. New statistics are calculated for each blob and predicted into the next frame using a Kalman filter. Then the system is ready for the next frame. For this to work it needs to be initialized. The model of the background is learned by having a few seconds of video where only the background is present. Both a contour method and a color method are used to create models of the human body. The hands, head and feet can be found by the contour algorithm when they are present in the silhouette. The color method are used to define new blobs which are needed in the initialization phase and when a blob reappears, e.g. after occlusion. The two methods are combined and respectively weighted according to their probability. Due to these two methods the Pfinder will always be able to recover a lost track after a few frames. The system has been used as a front end module in several applications at the Media lab.
Comments: This is a very nice system which is referred to in almost every human motion tracking paper. I would very much like to see it in action.

Title: Dynaman: A Recursive Model of Human Motion [100]
Author(s): C.R. Wren and A.P. Pentland
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1998
Published: Image and Vision Computing
Type: Paper
Key words: Tracking upper body, dynamics, kinematic model, hard constraints, Lagrangian dynamics, Spfinder, Kalman filters and Dynaman.
Summary: This work is about tracking the upper body of a human in 3D. A framework is defined for the dynamics of a kinematic model. Hard constraints are build into the framework, using Lagrangian dynamics, to handle e.g. kinematic limitations on a skeleton joint. The input to the framework is 3D positions of hands and face found by the Spfinder [7]. To handle the uncertainties in these data the framework is extended with the concept of soft constraints. This can be thought of as a force acting on the dynamics of the system. To increase the efficiency of the framework some of the dynamic aspects are replaced by Kalman filters. The 3D model of the human is predicted and projected into the two image planes, where it is used to

resolve ambiguities. The system runs at 30Hz using four computers.
Comments: An interesting paper, but it is bit tricky to read/understand the math. Its not clear how the positions of the elbows are found.

Title: Dynamic Models of Human Motion [101]
Author(s): C.R. Wren and A.P. Pentland
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1998
Published: International Conference on Automatic Face- and Gesture-Recognition
Type: Paper
Key words: Behavior models, same as for [100]
Summary: Same as [100] but with a section about behavior models. Different behavior models can be used in the pose estimation phase of the system.

Title: Perceptive Spaces for Performance and Entertainment [102]
Author(s): C.R. Wren *et al.*
Location: Vision and Modeling Group, Media Lab, MIT, USA
Year: 1998
Published: International Conference on Automatic Face- and Gesture-Recognition
Type: Paper (Overview)
Key words: Overview, Perceptual Spaces and applications
Summary: This paper gives an overview of the concept 'Perceptual Spaces' which are developed at MIT's media lab and all it's applications. First the interfaces are described and then the applications. The interface consists of the 'Interactive Virtual Environment' and the 'Instrumented Desktop' together with the Pfinder and Spfinder. The different applications or perceptive spaces are: SURVIVE (interface to 'Dome'), visually-animated characters (motion capture and kinematics), city of news (3D information navigation), virtual PAT (the aerobic trainer), KidsRoom, DanceSpace (the body motion controls the music), Improvisational Theater Space (where the audience can drive part of the play) and ALIVE (see [28]).
Comments: Good overview

Title: Parameterized Modeling and Recognition of Activities [103]
Author(s): Y. Yacoob* and M.J. Black**
Location: * = University of Maryland, ** = Xerox Palo Alto Research Center
Year: 1998
Published: International Conference on Computer Vision
Type: paper
Key words: modeling and recognition temporal activities, PCA and affine

transformation
Summary: They propose a method for modeling and recognizing human activities, e.g., walking in different directions and in different manners. Their approach is based on standard PCA combined with affine transformation to deal with time scaling and shift. They test their ideas using the method and code developed by Shanon Ju to obtain the eight parameters motion description of each human limb, see [53]. Three tests are carried out. One for recognizing walking from different views, one for recognizing four different kinds of walking and the last one for speech recognition. The recognition rates are not impressive, but they seem to be pretty satisfied!
Comments: Interesting, but the math is a bit hard to follow.

