

3D MODEL-BASED HUMAN MOTION CAPTURE

LAO WEI LUN
(B. Eng.)

A THESIS SUBMITTED
FOR THE DEGREE OF MASTER OF ENGINEERING
DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE

2005

Acknowledgements

I wish to sincerely thank my supervisors Dr. Alvin Kam, Dr. Tele Tan and Associate Professor Ashraf Kassim for their guidance, encouragement, support, patience, persistence and enthusiasm during the past two years. Their advices, ideas and suggestions on my research and thesis writing are invaluable. Whenever I consulted with them confused, I would afterwards become enlightened, inspired, and enthusiastic. I would also like to thank Dr. Yang Wang and Mr. Zhaolin Cheng for their kindly assistance and help.

I would like to express my deepest appreciation to my parents. Without their unlimited love, it is impossible for me to grow up and make progress ever since. Without the education and support coming from my family members, my development would never have reached this level.

Funding for my research work was made possible through generous grants from Institute for Infocomm Research (I²R). Thanks also for National University of Singapore (NUS) providing me the perfect opportunity to study. They help me fulfill my dream.

Sincerely I would also like to thank my wonderful friends who have, at every step of the way, supported me in the pursuit of the master degree.

Table of Contents

Summary.....	v
List of Tables.....	vii
List of Figures.....	viii
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Main contribution.....	2
1.3 Thesis outline.....	4
Chapter 2 Related Work on 3D Human Motion Analysis	5
2.1 Literature survey on human motion capture.....	5
2.1.1 Approaches without explicit models.....	5
2.1.2 Model based approaches.....	8
2.1.3 Tracking from multiple perspectives.....	12
2.2 Application.....	16
2.3 Motion capture systems.....	17
2.3.1 Magnetic systems.....	18
2.3.2 Mechanical systems.....	18
2.3.3 Optical systems.....	18
Chapter 3 An Overview of Our 3D Model-based Motion Capture System	21
3.1 Methodology	21
3.2 System overview.....	22
3.2.1 Summary.....	22

3.2.2 Camera network.....	22
3.2.3 Camera calibration model.....	23
3.2.4 3D puppet model construction	23
3.2.5 3D puppet pre-positioning.....	24
3.2.6 Model-based tracking.....	25
3.2.7 Data reporting.....	25
Chapter 4 Estimation of Focal Length Self-Calibration	26
4.1 Introduction.....	26
4.2 Related work.....	27
4.3 Background.....	29
4.4 Methodology.....	34
4.4.1 Linearisation of Kruppa’s equations.....	34
4.4.2 Algorithm.....	36
4.5 Experimental results.....	37
4.5.1 Experiments involving a synthetic object.....	38
4.5.2 Experiment involving real images.....	40
4.5.3 3D reconstruction of objects.....	42
4.6 Discussion and future work.....	43
4.7 Conclusion.....	45
Chapter 5 3D Modeling of Human Body	46
5.1 Introduction.....	46
5.2 Related work	47
5.3 Methodology.....	51

5.3.1 Image acquisition.....	52
5.3.2 Camera self-calibration.....	52
5.3.3 Dense correspondences.....	53
5.3.4 3D metric reconstruction.....	54
5.3.5 3D modeling building.....	56
5.4 Experimental results.....	56
5.5 Future work.....	60
5.6 Conclusion.....	63
Chapter 6 3D Human Model Tracking	64
6.1 Introduction	64
6.2 Methodology.....	64
6.2.1 Silhouette extraction	64
6.2.2 Human body model.....	68
6.2.3 Energy function.....	69
6.2.4 Model initialization.....	70
6.2.5 Motion parameter estimation.....	72
6.3 Experimental results.....	74
6.4 Future work.....	79
Chapter 7 Conclusion	81
Reference.....	83

Summary

Human motion capture (mocap) is recently gaining more and more attention in computer graphics and computer vision communities. The demand for a high resolution motion capture system motivates us to develop an unsupervised (i.e. no markers) video-based motion capture system with the aid of high quality 3D human body models.

In this thesis, a practical framework for a 3D model-based human motion capture system is presented. We focus our attention on the self-calibration and 3-D modeling aspects of the system. Firstly, an effective linear self-calibration method for camera focal estimation based on degenerated Kruppa's equations is proposed. The innovation of this method is that using the reasonable assumption that only the camera's focal length is unknown and that its skew factor is zero, the former can be obtained using a closed-form formula without the common requirement for additional motion-generated information. Experimental results demonstrate the robust and accurate performance of the proposed algorithm on synthetic and real images of indoor/outdoor scenes. Secondly, a novel point correspondence-based 3D human modeling scheme from uncalibrated images is proposed. Highly realistic 3D metric reconstruction is demonstrated on uncalibrated images through an automated matching process which does not require the use of any a priori information of or measurements on the human subject and the camera setup. Finally, an effective motion tracking scheme is developed using a novel scheme based on maximising the

overlapping areas between projected 2-D silhouettes of the utilised 3-D model and the foreground segmentation maps of the subject at each camera view.

List of Tables

Table 2.1	Application of motion capture techniques.....	17
Table 2.2	Pros and cons of different mocap systems.....	19
Table 4.1	Focal length estimation in an indoor scene.....	41
Table 4.2	Focal length estimation in an outdoor scene.....	42

List of Figures

Figure 3.1 Block diagram of system.....	22
Figure 4.1 Algorithmic block diagram	37
Figure 4.2 The synthetic object.....	38
Figure 4.3 Relative error of focal length estimation with respect to different Gaussian noise levels.....	39
Figure 4.4 Some images of the indoor scene.....	40
Figure 4.5 Some images of the outdoor scene.....	41
Figure 4.6 3D model reconstruction results (a) An original image of the box to be reconstructed; (b) Rendition of 3D reconstruction (left: side view; right: top view)..	43
Figure 5.1 Block diagram of the methodology of 3D human body modeling.....	51
Figure 5.2 Two images used for the reconstruction in experiment I.....	57
Figure 5.3 Epipolar line aligns with exact location of a feature point.....	58
Figure 5.4 Recovered 3D point cloud of the human body (experiment I).....	58
Figure 5.5 Reconstructed 3D human body model depicted in back-projected colour (experiment I).....	59
Figure 5.6 Two images used for the reconstruction in experiment II.....	59
Figure 5.7 Recovered 3D point cloud of the human body (experiment II).....	60
Figure 5.8 Reconstructed 3D human body model depicted in back-projected colour (experiment II).....	60
Figure 5.9 Example of the pre-defined human skeleton model.....	63
Figure 6.1 Setup of the cameras in the experiment.....	66

Figure 6.2 Silhouette extraction from three cameras.....	67
Figure 6.3 Human body model and the underlying skeletal structure.....	68
Figure 6.4 Measuring the difference between the image (left) and one view of the model (right) by the area occupied by the XORed foreground pixels.....	70
Figure 6.5 Initialization of the human body model.....	71
Figure 6.6 Results of full-body tracking.....	78
Figure 6.7 Free-view rendering of human motion (Frame 3).....	78
Figure 6.8 Free-view rendering of human motion (Frame 12).....	79

Chapter 1 Introduction

1.1 Motivation

Human motion capture was first encountered by Eadweard Muybridge in his famous experiments entitled *Animal Locomotion* in 1887. He is considered to be the father of motion pictures for his work in early film and animation. The study included recording photographs of the subjects, at discrete time intervals, in order to visualise motion. In 1973 psychologist Johansson conducted his famous Moving Light Display (MLD) experiments with the visual perception of biological motion [1]. He attached small reflective markers to the joint locations of human subjects and recorded their motion. The experiment became the first few steps into what is becoming an ever increasingly popular research area: human motion capture.

Human motion capture (mocap) can be defined as the process of recording a human motion event, modeling the captured movement and tracking a number of points, regions or segments corresponding to the movement over time. The goal of the process is to obtain a three-dimensional representation of the motion activity for subsequent analysis.

Mocap, as a research area, is receiving increasing attention from several research communities. Today there is a great interest in the topic of motion capture and the number of papers published in this subject area grows exponentially. Computer vision

researchers, on one hand, are interested in mocap to build models of real-world scenes captured by optical sensors. Computer graphics researchers, on the other hand, are looking at mocap as an attractive and cost-effective way of replicating the movements of human beings or objects for computer-generated productions.

Overall, the growing interest in human motion capture is motivated by a wide spectrum of applications involving automated surveillance, performance analysis, human computer interactions, virtual reality and computer generated animation. Automated surveillance provides the promise of unsupervised tracking of multiple subjects with intelligent detection of activities of interest. Performance analysis meanwhile is extremely useful in the clinical setting of physiotherapy and increasingly, in the field of movement analysis in sports. Understanding human computer interactions is the key in developing next generation man-machine interfaces which are natural and intuitive to use. Virtual reality applications meanwhile will be driven primarily by gaming where more enriched forms of interaction with other participants or objects will be possible by adding gestures, head pose and facial expressions as cues. Finally, computer generated animation, as we all know, is now a big and lucrative industry with its films depicting ever greater realism.

The increasing sophistication of the above applications is pushing the performance envelope of motion capture, specifically towards ever higher resolution. To address the demand for higher resolution motion capture systems, one needs to produce higher quality 3D models in a more automated way. These factors provide the essential motivation for the work presented in this thesis - the development of an un-

aided (i.e. no markers) video-based system that produces high resolution 3D human body models.

1.2 Main Contributions of Thesis

An outline of the main contributions of this thesis is as follows:

1. *A framework for practical optical motion capture is demonstrated*

A structure for practical 3D model-based motion capture is proposed and its implementation demonstrated. The structure comprises of three modules, namely calibration, modeling and tracking. The functionality of each module is defined and its implementation discussed in the thesis. The development tasks involved in the setup of an actual system based on this structure are also addressed. We believe that this motion capture framework provides useful pointers for practical industry implementation or for further research.

2. *A 3D human body modeling scheme based on camera focal length self-calibration is proposed*

We present a novel point correspondence-based scheme that creates accurate 3D shape models of a static human body from a pair of uncalibrated images. The method is based on the assumption that only the camera's focal length is unknown and that the skew factor is zero. The Kruppa's equations are decomposed into one quadratic and two linear equations. Thus the focal length of camera can be obtained in closed form. The advantage of our method is that no a priori information of the human subject or physical measurement on it is required. The procedure is simple, reliable and it

achieves realistic results. The automated matching process on the human body recovers point clouds that can be exported for editing, modeling or animation purposes.

1.3 Thesis Outline

This thesis consists of seven chapters, the organization of each is as follows:

Chapter 1 introduces the motivation, objective, main contributions and outline of the thesis to the readers. A survey of current related work is presented in chapter 2. Chapter 3 briefly explains the functional structure of the practical motion capture system developed. Each module of the structure will be discussed further in the subsequent chapters. Chapter 4 describes and evaluates a linear self-calibration method for camera focal length estimation based on degenerated Kruppa's equations. In chapter 5, we describe the integration of this novel self-calibration technique within the system in the process of developing a novel point correspondence-based scheme for dense 3D human body modeling. The performance of the system in executing human body parts tracking over an entire video sequence is shown in chapter 6. We conclude the thesis in chapter 7.

Chapter 2 Human Motion Capture: A Review

2.1 Literature Survey on Human Motion Capture

This literature survey attempts to present recent developments and current state in the field of body analysis by the use of non-intrusive optical systems. It shows that various mathematical body models are used to guide the tracking and pose estimation processes. In the following sections, we will briefly describe different methods that have been used to extract human motion information without and with explicit models. Tracking from multiple cameras setup is also described afterwards.

2.1.1 Approaches without explicit models

One simple approach to analyse human movements is to describe them in terms of movements of simple low-level 2D features that are extracted from regions of interest. This approach thus translates the problem of human motion analysis to one of joint-connected body parts identification and tracking. The tasks of automatically labeling body segments and locating their connected joints alone are highly non-trivial.

Polana and Nelson's work [2] is an example of point feature tracking. They assumed that the movements of arms and legs converge to those of the torso. Each monitored subject is bounded by a rectangular box, with the centroid of the bounding box being used as the feature to be tracked. Tracking could be done even when there are

occlusions between two subjects as long as the velocity of the centroids of the subjects could be differentiated. The advantage of this approach lies in its simplicity and its use of body motion information to solve the problem of occlusion occurrence during tracking. The approach is however limited by the fact that it considers only 2D translation motion; furthermore, its tracking robustness could be further improved by incorporating additional features such as texture, colour and shape.

Heisele et al. [3] used groups of pixels as basic units for tracking. Pixels are grouped through clustering techniques in a combined color (R, G, B) and spatial (x,y) feature space. The motivation behind the addition of spatial information is the added stability compared to if only colour features are used. Properties of the pixel groups generated are updated from one image to the next using k-means clustering. The fixed number of pixel groups and the enforcement of one-to-one correspondences over time make tracking straightforward. Of course, there is no guarantee that the pixel groups may remain locked onto the same physical entity during tracking but preliminary results of a pedestrian tracking experiment appear promising.

Oren et al. [4] used Haar wavelet coefficients as low-level intensity features for object detection in static images; these coefficients are obtained by applying a differential operator at various locations, scales and orientations on the image grid of interest. During training, one is to select a small subset of coefficients to represent a desired object, based on considerations regarding relative coefficient strength and positional spread over the images of the training set. These wavelet coefficients are then trained on a support vector machine (SVM) classifier. During detection, the SVM classifier

operates on features extracted from window shifts of various sizes over the image and makes decisions on whether a targeted object is present. However, the technique can be only applied to detect front and rear views of pedestrians.

Baumberg and Hogg [5], in contrast, applied active shape models to track pedestrians. Assuming the camera to be stationary, tracking can be initialised on foreground region which is achieved by background subtraction. Moreover, spatial-temporal control can be achieved using a Kalman filter.

Blob representation was used by Pentland and Kauth et al. [6] as the way to extract a compact, structurally meaningful description of multi-spectral satellite (MSS) imagery. Feature vectors of each pixel are first formed by concatenating spatial coordinates to its spectral components. These pixel features are then clustered so that image properties such as color and spatial similarity combine to form coherent connected regions, or “blobs”. Wren et al. [7] similarly explored the use of blob features. In their work, blobs could be any homogenous areas in terms of colour, texture, brightness, motion, shading or any combination of these features. Statistics such as mean and covariance were used to model blob features in both 2D and 3D. The feature vectors of a blob consist of spatial (x, y) and colour (R, G, B) information. A human body is then constructed by blobs representing various body parts such as head, torso, hands, and feet while the surrounding scene is modeled as texture surfaces. Gaussian distributions are assumed for both the human body and the

background scene blob models. For pixels belonging to the human body, different body part blobs are assigned using a log-likelihood measure.

Cheung et al. [8] meanwhile developed a multi-camera system that performed 3D reconstruction and ellipsoid fitting of moving humans in real time. Each camera is connected to a PC which extracts the silhouettes of the moving person in the scene. In this way, the 3D reconstruction is successfully achieved using shape from silhouette techniques. Ellipsoids become an effective tool to fit the reconstructed data.

2.1.2 Model based approaches

For model based approaches of human motion capture, the representation of the human body itself has steadily evolved from stick figures to 2D contours to 3D volumes as models become more complex. The stick figure representation is based on the observation that human motion is essentially the movement of the supporting bone structure while the use of 2D contours is directly associated with the projection of the human figure in images. Volumetric models, such as generalized cones, elliptical cylinders and spheres, meanwhile attempt to describe human body motion details in 3D and require far more parameters. Each of these approaches will be discussed as follows.

