Model-Based Recovery of Human Body Structure From Two-Dimensional Images

by

Yiqun Fu

B.Sc.E.E., Nanjing Institute of Posts & Telecommunications, 1988

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE (ENGINEERING SCIENCE)

in the School

of

Engineering Science

Simon Fraser University

July, 1991

ISBN 0-315-78185-8

APPROVAL

NAME: Yiqun Fu

DEGREE: Master of Applied Science (Engineering Science)

TITLE OF THESIS: **Model-Based Recovery of Human Body Structure From Two-Dimensional Images**

EXAMINING COMMITTEE:

Chairman: Dr. Vladimir Cuperman

____

Dr. Thomas W. Calvert
Senior Supervisor

____

Dr. Kamal Gupta
Supervisor

____

Dr. John Dill
Examiner

DATE APPROVED: August 12, 1991

PARTIAL COPYRIGHT LICENSE

Title of Thesis/Project/Extended Essay

"Model-Based Recovery of Human Body Structure from Two-Dimensional Images"

Author: ______________________

(signature)

Yiqun FU

(name)

August 14, 1991

(date)

ABSTRACT

The recovery of the three-dimensional (3-D) structure of the human body from a sequence of motion images is a class of human motion analysis which is receiving increasing attention due to its complicated nature and numerous potential applications.

Various technologies for recovering 3-D information from two-dimensional (2-D) images have been proposed by many authors. The primary difference between our work and that of others is that we are focusing on recovery from a sequence of *single view* images. We propose a model-based approach as opposed to the usual method, where the starting point is to identify the correspondences between two view images.

We describe our model-based interpretation system in terms of three components: image analysis, image interpretation, and recovery and display of the 3-D body structure. The image analysis component applies noise removing algorithms and edge detection. The image interpretation component matches image features to a well-defined human body model which has geometric constraints. The third dimension is inferred by a kinematic method. The model also allows the prediction of image features. Finally, the recovered 3D structure is displayed in a 3-D graphics window.

Results from the experiments are encouraging. The system tracks the body motion and presents an "understanding" of the movement. It also gives a tentative explanation for occluded body motion.

Acknowledgements

I would like to express my deep appreciation to all members of my thesis committee - Dr. Thomas W. Calvert, Dr. Kamal Gupta, for their support and valuable instruction. In particular, I am deeply indebted to my supervisor, Dr. Thomas W. Calvert, for his guidance and encouragement throughout my whole graduate study, and for his very helpful comments and suggestions on the preparation of the final draft of this thesis. I also thank Dr. John Dill for reading the thesis carefully and making thoughtful suggestions. Finally, my special thanks to my parents and my dear friend Deming Liu for all the support and good advice which they have given to me.

CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

In the context of this thesis, motion is defined as the relative movement between an observing camera and the objects(or the background) of interest. Motion is related to the structure of objects in space and their positions relative to the viewpoints. Motion analysis may reveal information unavailable in the stationary images. Although motion analysis is a relatively young field, researchers are highly motivated because of the extensive applications. These cover a broad range of fields including medicine, autonomous navigation, tomography, communications and television, dancing and choreography, meteorology, and animation. For example, the automatic analysis of scientific-graphic image sequences of the human heart is used to assess motility of the heart and is finding application in diagnosis and supervision of patients after heart surgery; the processing of sequences of images for the recognition and the tracking of targets is of immense interest to the department of defense of every country; and the computation, characterization, and understanding of human motion in the contexts of dancing and athletics is another field of endeavor receiving much attention. In meteorology, satellite imagery provides the opportunity for interpretation and prediction of atmospheric processes through estimation of the shape and motion parameters of atmospheric disturbances. These examples are all concerned with motion and time-varying imagery analysis.

We focus on the subject of human motion analysis, which is a challenging research area because the human body has a highly complex structure which can twist, bend, and rotate. The goal of this project is to work from a computer vision point of view and try to reveal 3-D information from a set of single view 2-D images.

In this chapter we discuss the different methods used for understanding and determining the 3-D structure of the human body. Methods employed by university and industry working in the human motion research area are introduced. We are especially interested in the computerized approaches for recovering 3-D information from stereo images or monocular images.

