Markerless Based Human Motion Capture: A Survey

Joseph Bray
Vision and VR Group
Dept Systems Engineering
Brunel University
Uxbridge UB8 3PH

In collaboration with Televirtual Ltd, Norwich, UK

Abstract

This literature survey attempts to recent developments and current state-of-the-art in the field of body analysis by the use of non-intrusive optical systems. Markerless based human motion capture describes the activity of analysing and expressing human motion in mathematical terms. The survey shows that the task of motion capture can be divided into a number of systematically distinct groups, initialisation, tracking, pose estimation and gesture recognition. Tracking and pose estimation are addressed in turn explaining their constituent parts and the relationships between them. The literature shows that various mathematical body models are used to guide the tracking and pose estimation processes. A review of these models and methods of coping with their inherent high parameter dimensionality is included. Finally a discussion of work performed and conclusions drawn is presented.

List of Abbreviations

ASM	**A**ctive **S**hape **M**odel
CCD	**C**harged **C**ouple **D**evice (camera)
COG	**C**entre **o**f **G**ravity
CONDENSATION	**Cond**itional **Dens**ity Propag**ation**
EKF	**E**xtended **K**alman **F**ilter
HCI	**H**uman **C**omputer **I**nteraction
HSI	**H**ue **S**aturation **I**ntensity
HMM	**H**idden **M**arkov **M**odel
IR	**I**nfra**r**ed
LED	**L**ight **E**mitting **D**iode
MLD	**M**oving **L**ight **D**isplay
PCA	**P**rinciple **C**omponent **A**nalysis
PDM	**P**oint **D**istribution **M**odel

1 Introduction

Human motion capture was first encountered by Eadweard Muybridge in his famous experiments entitled Animal Locomotion in 1887. Muybridge is considered to be the father of motion pictures for his work in early film and animation. Animal Locomotion was a study into the way in which animals and birds moved. The study included recording at discrete time intervals, photographs of the subjects in order to visualise motion. In 1973 psychologist Johansson conducted his now famous Moving Light Display (MLD) experiments, into the visual perception of biological motion [27]. Johansson attached small reflective markers to the joint locations of human subjects and recorded their motion. He asked subjects to identify known movements after being shown just the marker trajectories. These experiments where the first few steps into what is becoming an ever increasingly travelled path of research.

Motion capture is the analysis of a scene giving rise to some mathematical representation of the movement given by a human subject, or as Menache writes “Motion Capture is the process of recording a live motion event and translating it into usable mathematical terms by tracking a number of key points in space over time and combining them to obtain a single 3D representation of the performance.” [35]. In order to describe well, it is often best to give an example, imagining that your task is to observe a friend perform some action or gesture which requires movement, you must estimate and report the position of his joints or body parts as the task is performed. This is a trivial task for a human to achieve. The aim of motion capture technologies is to perform exactly that task.

Today there is a great interest in the topic of motion capture and the number of papers published in this subject area grows exponentially. The increased interest and research is due to a number of factors. Advances in silicon technology and the continually lowering cost of video capture and processing, have allowed easier access to the equipment required in the research. Equally the intellectual challenge from the demanding nature of the task has fuelled attention. Segmenting non-rigid objects which exhibit self-occluding motion, is an inherently difficult task. Compounded by the fact that the estimation of this motion is necessarily

required to be accurate since humans naturally detect discrepancies in acquired motion data. The number of potential applications arising from motion capture technologies is also a strongly influencing factor in recent interest. Outlined below are some of the potential uses of motion capture.

Advanced user interfaces	Social Interfaces Sign Language interpretation Gesture driven application interface
Model-Based Encoding (MPEG-4)	Low bit-rate video compression Low variable animation control
Motion Analysis	Clinical studies Choreography of dance/theatre Assisted sports training Content based indexing of TV footage
Smart surveillance systems	Indoor and outdoor scenes Gait recognition
Virtual Reality	Interactive virtual worlds Character animation Teleconferencing Virtual film/TV production Computer Games

Table 1 Potential Applications Of Motion Capture Technologies

1.1 Current and Potential Applications

Described briefly in the following sections are some of the current and potential applications of human motion capture technologies. What follows does not represent a comprehensive

study but a cross-section of general topics in which human motion capture technologies have been applied and topics in which motion capture technologies may be beneficial. The information outlined is summarised in Table 1.

1.1.1 Advanced user interfaces

A basic goal of Human Computer Interaction (HCI) is to construct more natural methods of communication between computer and man. Much research has been performed creating computer systems that are able to comprehend human methods of communication. One such method, speech, has received great attention from the research community and there have been many advances. However the problem has not yet been fully solved. A vision based computer interface capable of recognising movements or gestures could be used solely or in conjunction with other modal interfaces such as speech. In this sense the vision system is used to disambiguate sounds by recognising lip movements. Other potential uses of a vision based interface capable of gesture recognition are sign language comprehension and articulation control. Articulation control refers to driving software applications by the computers ability to understand predefined gestures performed by a user.

1.1.2 Model-Based Encoding (MPEG-4)

The MPEG group aims at encoding video streams efficiently, that allows for low-bite rate transmission over IP networks. The MPEG-4 group looks more specifically at the problem of compression of model based media. The MPEG–4 SNHC [1] group concerns it self with encoding hybrid Synthetic and Natural multimedia. Essentially the group looks at ways of compressing the amount of data needed to express fully a scene that is constructed from a mixture of real and computer-generated sources. A simple example of this is encoding the position of a face in a scene and transmitting only the changes that happen within the face

[1] http://130.187.2.100/mpeg4-snhc/index.html

region. A more complex example is encoding human mesh models into a form in which motion is propagated by the use of a few key parameters. Currently the method of transmitting sequences of character animation is to send an entire mesh for each change in motion.

1.1.3 Motion Analysis

There are many established uses of existing motion capture technologies. There is widespread use amongst the clinical and sports analysis arenas. Clinical studies require accurate motion knowledge for the diagnosis of locomotion difficulties in patients. Sports people use motion capture systems to record themselves in order to diagnose potential improvements in their performance. Systems have been proposed that would allow the creation of a database that would index video footage according to the type of motion within the clip.

1.1.4 Smart Surveillance Systems

Motion capture systems have been used to detect gait as a biometric. It has been proposed [45] that gait is sufficiently distinct to allow discrimination. Vision systems may also be capable of learning and distinguishing patterns of unusual behaviour. A system may be able to learn normal behaviour and signal an alarm when abnormal behaviour is detected. These systems would have to be robust enough to perform in real world unconstrained environments, but fortunately may only have to recognise a small motion vocabulary.

1.1.5 Virtual Reality

By far the most common application of human motion capture systems is computer-generated virtual character animation. Animators record the motion of a human actor and then propagate that motion through a virtual character. This gives rise to convincing motion being performed by the virtual character. These virtual characters are deployed in a wide range of uses from TV special effects to computer video games.

It has been shown that there are many areas of potential applications, and in general approaches differ with respect to target applications. These potential application have fuelled interest from commercial market places and correspondingly academic research.