2.1.2.1 Stick figure models

Lee and Chen [9] recovered the 3D configuration of a moving subject using its projected 2D images. The method is computationally expensive as it searches all

possible combinations of 3D configurations given the known 2D projection and requires accurate extraction of 2D stick figures. Their model uses 17 line segments and 14 joints to represent the features of the human head, torso, hip, arms and legs. There are at least seven more features on the head, corresponding to the neck, nose, two eyes, two ears and chin, etc. It is assumed that the lengths of all rigid segments and the relative location of the feature points on the head are known in advance. After the feature points of the head are determined, possible locations of feature points for the other subparts can be determined from joint to joint in a transitive manner.

Iwasawa et al. [10] described a novel real-time method which heuristically extracts a human body's significant parts (top of the head and tips of hands and feet) from the silhouette acquired from a thermal image. The method does not need to rely on explicit 3D human models or multiple images from a sequence, and is robust against changes in environmental conditions. The human silhouette, which corresponds to the human body area in the thermal image, is extracted by certain threshold before its center of gravity is obtained from a distance-compensated version of the image. The orientation of the upper half of the body (above the waist) is obtained based on the orientation in the previous frame. Significant points, namely the foot and hand tips and the head top are detected through a heuristic contour analysis of the human silhouette. Before processing can proceed, the very first frame needs to be calibrated. During calibration, the person needs to stand upright and keep both arms horizontal for significant body points to be extracted. For subsequent frames, main joint

positions are estimated based on detected positions of significant points which were obtained using a genetic algorithm based learning procedure.

2.1.2.2 2D Contour models

Leung and Yang [11] applied a 2D ribbon model to recognise poses of a human performing gymnastic movement. A moving edge detection technique is successfully used to generate a complete outline of the moving body. The technique essentially relies on image differencing and coincidence edge accumulation. Coincidence edges, namely edges of both the difference and the original image, capture the edges of moving objects. Faulty coincidence edges however appear when moving objects move behind stationary foreground objects. Effective tracking is used as the means to eliminate erroneous coincidence edges and to estimate motion from the outline of the moving human subject. The motion capture part consists of two major processes: extraction of human outlines and interpretation of human motion. For the first process, a sophisticated 2D ribbon model is applied to explore the structural and shape relationships between the body parts and their associated motion constraints. A spatial-temporal relaxation process is proposed to determine if an extracted 3D ribbon belongs to a part of the body or that of the background. In the end, a description of the body parts is obtained based on the structure and motion consistencies. This work is one of the most complete for human motion capture, covering the entire spectrum from low level segmentation to high level body part labeling.

2.1.2.3 Volumetric models

The disadvantage of 2D models above is its restriction on the camera's angle, so many researchers are trying to depict the geometric structure of human body in more detail using some 3D volumetric models. Rohr [12] applied eigenvector line fitting to outline the human image and then fitted the 2D projections into the 3D human model using a similar distance measure. In the same spirit, Wachter and Nagel [13] also attempted to establish the correspondence between a 3D human model connected by elliptical cones and a real image sequence. Both edge and region information were incorporated in determining body joints, their degrees of freedom (DOFs) and orientations to the camera by an iterated extended Kalman filter.

Generally, works at recovering body pose from more than one camera have met with more success while the problem of recovering 3D figure motion from single camera video has not been solved satisfactorily. Leventon et al. [14] used strong priori knowledge about how humans move. Their priori models were built from examples of 3D human motion and they showed that a priori knowledge dramatically improves 3D reconstructions. They first studied 3D reconstruction in a simplified image rendering domain where Bayesian analysis provided analytic solutions to figural motion estimation from image data. Using insights from this simplified domain, they operated on real images and reconstructed 3D human motions from archived sequences. The system accommodated interactive correction of 2D tracking errors, making 3D reconstruction possible even for difficult film sequences.

An important advantage of volumetric models is its ability to handle occlusion and obtain more significant data for action analysis. However, it is restricted to impractical assumptions of simplicity regardless of the body kinematics constraints, and has high computational complexity as well.

2.1.3 Tracking from multiple perspectives

The disadvantage of tracking human motion from a single view is that the monitored area is relatively small due to the limited field of view of a single camera. One strategy to increase the size of the monitored area is to mount multiple cameras at various locations around the area of interest. As long as the subject is within the area of interest, it will be imaged from at least one of the perspectives of the camera network. Tracking from multiple perspectives also helps solve ambiguities in matching when subject images are occluded from certain viewing angles. However, compared with tracking moving humans from a single view, establishing feature correspondence between images captured from multiple perspectives is more challenging. As object features are recorded from different spatial coordinates, they must be adjusted to the same spatial reference before matching is performed.

Recent work by Cai and Agarwal [15] relied on using multiple points along the medial axis of the subject's upper body as features to be tracked. These points were sparsely sampled and assumed to be independent of each other, thus preserving a certain degree of non-rigidity of the human body. Location and average intensity features of the points were used to find the most likely match between two

consecutive frames imaged from different viewing angles. Camera switching was automatically implemented based on the position and velocity of the subject relative to the viewing cameras. Using a prototype system equipped with three cameras, experimental results of humans tracking within indoor environments demonstrated qualified system performance with potential for real-time implementation. The strength of this approach lies in its comprehensive framework and its relatively low computational cost given the complexity of the problem. However, as the approach relies heavily on the accuracy of the segmentation results, more powerful and sophisticated segmentation methods are needed to improve performance.

Iwasawa et al. [16] used a different approach and proposed a novel real-time method for estimating human postures in 3D using 3 CCD cameras that capture the subject from the top, front and side. The approach was based on an analysis of human silhouettes which were extracted through background subtraction. The centroid of the human silhouette was first obtained followed by the orientation of the upper half of the body above the waist. A heuristic contour analysis scheme was then used to detect representative points of the silhouettes, from which the positions of the major joints were estimated using learning based algorithm. Finally, to reconstruct 3D coordinates of the significant points, the appropriateness of each point within the three camera views were evaluated; two views were then used to calculate its 3D coordinates by triangulation.

Promising results have recently been reported on the use of depth data obtained from stereo cameras for pose estimation [17] [18]. The first attempt at using voxel data obtained from multiple cameras to estimate body pose has been reported in Cheung et al. [19]. A simple six-part body model was used for the 3D voxel reconstruction. Tracking was performed by assigning the voxels in the new frame the closest body part from the previous frame and by re-computing the new position of the body part based on the voxels associated with it. This simple approach however cannot handle two adjacent body parts that drift apart or moderately fast motions. Mikic et al. [20] meanwhile presented an integrated system for automatic acquisition of human motion and motion tracking using input from multiple synchronised video streams. Video frames are first segmented into foreground and background, with the 3D voxel reconstructions of the human body shape in each frame being computed from the foreground silhouettes. These reconstructions are then used as input to the model fitting and tracking algorithms. The human body model used consists of ellipsoids and cylinders and is described using the twists framework, producing a non-redundant set of model parameters. Model fitting starts with a simple body part localisation procedure based on template fitting and growing, which uses a prior knowledge of average body part shapes and dimensions. The initial model is then refined using a Bayesian network that imposes human body proportions onto the body part size estimates. The tracker is exactly an extended Kalman filter that estimates model parameters based on measurements made on the labeled voxel data. A special voxel labeling procedure that can handle large frame-to-frame displacements was finally

used to ensure robust tracking performance. However, voxel-based approaches are restricted to their compulsory requirement of a large number of cameras.

The method presented by Carranza et al. [21] used a detailed body model for motion capture as well as for rendering. Tracking was performed by optimising the overlap between the model silhouette projection and input silhouette images for all camera views. The algorithm is insensitive to inaccuracies in the silhouettes and does not suffer from robustness problems that commonly occur in many feature-based motion capture algorithms. As the fitting procedure works within the image plane only, reconstruction of scene geometry is not required. Indeed, many marker-free video-based motion capture methods impose significant constraints on the allowed body pose or the tractable direction of motion; this system, in comparison, handles a broad range of body movements including fast motions. The motion capture algorithm also makes effective use of modern graphics processors by assigning error metric evaluations to the graphics board.

Grauman et al.'s work [22] involved an image-based method to infer 3D structure parameters using a multi-view shape model. A probabilistic “shape+structure” model was formed using the probability density of multi-view silhouette contours augmented with 3D structure parameters (the 3D locations of key points on an object). Combined with a model of the observation uncertainty of the silhouettes at each camera, a Bayesian estimate of an object's shape and structure was computed. Using a computer graphics model of articulated human bodies, a database of views augmented with the known 3D feature locations (and optionally joint angles, etc.) was

rendered. This is the first work that formulates a multi-view statistical image-based shape model for 3D structure inference. The work also demonstrates how image-based models can be learned from synthetic data, when available. The main strength of the approach lies in the use of a probabilistic multi-view shape model which restricts the object shape and its possible structural configurations to those that are most probable given the object class and the current observation. Thus even when the foreground segmentation results are poor, the statistical model can still infer the appropriate structure parameters. Finally as all computations are performed within the image domain, no model matching or search in 3D space is required.

To summarise the literature survey, human motion capture has come a long way and the knowledge frontier of this domain has advanced tremendously. It is however a fact that the state-of-the-art in human motion capture is still unable to produce a full-body tracker robust enough to handle real-world applications in real time. As a research area, 3D human motion capture and tracking is still far from being mature. Problems such as developing high resolution 3D human models, extracting precise joints position and analysing high-level motion remain largely unsolved.

2.2 Application

There are numbers of promising applications in the motion capture area in computer vision in addition to the general goal of designing a machine capable of interacting intelligently and effortlessly with a human-inhabited environment. The summary of the possible application is listed in Table 2.1.

Table 2.1 Application of motion capture techniques

General domain	Specific area
Virtual reality	<ul style="list-style-type: none"> - Interactive virtual worlds - Games - Virtual studios - Character animation - Teleconferencing (film, advertising, home-use, etc.)
“Smart” surveillance systems	<ul style="list-style-type: none"> - Access control - Parking lots - Supermarkets, department stores - Vending machines, ATMs - Traffic
Advanced user interfaces	<ul style="list-style-type: none"> - Social interfaces - Sign-languages translation - Gesture driven control - Signaling in high-noise environments (airports, factories)
Motion analysis	<ul style="list-style-type: none"> - Content-based indexing of sports video footage - Personalised training in golf, tennis, etc. - Choreography of dance and ballet - Clinical studies of orthopedic pat
Model-based coding	<ul style="list-style-type: none"> - Very low bit-rate video compression

2.3 Existing Motion Capture Systems

Nowadays, three main types of technology underlie most popular commercial human motion capture systems:

2.3.1 Magnetic systems

Magnetic motion capture systems use a source element radiating a magnetic field and small sensors (typically placed on the body of the subject being tracked) that report their position with respect to the source. These systems are multi-source and very complex. They can track a number of points at up to 100 Hz, in ranges from 1 to beyond 5 metres, with accuracy better than 0.25 cm for position and 0.1 degrees for rotation. The two main manufacturers of magnetic mocap equipments are Polhemus (www.polhemus.com) and Ascension (www.ascension-tech.com).

2.3.2 Mechanical Systems

The monitored subject typically wears a mechanical armature fitted to his body. The sensors in a mechanical armature are usually variable resistance potentiometers or digital shaft encoders. These devices encode the rotation of a shaft as a varying voltage (potentiometer) or directly as digital values. The advantage of mechanical mocap systems is that they are free from external interference from magnetic fields and light. The main manufacturer of mechanical mocap equipments is Polhemus (www.polhemus.com).

2.3.3 Optical systems

Existing optical mocap systems utilise reflective or pulsed-LED (infrared) markers attached to joints of the subject's body. Multiple infrared cameras are used to track the markers to obtain the movement of the subject. Post-processing and manual cleaning-up of the movement data are required to overcome errors (e.g. markers

confusion) caused by the tracker. The three main manufacturers of optical mocap equipments are Vicon (www.vicon.com), Peak Performance (www.peakperform.com) and Motion Analysis (www.motionanalysis.com).

The advantages and disadvantages of the three motion capture systems above are listed in Table 2.2.

Table 2.2 Pros and cons of different mocap systems

Systems	Advantages	Disadvantages
Magnetic Systems	<ul style="list-style-type: none"> • Position and rotation are easily measured; • Orientation in space can be determined; • No constraints on tracked subject. 	<ul style="list-style-type: none"> • Distortion proportional to distance from tracked subject; • Noisier data; • Prone to interference from external magnetic fields; • Encumbrance generated by magnetic markers.
Mechanical Systems	<ul style="list-style-type: none"> • Free from external interference from magnetic fields and light 	<ul style="list-style-type: none"> • No awareness of ground, so there can be no jumping, plus feet data tend to “slide”; • Need frequent calibration; • Does not have notion of orientation; • Highly encumbering and range of motion limiting
Optical Systems	<ul style="list-style-type: none"> • Subject free to move as there are no cables connecting body to equipment; 	<ul style="list-style-type: none"> • Prone to light interference; • Self-occlusion of reflective markers; • Offset of reflective markers

	<ul style="list-style-type: none"> • Multiple subjects can be measured at any one time; • Good realism of detected movements. 	<p>from joints and possibility of slippage;</p> <ul style="list-style-type: none"> • Long and expert manual intervention is needed and accuracy is not high enough.
--	---	--

It is interesting to note that human motion capture systems based on multiple cameras have yet to truly take off commercially and still remain in the realm of research for the time being. But as these systems possess most of the advantages of existing commercial systems with little or none of the disadvantages, they hold the greatest promise for flexible, scalable and high quality motion capture for the plethora of applications that are pushing the performance envelope of mocap systems as described in chapter 1.

Chapter 3 An Overview of Our 3D Model-Based Motion Capture System

3.1 Methodology

With the rapid advancement in the fields of 3D computer vision and computer graphics, we now consider the development of a markerless vision system that has the potential to augment present mocap systems.

The task of tracking motion is made more tractable if we can incorporate 3D shape models of the subject as prior knowledge to drive the tracking system. Used extensively in computer vision, this is a very powerful way to control a tracker's stability and robustness.

Our proposed system comprises the following components:

- i. *3D Puppet Model Building* – building a suitable skeleton model of the subject;
- ii. *Model Customization* - devising a technique to customise parameters of the generic puppet model to fit the subject of interest;
- iii. *Model Alignment* - designing an input interface to align the model with the positions of the subject's body parts within the initial video frame;
- iv. *Tracking* - developing a model-based tracker to capture the human motion;

- v. *Data Reporting* - implementing a data-reporting module that displays and analyses the captured motion.

3.2 System Overview

This section provides the overview of our proposed 3D model-based human motion capture system. Details will be further presented in from chapter 4 to chapter 6.

3.2.1 Summary

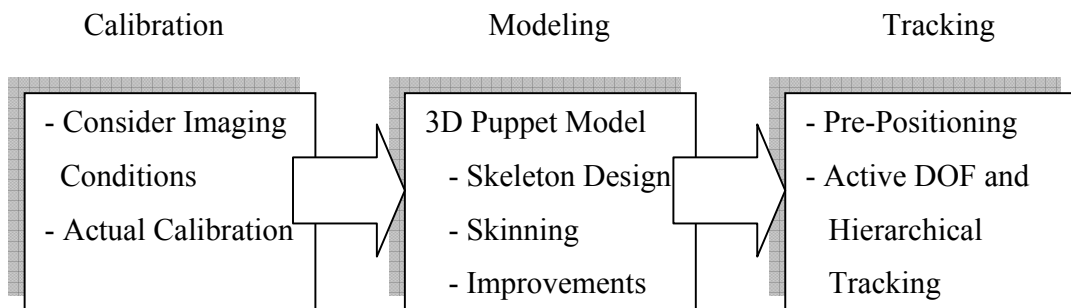


Figure 3.1: Block diagram of system

The block diagram of the proposed system is shown in Figure 3.1. Implementation of the system requires effective operation of three sub-systems:

- i. A calibration sub-system factoring in the imaging conditions.
- ii. A modeling sub-system which first builds a generic 3D puppet model and refining it based on anthropometrical data.
- iii. A tracking system which first pre-positions the puppet model during initiation and follows the subject's movements thereafter.