The remainder of this thesis is organized as follows. Chapter 2, is a literature review of methods for the measurement of 3-D structures from images and related research. In chapter 3 we discuss work in how, given a sequence of single view human movement images, the 3-D information from each feature joint can be recovered. The recovery is based on a model matching approach. In chapter 4, we describe the use of image preprocessing to remove the impulsive noise and white Gaussian noise in the image. Geometric feature edge detection and thinning is also introduced in this chapter. In chapter 5, we describe the human body model and its kinematic constraints, the recovery process which matches a geometric image model to a human body model, and predicts the appearance of feature points and explains occluded body parts. In chapter 6, we present the results and evaluate the whole system.

Finally, chapter 7 summarizes the thesis.

1.1 Human Motion Analysis From Instrumentation

Human body movement can be captured in real-time with special instrumentation. One of the many quantitative tools available for the study of human movement is to attach electrogoniometers to the different parts of the body.

To monitor the pattern of stepping during human locomotion, the C.A.R.S.-UBC electrogoniometer involves three orthogonal potentiometers at each joint. The potentiometers measure three cardinal plane rotations at the hips, knees and ankles bilaterally. The closure of switches by a metal contact attached to the subject's foot assists in determining heel strike and toe-off. The surface of the walkway is covered with copper strips which form switch contacts. These switches allow the contacts between the foot and the ground to be recorded[Hannah80].

Another promising approach, similar to the goniometers, involves a specially instrumented body-suit which gives the computer real-time analog signals proportional to the angle of each joint.

A different class of instruments is based on optoelectronics. Rows of light emitting diodes (LEDs) are attached to the lateral aspects of the neck, upper and lower limb segments, feet, and the pelvic girdles. The LEDs flash at a frequency of 20Hz. Small light bulbs at leg and arm levels are turned on by foot-floor contacts. Two cameras, with open shutters and winders, record the sequential displacements of the rows of lights as the subject walks in the field of view in semidarkness.

Such records are projected onto a digitizing tablet. Using a special stylus, the relative positions of the light bars are digitized and used to derive the angular changes at the angle, knee, hip, elbow, shoulder, thorax, and pelvis and the vertical and lateral displacements of the body.

Although instrument systems are expensive, they do not need sophisticated computer analysis. The great advantage is that the attachments are easy to place and of little encumbrance for the subject. However, they pose the important problem of their identification, which is difficult to extend over a large distance, as each sensing strip needs either its own switch or a wire connection to the controlling electronics; further, a resolution of one or two millimeters over a length of several meters may require many hundreds of wire connections[Perry90].

1.2 Interactive Computer-Aided Image Digitizing Analysis System

Compared to instrumentation, an interactive computerized image digitizing system is reasonably priced. It allows the experimenter to get the results quickly without requiring the human to wear different kinds of markers.

Generally, the analysis system utilizes a mouse or a stylus which is used to select significant points. The system works with one video frame at a time. Specific software performs direct linear transformations on the 2-D data from each camera and converts them into 3-D coordinates. The system connects the points, creats 3-D representation figures, and can automatically calculate the walking speed and other parameters[Perry90].

University of California at Davis has been doing research in human body movement for a rather long time. In a research project which helps Olympic swimmers in training, they developed a less cumbersome process that involves using high-speed cameras, and then projecting developed film onto a digitizing tablet. An electric puck or pen is used to input each set of joint coordinates into a minicomputer. Software can calculate angles, forces, and speeds of the motion for the swimmer's arms, legs, torso, and other body parts. It has been proved that swimmers benefit from the result of this analysis system. They no longer need to put pads with pressure sensors on their hands and legs[Perry90].

This kind of system has the advantage of freeing the experimental subject from wearing markers, however, it takes time for the analyst to put the coordinates into the system, especially when precise data are necessary.

1.3 Automatic Digitized Analysis System

Automatic digitizing systems, consisting of video cameras connected to personal computers or workstations, can capture simple motions in real time. These systems rely on reflective markers or infrared LEDs to identify significant body points. A controller wired to the LEDs and to the computer allows software to keep track of which spot is lit, and an infrared-sensitive camera detects the lights.

A system made by the Motion Analysis Corp. of Santa Rosa, California, is typical of the reflective method of data capture. As many as 30 reflective markers, are attached to the subject at points to be tracked. Up to four cameras, run through a video processor to a Sun workstation, track the markers at a rate of up to 200 hertz. If the motion is basically linear, like a walk or a run, the computer can be instructed to correlate markers to body parts, track these points by identifying pixels that exceed a certain brightness, and connect them to create stick figures for display and analysis[Perry90].

Automatic analysis systems have now become the most common approach to capture the actual movement patterns of a live subject because of their speed and accuracy. But there are still some problems left: the pattern recognition problems involved are quite difficult, since the human body joint looks different from every angle and changes with lighting conditions. When motion involves twisting, which can obscure markers or cause their paths to converge, the computer software may flounder.