1.2 An Overview of Current Areas of Research

Generally speaking, the literature can be divided in three parts; face analysis, body analysis and gesture analysis. A number of subcategories exists for each discipline. A brief account of each subcategory follows. Face analysis can be divided in two general subcategories. Firstly, face analysis regarding the recognition of faces from a scene, usually in conjunction with a security system. Secondly, the analysis of a face in order that face models can be built. These face models are then used in animations. Gesture analysis can be thought of as the recognition of a gesture given by either both the hands and arms or just the hands. Body analysis is the estimation of the pose of a human subject.

Generally the level of detail in which techniques attempt to capture motion is dependent on the scale of the exhibited movements. In the case of body analysis human motion capture is defined as the capturing of large scale body movements whilst ignoring smaller scale motion of facial features or digits.

Existing motion capture technologies are centred around three principle approaches; electromechanical, electromagnetic and optical tracking systems[2]. Electromechanical systems consist of body suits with measuring devices at fixed points. The measuring devices are constructed from potentiometers and sliding rods. When the human subject moves the motion is detected by small changes in the potentiometers. The suits give accurate results but are restrictive in their weight and freedom of movement. The major disadvantage of

[2] Refer to Appendix I

electromechanical suits is that they inhibit the motion that can be performed. A full range of human motion can not be expressed, in particular walking or other transitional motion.

The electromagnetic approach improves the range of motion that is possible to capture. Electromagnetic sensors are used at key locations upon the body to extract both the position and orientation of the sensor. The sensor registers differences in an electromagnetic field that is produced by an accompanying transmitter. The electromagnetic suits are lighter and more comfortable than the electromechanical suits but are still disadvantaged by wires attached to each sensor.

The least restrictive motion capture technology is optical based. Optical motion capture systems consist of pre-calibrated multi-camera set-ups and markers that are attached to key locations on the human body. Infrared reflective balls are tracked giving joint and limb locations. Infrared reflective systems are prone to misclassification of individual markers, giving erroneous data. This may be overcome by the use of frequency varying pulsating LED’s which can be labelled from the sequence in which the pulsating occurs. Both types of marker are lightweight and can be easily jogged or nudged as the actor expresses motion. Optical systems work on the triangulation of detected markers which requires that the subject be located in a predefined and usually small capture area.

All of the above approaches need specialist equipment and are restrictive to some degree. The feasibility of these approaches to the application areas defined above is limited. One can appreciate the obvious need for a motion capture system that does not require specialist restrictive equipment. They are also exclusively expensive and used only by professional motion capture companies.

It is now possible to describe the term markerless motion capture in full. Markerless motion capture refers to the task of full body motion capture without the need of markers or specialist suits attached to the body.

This literature survey attempts to review a representative sample of current research in the field of body analysis by the use of non-intrusive optical systems. The survey has described existing motion capture technologies and highlighted their weakness. An increase in research activity is noted together with a review of current and potential applications.

The remainder of this work outlines current research into markerless based human motion capture. Chapter 2 describes a succinct categorisation of authors approaches and sets a structure for the following chapters. Chapter 3 presents an overview of the human body models authors have utilised to assist in tracking (chapter 4) and pose estimation (chapter 5). Finally conclusions are drawn and possible avenues of further work discussed.

2 Research Methodology Categorisation

The number of recently published papers is indicative of the number of approaches to the common problem, amongst these approaches similar trains of thought arise. This survey groups the papers reviewed into a number of categories. It is important to choose a relevant and well structured classification scheme in order not to over or under-classify similarities and differences in approaches. It is important to have a scheme that is hierarchical yet simultaneously homogenous. There exists a number of classes under which previous work might be classified.

- 2D approaches versus 3D approaches
- model_based versus non model_based
- Type of model (stick figure, statistical)
- kinetic versus kinematics,
- sensor modality (visual light, IR light, range data, etc.)
- number of sensors
- mobile sensors versus stationer sensors
- tracking versus recognition
- pose estimation versus tracking
- pose estimation versus recognition
- applications
- one person versus multiple persons
- number of tracked limbs
- distributed versus centralised
- motion type assumptions (rigid, non_rigid, elastic, etc.).

However, none of the above categories is both detailed and broad enough and would fall foul to the complexity of the subject matter. Hence, many authors would not be able to be listed within the framework. Surveys have been published in recent years, each giving a particular flavour of classification. The surveys which already exist in the field tend to classify research somewhat differently. Cedras [12] gives an overview of different methods within motion extraction which can all be classified as belonging to optical flow or motion correspondence. Next a taxonomy for the human motion capture problem is given as: action recognition, recognition of the different body parts and body configuration estimation.

Aggarwal [2] gives a review based upon the type of motion, namely, articulated and elastic motion. The paper describes and delineates previous more general descriptions of motion types. The authors then proceed to classify both types of motion to approaches that use shape models against those that do not. In a later survey by Aggarwal [1] the same categorisation is used as in the work by Cedras [12] even though they use different labels for the three classes. The three classes are each divided into subclasses yielding a rather complex classification scheme. More recently, a survey by Gavrila [21] provides a general introduction to the topic with a special focus on different applications. His taxonomy consists of 2-D approaches without explicit shape models, 2D approaches with explicit shape models and 3-D approaches. Across these three classes he describes the approaches dealing with recognition. The surveys by Cedras [12], Aggarwal [1] and Gavrila [21] also offer robust classifications of the field. However, Moesland [37] offers the most logical and flexible taxonomy. He categorises the field not by approaches or techniques used, but by the stages that have to occur in order to solve the general problem of motion capture. He then subdivides these stages into approaches. This approach is similar to Gavrila [21] whom allows the cross classification of an approach, a paper can be located both in the recognition class as well as in one of the authors other three classes. This reduces the problem of classifying approaches which are located in-between classes or in more than one class. Moesland postulates that motion capture and approaches maybe categorised in a systematic way. The author looks at how researchers tackle the subject in a temporal sense. To illustrate; firstly the system is *initialised*, this includes choosing the correct models and performing background tasks such as camera calibration, or for example, finding a human in a static scene. Secondly, the human is *tracked* i.e. the human image is tracked from frame to frame. Thirdly, An *estimation of the pose* of the subject is performed in image I_1. Finally, *recognition* of motion is accomplished over a sequence of images $I_1,\ldots I_n$.