3.2.2 Camera network

The camera network comprises a set of fixed cameras (minimum three) that are arranged to maximize the view coverage of the subject. A PC is used to host the video capture card (which handles A/D conversion and image rendering) as well as the processing software.

Off-the-shelf standard CCD cameras are used in contrast with pulsed infrared cameras used by most existing mocap system. The frame capture rate can be set at 25 frames per second, unless there is an explicit need to capture at higher frame rates. From our experience, a sufficiently high shutter speed (at least 1/500) is needed to obtain crisp images of very rapid movements. A gen-lick circuit helps synchronise the multiple cameras.

3.2.3 Camera calibration

Camera calibration is needed before subsequent processing can take place. Information that is needed at this stage is the 3D-2D correspondences which can be obtained using the 3D extrinsic and intrinsic camera parameters. A pinhole camera model is assumed. Details will be described in section 4.4. In our proposed self-calibration scheme conventional calibration tools are no longer needed.

3.2.4 3D puppet model construction

The purpose of constructing a 3D puppet model is to precisely mimic the behavior of the subject so that robust quantitative data about the subject's movements can be obtained. The unique property of the proposed model is that it comprises of two

components: the underlying skeleton and the external skin. These two components interact with each other, producing an integrative model that best represents the subject.

The puppet model construction uses a generic human puppet provided by computer animation software, such as 3D Studio MAX, as its starting point. There are two approaches to further proceed. One alternative is to take anatomical measurements of the subject to parameterise the generic puppet model; this approach is called renormalisation. Renormalisation needs to be performed on both the skeleton and the skin. There are obvious problems with renormalising the skeleton component of the model because skeletons cannot be observed and measured directly, and thus intelligently estimated. The other alternative, the so-called image-based approach, uses a correspondence point scheme operating on multiple captured images of the subject to parameterise the 3D puppet model. This is a more flexible approach and the one we have chosen, details of which will be presented in chapter 5.

3.2.5 3D Puppet pre-positioning

This step is needed to initialise the positions of the different parts of the 3D puppet model to the positions of the subject's corresponding body parts. Pre-positioning is typically implemented in a manual or semi-automatic way. Automated tracking of the various body parts of the subject can only take place after proper pre-positioning is achieved for the first image frame. An interactive software interface is about to be developed to facilitate model pre-positioning in an intuitive way.

3.2.6 Model-Based Tracking

An articulated 3D human model is utilised to drive the body feature detection and movement tracking tasks. 3D tracking is then performed using an analysis by synthesis approach that guarantees stable and accurate performance over extended periods of time.

3.2.7 Data Reporting

3D rendering of the tracked skeleton and body surface should be performed with the following features displayed:

- i. Segments of the body joints and the centre of mass; for position, displacement, velocity, acceleration and orientation analysis.
- ii. Rotational motions, including their angular velocity and acceleration.
- iii. Orthopedic angles for all body joints; for analysis or graphing.
- iv. Forces and moments on each body joint.

Chapter 4 Self-Calibration for Focal Length Estimation

4.1 Introduction

Camera calibration is the first module of our proposed human motion capture system; its role is to estimate the metric information of the camera. In other words, the module attempts to establish the relationship between the camera's internal coordinate system and the coordinate system of the real world. It is therefore the logical first step for a calibrated motion capture system. Camera calibration can generally be achieved using two approaches: photogrammetric [23] and self-calibration [24].

A photogrammetric calibration approach uses a precise pattern with known metric information to calibrate a camera. This pattern is usually distributed over two or three orthogonal planes. Since the metric information of this pattern can be specified with precision, calibration could be done to a high degree of accuracy. Generation of such precise patterns, however, is often expensive and unfeasible for some applications. Furthermore, there may be cases where the camera metric information changes with time.

Camera self-calibration addresses this exact need. The approach automatically calibrates a camera without the need for any *à priori* 3D information. The main

principle is to extract a camera's intrinsic parameters (from which metric information could be derived) from several uncalibrated images captured by the camera. A self-calibration approach typically relies on two or more captured images as the input and produces the camera's intrinsic parameters as the output.

In this chapter, we propose a linear self-calibration approach for camera focal length estimation based on degenerated Kruppa's Equations. Compared with other linear techniques, this method does not require any a priori information generated from motion. By using the degenerated equations (one quadratic and two linear) and making reasonable assumptions that only the focal length of the camera is unknown and that its skew factor is zero, the focal length can be calculated from a closed form formula. We will demonstrate the accuracy of this approach through experimental results based on both synthetic and real images of indoor and outdoor scenes and its effectiveness for 3D object reconstruction.

4.2 Related Work

Traditional camera self-calibration is highly non-linear as the constraint for a camera's internal parameter matrix is quadratic [24]. As the solution for such non-linear optimisation often falls into local minima, conventional approaches for camera self-calibration are often unsatisfactory [25].

Due to the above difficulty, there have been attempts to perform self-calibration using controlled motions of a camera. In [26], for example, self-calibration relies on a pure translational camera motion. As this approach makes it possible to derive much

information through a long duration of sequence capture, the estimation of the camera calibration matrix can be fairly robust. Furthermore, as the equation that constrains the calibration matrix is linear, most general linear models can be used. In [27], self-calibration using both translational and rotational camera motions is considered. Since the implementation of the approach involves the use of a robot, the motion and orientation of the cameras can be accurately controlled. Self-calibration relying on pure rotational camera motion is described in [28]. In this case, point correspondences between two images are achieved through a conjugate of a rotation matrix, with the camera intrinsic parameter matrix being one such conjugating element. Consequently, eight point matches are enough to obtain the intrinsic parameter matrix.

Recently, some other efforts have been made to linearise the Kruppa's equations which constrain the camera intrinsic parameter matrix. In [29], Ma finds the constant scaling factor for Kruppa's equations for two special motion cases. When translation is parallel to the rotation axis, the constant scale factor is given by the two norms of the conjugate of normalised epipoles. The conjugate factor here is the fundamental matrix. In the other case when translation is perpendicular to the rotation axis, the scaling factor is determined by one of the non-zero eigenvalues of the product between the normalised epipoles and the fundamental matrix. These constant scaling factors help provide the linear constraint for the camera intrinsic parameters.

When some assumptions are made regarding the camera intrinsic parameters, closed form calibration equations could be obtained [30]. Making the reasonable assumptions that only the focal length of the camera's lens is unknown and that its

skew factor is zero, we can degenerate Kruppa's equations to one quadratic and two linear equations without any à priori information from camera motion. The methodology of this approach will be explained in fuller detail and with more rigorous derivation and evaluation, compared to brief explanations of the method in our previous publication in [31]. Complementing experimental results are critical discussions and analysis of future work followed by concluding remarks.

4.3 Background

The pinhole camera model is widely used in computer vision area. It maps 3D projective spaces to 2D ones. In this model, the camera frame is determined by a coordinate system whose origin is the optical center of a camera and with one axis (usually the z axis) being parallel to the optical axis. The other two axes (x and y axes) are on planes orthogonal to the z axis. The two axes of captured image's coordinate system are often assumed parallel to the x and y axes of the camera frame if optical distortion is ignored. Let a 3D point in the camera frame and its correspondent image be $M = (x, y, z)^T$ and $m = (u, v)^T$ respectively. The relationship between M and m is:

$$\frac{u}{x} = \frac{v}{y} = \frac{f}{z} \quad (4.1)$$

where f is the focal length of the camera's lens. In the same way, let $\tilde{M} = (x, y, z, 1)^T$ and $\tilde{m} = (u, v, 1)^T$ denote the 3D homogeneous coordinates of M and m respectively.

(4.1) can then be denoted as:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (4.2)$$

with s being a scaling factor and $A = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix}$ the camera intrinsic parameter matrix.

If the origin of the image coordinate system is not the image center and lens distortion

is taken into account, the intrinsic parameter matrix becomes $A' = \begin{bmatrix} \alpha f & \gamma & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix}$. In

this case, $[u_0 \ v_0]^T$ is called the *principal point*, the point of intersection between the optical axes and the image plane. α meanwhile is the *aspect ratio* determining the extent of unequal sampling along the u and v directions, while γ is the *skew factor* corresponding to a skewing of the coordinate axes. For a real CCD camera, it is unlikely that there is any unequal sampling or that the skew factor is anything other than 0. It is thus quite reasonable to assume that $\gamma = 0$ and $\alpha = 1$. In most cases where high precision is not needed, it is also safe to assume the image center to be the location of principal point. Note that if we first subtract u_0 and v_0 from the image coordinates, equation (4.2) still holds.

In the case of a binocular stereo system, the coordinate system is usually set on one of the two camera frames. Let t and R denote the 3D translation vector and the rotation matrix respectively of the other camera with respect to the first; the two camera projection matrices, P_1 and P_2 can be written as:

$$P_1 = A[I | 0] \quad P_2 = A[R | t]$$

where I is the 3×3 identity matrix, 0 is the 3D zero vector and A is the cameras' intrinsic parameter matrix as defined earlier.

According to epipolar geometry, the relation of a pair of projection of the same 3D point is given by:

$$\tilde{m}_2^T F \tilde{m}_1 = 0 \quad (4.3)$$

where F denotes the fundamental matrix, $[t]_{\times}$ the skew symmetric matrix of translation vector t , with $F = A^{-T} [t]_{\times} R A^{-1}$ if we assume that the two cameras have identical intrinsic parameters.

Following the above notation, we have $s_1 \tilde{m}_1 = P_1 \tilde{M}$ and $s_2 \tilde{m}_2 = P_2 \tilde{M}$ where s_1 and s_2 are scalar factors. The general form of the fundamental matrix in terms of the projection matrix then becomes $F = s [e_2]_{\times} P_2 P_1^+$ where s is a scalar factor, "+" denotes the pseudo inverse and $P_2 P_1^+ = A R A^{-1}$. Then it is easy to produce the Kruppa's equation as:

$$F C F^T = s [e_2]_{\times} C [e_2]_{\times} = s [e_2]_{\times} A A^T [e_2]_{\times} \quad (4.4)$$

where e_2 is the right epipole, and as a reminder, F the fundamental matrix and A the camera intrinsic parameter matrix. $C = A A^T$ represents the dual image of the absolute conic (DIAC).

Kruppa's equations are intimately connected with the absolute conic. The properties of this conic and its connection with calibration are now briefly introduced for the benefit of the reader. The absolute conic is a conic lying on the plane at infinity,

represented by the equation $x^2 + y^2 + z^2 = 0$. The absolute conic does not contain any points with real coordinates as it is composed entirely of complex points. The image of the absolute conic in an image is however representable by a real symmetric 3×3 matrix.

Let a point x be $(x, y, z)^T$. Its homogeneous coordinate system point $(x, y, z, 0)^T$ is on the absolute conic if and only if $x^T x = 0$. Consider a camera projection matrix $P = A[R | -Rt]$. The point $(x, y, z, 0)^T$ on the absolute conic maps to $u = P(x, y, z, 0)^T = ARx$. Thus, $x = R^T A^{-1} u$, and the condition $x^T x = 0$ becomes $u^T A^{-T} R R^T A^{-1} u = u^T A^{-T} A^{-1} u = 0$. Thus, a point u is on the image of the absolute conic if and only if it lies on the conic represented by the matrix $A^{-T} A^{-1}$. In other words, $A^{-T} A^{-1}$ is the matrix representing the image of the absolute conic. Taking inverses (dual conics) reveals that AA^T is the dual image of the absolute conic. We will denote AA^T by C . If C is known then the calibration matrix A may be retrieved by Choleski factorization.

We have shown how the calibration matrix A may be retrieved if the matrix C representing the DIAC is known. Conversely, if A is known, then $C = AA^T$ depends only on the calibration matrix, and not on the orientation R or the position t of the camera. The DIAC is fixed under Euclidean motions of the camera.

Equation (4.4) gives the general constraint for camera intrinsic parameter matrix. Since C is a symmetric matrix, (4.4) provides six equations to compute C . However, at most only two equations are independent. These six equations are quadratic in

terms of C and thus fourth order in terms of elements of A . The common way to compute A is to first obtain C by iterative estimation and to obtain A next by Cholesky decomposition.

The objective of using the above derivation is to determine which two equations are independent. Hartley [32] decomposes Kruppa's equations using singular value decomposition (SVD) into two equations.

Suppose the singular value decomposition (SVD) of the fundamental matrix is $F = U\Sigma V^T$ where $U = [u_1 \ u_2 \ u_3]$, $V = [v_1 \ v_2 \ v_3]$, and $\Sigma = \text{diag}(a, b, 0)$, a and b are the two singular values of F . Since the right epipole $e_2 = \text{null}(F^T) = u_3$, the null space of F^T , then $[e_2]_{\times} = UMU^T$, where M is the skew symmetric matrix for $[0 \ 0 \ 1]^T$.

Hence after some mathematical manipulations, the Kruppa's equations (4.4) becomes

$$\Sigma V^T C V \Sigma = s M U^T C U M \quad (4.5)$$

The LHS of (4.5) is

$$\Sigma V^T C V \Sigma = \begin{bmatrix} a & & \\ & b & \\ & & 0 \end{bmatrix} V^T C V \begin{bmatrix} a & & \\ & b & \\ & & 0 \end{bmatrix} = \begin{bmatrix} a^2 v_1^T C v_1 & a b v_1^T C v_2 & 0 \\ a b v_1^T C v_2 & b^2 v_2^T C v_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (4.6)$$

The RHS of (4.5) is

$$s M U^T C U M = s [u_2 \ -u_1 \ 0]^T C [u_2 \ -u_1 \ 0] = s \begin{bmatrix} u_2^T C u_2 & -u_2^T C u_1 & 0 \\ -u_1^T C u_2 & u_1^T C u_1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (4.7)$$

Only four elements in the matrix at the right end of both (4.6) and (4.7) are nonzero and (4.6) is a symmetry matrix, so (4.5) gives only two equations for constraining A .

The final result is:

$$\frac{a^2 v_1^T C v_1}{u_2^T C u_2} = \frac{b^2 v_2^T C v_2}{u_1^T C u_1} = \frac{ab v_1^T C v_2}{-u_2^T C u_1} \quad (4.8)$$

In this chapter, equation (4.8) is fundamental in calibrating the camera. Since (4.8) provides two independent equations and the number of unknown parameters in C is five, at least three images are needed to obtain C .

4.4 Methodology

4.4.1 Linearisation of Kruppa's equations

The simplified form of Kruppa's equations may be generated based on equation (4.8). The common method to obtain C is to initially estimate its parameters, then conduct non-linear optimisation methods to obtain more accurate results. Although equation (4.8) provides neat constraints on the camera's intrinsic parameter matrix, it is not easy to solve for these parameters as multiple solutions exist for these quadratic equations. To illustrate: generally, three fundamental matrices or three images are needed to fully calibrate a camera. However, these three images represent six quadratic constraints. It is difficult to know whether these six constraints are independent. Even if they are actually independent, solutions from any five of the six constraints could lead to a total of $2^5 = 32$ possible solutions. We have to eliminate spurious solutions one by one. Thus it is not a particular promising approach.