1.3.1 Problem With Single View Human Motion Analysis

The systems described above usually require two views or three views for any meaningful human body movement analysis. But in ordinary life, most pictures of physical object are taken from a single view. This makes analysis more difficult. People are giving increasing attention to the use of computer vision technology to derive 3-D human motion from single view images.

Various methods have been proposed to determine the 3-D structure of objects from images obtained via a single camera. The state of the art technologies are shape from shading, shape from texture, shape from optical flow, and shape from contour. Some techniques are more mature than others, but they seldom have been used in the human motion area. The difficulties mainly come from the human body itself. The human body has a very complicated, always-varying structure, especially when the body is in a twist, or jump movement.

Badler [O'RourkeBadler80] worked on the human motion analysis problem about ten years ago. A model-driven analysis was proposed to track the motion in three dimensions. But his emphasis was on the constraint propagation, in which he attempted to interpret low-level knowledge in accordance with a world model of the object.

Similar work was done by Lee [LeeChen85]. He proposed a binary interpretation tree to determine the final feasible body structures of a person from a single view. Physical and motion constraints were used to prune the interpretation tree. But his interpretation was based on: (1) at least six feature points on the head and a set of body joints available on the image plane, and (2) knowledge of the geometry of the head and lengths of body segments formed by joints. This approach has some unsolvable constraints in practical applications.

During this thesis work, we will investigate the problem of recovering the 3-D human body structure from a single view. A model-based recovery system is proposed which is able to track and explain the human body movement.

Chapter 2

Literature Review

In principle, it is impossible to capture complete 3-D information directly from a single monocular image. However, it is possible to use certain cues to partially deduce the third dimension. Vision is not only a general single action, it is also a sequential recursive and cyclical process of alternating information gathering and decision making. The 3-D structure of the environment, together with the laws of optics and photometry, determine the structure of the 2-D visual field. Given something known about a scene, a monocular image or an image sequence may often provide extra 3-D information. Under certain assumptions, it is often possible to perform the inverse mapping from 2-D images to a 3-D model.

Although the information processing involved in analyzing a sequence of a single view images presents a serious technical problem, even today, researchers have made progress in interpreting 3-D information from 2-D images. The solutions appear to follow one of several distinct methods: determine 3-D structure from stereo vision, from optic flow, from shading, from texture, and from contour.

2.1 Determine 3-D Structure From Stereo Vision

In the stereo paradigm, stereo matching methods involve two or more images. The matching is implemented using the low-level image features, such as image intensities or image edge points. Given the relative geometry of the two cameras (eyes) that acquired the images, simple trigonometry determines the depth of the matched features in two images.

Martin Herman[Herman84] proposed his method which matches structural features, i.e., different kinds of junctions extracted from the two images. The process contains three steps:

(1) Extracting line features from images;

(2) Extracting different kinds of junctions, for instance, junctions of 'L' shape which consists of two line features; and

(3) Finding potential matches between the corresponding junctions in the two images.

In his method, knowledge about the analyzed scenes is also involved. For instance, the roofs of building tend to be parallel to the ground plane, while the walls tend to be perpendicular to this plane. Face boundaries visible in both images are selected for matching rather than those which may be occluded in one of the images. L-junctions and Arrow junctions are initially assumed to lie on a horizontal scene plane. When they are found in one image, the shape and orientation of the corresponding junction in the other image can therefore be predicted by using this knowledge. Each junction in the first image is associated with a set of potentially matching junctions in the other images.

Referring to the problem of how many matched points are enough to derive the 3-D structure of the model, Rogers and Adams[RogersAdams76] showed that at least six object points were needed to determine their 3-D coordinates from a single 2D view. Later Ullman[Ullman76] points out that one can determine the exact model of a nonplanar structure over four points, from three views of this structure, using the assumption of parallel projection to model this imaging process. Roach and Aggarwal[RoachAggarwal80] studied the recovery method based on five noncoplanar points over two views under perspective projection and gave a general formulation for the relationship between the number of points and views.

Stereo vision is only useful for objects within a restricted portion of the visual field and a limited range of depths for any given degree of eye vergence, and is never useful for distant objects. At any moment, most parts of a scene will be outside of the limited fusional area, so the stereo vision system will fail to give a solution.