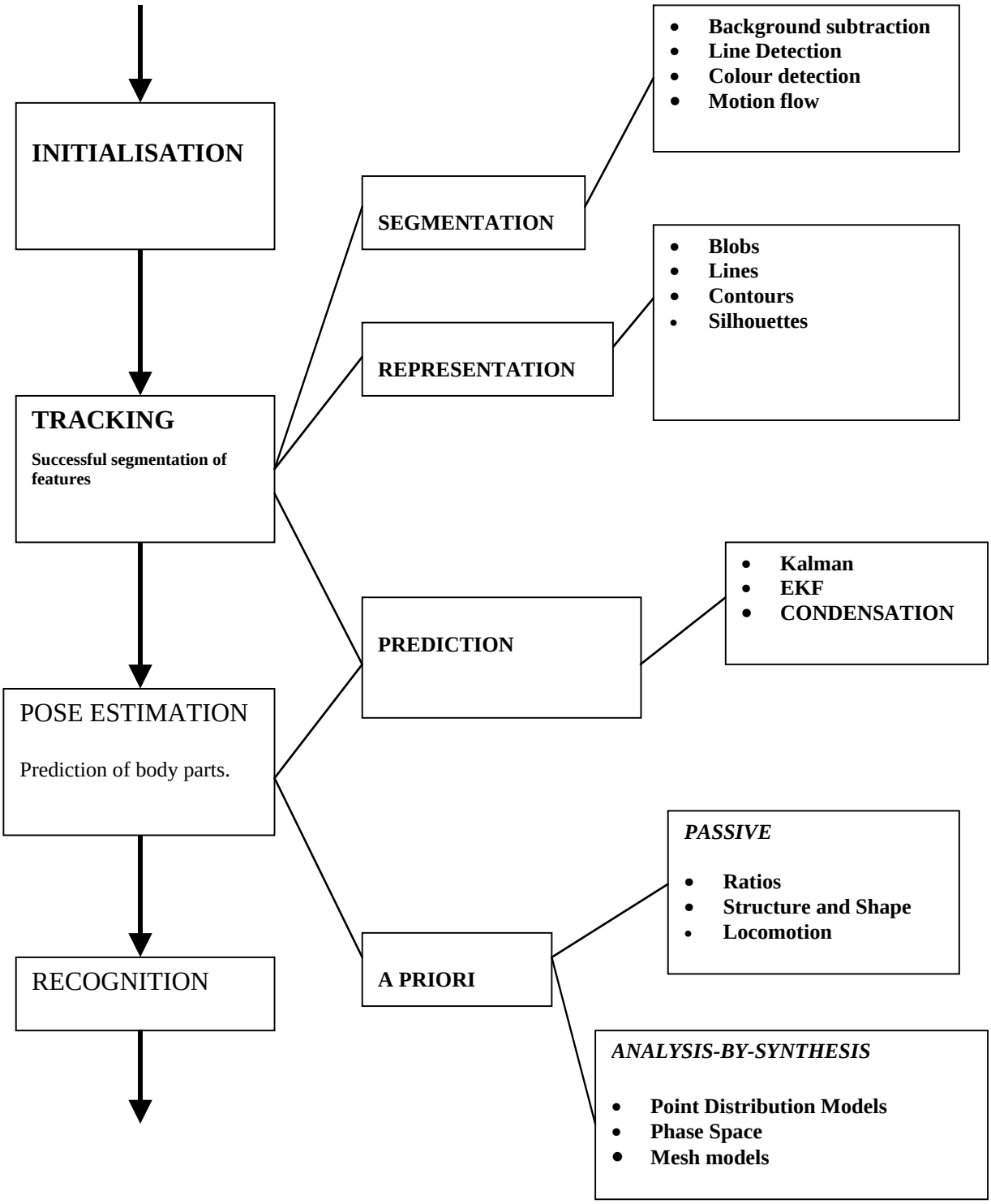
Figure 1 shows an overview of the classification scheme from a systems requirement perspective. Each stage in the human motion task is represented down the left hand side of the figure. An approach may not necessarily consist of each stage nor do the stages necessarily have to be performed in order of the flow depicted. Indeed many authors

feedback and forward information from stage to stage. The scope of this report will consist mainly in the tracking and pose estimation stages.

The approaches to human tracking reviewed in this report vary from complex detection to segmenting using image subtraction, the aim of tracking is to achieve consistent motion and robust detection of the same human in a scene. Having achieved this detection one would like to determine the pose of the subject, again this may lead to very accurate estimations of joint locations in space or an estimate of general pose. The results and accuracy usually depend on the application that requires the motion data.

This chapter has discussed research taxonomies utilised by other authors. It infers that the approach by Moesland [37] is the most encompassing and flexible and lays down the framework on which the remainder of this report will follow.

Figure 1 Systematic Classification Scheme



3 Models

It should be noted that the use of body models is very common in the literature, most authors employ some model or abstraction of the human form at some level. This is because it is often beneficial to guide the tracking process by using some *a priori* information about the locomotion of a human. Considering that the required data from a motion capture system is a description of locomotion, usually denoted by joint positions over time, or movements of some human body model over time; it is usual to guide pre-processes such as pose estimation and tracking by the plausible states of a body model. This highly defined representation of motion is exactly what is required by most applications of motion capture. Chapter 5 discusses pose estimation and how it can be assigned three subcategories. Those approaches that use *a priori* models for guidance, those that use *a priori* models for constraint and those that do not use *a priori* models.

Human body models which contain *a priori* information are used in two ways; actively and passively. Passive models use information to guide the tracking and pose estimation stages to give more convincing results, but are not confined to the model. Active models usually have a set of predefined and allowable postures which are represented by individual states of the model, information is drawn form the image and tracking processes and used to select the most convincing predefined pose.

The simplest representation of a human body is the stick figure, this concept first appears in the literature after the founding work of Johansson [27]. Each node represents a stylised joint position and the lines between nodes represent bones. This representation forms the fundamental data representation for most commercial motion capture systems. The positions of the nodes may be given as 3D points in Euclidean space but are more usually given as relative transformations from parent nodes. The stick model has an in-built hierarchy. Motion simulation is achieved by having a subject adopt a starting or initialisation pose and then performing local transformations of joint positions to give global motion. The deficiency of such a model is the inherent difficulty of matching image features to model configuration. i.e.

the task of matching image areas to nodes of the model. In order to facilitate such a task authors “flesh out” the “bones” of the model. Leung [34] used ribbons to represent 2-D contours and constructed silhouettes using these contours. They reported progress on the general problem of segmenting, tracking and labelling of human body parts from the silhouette of a human. Their basic body model, consisted of 5 U shaped ribbons and a body trunk, various joint and mid points, plus a number of structural constraints, such as support. Support was defined as a notion of balance under gravity. In addition to the basic 2D model, view based knowledge is defined for a number or generic human postures to aid the interpretation process. The segmentation of the human silhouette is performed by detecting moving edges. Another method of “fleshing out” the body models is described by Akita [3] in which both stick figures and cone approximations where integrated and processed in a course to fine manner. The author utilises the idea of using shape primitives around a stick figure body model. The use of primitive shapes facilitates processing since the shape is easy to synthesis and requires less parameters than for example a polygonal mesh. Akita’s work employed a key frame sequence of stick figures indicating the approximate order of the motion and spatial relationships between the body parts. A further complexity of body model was used in the work of O’Rourke [39] in the form of 24 rigid segments and 25 joints; the segments were overlapping and spherical. In continuation of body model complexity, superquadrics have been utilised. Superquadrics are generalisations of ellipsoids which have additional “squareness” parameters along each axis, the benefit of these models is that they exhibit better shape modelling near the joints. Full polygonal mesh models gained from PET scans have been used, but allow no generalisation for tracking [33].