If we make the reasonable assumptions that only the focal length of the camera's lens is unknown (but constant) and that its skew factor is zero, most of these complications disappear. Using these assumptions and with simple coordinate transformations, Kruppa's equations in (4.8) can be further linearised. Specifically,

we transform the image coordinates to reflect our assumptions of the aspect ratio being one and the principal point being at the image center. The new fundamental matrix is now $F' = T^{-T}FT^{-1}$, where F' and F are the new and the original fundamental matrices respectively, and T the transformation matrix. Consider the singular value decomposition of F' , i.e. $F' = USV^T$, equation (4.8) now yields:

$$\frac{a'^2 v_1^T \text{diag}(f^2, f^2, 1)v_1}{u_2^T \text{diag}(f^2, f^2, 1)u_2} = \frac{b'^2 v_2^T \text{diag}(f^2, f^2, 1)v_2}{u_1^T \text{diag}(f^2, f^2, 1)u_1} = \frac{a'b'v_1^T \text{diag}(f^2, f^2, 1)v_2}{-u_2^T \text{diag}(f^2, f^2, 1)u_1} \quad (4.9)$$

where a' , b' are the two singular values of F' and f the focal length of the camera, u_i and v_i are the i th and j th column of U and V respectively.

Expanding the above equations, we further obtain:

$$\frac{a'^2 (v_{11}^2 f^2 + v_{12}^2 f^2 + v_{13}^2)}{u_{21}^2 f^2 + u_{22}^2 f^2 + u_{23}^2} = \frac{b'^2 (v_{21}^2 f^2 + v_{22}^2 f^2 + v_{23}^2)}{u_{11}^2 f^2 + u_{12}^2 f^2 + u_{13}^2} = \frac{a'b'(v_{11}v_{21}f^2 + v_{12}v_{22}f^2 + v_{13}v_{23})}{u_{11}u_{21}f^2 + u_{12}u_{22}f^2 + u_{13}u_{23}} \quad (4.10)$$

Where u_{ij} and v_{ij} are the j th element of the vector u_i and v_i respectively.

Because of the orthogonality of U and V , the three fractions are rewritten as:

$$\frac{a'^2 (1 - v_{13}^2) f^2 + a'^2 v_{13}^2}{(1 - u_{23}^2) f^2 + u_{23}^2} = \frac{(1 - u_{13}^2)(1 - v_{23}^2) f^2 + b'^2 v_{23}^2}{(1 - u_{13}^2) f^2 + u_{13}^2} = -\frac{a'b'v_{13}v_{23}}{u_{23}u_{13}} = s \quad (4.11)$$

where s is a constant scalar factor. After rearranging equation (4.10), we obtain the same results of Sturm's work [33] as shown in the following two linear equations:

$$f^2 [a'u_{13}u_{23}(1 - v_{13}^2) + b'v_{13}v_{23}(1 - u_{23}^2)] + u_{23}v_{13}(a'u_{13}v_{13} + b'u_{23}v_{23}) = 0 \quad (4.12)$$

$$f^2 [a'v_{13}v_{23}(1 - u_{13}^2) + b'u_{13}u_{23}(1 - v_{23}^2)] + u_{13}v_{23}(a'u_{13}v_{13} + b'u_{23}v_{23}) = 0 \quad (4.13)$$

and one quadratic equation:

$$f^4[a'^2(1-u_{13}^2)(1-v_{13}^2)-b'^2(1-u_{23}^2)(1-v_{23}^2)] + f^2[a'^2(u_{13}^2+v_{13}^2-2u_{13}^2v_{13}^2)-b'^2(u_{23}^2+v_{23}^2-2u_{23}^2v_{23}^2)] + (a'^2u_{13}^2v_{13}^2-b'^2u_{23}^2v_{23}^2) = 0 \quad (4.14)$$

From the above derivation, it is clear that when only the focal length is unknown, the Kruppa's equations can be further decomposed into one quadratic and two linear equations. Here we denote three parameters in the quadratic equation (4.14) are

$$c_1 = a'^2(1-u_{13}^2)(1-v_{13}^2) - b'^2(1-u_{23}^2)(1-v_{23}^2) \quad ,$$

$$c_2 = a'^2(u_{13}^2+v_{13}^2-2u_{13}^2v_{13}^2) - b'^2(u_{23}^2+v_{23}^2-2u_{23}^2v_{23}^2) \quad \text{and} \quad c_3 = a'^2u_{13}^2v_{13}^2 - b'^2u_{23}^2v_{23}^2$$

respectively. If $c_1^2/(c_2^2+c_3^2) < 0.001$, we can calculate the focal length from linear equations (4.12) and (4.13). Otherwise, the focal length may be thereby calculated by solving the quadratic equation (4.14). The spurious solution can be singled out due to the truth that the focal length should be positive and within certain range.

4.4.2 Algorithm

The block diagram of the developed algorithm is shown in Figure 4.1. It basically comprises three steps: feature detection, robust matching estimation and self-calibration. As the first step toward image correspondence, Harris corner detection [34] is applied on a pair of images to detect feature points of interest. Next, a robust matching estimation technique [35] is implemented with the aim of finding sufficient corresponding points between the two images. The routine consists of cross-correlation matching, relaxation, RANSAC fitting and Least Median Square (LMedS) optimisation [36] for epipolar geometry estimation. These correspondences can also

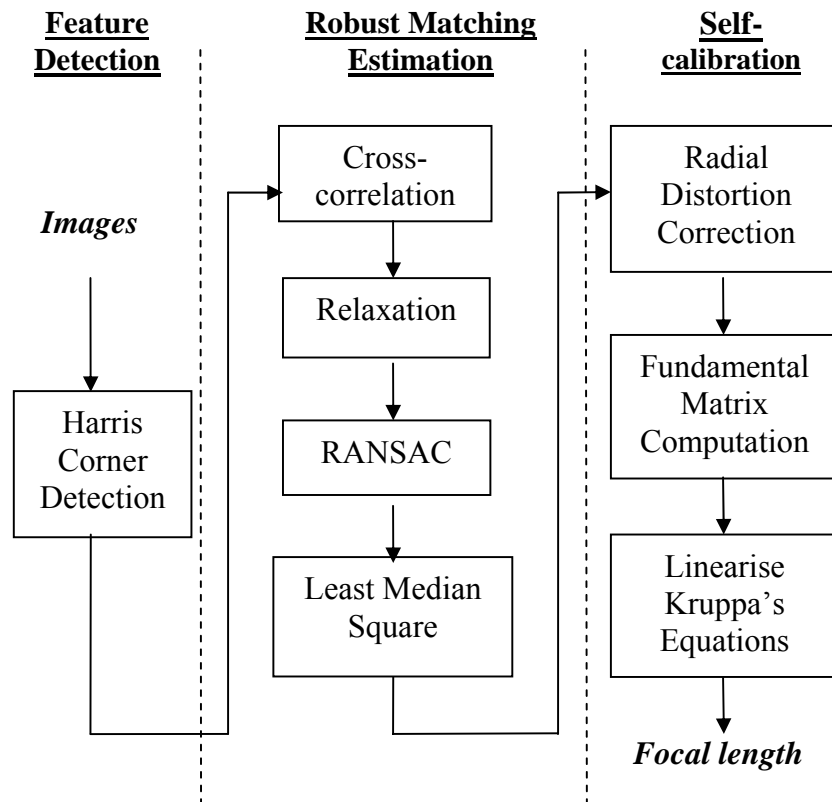


Figure 4.1 Algorithmic block diagram

be established manually especially when the base line of the two images is long. For self-calibration, it is important to note that if radial distortion is significant, it should be corrected at this stage. After any needed distortion correction, the fundamental matrix for each pair of images is obtained. Lastly, the focal length of the camera is estimated using the linearised Kruppa's equations.

4.5 Experimental Results

We have carried out a large number of experiments to study the performance of the algorithm and examine its robustness and accuracy. Section 4.5.1 presents the results

of experiments using a synthetic object. Experiments on real images are conducted next and their results are reported in Section 4.5.2. Finally, the utility of the algorithm in performing a 3D reconstruction of a real object is demonstrated in Section 4.5.3.

4.5.1 Experiments involving a synthetic object

4.5.1.1 Synthetic object and image

The configuration of this experiment is shown in Figure 4.2. The synthetic ‘object’ is a composition of points on two planar grids at a 135° angle with each other. There are a total of 120 points within each grid. The object is placed at a distance of 1000 units from the camera center. The ground truth of the camera's intrinsic parameters is: $f = 600$, $u_0 = 320$, $v_0 = 240$, and the skew factor $\gamma = 0$.

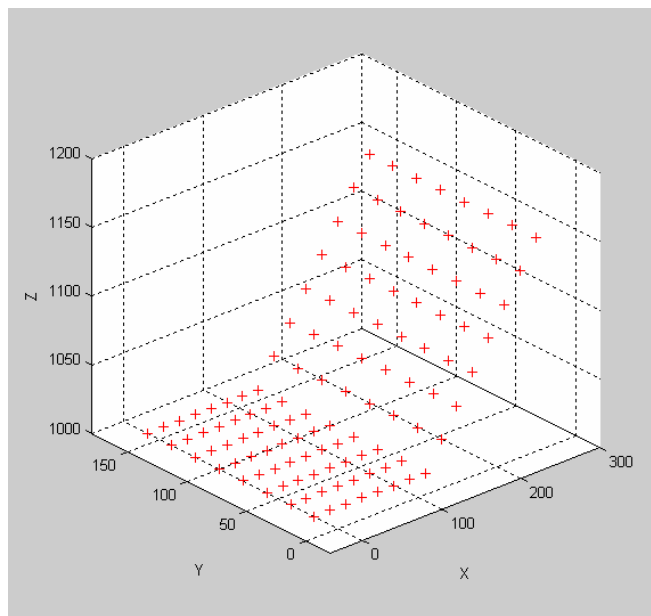


Figure 4.2 The synthetic object

4.5.1.2 Performance with respect to different levels of Gaussian noise

It can be shown that the coplanarity of optical axes is the singularity of the two views calibration algorithm [33]. In our design therefore, the optical axes of our two-camera system are never exactly coplanar. For this experiment, the image coordinates of the grid points are perturbed by independent Gaussian noise with mean of 0 and a standard deviation of σ pixels. σ varies from 0.1 to 2.0 pixel. For each noise level, a total of 100 trials are performed; in other words, there are 50 calibration results for each σ value. The average of these 50 estimations is then taken as the estimated calibrated focal length with respect to the noise level. Finally, this estimated focal length value is compared with the ground truth for further analysis.

The relation between the relative error of the focal length estimation and noise level is shown in Figure 4.3. It can be seen that the relative error of focal length is quite low (mostly around 2%) and it increases slowly with the noise level.

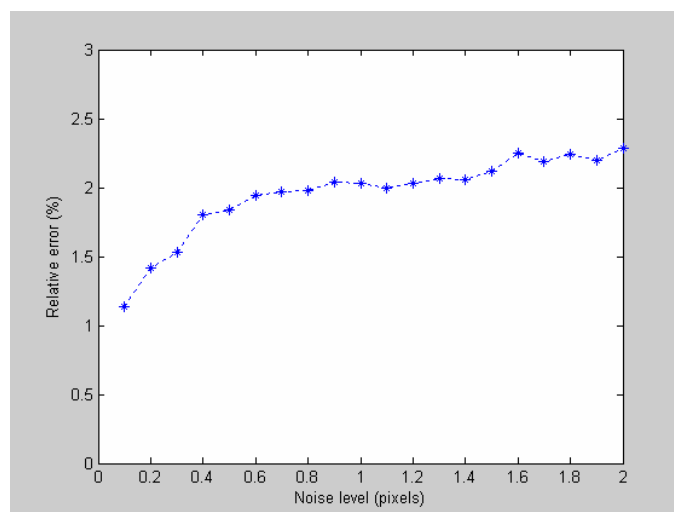


Figure 4.3 Relative error of focal length estimation with respect to different Gaussian noise levels

4.5.2 Experiment involving real images

4.5.2.1 Camera setup

A Canon A75 digital camera is used to capture images at 640×480 resolution. The camera is first calibrated using a photogrammetric calibration algorithm [23] with four images of a highly accurate calibration grid. Using this approach, the calculated camera focal length of 692 pixels is employed as the “ground truth” for the following experiments.

4.5.2.2 Calibration with real images of an indoor scene

For this experiment, one cup and one box are placed together. We can move the camera close to the objects in order to capture enough features. A total of four images of the subjects are taken. Two of them are shown in Figure 4.4 with the calibrated results presented in Table 4.1. Entries in the “image pair used” column in Table 4.1 (and the subsequent Table 4.2 is in the same layout) refer to particular image pairs used as inputs to the calibration algorithm. From the results in Table 4.1, we can see that the maximum relative error is about 32 pixels or about 4.6%.



Figure 4.4 Some images of the indoor scene

Table 4.1 Focal length estimation in an indoor scene

Focal length ground truth	Image pair used				
	1&2	1&3	1&4	2&4	3&4
692.0	670.1	702.5	660.4	680.4	708.1

4.5.2.3 Calibration with real images of outdoor scene

We used the same camera setting to take four images of an outdoor scene. Two of them are shown in Figure 4.5; the scene being a building. The focal length estimated results based on the scene are presented in Table 4.2. Although the same camera setting is used, the outdoor scene is much more complex than the earlier indoor one. Hence we can see that the maximum relative error is up to about 77 pixels, or about 11.1%. In Figure 4.5, we find that the building is of a large depth variety. It may lead to serious problems for matching (A little displacement in an image contributes to a great a displacement in the 3D world).



Figure 4.5 Some images of the outdoor scene

Table 4.2 Focal length estimation in an outdoor scene

Focal length ground truth	Image pair used				
	1&2	1&3	1&4	2&3	2&4
692.0	721.5	696.4	631.9	719.5	769.1

Therefore, variation in object depths is one main factor affecting the accuracy of the calibrated results.

4.5.3 3D reconstruction of objects

The application of the self-calibration focal length estimation algorithm can be demonstrated by using it to perform a 3D reconstruction of a paper box. Firstly, we implement the techniques described in [37] to recover the scene's structure. A triangular mesh is then semi-automatically adjusted to the reconstructed 3D points and used to create a textured VRML model.

The final 3D rendition of the reconstructed object is shown in Figure 4.6. Due to the sparseness of extracted points of interest, there are some triangular meshes across the ridge of the box. Some visual artefacts on the ridge are inevitably produced. Despite these limitations, the 3D reconstruction is qualitatively correct. For example, the top view of the reconstructed result as depicted in figure 4.6(b) indicates that the two planes of the box form an approximate 90-degree angle corresponding to the truth.



(a) An original image of the box to be reconstructed



(b) Rendition of 3D reconstruction (left: side view; right: top view)

Figure 4.6 3D model reconstruction results

4.6 Discussion and Future Work

In the above experiments, we first demonstrate the robustness of our self-calibration focal length estimation algorithm on a synthetic image in the presence of Gaussian noise. Next, we show that the algorithm can be used for both real indoor and outdoor scenes although performance of the algorithm for the latter suffered some degradation. Although the overall results are probably not as good as those obtained using traditional photogrammetric calibration method involving calibration grids [23], the

accuracy of the calibrated estimation is still reasonable given its ease of implementation. We believe that the algorithm will help increase the possibilities for applications requiring automatic structure from motion. Finally, we demonstrated the algorithm's application for 3D object reconstruction.

As for future work in the area, firstly, although the self-calibration algorithm works well, embedding bundle adjustment techniques [38] in our algorithm could increase the estimation accuracy of the camera's intrinsic parameters. To ensure stable results, singularities in the case of coplanar optical axes must be avoided. Automated detection and prompting when two input images are close to generating generic singularities [33] could be added. One practical solution to avoid singularities is: after taking the first image, face the camera to the same point in the scene and tilt the camera slightly upwards or downwards before capturing the second image.

Secondly, the number of correspondence matches obtained using epipolar geometry estimation is still limited. One can however perform dense matching after the epipolar geometry is established. This is an important future work for more realistic 3D object reconstruction.