2.2 Shape Recovery From Optical Flow

Optical flow is the distribution of apparent velocities of movement with brightness patterns in an image. Optical flow can arise from relative motion of objects and the viewer. There is a relationship between optical flow changes in the image plane and the velocities of objects in the 3-D world. In the optic flow paradigm, two or more images are needed to compute the image velocity of corresponding scene points. If the camera's motion and imaging properties are known, we can use simple trigonometry to convert velocity measurements of the corresponding pc.nts in the image to depths in the scene[HornSchunck81].

There are two broad classes of methods to compute optical velocities: feature based and gradient based. In feature based methods, matching is the main operation. It generally provides a process that tracks characteristic brightness patterns from frames for a time-ordered sequence of images. The optical flow patterns are matched to get the corresponding points in the images. Gradient-based techniques rely on an equation that relates optical velocities to spatial and temporal changes in the image:

$$\frac{\partial f}{\partial x}u + \frac{\partial f}{\partial y}v + \frac{\partial f}{\partial t}t = 0 \tag{2.1}$$

where f is the image brightness function, t is the time, and u and v are the x and y components of optical velocity. To solve the equation above, we may need additional constraints:

(1) The surface being imaged is flat;

(2) The incident illumination is uniform across the surface, which assures us that the brightness at a point in the image is proportional to the reflectance of the surface at the corresponding point on the object;

(3) Reflectance varies smoothly and has no spatial discontinuities, which assures us that the image brightness is differentiable;

Since the motion of the brightness patterns in the image is determined directly by the motion of corresponding points on the surface of the object, object structure can be derived when the optical velocities are calculated.

2.3 Shape Recovery From Shading

The process of recovering surface orientation from image shading has been primarily studied by B. K. P. Horn and his colleagues at MIT[Horn77]. His method is explained as follow: The intensity I at a point (x, y) in the image plane is a function of the corresponding surface normal, and can be formed as the solution of a non-linear first-order partial differential equation in two unknowns. This equation can be solved using a modified characteristic strip-expansion method, assuming the surface is smooth. Horn introduced a reflectance map R(p, q) which represents the relationship between surface orientation and surface brightness. The shape of the object can be described by its height, z, above the xy-plane:

$$p = \partial z/\partial x \qquad and \qquad q = \partial z/\partial y \tag{2.2}$$

The pq-plane is referred to as *gradient space*, since every point in it corresponds to a particular surface gradient. Each point in the gradient space is associated with the brightness of a surface patch with the specified orientation.

The reflectance map can be obtained experimentally and can also be determined theoretically if the surface-reflectance is known as a function of the incidence, emittance, and phase angles. Once the brightness, I(x, y), is known at a point, we can get the possible surface orientations at that point, since the brightness I(x, y) restricts the possible surface orientations at the corresponding point on the surface of the object. This constraint is expressed:

$$I(x,y) = R(p,q) \tag{2.3}$$

Some additional constraint is necessary for a unique solution to the equation above. That usually comes from the assumption about the class of surfaces. Ikeuchi and

Horn [IkeuchiHorn80] proposed an iterative method for computing shape from shading using occluding boundary information based on this theory. They applied this method to analyze scanning microscope pictures and other applications.

2.4 Recovering Shape Orientation From Texture

Natural texture provides an important source of information about the local orientation of visible surfaces. Assumptions, together with the constraints imposed by projective geometry, could be applied to recover the shape, since the different distortions effects of the projection apply to different properties of the texture.

The recovery of surface orientation using texture was first investigated by J.J.Gibson[Gibson66], who studied the perspective projection of a ground plane. Assuming the plane to be covered with elements of a uniform density, and that those elements' projections could be identified and counted, he observed that under these assumptions, the gradient of texture density specifies surface orientation, where texture density was defined as the number of elements per unit area in the image. Gibson proposed the density gradient as the primary basis for surface perception by humans. This theme has since been pursued extensively. Subsequent work has largely accepted Gibson's premises and concentrated on geometric research using textures of known uniform properties.

Image texture gradients on oblique photographs can be used to estimate the surface orientation of the observed 3-D object. The first work of this kind was done by Charton and Ferris [ChartonFerris79]. They made use of the surface orientation of the surface over the object. The basis of the method was an analysis that related surface slant to the texture gradient in the perspective projection image. They measured the number of texture elements in a line by measuring the number of changes in brightness along the line. The number of changes in brightness was the number of relative extrema.