4 Tracking of Humans

Tracking is the task of correctly identifying a target from sequence of frames. The problem with tracking humans in a scene is that they exhibit non-rigid articulated motion, the appearance of the target changes over time. Also the target may be partially or wholly occluded over a sequence of frames or may perhaps self-occlude. This implies that tracking of humans is a difficult task. However, generally there exists three aspects to achieving this goal.

1. The principle stage is to segment the human (or part of) from the background.
2. It is then beneficial to reduce the complexity of the ensuing data by representing it in more manageable form.
3. Finally to deploy some model of motion between frames. The general assumption is that motion between frames is small, this allows the prediction of the new feature position by use of algorithms such as the Kalman Filter [30].

4.1 Segmentation

At its principal level, tracking requires some target to be identified from the scene as a whole. In computer vision terms, identification of a target from a scene can be thought of as the segmentation of a target (foreground) from clutter (background). Many types of segmentation techniques present in the literature have been used. Generally these methods are then followed by morphological operators to reduce noise, and improve the results.

4.1.1 Motion Data

The simplest method for detecting change between a sequence of images is to use a technique called image differencing or background subtraction. Two consecutive images I_{t-1} and I_t are taken from a sequence, where the camera angle hasn't changed significantly between frames.

Then image I_{t-1} is subtracted from I_t and the resultant image contains only information about the differences between those two frames. Since a fixed background is assumed, the only differences between frames should be any target that has moved. Unfortunately white Gaussian noise is also present [17]. However, as mentioned these techniques are highly susceptible to noise and require that no change, including lighting changes, occur from frame to frame. Lighting conditions are paramount in this type of segmentation. Outdoor scenes with lighting from the sun suffer dramatic lighting changes if the suns rays are blocked momentarily by a cloud, equally, fluorescent lights flicker normally at 50 Hz causing interference with the sampling devices within the CCD, and consequentially make luminosity changes in a scene. These lighting changes are manifested as changes (read movement) in the scene causing any simple background subtraction process to fail. These algorithms will also suffer from ghosting as motion detection algorithms detect the region of a scene exposed by movement as occupied. A method of overcoming these drawbacks is the lighting invariant background described by Wren [47] and is often referred to in the literature as Pfinder. It is possible to model the scene as a static background and dynamic foreground by building a background model of the variations of intensity such that each pixel has a Gaussian distribution, the foreground is modelled as a number of blobs each sharing statistically similar colour and spatial properties. Statistical texture properties of the background are observed over an extended period of time are used to construct a model. This model is used to decide which pixels in an input image fall into the background class. Essentially Pfinder uses colour and spatial information to segment.

Lee [32] uses motion and colour, to segment face regions in real scenes. Initially a velocity vector field of the face is extracted and then thresholded to show regions of clear motion, this information is used in conjunction with hue space thresholding. Davis assumes that the background is stationary or at most varying slowly, but that the person is moving [14].

Tracking of skin tones is common in the literature because of its obvious advantages, trackers that make use of head and hand colours make the assumption the target is wearing clothes with long sleeves and that hands and face are in the most part visible. Raja [41] describes a

system for building Gaussian models of skin tones, which according to the authors occupies a relative small distribution in HSI space.

Optical flow is used in the literature for both 2D and 3D approaches [36][40][46] (for a more detailed review of segmentation of image features, the interested reader is directed to footnote[3]). The obvious advantage of using motion flow segmentation algorithms is the data acquired may be useful in higher-level processing. If the velocity of a 2D target in a scene target is known, this data may be beneficial in calculating motion for the 3D configuration of the human body model.

The extension to this is background subtraction in stereo images; a reference map is created off-line that represents colour similarities in stereo views. Any disparity from this model in runtime indicates the target, Ivanov [25]. However in order to build the model a laser range finder and stereo calibration must be used.

4.1.2 Appearance Data

Apart from human forms being detected in a scene from their motion it may be possible to segment based upon their appearance. Making the assumption that the human in the scene is of sufficiently different appearance to the background or surrounding objects it is possible to threshold for that difference. The process of thresholding on colour hue is known as croma-keying (or blue screening) and is well utilised in the television broadcast and post production industries, a common example being the weather reports for news programs. Essentially by making the colour of the background uni-chrome (usually blue because it is most chromatically opposite to skin colour) the target may be segmented by selecting all non blue pixels. Authors have dressed subjects in coloured garments so that they can be thresholded

[3] For a review, see http://www.fmrib.ox.ac.uk/~steve/review/review/node5.html.

from the scene. Individual body parts may be coloured differently for easier segmentation and labelling [22]. The same idea may be used in thermal images where hot bodies or shadows from projected infrared light behind the subject may be detected [15]. The key element in this approach is to enhance the difference between foreground and background. In order to guarantee a uniform colour for background that is not susceptible to shadows or light changes special reflective cloth and light emitters are used[4]. As described earlier other segmentation techniques based on statistical representations may be used. In general, colour is used to segment, a foreground model is generated by describing a range of colours that do not belong to the background distribution. However, spatial information may be included with a statistical description of a pixel, or image segment [47]. Both background and foreground models are updated with a temporal delay to counteract changes in the scene. This gives rise to the problem of how often to update the models and the complexity required from the model in order to perform well against the processing time required to update the multivariate Gaussian statistics.

4.2 Low Level and High Level

In computer vision terms low level processing refers to operations performed upon the grey level or colour pixels value in a frame. No knowledge is required or gleaned from these operations, there is no need to know whether a pixel is belonging to a human or background etc. However higher level representations are required from low level information. The data which low-level operations manufacture can be further processed and amalgamated until some higher level representation can be formed. This is known as the bottom-up approach. Levels of data representation chosen by the authors reviewed are imposed to make the matching of low-level data to human body models easier. The representation of data chosen usually coincides with the human model the researcher is maintaining. They directly

[4] http://www.radamec.co.uk

correspond to the models described in the previous chapter. i.e. Silhouettes [7][8][15][18][19][26][29][47][48]. Blobs and multiple blobs [4][47]

Bregler [9] reports a multilevel approach. One makes the assumption that while performing an action or gesture, the majority of the human body segments are in motion most of the time. The framework which consists of four levels, delays hard decision thresholds by using higher level statistical models and temporal context for adaptive feedback. The first level is the segmentation of optical flow, that is spatio-temporal gradients and colour in HSI values. The next level sees these segmentations group into coherent blobs, each blob is represented with a probability distribution over coherent motion, colour and spatial "support regions". These blobs are then grouped to linear stochastic dynamic models. Finally these dynamic models correspond to the emission state of a Hidden Markov Model (HMM).

Once lower level pixel information has been farmed into mid-level constructs, e.g. blobs, lines etc. they represent a feature that may be tracked through frames. These features should be robust enough not to change vastly between frames. Essentially there exists enough cross-frame similarity to use tracking algorithms.