In addition to future work of somewhat incremental nature above, one can also question the fundamental assumptions made by the algorithm. For example, the focal length of the camera may not be constant. For added adaptivity and intelligence, camera zoom and the different camera focus operations need to somehow be taken into account during the calibration process. Finally, one should not forget that 3D modeling, not calibration, is the ultimate objective. Problems should thus be

evaluated from a systematic perspective: this includes considering the interdependency of tasks such as image feature extraction, correspondence matching, camera self-calibration, structure from motion and dense model reconstruction.

4.7 Conclusion

This chapter presents and evaluates a new linear approach of self-calibration for camera focal length estimation. The method is based on the reasonable assumptions that among the camera's intrinsic parameters, only the focal length is unknown and its skew factor is zero. In this case, the Kruppa's equations, which are popularly used to self-calibrate a camera, are shown to decompose into one quadratic and two linear equations. The first advantage of these equations is that they may produce closed form solutions. The second advantage is that the common requirement of à priori information generated by camera motion is no longer needed. Our experimental results demonstrate the robust and accurate performance of the proposed algorithm on a synthetic image and real images of indoor/outdoor scenes. Finally, the algorithm's application for 3D object reconstruction is shown. In the next chapter, we will apply techniques developed here for 3D human body modeling.

Chapter 5 3D Human Body Modeling

5.1 Introduction

Recent advances in *human body modeling*, which is evolving fast, promises to open a wide variety of new applications, particularly those that require 3D information of the human body. Examples of such applications [39] include fitting of virtual clothes, anatomical medical diagnosis and drug therapy assessment, virtual actors in film and video post-production, ergonomic workspace design and many others. In short, 3D human body modeling has great potential in many applications that benefit from a digital 3D model of a human being.

The generation of 3-D human body models from uncalibrated image sequences remains a challenging problem although it has been actively investigated in recent years. An area of particular interest is the modeling of real human individuals. In this chapter, we present a method for 3-D reconstruction of static human body parts using images acquired from a single digital camera. Based on the focal length self-calibration approach described in chapter 4, a point correspondence-based scheme to build a dense 3D human body model is demonstrated. The method involves the following steps: image acquisition, camera self-calibration, dense point correspondence, metric reconstruction and 3D model building which is visualised as a 3-D point ‘cloud’. The goal of the work presented here is to extract complete 3-D

surface information of a human being without any à priori information about the camera system or the person involved.

5.2 Related Work

In computer graphics there is a relentless pursuit of ever more realistic modeling of human body geometries and human motions for applications [39] like gaming, virtual reality and computer animation that demand highly realistic Human Body Models (HBMs). At present, the process of generating realistic human models is still very human labour intensive and so their application is therefore currently limited to the lucrative movie industry where HBMs' movements are predefined, well studied and painstakingly manually produced. Fully automatic rendition of highly realistic and fully configurable HBMs is still an open research problem. A major constraint involved is the computational complexity to produce realistic models with 'natural behaviors'.

Lately, a computer vision approach [40] [41] is increasingly being used for automatic generation of HBMs, processing video captured image sequences by incorporating and exploiting prior knowledge of human appearance. In contrast to computer graphics, computer vision approaches concentrate more on efficient rather than accurate models for human body modeling. The challenge is to improve the accuracy of the computer vision based HBMs with the objective of realizing fully automated rendition of highly realistic and fully reconfigurable HBMs.

Different 3D representations and mathematical formalisms have been proposed in the literature to model both the structure and movements of a human body. An HBM can be generally represented as a chain of rigid bodies, called links, interconnected to one another by joints. Links are typically represented as sticks [42], polyhedrons [43] generalised cylinders [44] or superquadrics [45]. A joint interconnects two links by means of rotational motions about the axes of rotation. The number of independent rotation parameters will define the degrees of freedom (DOF) associated with a given joint.

In summary, the different approaches to 3D human modeling are essentially based on three main modes of data capture: laser scanner based systems, computer graphics based systems and image-based systems.

The most traditional mode of 3D human modeling typically utilise *3D laser scanners* like Vitus, WB4 and 3-D Full Body [46] [47]. Partial-body or full-body scanners are commonly used in different applications. They are particularly useful in producing anthropometrical data on the person involved, for example to support the work of designers in the automotive industry by optimising the interior of the vehicle from an ergonomic point of view. People have also been laser scanned for tailor-made clothing and some sculptors are known to use laser scanned data when creating their work of art. There is also a wide range of other applications where full-body scanners are used. 3D laser scanning has the advantages of being easy to control, user-friendly and able to generate highly accurate models of the human body. Additionally, the surface appearance (colour and texture) of the subject can be also obtained. The main

disadvantage of laser-based systems lies in the massive amount data generated for each scan. The huge computational costs involved in rendering and animating the captured/scanned individual make most applications unfeasible, thereby severely constraining its general applicability.

Computer graphics based computer animation software, such as Maya, 3D Studio Max and Autocat that implement the modeling of a wide range of objects including the human body, is widely available. A full suite of polygons, NURBS and subdivision surface modeling tools can be utilised to obtain high resolution human body models. Smooth 3D meshes and the various underlying human skeleton models empower these software tools to create, edit, render and animate human body models. However, as most systems do not directly integrate information acquired from actual objects or individuals, the generated models lack realism.

In contrast with the above two systems, *image-based systems* were typically designed to generate novel new views of a real scene from camera captured input images of a particular view. However in the last few years, these systems are becoming increasingly popular as cost effective and flexible systems for 3D human body modeling.

Carranza et al. [48] utilised a human shape model that is adapted to the observed person's outline. Their scheme employs a shape-from-silhouette approach while avoiding the visual disturbing geometry errors in the form of phantom volumes or

quantisation artifacts. By employing multi-view texturing during rendering, time-dependent changes in the human body surface can be reproduced with high fidelity. However, they do not incorporate explicit lambertian reflection properties when generating textures from the input video images.

Remondino [49] meanwhile worked on analysing uncalibrated image sequences and creating 3-D shape models of static human bodies. A photogrammetric approach to extract camera calibration parameters is used. The proposed bundle adjustment with self-calibration is a powerful tool for calibration and systematic error compensation. However, as bundle adjustment requires both intrinsic and extrinsic parameters of the camera system as initial values, at least four points on the human body have to be chosen to compute approximations of the external orientation of the cameras.

Using image-based visual hulls from multiple cameras, Wuermlin et al. [50] reconstructed a point-based representation of a person. Their proposed 3D Video Recorder methodology is a powerful framework for generating three-dimensional video. The 3D video concept is founded on point primitives that are stored in a hierarchical data structure. Limitations however include the poor quality of the underlying surface representation and the lack of precision of the reconstructed normal. Photometric calibration of the cameras for their point merging scheme might also improve the texture quality of their results.

In [51], a new technique is introduced for automatically building recognisable, dynamic 3D models of human individuals. A set of multi-view colour images of a person is captured from the front, sides and back by one or more cameras. Model-based reconstruction of shape from silhouettes is used to transform a standard 3D generic humanoid model to approximate the person's shape and anatomical structure. Realistic appearance is achieved through colour texture mapping from the multi view images. The results show the reconstruction of a realistic 3D facsimile of the person suitable for animation in a virtual world.

Despite the above efforts, 3D human modeling from uncalibrated images remains a challenging task. We, contributing to the advancement of knowledge in this frontier, propose a low-cost, robust and automatic scheme to generate realistic 3-D models of static human bodies from uncalibrated image sequences. The resulting 3-D point clouds can be easily exported to create a realistic surface model of an individual using 3D processing software. This methodology differs from current state-of-the-art work in the simplicity of its operations while preserving the overall quality of the final reconstruction.

5.3 Methodology

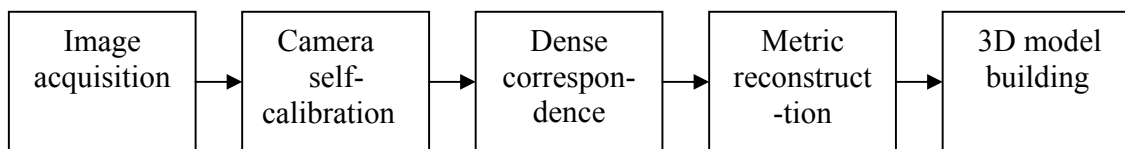


Figure 5.1 Block diagram of the methodology of 3D human body modeling

The methodology for 3D human body modeling is summarised in Figure 5.1 and described in detail in the following sub-sections.

5.3.1 Image acquisition

A sequence of images (at least two images) covering different views of an individual's body is acquired with an off-the-shelf digital camera. There should be some degree of overlap between adjacent images of the captured sequence. The camera's intrinsic parameters are assumed to be fixed during the acquisition process.

5.3.2 Camera self-calibration

Precise 3D information from images can only be extracted when the camera is accurately calibrated. If the aspect ratio of the camera is assumed to be one, and even if its focal length is unknown, the epipolar geometry (specially the fundamental matrix F) and the camera's intrinsic parameters may be calculated using the proposed focal length self-calibration approach described in chapter 4. Once the fundamental matrix has been obtained, it is used to establish a new set of correspondences using a correlation based approach that takes into account the recovered epipolar geometry. The matching approach that has been developed for human body modeling is a slightly modified version of the generic matching process described in the last chapter. In order to find possible matching partners in the second image for a corresponding feature point in the first image, the search should not deviate too far from the epipolar line in the second image.

5.3.3 Dense correspondences

The typically small number of correspondence points is far from sufficient to construct a high resolution 3D human body model. Obtaining dense reconstruction could be achieved by interpolation, but this does not yield satisfactory result in practice. If some salient features are missed during the point matching process, they will not appear during the reconstruction, degrading the final result significantly. This problem may be solved using algorithms that can estimate correspondences for almost every point in the images [52] [53].

The key to having sufficient matching points over a pair of images is to start with promising ‘seed points’. These seed points can be manually selected, generated semi-automatically (defining them only in one image) or generated in a fully automated way. The manual mode is used for very difficult cases where the automatic modes are known to fail. In the semi-automated mode, seed points will be manually selected only for the first image; corresponding points in subsequent images are established automatically by searching for the best matching results along the epipolar line. This mode is most suitable for typical cases of static surface measurement: the process is sufficiently fast and provides freedom in choosing the best initial seed points. The fully automatic mode is needed for cases of dynamic surface measurement from multi-image video sequences, where the number of image sets processed is large.

Starting from the seed points, sets of corresponding points grow automatically till the image is divided into polygonal regions. Point correspondence is based on the following scheme: starting from the location of the seed point, search is next performed for points located on a horizontal offset in the current image with

corresponding points in subsequent images using least squares matching. If the quality of the match is good, the location offset process continues horizontally until it reaches the region boundaries; if the quality of the match is not satisfactory, the algorithm automatically changes some parameters (e.g. smaller location offsets, moving vertically instead) before continuing with the matching process. The covering of the entire polygonal region of a seed point is thus achieved by a sequence of horizontal and vertical offsets. The process is the same for each polygonal region within an image.

The complete matching process (definition of seed points, automatic matching) is flexible and can be performed without prior knowledge of camera orientation and calibration information. This functionality can be useful if the information is not accurate enough or even unknown. Obviously, in these cases, the robustness of the matching result will decrease somewhat but remains satisfactory overall.

5.3.4 3D metric reconstruction

After dense matching points are obtained, we may proceed to conduct the 3D reconstruction. The geometry of the real world is Euclidean. However, when we move from the 3-D world to 2-D images, depth is lost. Without some control points in the Euclidean space, there is no way to fully recover the Euclidean structure [38]. However, in many applications, it may not be essential that absolute geometry (i.e., the exact dimension and structure) of the real world be recovered. In fact, it might be sufficient to have simpler reconstructions of the world, accurate up to a scaling constant, i.e. a metric reconstruction.

Once the intrinsic parameters of the camera and feature matching points over a pair of images are given, 3D metric reconstruction can be implemented [38]. In the 3D metric structure, not only parallelism but also angles and ratios of lengths are preserved. Hence the structure is very similar to the real world; only the dimension of the scene is missing.

When the intrinsic parameter matrix A is known, the fundamental matrix can be reduced to the essential matrix E , which is the ‘specialisation’ of the fundamental matrix F . The relation between E and F is given by:

$$E = A^T F A \quad (5.1)$$

The essential matrix has the property that $\text{rank}(E)=\text{rank}(F)=2$. The SVD of E takes the form of $U \text{diag}(1,1,0) V^T$, where U and V are two orthogonal matrices. Consider two cameras of a stereo rig, the first camera matrix is denoted as $P = K[I | 0]$. There are then four possible choices for the second camera matrix P' :

$$[UWV^T | u_3] \text{ or } [UWV^T | -u_3] \text{ or } [UW^T V^T | u_3] \text{ or } [UW^T V^T | -u_3],$$

where $W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, and u_3 is the third column of U .

For the four possible forms of P' above, only one of them can be used to produce reconstructed points in front of both the cameras. Trial testing with a single point would identify the correct camera matrix. This camera matrix solution leaves only one ambiguity: the scale of translation. In other words, knowledge of the camera's intrinsic parameters enables us to obtain the metric structure of the reconstructed

scene.

Suppose that a point x in R^3 is visible in two images with its projections in these images denoted by u and u' respectively. Besides, the two camera matrices P and P' , and u and u' of each point x involved in the dense correspondence dataset are known. From these data, the two rays in space corresponding to the two image points can be computed. The triangulation problem [37] is to find the intersection of the two rays, i.e. the 3D point x in space. At first sight, this seems to be a trivial problem, since finding the intersection between two lines in space is nothing difficult. Unfortunately, in the presence of noise, these rays cannot be guaranteed to cross; we need to develop robust solutions under some assumed noise model. The algorithm in [36] is employed here to solve the triangulation problem and produce the resulting reconstructed 3D point.

5.3.5 3D modeling building

After obtaining ‘clouds’ corresponding to masses of reconstructed 3D points, we are able to build the 3D model of the human body. Exact visualisation with correct pixel colour may also be generated if each point of the model is re-projected to the image concerned before the corresponding colour of the image point is obtained.

5.4 Experimental Results

A Canon A75 digital camera is used to capture images at 640×480 resolution. Based on our self-calibration approach described in chapter 4, the calculated camera focal

length is 690 pixels. Using the same previous assumptions, the camera intrinsic

$$\text{matrix is } K = \begin{bmatrix} 690 & 0 & 320 \\ 0 & 690 & 240 \\ 0 & 0 & 1 \end{bmatrix}.$$

In experiment I, the images used for the reconstruction are shown in Figure 5.2. Following the procedure outlined in section 5.3, 137 feature points are automatically extracted and used for point correspondence. The related epipolar geometry can thus be obtained. A sample feature point is shown in green in the left image of figure 5.3; based on the calculated fundamental matrix, the corresponding epipolar line is drawn as shown in the right image. The fact that it passes through the exact location of the feature point provides good indication of the accuracy of the fundamental matrix calculation.



Figure 5.2 Two images used for the reconstruction in experiment I

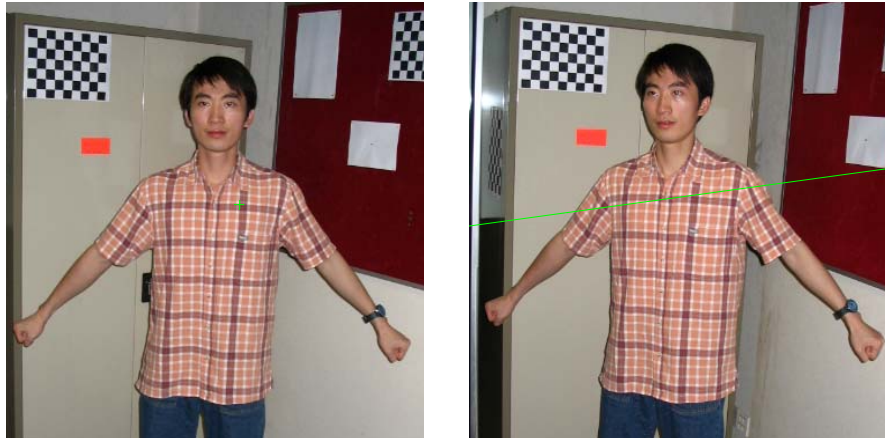


Figure 5.3 Epipolar line aligns with exact location of a feature point

Figure 5.4 meanwhile illustrates two different viewpoints of the reconstruction 3D model of the human subject (2062 3D points).