Other work that relates to the recovery of surface orientation from texture includes that of Kender[Kender80], who described an aggregation Hough-related transform by grouping together the edge direction associated with the same vanishing point. An edge direction E = (E_x, E_y) at position P = (P_x, P_y) has coordinates T = (T_x, T_y) in the transformed space where T = E$\star$P.

These methods can work well; however, natural textures are so unpredictable that no attempt to model their geometry precisely has much chance of success.

2.5 Shape Recovery From Contours

A surface contour is the image of a curve across a physical surface, such as the edge of a shadow cast across a surface, a gloss contour, wrinkle, seam, or pigmentation marking. The contours describe the surface shape along the boundary when the surface is smooth. Contour is an important resource for recovering object structure. However, from one view of a planar contour there exists infinite solutions for shape, so certain assumptions about the geometry of the curves must be made if we are to use them to infer surface shape. It is also clear that geometric constraints need to be considered, which determine the properties of various types of real physical curves. Combining a generic surface description with a model of image formation, the method consists of three steps:

(1) The contours are extracted and the relationships among them established;

(2) Among these contours, the ones which form a desired configuration, are selected;

(3) The selected contours are combined with constraints that come from the image formation process, in order to be interpreted in terms of discontinuities in surface orientation. The result is that the number of possible orientations of the associated scene surfaces is considerably decreased.

There are several different aspects to the problem of recovering shape from contours. We will emphasize the interpretation of lines and points that recovers the shape of the object.

2.6 Depth Recovery From Line Drawing Interpretation

Line drawings can result from image segmentation and contour analysis. Line drawings connected often give cues to recovering depth information of the imaged object. Various methods have been proposed for recovering 3-D structure of an object by analyzing the line drawings.

Significant work was done by Guzman [Guzman68] on the segmentation of bodies in a scene containing polyhedra. He first defined types of junctions consisted of several line drawings and developed many heuristics concerning probable association of regions suggested by each junction type. His SEE program accepted a line drawing and produced output lists identifying and describing the bodies present in the scene by using a set of heuristic rules related to the types of vertices. The basic idea behind SEE was to make global use of information collected locally at each vertex: SEE combined different kinds of strong or weak evidence to make reliable global judgements. The results were successful in decomposing even rather complex scenes of polyhedra. But his heuristics were very ad hoc and his program was intended only to partition the scene into bodies and provide this as input to a recognizer which might derive 3-D descriptions.

Huffman [Huffman71] and Clowes [Clowes71] stressed that the relationship between the scene and the image needs to be made explicit. Huffman classified lines that were the projections of edges into 3 types: convex, concave and occluding. Assuming that all images were taken from a general position, he showed that for a trihedral world, junctions could be catalogued into only 12 possible types. The consistent labeling of the lines in an image uniquely corresponded to a particular 3-D scene. If a picture had no possible labeling, it was impossible to realize it. Clowes determined a consistent interpretation by a search space technique.

Another attempt at reconstructing curved bodies was made by Freeman and Loutrel [FreemanLoutrel67]. For 3-D bodies with vertices formed by three faces, a cyclic-order property was defined. The property augmented the grammatical rules that govern the possibility or impossibility of the existence of 3-D bodies corresponding to 2-D line-structure projections.

The work described above is all aimed at getting more information about the object itself with rules employed to interpret the line drawing of an object. They do not recover the depth from the scene directly, but from the assumptions combined with knowledge about the object.

2.6.1 Model-Based Image Interpretation

Another model-based line interpretation approach has been proposed to recover the depth. It is quite different from the methods described above, where most of the initial work was based on low-level image processing and information extraction. The depth information is inferred by matching the 2-D images in terms of an object model.

Roberts[Roberts65] did pioneering work on interpreting line drawings from a photograph. After the line drawings were found, they were matched to a model and the 3-D information was inferred from the images. His matching process consisted of the selection of junctions corresponding to the vertices of the object, and the matching of the junctions with the vertices predicted by the object model. He proposed that for verification of match, at least seven point-to-point correspondences should be required for object and model parameters. Since his program was able to predict other views of the scene, it marked a significant break from pattern recognition by emphasizing descriptions of the objects present in a scene and the spatial relationships between them.

Other significant work was done by O'Rourke and Badler[O'RourkeBadler80], who have long been involved in human body movement research. In this approach, they advocated a model-based system for the human motion interpretation. Body feature position identification and localization were based on a matching between the image features and the human body model. They represented an articulated model of the human body by 24 segments and 25 joints. The shape presentation of each segment was in turn characterized by a set of spheres. The system was structured as a closed loop between a high level component, (i.e., the predication), and a low level component, (i.e., image analysis). The center of the proposed approach was a human model. The input to the image analysis component was a list of 3-D regions where various body features are predicted to appear; a search for the actual location of a feature was conducted by matching the model within the area predicted for that feature. The 2-D location boxes of extracted features are fed back through a constraint network to refine the 3-D best-guess rectangular boxes of features. The constraints checked included distance measures, angle limits, and collision detections. They were an important aid to the interpretation of motion.