4.3 Temporal Tracking

In order to facilitate location of an image feature in a frame, knowledge about the underlying motion model is used. Wren [47] segmented a scene into distinct regions called blobs. The motion of these blobs was then modelled and predicted using a Kalman filter [30] under the assumption of simple Newtonian dynamics.

The general notion of a Kalman Filter is that given some moving feature in a scene and a corresponding motion model, it is possible to predict the position of the feature. The area of prediction is initially quite large. If a local search around the predicted location is performed and the feature found, then the information of the new measurement is used to update and improve the prediction mechanism. Kalman filters are used widely [6][9][17][24][36] but are

generally only effective (and designed for) motion that can be described by linear equations. The principle assumption of a Kalman filter is that its measurement equations are linearly stochastic difference equations. However, since most motion exhibited by humans in a scene is non-rigid and articulated, the Kalman filter does not cope well with the non-linearities produced. In general these non-linearities are a composition of various non-linear rotation matrices and perspective mappings and in the extreme case, end points and self-collision. End points are described as the motion exhibited when a joint locks, halting otherwise fluid rotational motion rapidly. If a linearisation about the inherent Gaussian distributions occurs it may be possible to linearise some of the observed non-linearities. This approach is named the Extended Kalman Filter (EKF). However this filter can be difficult to implement and is computationally expensive due to the calculations of the Jacobian matrices required. Julier's Unscented filter [28] is reported to be more suitable, efficient and more easily implemented than the extended Kalman filter.

4.3.1 Multiple Hypothesis Tracking

The problem of tracking in dense visual clutter is challenging. Kalman filtering is inadequate because it is based on Gaussian densities which being unimodal cannot represent simultaneous multiple hypothesis. The CONDENSATION [24] algorithm uses factored sampling, previously applied to the interpretation of static images, in which the probability distribution of possible interpretations is represented by a randomly generated set. Condensation uses learned dynamical models, together with visual observations, to propagate the random set over time. The result is highly robust tracking of agile motion.

The major drawback of using the Condensation algorithm is outlined and partially conquered by continuing work from the original authors. Deutscher [16] proposes the annealed particle filter, which in principle deals with reducing the number or "particles" or hypothesis in the multi-hypothesis tracking. Despite increasing the efficiency of the CONDENSATION algorithm by a factor of 10 the approach is still far from real-time.

5 Pose Estimation

The goal of tracking and pose estimation is a description in terms of some model of the posture of a human subject. Approaches to 3_D articulated motion use parameterised human body models, this has the advantage that each state represents a physically valid pose thus taking advantage of all prior 3_D knowledge and relying as little as possible on error prone 2_D image segmentation. In this sense all pose estimators use models at some level of their processing. However, a distinction can be drawn between approaches that engage *a priori* models and those that do not. The major difference between the two methodologies is in establishing feature correspondence between consecutive frames.

5.1 How to get Your Man without Finding His Body Parts

When no *a priori* models are available, correspondence between successive frames is based upon the prediction or estimation of features related to position, velocity, shape, texture and colour. The mathematical constructs used to predict features in a frame may have some level of knowledge about human movement from a 2D view. Human locomotion can then be described in statistical terms, derived from low-level features or by simple heuristics. This is the technique Polana [40] refers to as "How to get Your Man without Finding His Body Parts". They assume that the task of segmenting body parts to fit to some higher level models requires that the body parts be segmented and recognised, and that this task requires some pre-recognition. They assume that image features may give rise to motion patterns in 2D if the action performed by the human is cyclic. It is assumed these motions are specific enough to determine human motion. Kakadiaris [29] proposes a Human Body Part Identification Strategy and recover all the body parts of a moving human based on the spatio-temporal analysis of its deforming silhouette. They start by detecting a single blob and then updating the blob model to include body parts as the subject begins to move. However, this approach suffers from self-occlusion which is tackled by the use of multi-cameras; Iwasawa [26] used multiple views, the author took 3 orthogonal views and subtracted the silhouette by background subtraction. A Centre Of Gravity (COG) is calculated from each using a distance

transform, which allows for the weighting of arms and extremities to be lessened. The joint positions are estimated using a genetic algorithm. The approach still suffers from self-occlusion (in the sense of a silhouette) even though multiple views have been used to tackle the problem. Fujiyoshi [20] used silhouettes, COG and distance transform from the COG to the silhouette contour. An approximation to image skeletonisation is performed yielding star shapes. The angles that are produced within the star shapes are analysed in the Fourier domain in order to gain rotation independent examples. The representation in the frequency domain are used to perform gait recognition. This methodology is more suited to outdoor surveillance as it isn't clear how any precise estimate of body parts would be achieved.

Moving from silhouettes to other forms of representation, Pfinder is encountered in the literature [47] and its extension to the 3D case, sPfinder. SPfinder Azarbayejani [5] uses the centroids of blobs as point representation. Another line of research involves statistical shape models or Active Shape Models (ASM) to detect and track the contours of hands Cootes [13], and humans (silhouette), Baumberg [6] describes human silhouettes in terms of B-Splines for the contour of the human then maps their parameters into "contour space" where a single point in this space represents an individual contour. The dimensionality of the space is dependent on the number of parameters and the number of training examples. The dimensionality of the space can then reduced by Principal Component Analysis (PCA) of the clusters of points in this space. The largest deformations are contained within the first few eigenvectors, see Bowden [8] for further discussion. The *a priori* and non *a priori* methodologies can be combined for verification.

5.2 A priori Models of Human Motion, Articulation and Biophysics

The methodologies outlined above suffer in establishing feature correspondence between consecutive frames. In order to ease this problem, models of human form and motion are applied. Indeed most methods for the motion analysis of human body parts apply predefined models for feature correspondence and body structure recovery. Amongst those

methodologies that use models in their approach are those that use their models in a passive guiding sense rather than their use in a highly constrained manner.

5.2.1 The use of passive a priori models, for constraint guidance

A priori knowledge of the human may be used without building a restrictive model with which to report observed motions as states. Information of appearance may be utilised to guide a tracking system, for example information about the relative height of a human may help to guide the tracking process and in the same sense information about the relative ratios of human limbs helps predict pose [23].

A coarse representation where information is used to track figures from camera to camera is described by Cai [11]. Figures are defined by Bayesian classification schemes of the spatio-temporal information along the "midline" of a human image. Authors have used *a priori* information coded into algorithms as parameters without constructing models *per se*. Leung [34][5] applied a sophisticated 2D-ribbon model to explore the structural and shape relationships between the body parts and their associated motion constraints. The user movements are assumed to be known and are used to label the different body parts thereby finding the pose of the user. This notion of having *a priori* knowledge of motion is also used by Akita [3]. Pioneering work by O'Rourke [39] in the field of model_based 3D human motion analysis, describes how the author applied a very elaborate volumetric model (see chapter 3). A co-ordinate system was embedded in the segments along with various motion

[5] Their paper also gives a good review of approaches executed prior to 1993.

constraints of human body parts. The system consists of four main processes: prediction, simulation, image analysis, and parsing. First, the image analysis phase accurately locates the body parts based on the previous prediction results. After the range of the predicted 3D location of the body parts becomes smaller, the parser fits these location_time relationships into certain linear functions. Then, the prediction phase estimates the position of the parts in the next frame using the determined linear functions. Finally, a simulator embedded with extensive knowledge of the human body translates the prediction data into corresponding 3D regions, which is verified by the image analysis phase in the next loop.