Title: A New Robust Real-time Method for Extracting Human Silhouettes from Color Images [104]
Author(s): M. Yamada*@, K. Ebihara* and J. Ohya@
Location: *=ATR Media Integration & Communications Research Lab., Soraku-gun, Japan. @=Kanazawa Institute of Technology, Nonoichi, Japan.
Year: 1998
Published: International Conference on Automatic Face- and FG Gesture-Recognition
Type: Paper
Key words: Silhouette image, color images and YIQ space
Summary: This paper describes a method for segmenting a human silhouette using color images. From the silhouette image the major joint positions are found as in [50]. The system is tested, as in [50], using a Kabuki actor. The segmentation is carried out in the following manner. The color images are transformed into YIQ space to avoid the lighting problems. Then the first several color images of a static background are averaged to find the statistic of each pixel. Each pixel in a new image is classified using the statistic as belonging to the background or to a new object.
Comments: Seems to be a nice way of segmenting a silhouette.

Title: Skill Recognition [105]
Author(s): M. Yamamoto, T. Kondo, T. Yamagiwa and K. Yamanaka
Location: Department of Information Engineering, Niigata University, Japan
Year: 1998
Published: International Conference on Automatic Face- and Gesture-Recognition
Type: Paper
Key words: Skill recognition, motion data, kinematic framework and human model
Summary: This work is about recognizing the skills of a skier. The idea is to have a couple of skiers running on in-liners and then obtain the motion

parameters of the skiers. These data are used to recognize the skiers skiing skills. The motion data are obtained using a gradient-based method within a kinematic framework, described elsewhere [106]. This method is based on a human model which is fitted to the image data in each frame (interactively in the first frame). The motion parameters of this model are then evaluated in order to determine the skiing skills of a person. Two types of information are evaluated: synchronization of both legs and smoothness of actions. The former is evaluated using a correlation method while the latter is evaluated using the lower frequencies in the Fourier spectrum. The system is tested on four persons and the skill recognition of the system is the same as when done by a human.
Comments: Nice application, but way to few test samples, only four.

Title: Human Motion Analysis Based on A Robot Arm Model [106]
Author(s): M. Yamamoto and K. Koshikawa
Location: Computer Vision Section, Electrotechnical Laboratory
Year: 1991
Published: Conference on Computer Vision and Pattern Recognition
Type: Paper
Key words: Motion tracking, optical flow, human arm modeling and Jacobian matrix
Summary: This work deals with a model based method for analyzing human body motion. The idea is to find the motion parameters of a human body part by calculating the optical flow of several points within this body parts. An equation is set up based on the idea that a small movement in the position of a point is, through the Jacobian matrix, related to a small movement in the parameters of a body part. This together with optical flow is used to set up a set of equations which are used to calculate the parameters of a body part based on a set of image points. The system needs to be initialized which is done by a geometrical modeler SOLVER. A test is shown where the right arm of a human is being tracked. The result is overlayed on the input image sequence.
Summary: Very short paper, 2 pages. Not many details. It is not clear how the optical flow is calculated. They don't use the model in a standard model-based approach. They only use it in the initialization where the relationship between the flow change and the body parameters is determined.

Title: 3D Region Graph for Reconstruction of Human Motion [107]
Author(s): C. Yaniz, J. Rocha and F. Perales
Location: University of the Baleric Islands, Spain
Year: 1998
Published: ECCV '98. Workshop on Perception of Human Motion
Type: Paper
Key words: Human motion capture, two orthogonal views, silhouette and

graph matching
Summary: A paper about a graph based method for human motion capture. They use two orthogonal views to detect silhouette images of a person wearing tight fitting clothes on a simple background. The silhouettes are analyzed and decomposed into regular regions - which are quasi-parallel lines close to each other - and singular regions (the rest). A skeleton is generated and converted into a graph where the edges correspond to the regular shapes and the nodes correspond to the singularities. They now have a 2D graph for each view. By using epipolar geometry they can guess on the location of the singularities in both views and use that, together with some rules to generate possible 3D graphs. A human model is defined to introduce some constraints which can help to determine which of the 3D graphs is the correct one. The constraints are: aspect constraints of individual parts, width constraints among parts and length constraints among parts. They test the system by letting a person following a line and measure the error between the estimated position and the 'true' position. The result is 4.27 degrees or RMS=1.6cm. In the future a Kalman filter, color and optical flow will be used to improve the system. The frame rate is 1.2 cpu seconds per frame and the user must have his appendage, body parts, visible on the silhouette images most of the time.
Comments: It is early work done by a Ph.D.-study, Yaniz, and he seems to be on the right track. This paper is more on the 'stereo'-correspondence problem than on the segmentation problem.