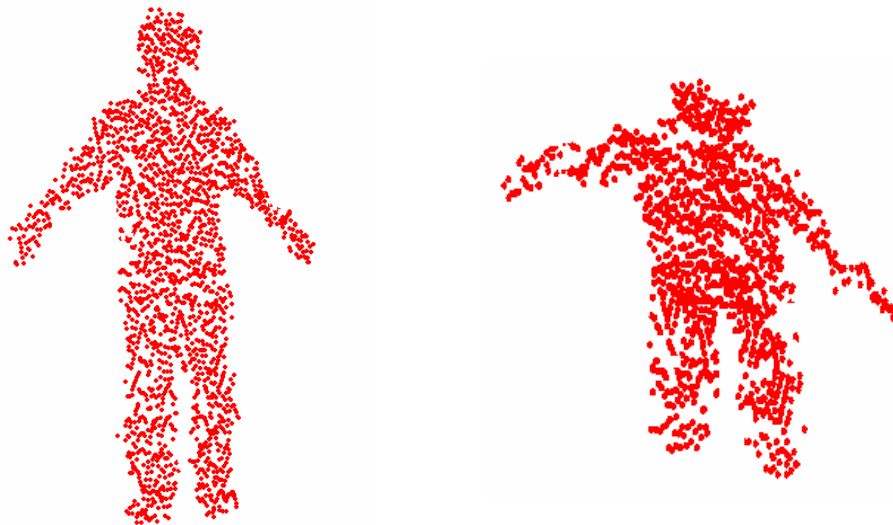


Figure 5.4 Recovered 3D point cloud of the human body (experiment I)

For realistic visualisation, each 3D point of the cloud mass is back projected onto the first image of the pair to get the corresponding pixel color. Thus we are able to depict the 3D human body model in full colour as shown in Figure 5.5.



Figure 5.5 Reconstructed 3D human body model depicted in back-projected colour (experiment I)

These steps are repeated for experiment II using a different subject in another setting. The two images, shown in Figure 5.6, are acquired using the same camera. A total of 185 data points are found and used for feature matching in this experiment. The resulting 3D reconstructed model (2275 3D points) is shown in Figure 5.7 with its visualisation in full colour presented in Figure 5.8.



Figure 5.6 Two images used for the reconstruction in experiment II

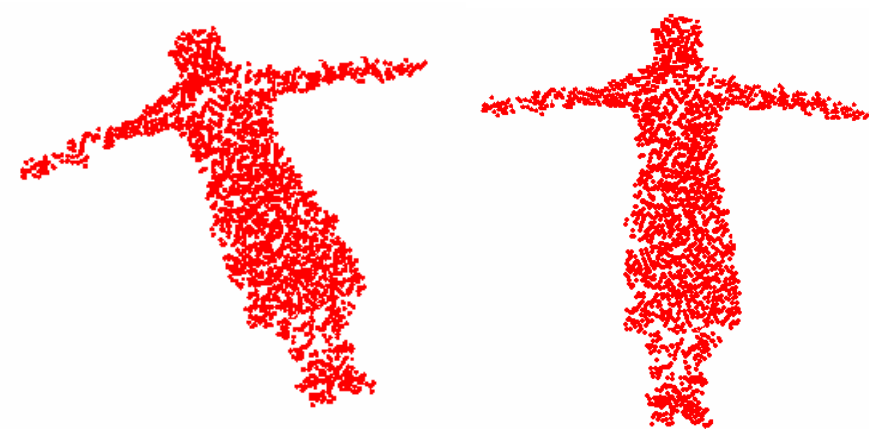


Figure 5.7 Recovered 3D point cloud of the human body (experiment II)



Figure 5.8 Reconstructed 3D human body model depicted in back-projected colour (experiment II)

5.5 Future Work

To refine the 3D human body model, additional processing can be applied. For example, photogrammetric bundle adjustment with self-calibration [54] can help ensure more accurate camera calibration results compared with those obtained here. This would translate to more realistic 3D reconstruction results. The technique makes

use of the collinearity model, i.e. the fact that a point in object space, its corresponding point in the image plane and the projective center of the camera, all lie on a straight line. The standard form of these collinearity equations is:

$$x - x_0 = -c \cdot \frac{r_{11}(X - X_0) + r_{21}(Y - Y_0) + r_{31}(Z - Z_0)}{r_{13}(X - X_0) + r_{23}(Y - Y_0) + r_{33}(Z - Z_0)} = -c \cdot \frac{U}{W} \quad (5.2)$$

$$y - y_0 = -c \cdot \frac{r_{12}(X - X_0) + r_{22}(Y - Y_0) + r_{32}(Z - Z_0)}{r_{13}(X - X_0) + r_{23}(Y - Y_0) + r_{33}(Z - Z_0)} = -c \cdot \frac{V}{W} \quad (5.3)$$

where:

x, y are the point image coordinates;

x_0, y_0 are the image coordinates of the principal point;

c is the camera constant;

X, Y, Z are the object points in real world coordinates;

r_{ij} is the element of the orthogonal rotation matrix R between image and object.

The two collinearity equations above are first formed for each of the salient feature points found using the technique described in section 5.3.2; a system of equations is therefore built. These equations are non-linear with respect to the unknowns. In order to solve them with the least squares method, they must be linearised, thus requiring approximations. However, this approach requires information on both the camera's internal and external parameter, which translates to approximations of the extrinsic parameters. To facilitate this, reference salient points on the human body or background have to be measured in order to obtain an approximation of the external

orientation of the cameras. In summary, photogrammetric bundle adjustment may ensure more accurate camera calibration results but a price has to be paid in the form of additional measurements using reference points.

Two other aspects of our work concerning the modeling of the reconstructed 3D point cloud need to be further investigated:

1) *Generation of a polygonal surface*: From the unorganized 3D data, a non-standard triangulation procedure is required. Algorithms that generate a correct triangulation and surface models allow editing operations, like point holes filling or polygon corrections.

2) *Fitting a predefined 3D human skeleton model*: this procedure does not usually require the generation of a surface model and the reconstructed 3-D point cloud is used as basis for the fitting process. Figure 5.9 shows one example of a predefined 3D human skeleton model which we have implemented on the 3D Studio Max platform. Various techniques are needed to fit the reconstructed 3D data points onto appropriate locations along the skeleton model.

Finally, we have only considered the case that two viewpoints are taken into account. If more images are involved, it is possible to generate more accurate 3D models with more sophisticated processing. Moreover, cases that the camera is still but the subject moving and that both camera and subject are moving also need to be further investigated.

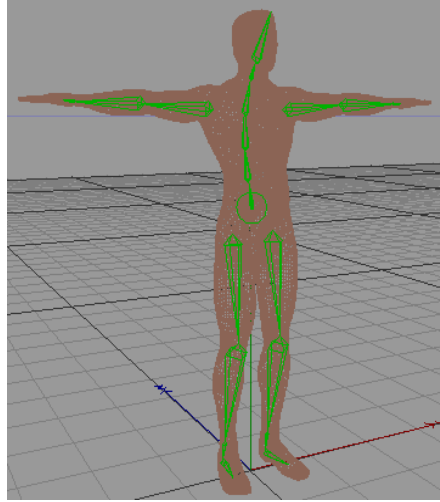


Figure 5.9 Example of the pre-defined human skeleton model

5.6 Conclusion

3D human body modeling from uncalibrated image sequences forms the second module in our overarching human motion capture system. Based on self-calibration approach for focal length estimated described in chapter 4, we develop a point correspondence-based scheme to build a dense 3D human body model of the captured individual. This method basically comprises five steps of image acquisition, self calibration, dense point correspondence, metric reconstruction and 3D model building.

Experimental results demonstrate the robustness and accuracy of the approach. Its simplicity makes it an attractive alternative for a number of potential applications. Moreover, it may be of high usefulness for the implementation of the third module of the proposed motion capture system: 3D human motion tracking.

Chapter 6 3D Human Motion Tracking

6.1 Introduction

At the heart of human motion capture is motion tracking. When we develop trajectories of key joints or establishing correspondence between semantically meaningful points on the human body in different views or frames, region-based approach with points and feature tracking may lead to an attractive and satisfactory solution. This 3D model-based module aims to track the body motion of the recorded person over time. With the aim of the works done in the previous chapters, a silhouette-based scheme is proposed in this chapter. After an initialisation step, the body pose parameters that maximize the overlap between projected model silhouettes and input foreground silhouettes are estimated for every time step. The idea will be further investigated and verified in the near future.

6.2 Methodology

6.2.1 Silhouette extraction

The inputs to the motion parameter estimation are silhouette images of the moving person from the background pixel. From a sequence of video frames without a moving subject, the mean and standard deviation of each background pixel in each color channel are computed [55]. If a pixel differs in at least one color channel by more than an upper threshold from the background distribution, it is classified as

certainly belonging to the foreground. If its difference from the background is smaller than a lower threshold in all channels, the pixel is classified as certainly background. All other pixels are considered potential shadow pixels.

For stationary cameras, we can use the static elements in a scene to help us discriminate foreground and background objects in a given image. Since the visible portions of the model will necessarily be in the foreground, background subtraction is an excellent way to prune the parameter space. Thus results improve with the incorporation of background subtraction [56].

For our proposed processing, background images are first separately captured. As a matter of fact, such background images could be generated automatically given enough footage of a scene; for each pixel, one could compute the median intensity value over time and consider this to be the background. In either case mentioned above, we can test whether any pixel in the source image, including foreground subjects, differs significantly from the corresponding pixel in the background image. The background subtraction criterion presumes that the areas of difference between the source and background images correspond with some portion of the model projection. Given that presumption, a background image can easily rule out many model configurations.

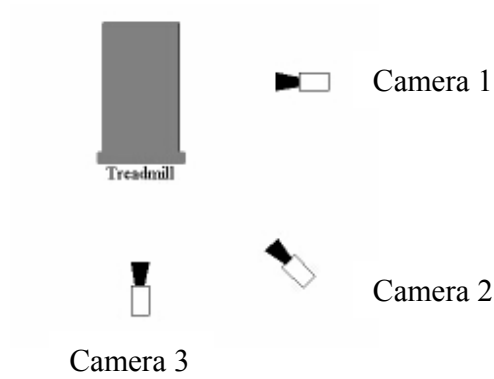
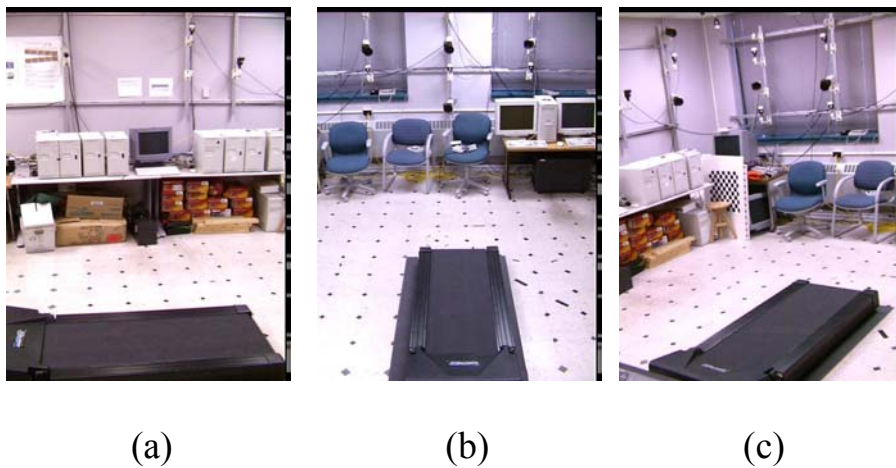


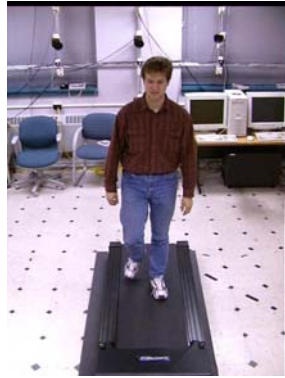
Figure 6.1 Setup of the cameras in the experiment

In our test experiment, CMU MoBo Database [57] is applied to capture multi-view motion sequences of human body motion. In the test video, a subject (one person) walks on a treadmill positioned in the middle of the room. A total of three synchronized cameras are used to capture the motion. The setup of the cameras is depicted in Figure 6.1. The resulting color images have a resolution of 640x480. The sequence is recorded at 20 frames / second.





(d)



(e)



(f)



(g)



(h)



(i)



(j)



(k)



(l)

Figure 6.2 Silhouette extraction from three cameras

The silhouette extraction and background subtraction are the initial processing for the implementation of motion tracking. Figure 6.2 shows the results of our experiment using CMU MoBo Database. In Figure 6.2, images (a), (b) and (c) show the background images captured from three different cameras; Images (d) , (e) and (f) show the first of many input images from three cameras involved in the dataset; Images (g) , (h) and (i) show the results of per-pixel background subtraction from three different cameras, the foreground objects are extracted in this case; Images (j) (k) and (l) represent the silhouette images for the different cameras. If a pixel differs in at least one color channel by more than an upper threshold from the background image, it is classified as certainly belonging to the foreground. If its difference from the background is smaller than a lower threshold in all channels, the pixel is classified as certainly background. Silhouette quality can be improved via subsequent morphological dilate and erode operations [58]

6.2.2 Human body model

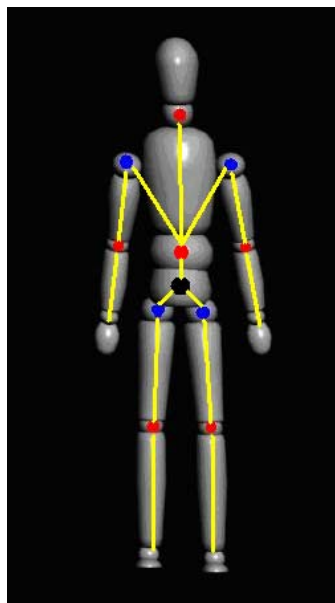


Figure 6.3 Human body model and the underlying skeletal structure

The 3D body model (5600 triangles and 2894 vertices) used in the experiment is shown in Figure 6.3. It is a generic model consisting of a hierarchic arrangement of 29 body segments (head, upper arm, torso etc.). The model's kinematics is defined via an underlying skeleton consisting of 11 joints connecting bone segments. Spheres in Figure 6.3 indicate joints and the different parameterisations used: blue sphere is 3 DOF ball joint and red sphere is 1 DOF hinge joint. The black sphere indicates the global position and orientation of the whole body model. Different joint parameterizations are used in different parts of the skeleton. Each limb, i.e. complete arm or leg, is parameterized via four degrees of freedom. This limb parameterization is chosen because it is particularly suitable for an efficient search of its parameter space which we will describe in Section 6.2.5. At every time instant, 24 parameters are needed to completely define a body pose. The surface of each body segment is represented by a closed triangle mesh.

6.2.3 Energy function

The error metric used to estimate the goodness of fit of the body model with respect to the video frames computes a pixel-wise exclusive-or operation between the image silhouette and the rendered model silhouette in each input camera view [59]. This process is demonstrated in Figure 6.4. The silhouettes are computed for the image and for one view of the 3D model and combined afterward. The energy function value is the sum of the non-zero pixels for every camera view after this pixel-wise boolean operation [59]. This error metric can thereby efficiently drive the model fit over a series of time steps.