5.2.2 Active Model Binding Using Model Configurations as Motion Descriptors

It can be seen in the literature that many authors [7][17][19][22][36][43] choose to employ parameterised models which includes *a priori* information, of joint constraints, and have the advantage that each state represents a physically valid pose. In this manner it is possible to maximise the use of underlying 3-D knowledge and minimise the use of error prone 2_D image segmentation.

Approaches using such parameterised models Di Bernado [17], Rehg [42], update pose by inverse kinematics. The state space maps onto image space by a non-linear measurement equation that takes into account the coordinate transformations at the joint locations and the 3_D to 2_D projection. Inverse kinematics involves inverting the mapping from state to image space to obtain changes in state parameters that minimise the residual between projected model and image features. The procedure involves a linearisation of the measurement equation, as defined by the Jacobian matrix, and a gradient_based optimisation scheme.

The inverse kinematics approach can also be taken with multiple cameras when no feature correspondence between cameras is assumed. One simply concatenates the residuals that occur between projected model and images features in all available camera views; see for example [43]. Another approach, using parameterised models, does not attempt to invert a non-linear measurement equation. Instead, it uses the measurement equation directly to

synthesise the model and uses a fitting measure between synthesised and observed features for feedback; see [22][44].

Two new elements have been introduced here, firstly matching between projected state features and image features and secondly, gradient based optimisation. The following paragraphs describe both topics very briefly.

5.2.3 Analysis by Synthesis

Analysis-by-synthesis is the term given to this process of analysing a scene by comparing its appearance to a model of that scene. The advantages of this approach is that it is no longer required to invert a non-linear measurement equations which is computationally expensive and in the case of human models, prone to singularities [38].

An example can be given of analysis by synthesis, from the subject of Point Distribution Models (PDM). PDM's may be used to represent silhouettes of human shape. In brief, a series of control points are placed around a silhouette in the scene. Each of these control points has a position in the 2D scene denoted usually by the coordinates of its pixel x,y the sequence of these points can be seen as a vector. This vector can be described as a point in a space that has the same dimensionality as the number of parameters used to describe the vector. If a silhouette model were to be constructed of a human it would create a space that contains a number of points each representing a silhouette of a given known pose. The approach known as analysis by synthesis is a comparison made between low level image features and high level, semantic, representations. Model configurations are mapped into low-level data representations. Usually low level image processing techniques are preferred, giving rise to blobs, lines, etc. The descriptions of these blobs are compared to a states of the model in question, by generating synthetic images of how the model would appear in the image plane. It is possible to see that the representations are met midway between the low-level pixel values and the higher level semantically coherent mathematical models. The approach allows a continuous feedback loop from higher to lower levels, which helps with the robustness of both elements.

An appropriate example of this approach is given by Di Bernardo [17] who used the analysis-by-synthesis approach. Di Bernado used a model in which the arm is modelled as two circular cones. The paper described used a number of assumptions about the allowable deformation of the arm. Real images are segmented and blurred against a dark background giving binary images. These images are compared to projected synthetic arm images and an error function is calculated between the real and generated images to update the model pose. Locomotion is predicted to ease the burden on the error function by the use of EKF. A similar approach [44] makes the assumption that the subject is walking parallel to the image plane. The target is detected and represented by a bounding box. Within this box contours are identified and each of them in turn are matched against the model. A search is performed through the set of all possible contour position and the pose that attains the least error weighting is assumed to be the correct pose. A reduction of the search space is performed by using a Kalman filter to predict the pose in the next frame. Similar work has been performed by Rehg [42] but only in the field of hand tracking.

5.2.4 Model Dimensionality and Search Strategies

Point correspondences between model and scene are not necessarily required. There exists many standard techniques used widely in the vision community to perform error minimisation or cost function minimisation. For example global search strategies based on genetic algorithms, neural nets, etc. Kuch used a greedy search strategy based on perturbation of individual state parameters [31]. Gavrila used a local search based on best first search [22]. The high dimensional search space which results from recovering whole body pose, and its corresponding high degrees of freedom model, necessitates in this work a decomposition technique. The pose recovery is done successively for torso, arms and torso twist, and legs.

Comparing the above greedy gradient_based inverse kinematics approaches with the non greedy combinatoric search approaches, it can be observed that the former have the advantage that they exploit gradient cues in the vicinity of minima and therefore are computationally more efficient [43]. The use of gradient based minimisation techniques afford a certain luxury that, as the parameters chosen approach the correct (or best) state of

the model, a natural snapping to this state occurs and reduces the computation involved with inverse kinematic approaches (namely the complex Jacobian computations).

Nevertheless concern is justified that a gradient_based search algorithm might get drawn to a local minimum giving rise to a sub-optimal estimate of model configuration or in the worst case a completely wrong solution. This would occur often since the measurement equations are highly non-linear and the sampling ratio at which one obtains image measurement is relatively low for fast movement such as locomotion and gesticulation. Furthermore, measurements are typically noisy and can be incorrect altogether, e.g. when corresponding the features with the wrong body parts.

A non_greedy search method also promises to be more robust over time; if it fails to find a good solution at time t, there is still a chance that it may recover at time t + 1 if the search area is sufficiently wide. A combination of non-greedy search followed by a gradient_based technique is probably a good compromise between robustness and efficiency [21].

6 Discussion

It is apparent from the literature that an all encompassing general purpose body tracker is far from feasible, it would be more beneficial at these early stages of the research field to design and implement system that is fully functional and robust under a specific application.

It has been shown that many works suffer from self-occlusion. This is the event in which a human target restricts the view of a body part by placing another body part between it and the camera. The simplest and most robust solution to this is to use more cameras. Whilst this answer doesn't alleviate the problem entirely it greatly reduces occurrence. Therefor it may be beneficial to employ some design that is scalable regarding the number of cameras present. Future work may propose to navigate a method that is functional for a single camera, performs faster with better results for 2 cameras and better still for 3 cameras. Again suitability of an n-camera approach is dependent on the final application, it would for example be impossible to use multi-camera approaches for next generation motion capture from old video footage.

The amount of pre-processing and initialisation would be dependent on the application in mind. The more pre-processing which is performed the more robust, efficient and accurate the final application. Generic models, could be constructed that would perform equally for any human subject regardless of size or shape. Calibration of cameras may be performed beforehand.

The system should be reliant solely on a cheap non-specific camera set up. Ideally a sub £100 web-cam should be entirely suitable. This would allow a potential application market of home computer users. Infrared and other costly specialist camera equipment would inhibit any widespread use because of the price of manufacture.