Title: A Model Based Approach in Extracting and Generating Human Motion [108]
Author(s): J.Y. Zheng and S. Suezaki
Location: Kyushu Institute of Technology, Iizuka, Japan
Year: 1998
Published: International Conference on Pattern Recognition
Type: Paper
Key words: Human model, personal model, key frames and analysis-by-synthesis
Summary: They present a system for extracting and generating human motion. They use a model-based approach and one camera. The idea is as follows. They have developed a tool which can be use to control the limbs of a human model. They can control the size, the texture, the rotation parameters, or more generally the pose parameters. They take a video-sequence of a human, e.g. an Olympic diver, and find an image where the limbs of the human are well defined. The generic human model is then tuned until it fits the human in the video, yielding a personal model of the human in the video. They do the same for key frames, which are defined to be frames where a dramatic move starts or ends, of the entire video sequence using the personal model. To find the poses between the key frames a correlation

method is used. Based on a key frame the possible body poses are predicted and the correlation carried out. The best match determines the body pose. Then this is done for the entire video sequence and a complete set of pose parameters are obtained. This set can then be used to make another human model carry out the same movements.
Comments: A good and useful system for entertainments and also a nice way to obtain the body pose between key frames. It is, however, off-line processing.

Bibliography

[1] J.K. Aggarwal and Q. Cai. Human Motion Analysis: A Review. In *Workshop on Motion of Non-Rigid and Articulated Objects*, Puerto Rico, USA, 1997.

[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*, pages 2–14, Austin, Texas, USA, 1994.

[3] K. Akita. Image Sequence Analysis of Real World Human Motion. *Pattern Recognition*, 17(1):73–83, 1984.

[4] P. Allard, I.A.F. Stokes, and J.P. Blanchi, editors. *Three-Dimensional Analysis of Human Movement.* Human Kinetics, 1995.

[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] A.M. Baumberg and D.C. Hogg. An Efficient Method for Contour Tracking using Active Shape Models. Technical Report 94.11, University of Leeds, UK, 1994.

[9] 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, Oregon, USA, August 1996.

[10] A.P. Bernat, J. Nelan, S. Riter, and H. Frankel. Security Applications of Computer Motion Detection. *Applications of Artificial Intelligence V*, 786, 1987.

[11] 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, Texas, USA, 1994.

[12] A.F. Bobick. Computers Seeing Action. In *British Machine Vision Conference*, Edinburgh, Scotland, September 1996.

[13] A.F. Bobick. Movements, Activity, and Action: The Role of Knowledge in the Perception of Motion. In *Royal Society Workshop on Knowledge-based Vision in Man and Machine*, London, England, February 1997.

[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] C. Bregler. Learning and Recognizing Human Dynamics in Video Sequences. In *Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997.

[17] Q. Cai and J.K Aggarwal. Tracking Human Motion Using Multiple Cameras. In *International Conference on Pattern Recognition*, 1996.

[18] Q. Cai, A. Mitiche, and J.K. Aggarwal. Tracking Human Motion in an Indoor Environment. In *International Conference on Image Processing*, 1995.

[19] L. Campbell and A. Bobick. Recognition of Human Body Motion Using Phase Space Constraints. In *International Conference on Computer Vision*, Cambridge, Massachusetts, 1995.

[20] 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.

[21] C. Cedras and M. Shah. Motion-Based Recognition: A Survey. *Image and Vision Computing*, 13(2), March 1995.

[22] O. Chomat and J.L Crowley. Recognizing Motion Using Local Appearance. In *International Symposium on Intelligent Robotic Systems*, University of Edinburgh, 1998.

[23] 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.

[24] C. Christensen and S. Corneliussen. Visualization of Human Motion using Model-based Vision. Technical report, Laboratory of Image Analysis, Aalborg University, Denmark, January 1997.

[25] 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.

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

[27] A. Cretual, F. Chaumette, and P. Bouthemy. Complex Object Tracking by Visual Servoing Based on 2D Image Motion. In *International Conference on Pattern Recognition*, 1998.

[28] 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.

[29] J.W. Davis and A. Bobick. The Representation and Recognition of Action Using Temporal Templates. In *Conference on Computer Vision and Pattern Recognition*, 1997.

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

[31] J.W. Davis and A. Bobick. Virtual PAT: A Virtual Personal Aerobics Trainer. Technical Report 436, MIT Media Lab, 1998.

[32] 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.

[33] Luc E. 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.

[34] 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.

[35] H. Fujiyoshi and A.J. Lipton. Real-Time Human Motion Analysis by Image Skeletonization. In *Workshop on Applications of Computer Vision*, 1998.

[36] D.M. Gavrila. The Visual Analysis of Human Movement: A Survey. *Computer Vision and Image Understanding*, 73(1), January 1999.