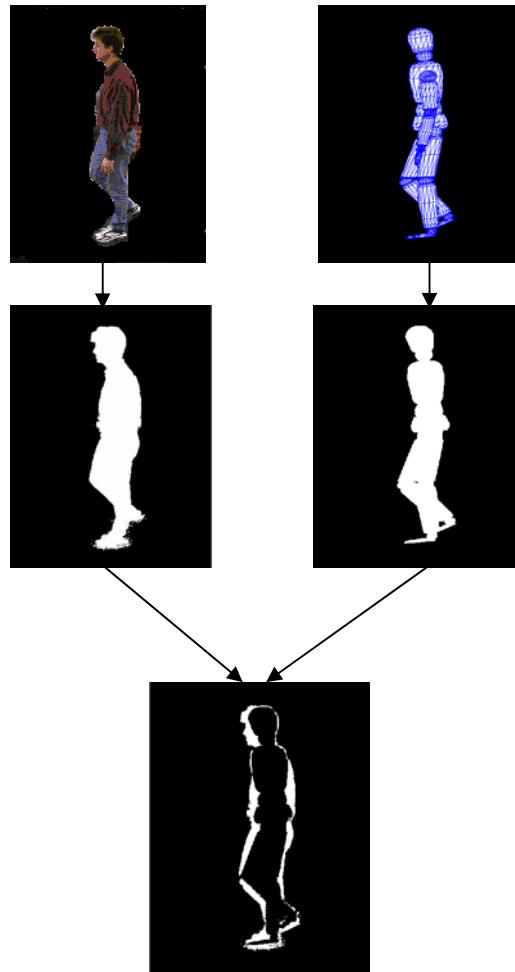


Figure 6.4 Measuring the difference between the image (left) and one view of the model (right) by the area occupied by the XORed foreground pixels

6.2.4 Model initialisation

The motion capture is initialised using a set of silhouette images that show the human body in an initialisation pose. The ideal initialisation pose is one in which both the arms and legs are bent, allowing for simple identification of elbow and knee locations. From these silhouettes, a set of scaling parameters as well as a set of pose parameters is computed. The global model position can be chosen that produces the best fit according to the error measure described in Section 6.2.2. The fit can be improved by

optimizing over the pose parameters and joint scaling parameters in an iterative process that employs the same error measure. After initialisation, all scaling parameters are fixed, and continuous tracking is performed for all subsequent time steps.

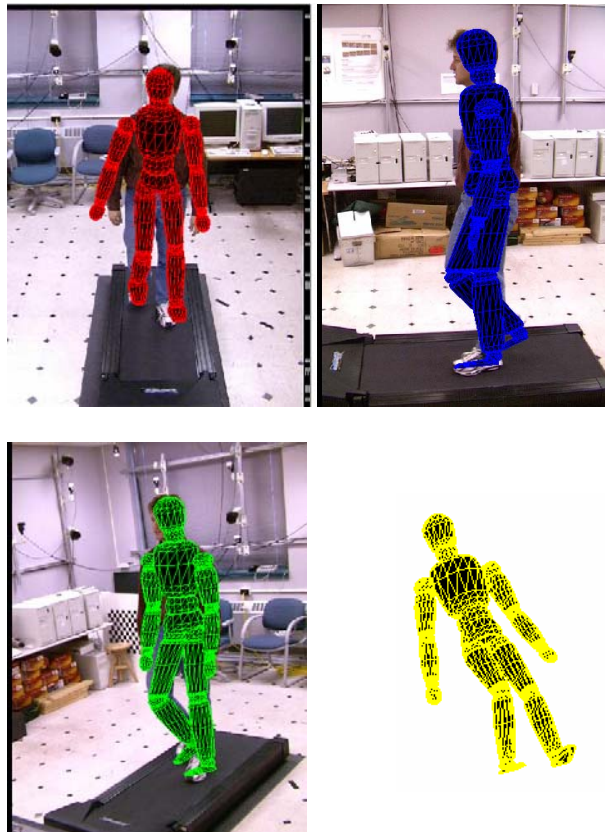


Figure 6.5 Initialisation of the human body model

In our experiment, the initial body configuration is only performed manually for the time being. The initial pose of the model is shown in Figure 6.5. The 3D human model (its textured version may be available from the further result of chapter 5) is superimposed in the foreground object presented in different images. On the platform of 3D Studio Max, three viewpoints are generated to show the initialisation of the human body. In this way, the 3D position and orientation of different parts of the

body can be therefore extracted. The new free-view human model is also rendered in the lower right image in Figure 6.5. This step would be the basis for the human body tracking over the video sequence. 3D reconstruction of the human body motion may be thus implemented.

6.2.5 Motion parameter estimation

After the initialisation, the model parameters for each time instant may be computed using an exhaustive minimization approach based on the previously described energy function. A straightforward approach would be to apply Powell's method [60] to optimise over all degrees of freedom in the model simultaneously. This simple strategy, however, exhibits some of the fundamental pitfalls that make global optimisation infeasible. In the global case, the goal function reveals many erroneous local minima. Fast movements between consecutive time frames are almost impossible to resolve correctly. For every new time step, the optimisation uses the result from the previous frame as a starting point. For fast moving body segments, there will be no overlap between the starting model pose and the current time frame, and no global minimum can be found. Another problem may arise if one limb moves very close to the torso.

To make the tracking procedure robust against these problems and to enable it to follow complex motions, we may split the parameter estimation into a sequence of optimisations on subparts of the body.

Temporal coherence can be exploited during the computation of the motion parameters. Starting from the body pose in the previous time step, the global translation and rotation of the model root are computed. The rotations of head and hip joints are then independently computed using an identical optimisation procedure.

With the main body aligned to the silhouettes, the poses of the two arms and two legs can be found with independent optimisation steps. The final step in the sequence of optimisations is the computation of hand and foot orientation by optimizing over their local parameter space.

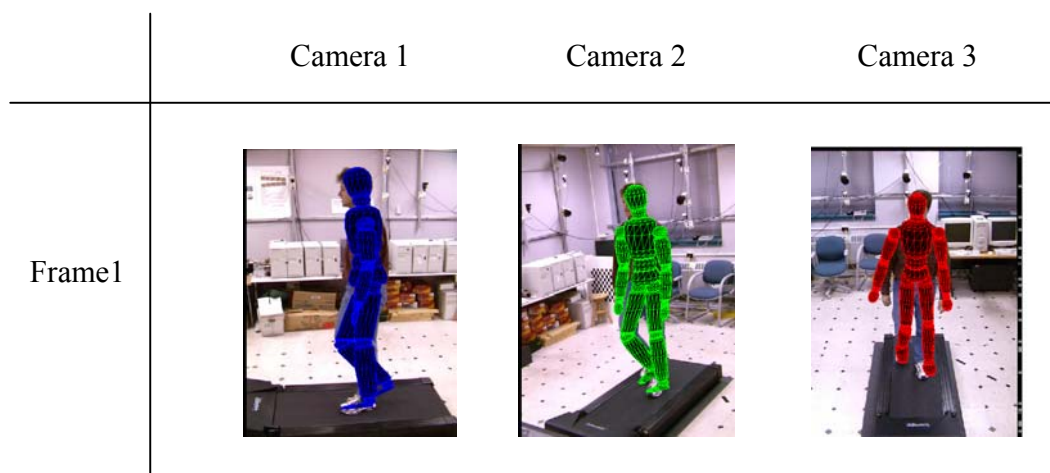
Here we take the limb for example. Suppose the limb of the human model is determined by four degrees of freedom (DOFs), as shown in Figure 6.3, fitting an arm is a four-dimensional optimisation problem. The limb fitting employs the following steps. The parameter space is efficiently constrained by applying a global search on the four-dimensional parameter domain. The search samples the parameter space regularly and tests each sample for representing a valid arm pose. A valid pose is defined by two criteria. First, the wrist and the elbow must project into the image silhouettes in every camera view. Second, the elbow and the wrist must lie no less than certain distance from the axis of the bounding box defined around the torso segment of the model. For all detected valid poses, the error function is evaluated, and the pose possessing the minimal error is used as starting point for a downhill optimisation procedure [60]. The arm pose at the current time instant is the result of the downhill optimisation procedure. For all four arm parameters, the search space for

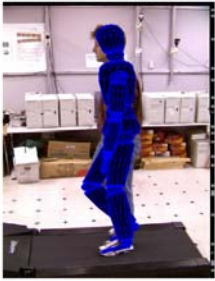
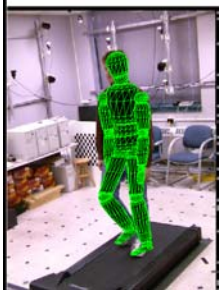


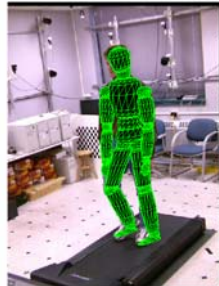



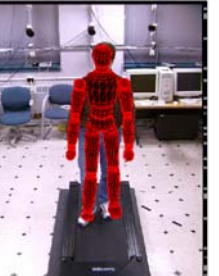
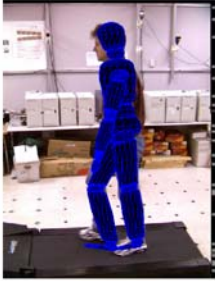


valid poses is adapted to the difference in the parameter values observed during the two preceding time steps, implicitly including the assumption of a smooth arm motion into the fitting procedure.


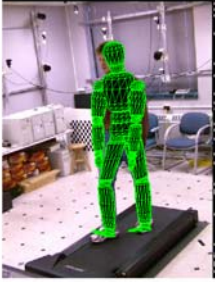



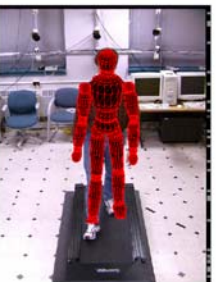

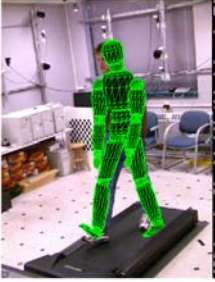


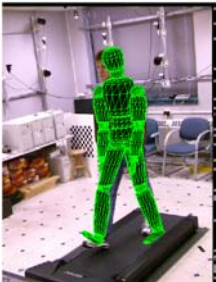

In summary, the proposed overall silhouette-based motion parameter estimation has several advantages. The algorithm is not tied to any specific body model. More complex parameterizations or different surface representations could easily be used. Furthermore, the algorithm may scale to higher input image resolutions. Model fitting can be applied to lower resolution versions of the video frames by means of an image pyramid. On the whole, the proposed fitting procedure exhibits a high degree of robustness and efficiency and yet is comparably simple.






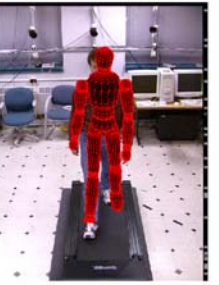


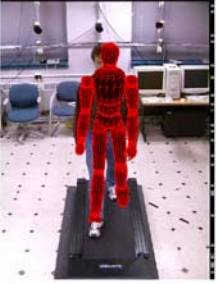



6.3 Experimental Results

Based on the methodology mentioned in section 6.2, full-body tracking is implemented with the results demonstrated in Figure 6.6.



	Camera 1	Camera 2	Camera 3
Frame2			
Frame3			
Frame4			
Frame5			

	Camera 1	Camera 2	Camera 3
Frame6			
Frame7			
Frame8			
Frame9			

	Camera 1	Camera 2	Camera 3
Frame10			
:			
Frame11			
Frame12			
Frame13			

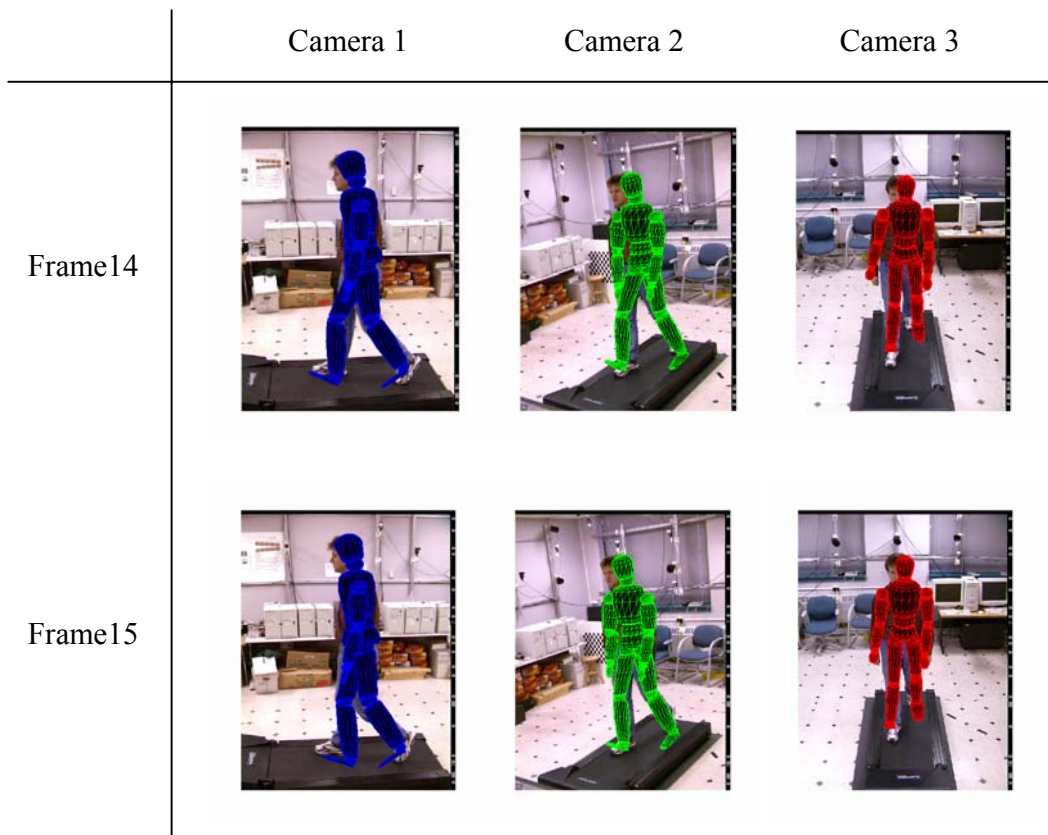


Figure 6.6 Results of full-body tracking

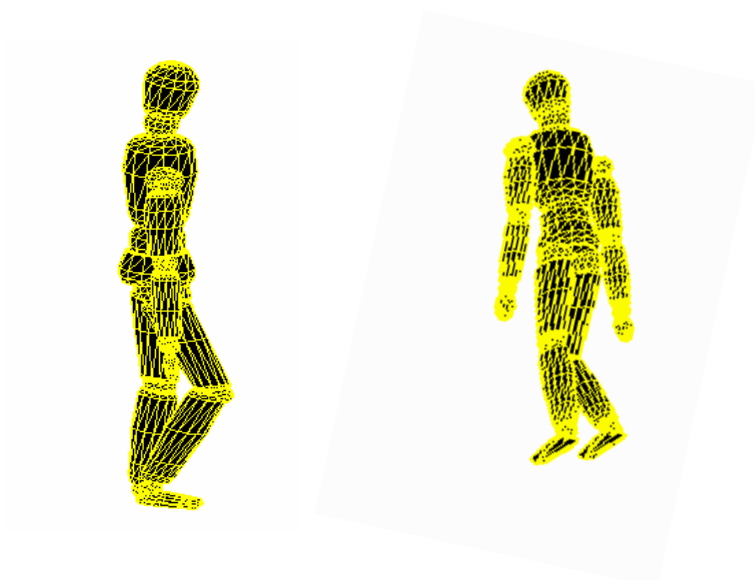


Figure 6.7 Free-view rendering of human motion (Frame 3)

With the obtained tracking results, we may render the animation of the walking process. Two individual free-view human models for both Frame 3 and Frame 12 are demonstrated in Figure 6.7 and Figure 6.8 respectively.