Restrictive, coloured and specialist equipment may be worn but is essentially against a markerless theme, it would be infeasible in the light of this document to replace restrictive markers for restrictive clothing.

The fixed background assumption, or fixed camera angle assumption, allows the use of background subtraction models that can be pre-processed, which would give rise to robust segmentation of motion. This gives only those parts of the human that are actually moving, granted that this is exactly the information which is required but it needs to be contextualised with respect to the body. The connectedness of human shape may be utilised to overcome ghosting from updating background models. Temporal updating of background models has an unwanted side effect, that when a subject moves the background which is revealed is no longer seen as part of the background and shown as a moving target. This unwanted ghosting effect may be answered by not updating the background model. Of course the disadvantages of not doing so is that objects that have changed in the background scene are encoded as a human target. A technique of connectedness might be applied or the assumption that non-connected, non-human temporally persistent shapes in the background are updated into the background model.

Within the segmented region, it is possible to glean other clues; colour, edge and therefore velocity within a “windowed” area that is human shape (silhouette). It is at this point, issues of a single target may be addressed. Ubiquitous throughout the literature with possibly the exception of the work by Isard [24] is the small motion assumption.

It is highly beneficial to utilise a human model where *a priori* information is encoded. These models should be multileveled to aid in the process of creating the mappings between low-level and higher-level, model-state space The problem that these models encounter is the high dimensionality of the spaces that they occupy and high computational expense. There are a number of methods to cope with the dimensionality of these spaces. This document has mentioned previously that many authors include biophysical information in their models by limiting the range of movement, some introduce further constraints and assumptions [44] by

searching for phase of the walking cycle. Constant angle of view is also assumed. [10][16][17][43]

A generic optical human motion capture system must be able to track unconstrained motion. One must ponder the question of diminishing returns in statistically constraining a model space and the differences between combinatoric search and greedy gradient descent approaches. There exists a seemingly infinite complexity of analysis one can perform on these data sets. A full comparative study does not exist within the literature.

Which method of comparison between state and image space should be employed? There are essentially two approaches used.

1. Project image features into state space and perform a greedy gradient based search of the nearest configuration of the model to the projected features.

2. Project the state space out into the image feature space and search combinatorically through the projections of the state space for the state which minimises some error metric.

Gavrila [21] suggested that a combination of the two would be optimal. Using combinatoric search to initialise and constrain the greedy gradient based approach would be beneficial. Also the state space would be heavy with constraints and information regarding general human locomotion. It is not fully addressed at this moment whether it is plausible to constrain the non-linear measurement equation that occurs from image to state space, enough to employ linear methods.

7 Future Work

The holy grail of markerless motion capture is a system that can interpret accurately the motion of a human wearing clothing of any description under varying lighting conditions with a camera that is moving and tracking the subject. The system must work in real time and provide feedback regarding the visual accuracy to its users. This system may even try to recognise parts of the motion as identifiable known gestures, and be able to label or interpret accordingly. At the current stage this goal is yet to be achieved.

Significant work has been achieved, but as yet no markerless motion capture system exists that fully encapsulates all aspects of motion capture. The nature of the difficulties involved with optical motion capture has led researchers to produce systems that describe general estimated motion, rather than accurate joint locations through time.

Researchers relax requirements or impose certain assumptions based on the particular application of motion capture on which they are focussing as a final application. Since this report is in collaboration with Televirtual further discussion is needed to identify exact application desires.

This review has been intentionally narrow in its focus. The topics discussed are directed thematically towards the construction and functionality of a vision system. The survey addressing issues in a “how to” manner. It is important to reiterate the necessity in planning research in markerless motion capture of targeting the desired application.

8 Conclusion

This literature survey has attempted to review a representative sample of current research in the field of body analysis by the use of non-intrusive optical systems. It has described fully the term markerless based human motion capture. It has given a brief introduction to a range of potential applications and noted that this large application market has fuelled commercial and academic interest.

The survey chose a taxonomy based on four identified stages involved with markerless based motion capture, namely; initialisation, tracking, pose estimation and gesture recognition. The focus of the report fell on tracking and pose estimation. Tracking and pose estimation were addressed in turn explaining their constituent parts and relationships between them.

The literature shows that various models are used to guide the tracking and pose estimation processes, a review of these models and methods of coping with their high dimensionality was included. The survey has suggested that further discussion regarding the exact application desired would be beneficial to guide and constrain the research therefore giving better chances of success.

9 Appendix I Motion Capture World Industry Review

9.1 Motion Capture Equipment and Software Manufacturers

www.actisystem.fr
www.adaptiveoptics.com
www.arielweb.com
www.bts.it
www.biomechanics-inc.com
www.dnasoft.com
www.hpt-biolink.com
www.mikromak.com
www.mie-uk.com
www.motionanalysis.com
www.ndigital.com
www.peakperform.com
www.phasespace.com
www.ptiphoenix.com
www.qualisys.com
www.vicon.com
www.x-ist.de

9.2 Electromagnetic Tracking Systems

www.ascension-tech.com
www.polhemus.com
www.skilltechnologies.com

9.3 Electromechanical Tracking Systems

www.analogous.com
www.didi.com
www.puppetworks.com

9.4 Hand Tracking Devices

www.5dt.com
www.infusionsystems.com
www.virtex.com

9.5 Other Tracking Systems

www.character-shop.com
www.charnadyn.com

9.6 Motion Capture Service Bureaux and Virtual Content Providers

www.3dcreations.com
www.acclaim.net
www.audiomotion.com
www.biovision.com
www.electrashock.com
www.filmeast.com
www.smstudios.com/fl/fl_index.html
www.moves.com
www.lamb.com
www.medialab3D.com
www.metro3D.com/motion.htm
www.motek.org
www.beam.com.au/mcm
www.pactitle.com
www.pyros.com
www.protozoa.com
www.quantumworks.com
www.rainbo.com
www.sing.com
www.televirtual.com
www.virtualceleb.com
www.waxworks.com

10 References

[1] Aggarwal, J. K. and Q. Cai (1999). “Human Motion Analysis: A Review.” Computer Vision and Image Understanding **73**(3): 428-440.

[2] Aggarwal, J. K., Q. Cai, *et al.* (1998). “Nonrigid Motion Analysis: Articulated and Elastic Motion.” Computer Vision and Image Understanding **70**(2): 142-156.

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

[4] Azarbayejani, A. and A. Pentland (1996). Realtime Selfcalibrating Stereo Person Tracking Using 3d Shape Estimation from Blob Features, M.I.T Media Lab.

[5] Azarbayejani, A., C. Wren, *et al.* (1996). Realtime 3d Tracking of the Human Body, M.I.T Media Laboratory Perceptual Computing Section.

[6] Baumberg, A. M. and D. C. Hogg (1994). An Efficient Method for Contour Tracking Using Active Shape Models. Leeds, University of Leeds, UK. **94.11**.

[7] Bottino, A., A. Laurentini, *et al.* (1998). Towards Non-Instrusive Motion Capture. ACCV, Hong Kong.