[37] 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, USA, 1996.

[38] 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.

[39] 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, Massachusetts, 1995.

[40] 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.

[41] 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 *Lecture Notes in Artificial Intelligence 1537. Modeling and Motion Capture Techniques for Virtual Environments*, 1998.

[42] Y. Guo, G. Xu, and S. Tsuji. Tracking Human Body Motion Based on a Stick Figure Model. *Journal of Visual Communication and Image Representation*, 5(1):1–9, 1994.

[43] I. Haritaoglu, D. Harwood, and L.S. Davis. Ghost: A Human Body Part Labeling System Using Silhouettes. In *International Conference on Pattern Recognition*, 1998.

[44] 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.

[45] B. Heisele and C. Wohler. Motion-Based Recognition of Pedestrians. In *International Conference on Pattern Recognition*, 1998.

[46] A. Hilton and T. Gentils. Popup People: Capturing human models to populate virtual worlds. In *Siggraph*, 1998.

[47] D. Hogg. Model-Based Vision: A Program to See a Walking Person. *Image and Vision Computing*, 1(1), February 1983.

[48] D.C. Hogg. *Interpreting Images of a Known Moving Object.* PhD thesis, University of Sussex, UK, 1984.

[49] 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, USA, 1997.

[50] 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.

[51] 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.

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

[53] 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, Vermont, USA, 1996.

[54] I. Kakadiaris and D. Metaxas. Vision-Based Animation of Digital Humans. *Computer Animation*, pages 144–152, 1998.

[55] I.A. Kakadiaris and D. Metaxas. 3D Human Body Model Acquisition from Multiple Views. In *International Conference on Computer Vision*, pages 618–623, Cambridge, Massachusetts, June 20-23 1995.

[56] 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.

[57] Y. Kameda and M. Minoh. A Human Motion Estimation Method Using 3-Successive Video Frames. In *International Conference on Virtual Systems and Multimedia*, 1996.

[58] 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.

[59] 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.

[60] 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, Maryland, February 1992.

[61] H.J. Lee and Z. Chen. Determination of 3D Human Body Posture from a single View. *Computer Vision, Graphics, and Image processing*, 30:148–168, 1985.

[62] H.J. Lee and Z. Chen. Knowledge-Guided Visual Perception of 3-D Human Gait from a Single Image Sequence. *Transactions on Systems, Man, and Cybernetics*, 22(2), March/Apri 1992.

[63] F. Lerasle, G. Rives, and M. Dhome. Human Body Limbs Tracking by Multi-ocular Vision. In *Scandinavian Conference on Image Analysis*, Lappeenranta, Finland, 1997.

[64] M.K. Leung and Y.H. Yang. A Region Based Approach for Human Body Motion Analysis. *Pattern Recognition*, 20(3):321–339, 1987.

[65] M.K. Leung and Y.H. Yang. Human Body Motion Segmentation in a Complex Scene. *Pattern Recognition*, 20(1):55–64, 1987.

[66] M.K. Leung and Y.H. Yang. First Sight: A Human Body Outline Labeling System. *Transactions on Pattern Analysis and Machine Intelligence*, 17(4), April 1995.

[67] 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.

[68] W. Long and Y.H. Yang. Log-Tracker: An Attribute-Based Approach to Tracking Human Body Motion. *Pattern Recognition and Artificial Intelligence*, 5(3):439–458, 1991.

[69] Y. Luo, F.J. Perales, and J.J. Villanueva. An Automatic Rotoscopy System for Human Motion based on a Biomechanic Graphical Model. *Computers & Graphics*, 16(4), 1992.

[70] D. Meyer, J. Denzler, and H. Niemann. Model Based Extraction of Articulated Objects in Image Sequences. In *Fourth int. conf. on Image Processing*, 1997.

[71] T.B. Moeslund. Computer Vision-Based Human Motion Capture - A Survey. Technical report, Laboratory of Image Analysis, Aalborg University, Denmark, 1999.

[72] S. Moezzi, A. Katkere, D.Y. Kuramura, and R. Jain. Reality Modeling and Visualization from Multiple Video Sequences. *Computer Graphics and Applications*, November 1996.

[73] T. Molet, R. Boulic, and D. Thalmann. A Real Time Anatomical Converter for Human Motion Capture. In *EUROGRAPHICS Int. Workshop on Computer Animation and Simulation*, Poitier, France, 1996.

[74] 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.