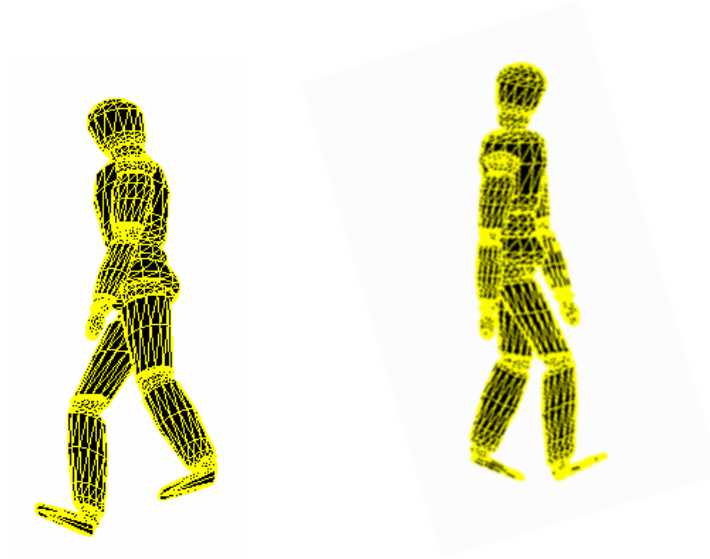


Figure 6.8 Free-view rendering of human motion(Frame 12)

6.4 Future Work

This chapter presents the initial work in the human body tracking module in the whole motion capture system. The framework for the silhouette-based motion parameter estimation has been proposed. More experiments are about to be done in the near future. The common tracking algorithm like Kalman Filter [61] or Condensation [62] will also be imbedded in the procedure if needed.

We have to admit that there are quite a few potential technical difficulties available. One of the most limiting characteristics in the video-based analysis system is the difficulty of initialisation. The current approach to 3-D reconstruction and tracking

requires a very accurate estimate of 3D position across multiple views. There exist no algorithms available today that can perform this task with sufficient regularity, reliability, and exactness. This initialisation is required not only for the sake of generating a consistent 3D point set, but also for building a semantically meaningful structure for the underlying human body.

Incremental improvements to the tracking and further recognition algorithms may be possible; however, the greatest potential for future work is in the extension of the higher-level and more complicated activities and events. Even within the framework of 3D motion tracking, there is still substantial room for contribution simply by considering additional applications. Of particular interest may be the modeling of long-term interactions between multiple individuals or between individuals and their environment.

Chapter 7 Conclusion

The thesis presented the development of a 3D model-based human motion capture system. A framework for practical optical motion capture was demonstrated. Basically, the whole system comprises three modules: calibration, modeling and tracking. In addition to the functionality of each subpart, the engineering tasks involved in the setup of the system were also addressed and evaluated.

The thesis mainly focused on the calibration and 3D human body modeling subsystems while we also initialized the work on motion tracking part. An effective approach for camera's focal length calibration was proposed. The approach assumes only the camera's focal length is unknown and constant. The Kruppa's equations are thus able to be decomposed as two linear and one quadratic equations. In this case the closed form solution for camera calibration, without additional motion-generated information, was successfully obtained. The proposed algorithm could be implemented in an automatic way and it achieved robust and accurate performance on synthetic and real images of indoor/outdoor scenes. We also succeeded in developing a point correspondence-based modeling scheme to build a dense 3D shape model of a static human body from uncalibrated images. The automated matching process on the human body is able to implement highly realistic 3D metric reconstruction. In addition, no à priori information or measurement of the human subject and the camera setup is required. Finally, we presented a silhouette-based scheme in the motion

tracking module. The fit energy function may effectively drive the 3D human model to fit the exact position over time. Highly accurate motion tracking was successfully performed.

The work presented here is not the end. Our final objective is to analyse the human body kinematics from multiple viewpoints using a high resolution 3D articulated human body model. To address the demand for a higher resolution motion capture system, it is required to produce high quality 3D shape model in a more automatic and realistic mode. The silhouette-based tracking module will be also further investigated to provide accurate human motion property. In the future, we will try to integrate the three modules, i.e. calibration, modeling and tracking, in a seamless video-based human motion capture system. The system can then applied in various applications such as surveillance, performance analysis, virtual reality, human computer interactions and so on.

References

- [1] G. Johansson, Visual Perception of Biological Motion and a Model for Its Analysis, Perception Psychophysics, Vol.14(2), pp.201-211, 1973

- [2] R. Polana and R. Nelson, Low Level Recognition of Human Motion, In Proc. of IEEE Workshop on Motion of Non-rigid and Articulated Objects, Austin, 77-82, 1994

- [3] B. Heisele, U. Kressel and W. Ritter, Tracking Non-rigid, Moving Objects Based on Color Cluster Flow, in Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, 257-260, 1997

- [4] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna and T. Poggio, Pedestrian Detection Using Wavelet Templates, in Proc. of IEEE Conference on Computer Vision and Pattern recognition, San Juan, 19-199,1997

- [5] A. Baumberg and D. Hogg, An Efficient Method for Contour Tracking Using Active Shape Models, in Proc. of IEEE Workshop on Motion of Non-rigid and Articulated Objects, Austin, 194-199,1994

- [6] R. Kauth, A. Pentland, G. Thomas, Blob: an unsupervised clustering approach to spatial preprocessing of mss imagery, 11th International Symposium on Remote Sensing of the Environment, Ann Harbor MI, 1977

- [7] C. R. Wren, A. Azarbayejani, T. Darrell and A. P. Pentland, Pfinder: Real-time Tracking of the Human Body, IEEE Trans. Pattern Analysis and Machine Intelligence vol.19(7), 780-785, 1997

- [8] G. Cheung, T. Kanade, J.-Y. Bouguet and M. Holler, A Real Time System for Robust 3D Voxel Reconstruction of Human Motions, in Proc. of IEEE Conf. Computer Vision and Pattern Recognition, vol.2, 714-720, 2000
- [9]H. J. Lee and Z. Chen, Knowledge-guided Visual Perception of 3D Human Gait from a Single Image Sequence, IEEE Trans. Systems, Man, Cybernetics, 336-342, 2002
- [10] S. Iwasawa, K. Ebihara, J. Ohya and S. Morishima, Real-time Estimation of Human Body Posture from Monocular Thermal Images, in Proc. Conf. on Computer Vision and Pattern Recognition, 1997
- [11] M. K. Leung and Y. H. Yang, First sight: A Human Body Outline Labeling System, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 17(4), 359-377, 1995
- [12] K. Rohr, Towards Model-based Recognition of Human Movements in Image Sequences, CVGIP: Image Understanding vol. 59(1), 94-115, 1994
- [13] S. Wachter and H.-H. Nagel, Tracking Persons in Monocular Image Sequences, Computer Vision and Image Understanding vol. 74, 174-192, 1999
- [14] M. E. Leventon and W. T. Freeman, Bayesian Estimation of 3D Human Motion from an Image Sequence, Technical report, 1998
- [15] Q. Cai and J. K. Aggarwal, Automatic Tracking of Human Motion in Indoor Scenes Across Multiple Synchronized Video Streams, in Proc. International Conference on Computer Vision, Bombay, India, 1998
- [16]S. Iwasawa, J. Ohya, K. Takahashi, T. Sakaguchi, K. Ebihara and S. Morishima, Human Body Postures from Trinocular Camera Images. IEEE International Conference on Automatic Face and Gesture Recognition, 426-331, 2000
- [17] M. Coell, A. Rahimi and M. Harville and T. Darrell, Articulated-pose Estimation Using Brightness and Depth Constancy Constraints. In IEEE International Conference on Computer Vision and Pattern Recognition, 438-445, USA, 2000

- [18] R. Plankers and P. Fua, Tracking and Modeling People in Video Sequences, Computer Vision and Image Understanding, Vol.81, 285-293, 2001
- [19] G. Cheung, T. Kanade, J. Bouguet and M. Holler, A Real Time System for Robust 3D Voxel Reconstruction of Human Motions. In IEEE International Conference on Computer Vision and Pattern Recognition, 714-720, USA, 2000
- [20] I. Mikic, M. Trivedi, E. Hunter and P. Cosman, Human Body Model Acquisition and Tracking Using Voxel Data, International Journal of Computer Vision vol. 53(3), 199-223, 2003
- [21] J. Carranza, C. Theobalt M. A. Magnor and H.-P. Seidel, Free-viewpoint Video of Human Actors, ACM Transactions on Graphics, 569-576, 2003
- [22] K. Grauman, G. Shakhnarovich and T. Darrell, Inferring 3D Structure with a Statistical Image-based Shape Model, In IEEE International Conference on Computer Vision and Pattern Recognition, 714-720, USA, 2003
- [23] O. Faugeras. Three-Dimensional Computer Vision: a Geometric Viewpoint, MIT press, 1993
- [24] S. J. Maybank and O. D. Faugeras, A Theory of Self-Calibration of a Moving Camera, International Journal of Computer Vision, Vol.8(2), 123-152, 1992
- [25] S. Boughoux, From Projective to Euclidean Space under any Practical Situation, a Criticism of Self-calibration, In Proceedings of IEEE conference on Computer Vision and Pattern Recognition, 790-796, 1998
- [26] L. Dron, Dynamic Camera Self-Calibration from Controlled Motion Sequences, In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 501-506, 1993

- [27] F. Du and M. Brady, Self-calibration of the intrinsic parameters of cameras for active vision systems, In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 477-482, 1993
- [28] R. I. Hartley, Self-calibration of Stationary Cameras, International Journal of Computer Vision, Vol.22(1), 5-23, 1997
- [29] Y. Ma and R. Vidal, Kruppa's Equation Revisited: Its Renormalization and Degeneracy, In Proceedings of European Conference on Computer Vision, 2000
- [30] M. Pollyfeys, R. Koch and L. van Gool, Self-calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters, In Proceedings of International Conference on Computer Vision, 90-95, 1998
- [31] W. Lao, Z. Cheng, A. Kam, T. Tan and A. Kassim, Focal Length Self-calibration Based on Degenerated Kruppa's Equations: Method and Evaluation, In Proceedings of IEEE International Conference on Image Processing, 3391-3394, 2004
- [32] R.I. Hartley, Kruppa's Equations Derived from the Fundamental Matrix, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 19(2), 133-135, 1997
- [33] P. Sturm, On Focal Length Calibration from Two Views, In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, Vol.2, 145-150, 2001
- [34] C. Harris and M. Stephens, A Combined Corner and Edge Detector, Fourth Alvey Vision Conference, 147-151, 1988
- [35] Z. Zhang, R. Deriche, O. Faugeras and Q.-T. Luong, A Robust Technique for Matching Two Uncalibrated Images through the Recovery of the Unknown Epipolar Geometry, Artificial Intelligence Journal, Vol.78, 87-119, October 1995
- [36] P. Rousseeuw, Robust Regression and Outlier Detection, Wiley, New York, 1987

- [37] R. I. Hartley and P. Sturm, Triangulation. In DARPA Image Understanding Workshop, Monterey, CA, 957-966, 1994
- [38] R. I. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2000
- [39] P. Ratner, 3-D Human Modeling and Animation, John Wiley & Sons, 1998
- [40] T. B. Moeslund and E. Granum, A Survey of Computer Vision-Based Human Motion Capture, Computer Vision and Image Understanding, Vol. 81, 231-268, 2001
- [41] R. Plankers and P. Fua, Tracking and Modeling People in Video Sequences, Computer Vision and Image Understanding, Vol. 81 (3), 2001
- [42] C. Barron and I.Kakadiaris, Estimating Anthropometry and Pose from a Single Camera. IEEE International Conference on Computer Vision and Pattern Recognition. Hilton Head Island, SC, 2000
- [43] M. Yamamoto, A. Sato, S. Kawada, T. Kondo and Y. Osaki, Incremental Tracking of Human Actions from Multiple Views. IEEE International Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, 1998
- [44]I. Cohen, G. Medion and H. Gu, Inference of 3D Human Body Posture from Multiple Cameras for Vision-based User Interface. World Multiconference on Systemics, Cybernetics and Informatics, USA, 2001
- [45] D. M. Gavrila and L. Davis, 3D Model-based Tracking of Humans in Action: A Multi-view Approach, IEEE International Conference on Computer Vision and Pattern Recognition, San Francisco, CA, 1996
- [46] K. M. Robinette and H. Daanen, Lessons Learned from CAESAR: A 3-D Anthropometric Survey, IEA conference, 2003

- [47] B. Bradtmiller and M.E. Gross, 3-D Whole Body Scans: Measurement Extraction Software Validation, Proceedings of the SAE International digital Human Modeling for Design and Engineering Internaitonal Conference and Exposition, The Netherlands, 1999
- [48] J. Crranza, C. Theobalt, M. A. Magnor and H.P. Seidel, Free-Viewpoint Video of Human Actors, ACM Transactions on Graphics, Vol.22(3), 569-577, 2003
- [49] F. Remondino, Human Body Reconstruction from Image Sequences, DAGM 2002, 50-57
- [50] S. Wuermlin, E. Lamboray, O. Staadt and M.Gross, 3d Video Recorder, In Proceedings of Pacific Graphics, IEEE Computer Society Press, 325-334, 2002
- [51] A. Hilton, D. Beresford, T. Gentils, R. Smith , W. Sun, J. Illingworth , Whole-body Modeling of People from Multiview Images to Populate Virtual Worlds, The Visual Computer, Vol.16, Springer Verlag, Berlin, 411-436, 2000
- [52] R. Koch, M. Pollefeys and L.V. Gool, Multi Viewpoint Stereo from Uncalibrated Video Sequence, Proceeding of European Conference on Computer Vision, Germany, 1998
- [53] M. Pollefeys, R. Koch and L.V. Gool, Self-Calibration and Metric Reconstruction in spite of Varying and Unknown Intrinsic Camera Parameters, International Journal of Computer Vision, 1998
- [54] A. Grun, H. Beyer, System Calibration through Self-Calibration, Calibration and Orientation of Cameras in Computer Vision, Vol. 34, 163-193, Springer, New York, 2001
- [55] K. Cheung, T. Kanade, J.-Y. Bouguet and M. Holler, A Real Time System for Robust 3D Voxel Reconstruction of Human Motions, In Proceedings of Computer Vision and Pattern Recognition, Vol.2, 714-720, 2000
- [56] J. Deutscher, B. North, B. Bascle, and A. Blake, Tracking Through Singularities and Discontinuities by Random Sampling. Vol. 2, 1144-1149, 1999

- [57] J. Shi and R. Gross. The CMU Motion of Body (MoBo) Database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, June 2001
- [58] R. Jain, R. Kasturi and B. Schunck, Machine Vision, McGraw-Hill, 1995
- [59] H. Lensch, W. Heidrich and H.P.1 Seidel, A Silhouette-based Algorithm for texture Resistration and Stitching, Graphical Models Vol.64(3), 245-263, 2001 Lensch et al. 2001
- [60] W. Press, S. Teukolsky, W. Vetterling and B. Flannery, Numerical Recipes, Cambridge University Press, 1992
- [61] E.Brookner, Tracking and Kalman Filtering Made Easy, John Wiley & Sons, 1998
- [62] M.Isard and A.Blake, Condensation: Conditional Density Propagation for Visual Tracking, International Journal on Computer Vision, Vol. 28(1), 1998