[8] Bowden, R., T. Mitchell, *et al.* (1998). “Reconstructing 3d Pose and Motion from a Single Camera View.” BMVC.

[9] Bregler, C. (1997). Learning and Recognising Human Dynamics in Video Sequence. Proceedings IEEE Conference on CVPR, SanJuan, Puerto Rico.

[10] Bregler, C. (1997). Video Motion Capture. Berkeley, Computer Science Division, University of California.

[11] Cai, Q. and J. K. Aggarwal (1999). "Tracking Human Motion in Structured Environments Using a Distributed-Camera System." Ieee Transactions on Pattern Analysis and Machine Intelligence **21**(11): 1241-1247.

[12] Cedras, C. and M. Shah (1995). Motion Based Recognition: A Survey. IEE Proceedings, Image and Vision Computing.

[13] Cootes, T. F., D. Cooper, *et al.* (1995). "Active Shape Models - Their Training and Application." Computer Vision and Image Understanding. **61**(1): 38-59.

[14] Davis, J. W. and A. Bobick (1997). "The Representation and Recognition of Action Using Temporal Templates." International Conference on Computer Vision and Pattern Recognition.

[15] Davis, J. W. and A. F. Bobick (1998). "A Robust Human-Silhouette Extraction Technique for Interactive Virtual Environments." Lecture Notes in Artificial Intelligence **1537**: 12-25.

[16] Deutscher, J., A. Blake, *et al.* (2000). "Articulated Body Motion Capture by Annealed Particle Filtering." Proc. Conf. Computer Vision and Pattern Recognition (CVPR).

[17] Di Bernado, E., L. Goncalves, *et al.* (1996). "Monocular Tracking of the Human Arm in 3d: Real-Time Implementation and Experiments." Internation Conference on Pattern Recognition.

[18] Fua, P. and C. Miccio (1999). “Animated Heads from Ordinary Images: A Leastsquares Approach.” Computer Vision and Image Understanding **75**(3): 247-259.

[19] Fua, P., R. Plänkers, *et al.* (1999). “From Synthesis to Analysis: Fitting Human Animation Models to Image Data.” Computer Graphics Interface.

[20] Fujiyoshi, H. and A. J. Lipton (1998). Real-Time Human Motion Analysis by Image Skeletonization. IEEE Workshop on Applications of Computer Vision (WACV).

[21] Gavrila, D. M. (1999). “The Visual Analysis of Human Movement: A Survey.” Computer Vision and Image Understanding **73**(1): 82-98.

[22] Gavrila, D. M. and L. S. Davis (1996). 3-D Model-Based Tracking of Humans in Action: A Multi-View Approach. Proc. IEEE CS conf. on CVPR, San Francisco.

[23] Haritaoglu, I., D. Harwood, *et al.* (1998). “W 4 : Who? When? Where? What? A Real Time System for Detecting and Tracking People.” International Conference on Face and Gesture Recognition,.

[24] Isard, M. and A. Blake (1998). “Condensation - Conditional Density Propagation for Visual Tracking.” Int. Journal Computer Vision.

[25] Ivanov, Y., A. Bobick, *et al.* (1997). Fast Lighting Independent Background Subtraction, M.I.T Media Laboratory Perceptual Computing Section.

[26] Iwasawa, S., J. Ohya, *et al.* (1998). “Real-Time Estimation of Human Body Postures from Trinocular Images.”.

[27] Johansson, G. (1973). “Visual Perception of Biological Motion and a Model for Its Analysis.” Perception Psychophysics **14**(2): 201-211.

[28] Julier, S. (1995). A New Approach for Filtering Nonlinear Systems. American Control Conference.

[29] Kakadiaris, I. A. and D. Metaxas (1995). 3d Human Body Model Acquisition from Multiple Views. Proceedings of the Fifth International Conference on Computer Vision, Boston, MA.

[30] Kalman, R. E. (1960). “A New Approah to Linear Filtering and Prediction Problems.” Basic Engineering: 35-45.

[31] Kuch, J. J. and T. S. Huang (1995). “Vision Based Hand Modeling and Tracking for Virtual Teleconferencing and Telecollaboration.” ICCV: 666-671.

[32] Lee, C. H., J. S. Kim, *et al.* (1996). “Automatic Human Face Location in a Complex Background Using Motion and Colour Information.” Pattern Recognition **29**(11): 1877-1889.

[33] Lerasle, A. F., G. Rives, *et al.* (1997). “Human Body Limbs Tracking by Multiocular Vision.” Scandinavian Conference on Image Analysis.

[34] Leung, M. K. and Y.-H. Yang (1995). “First Sight: A Human Body Outline Labeling System.” IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(4): 359-377.

[35] Menache, A. (1999). Understanding Motion Capture for Computer Animation and Video Games, Morgan Kaufmann.

[36] Meyer, D., J. Denzlar, *et al.* (1997). “Model Based Extraction of Articulated Objects in Image Sequences.” Fourth Int. Conf. on Image Processing.

[37] Moesland, T. (1999). Computer Vision-Based Human Motion Capture - a Survey. Aalborg, University of Aalborg.

[38] Morris, D. and J. M. Rehg (1998). Singularity Analysis for Articlulated Object Tracking. CVPR, Santa Barbara, CA.

[39] O'Rourke, J. and N. Badler (1980). “Model-Based Image Analysis of Human Motion Using Constraint Propagation.” IEEE Transactions on Pattern Analysis and Machine Intelligence **2**(6): 522-535.

[40] Polana, R. and R. C. Nelson (1994). Low-Level Recognition of Human Motion (or How to Get Your Man without Finding His Body Parts. IEEE Computer Society Workshop on Motion of Nonrigid and Articulate Objects, Austin, TX.

[41] Raja, Y., S. McKenna, *et al.* (1998). Tracking and Segmenting People in Varying Light Conditions Using Colour. FG.

[42] Rehg, J. M. and T. Kanade (1994). Digiteyes: Vision-Based Human Hand Tracking. European Conf. on Computer Vision, Stockholm, Sweden.

[43] Rehg, J. M. and T. Kanade (1995). Visual Tracking of Self-Occluding Articulated Objects. Proc. Fifth International Conf. on Computer Vision, Boston, MA.

[44] Rohr, R. (1997). Human Movement Analysis Based on Explicit Motion Models. Motion-Based Recognition. M. Shah and R. Jain, Kluwer Academic Publishers. **8:** 171-198.

[45] Shutler, J. D., M. S. Nixon, *et al.* (2000). Statistical Gait Recognition Via Velocity Moments. IEE Colloquium: Visual Biometrics.

[46] Vedula, S., S. Baker, *et al.* (1999). Three-Dimensional Scene Flow. Proceedings of the 7th International Conference on Computer Vision.

[47] Wren, C., A. Azarbayejani, *et al.* (1997). "Pfinder: Real-Time Tracking of the Human Body." IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(7): 780-785.

[48] Yaniz, C., J. Rocha, *et al.* (1998). 3d Regular Region Graph for Reconstruction of Human Motion. ECCV 98 Workshop on Perception of Human Action.