[75] A. Nakazawa, H. Kato, and S. Inokuchi. Human Tracking Using Distributed Video Systems. In *International Conference on Pattern Recognition*, 1998.

[76] P.J. Narayanan, P.W. Rander, and T. Kanade. Constructing Virtual Worlds Using Dense Stereo. In *International Conference on Computer Vision*, Bombay, India, Januar 1998.

[77] S.A. Niyogi and E.H. Adelson. Analyzing and Recognizing Walking Figures in XYT. In *Conference on Computer Vision and Pattern Recognition*, 1994.

[78] P. Nordlund. *Figure-Ground Segmentation Using Multiple Cues*. PhD thesis, Kungl Tekniska Hogskolan, Sweden, 1998.

[79] 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.

[80] J. O'Rourke and N.I. Badler. Model-Based Image Analysis of Human Motion Using Constraint Propagation. *Transactions on Pattern Analysis and Machine Intelligence*, 2(6), November 1980.

[81] A. Pentland. Interactive Video Environments and Wearable Computers. In *International Workshop on Automatic Face-and Gesture-Recognition*, Zurich, Switzerland, 1995.

[82] A. Pentland. Machine Understanding of Human Action. In *Int'l Forum on Frontier of Telecommunication Technology*, Tokyo, Japan, November 1995.

[83] A. Pentland. Smart Rooms, Smart Clothes. In *International Conference on Pattern Recognition*, 1998.

[84] 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.

[85] R. Polana and R. Nelson. Low Level Recognition of Human Motion. In *Workshop on Motion of Non-Rigid and Articulated Objects*, Austin, Tx, USA, October 1994.

[86] K. Rohr. *Human Movement Analysis Based on Explicit Motion Models*, chapter 8, pages 171–198. Kluwer Academic Publishers, Dordrecht Boston, 1997.

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

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

[89] A. Shio and J. Sklansky. Segmentation of People in Motion. In *Workshop on Visual Motion*, pages 325–332, October 1991.

[90] C. Sul, K. Lee, and K. Wohn. Virtual Stage: A Location-Based Karaoke System. *Multimedia*, 5(2), 1998.

[91] 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.

[92] T. Tsukiyama and Y. Shirai. Detection of the Movements of Persons from a Sparse Sequence of TV Images. *Pattern Recognition*, 18(3/4):207–213, 1985.

[93] M. Turk. Visual Interaction with Lifelike Characters. In *International Conference on Automatic Face- and Gesture-Recognition*, Killington, VT, USA, 1996.

[94] A. Utsumi, H. Mori, J. Ohya, and M. Yachida. Multiple-View-Based Tracking of Multiple Humans. In *International Conference on Pattern Recognition*, 1998.

[95] 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, USA, 1997.

[96] J. Wang, G. Lorette, and P. Bouthemy. Analysis of Human Motion: A Model-Based Approach. In *Scandinavian Conference on Image Analysis*, 1991.

[97] 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.

[98] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland. Pfinder: Real-Time Tracking of the Human Body. In *International Conference on Automatic Face- and Gesture-Recognition*, Killington, Vermont, USA, 1996.

[99] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland. Pfinder: Real-Time Tracking of the Human Body. *Transactions on Pattern Analysis and Machine Intelligence*, 19(7), July 1997.

[100] C.R. Wren and A.P. Pentland. Dynaman: A Recursive Model of Human Motion. In *Image and Vision Computing*, 1998.

[101] C.R. Wren and A.P. Pentland. Dynamic Models of Human Motion. In *International Conference on Automatic Face- and Gesture-Recognition*, Nara, Japan, 1998.

[102] C.R. Wren *et al.* Perceptive Spaces for Performance and Entertainment. In *International Conference on Automatic Face- and Gesture-Recognition*, Nara, Japan, 1998.

[103] Y. Yacoob and M.J. Black. Parameterized Modeling and Recognition of Activities. In *International Conference on Computer Vision*, Bombay, India, 1998.

[104] 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*, 1998.

[105] M. Yamamoto, T. Kondo, T. Yamagiwa, and K. Yamanaka. Skill Recognition. In *International Conference on Automatic Face- and Gesture-Recognition*, Nara, Japan, 1998.

[106] M. Yamamoto and K. Koshikawa. Human Motion Analysis Based on A Robot Arm Model. In *Conference on Computer Vision and Pattern Recognition*, 1991.

[107] 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.

[108] J.Y. Zheng and S. Suezaki. A Model Based Approach in Extracting and Generating Human Motion. In *International Conference on Pattern Recognition*, 1998.