2.7 Two Model-Based Vision Systems

In this section, we introduce two model-based vision systems that have been used in image understanding. Given models of the different objects, they have the potential of automatically calculating the positions and orientations for corresponding objects. These two systems have provided improved algorithms for 3-D information recovery.

The ACRONYM system [Brooks83] was the first model-based vision system which succeeded in using a general symbolic constraint solver to calculate bounds on viewpoint and model parameters from image measurements. Matching was performed by looking for particular sizes of elongated structures in the image space(coded ribbons) and matching them to potentially corresponding parts of the model. The bounds given by the constraint solver tree were then used to check the consistency of all potential matches of ribbons to object components. While providing an influential and very general framework, the actual calculation of bounds for such general constraints was mathematically difficult and approximations had to be used that did not lead to exact solutions for a viewpoint. In practice, prior bounds on a viewpoint were required which prevented application of the system to full 3-D ranges of viewpoints.

In 1987, David G. Lowe[Lowe87] implemented a computer vision system that could recognize 3-D objects from unknown viewpoints in single gray-scale images. Unlike most other approaches, the recognition was accomplished without any attempt to reconstruct depth information bottom-up from visual input. Instead, three other mechanisms that can bridge the gap between the 2-D image and knowledge of 3-D objects were used. First, a process of perceptual organization was used to form groupings and structures in the image that are likely to be invariant over a wide range of viewpoints. Second, a probabilistic ranking method was used to reduce the size of the search space during model-based matching. Finally, a process of spatial correspondence brought the projections of a 3-D model into direct correspondence with the image by solving for the unknown viewpoint and model parameters. A high level of robustness in the presence of occlusion and missing data was achieved through full application of a viewpoint consistency constraint. It was argued that similar mechanisms and constraints form the basis for recognition in human vision.

2.8 Two Approaches in Recovering 3-D Structure

Although many methods have been developed to recover 3-D information from 2-D images, basically they can be divided into two categories: bottom-up and top-down . A bottom-up approach relies on low-level image processing. Usually low-level image processing which is carried out for image interpretation can be roughly divided into three phases: (a) moving body parts are separated from the background; (b) the moving body part features are then labeled; (c) motion verbs are assigned to the movement. These steps may be slightly changed in different applications.

Kanade[Kanade81], in the first part of his method for recovery of the 3-D shape of an object from a single view, focused on line-labeling according to the junction dictionary. Each junction label has attached to it information on the constraints in the gradient space that should be satisfied by the surfaces incident at the junction. In this approach, he had a detailed line finding method based on edge detection and line linking, and then assigned a junction label to each junction one by one. His method represented a traditional bottom-up approach: first the primitives and relations are extracted and then a preliminary object description is constructed.

A different method was provided by O'Rourke and Badler[O'RourkeBadler80]. They did not preprocess the image to segment it into regions or detect edges, and no low-level processing was performed on the whole image. Preprocessing was only done when needed, and only within the area predicted for a particular feature. In this method a model with feature detectors was applied to match structure in the predicted image region. The matching is complete when the features have been localized to a small enough region. The output is a list of 3-D regions where the various body features had been found. This represents a top-down approach: the model was used to predict or anticipate future position of the body using processing which started from the model description.

2.9 Project Objective

For the project described in this thesis, we would like to implement a model-based vision system capable of tracking human body parts in motion and describing the body structure in 3-D coordinates. The input to the system is a sequence of single view, grey-level images. The output is a list of 3-D human body data structure. We would like to build a system which has the ability of:

(a) Identifying the feature points attached to the body;

(b) Tracking human body parts in a set of motion images;

(c) Inferring the 3-D human body position from a sequence of single view images, and;

(d) Estimating the position of occluded human body parts.

Chapter 3

General Approach

Much of the current work in interpreting single view images is based on trying to extract maximal information from an image without using any knowledge about the objects being viewed. Usually the techniques are based on physical considerations concerning the image producing process. Other approaches, those principally proposed by David Marr and his students[Marr81], start from the physiological view, and use algorithms which extract information including identification of surfaces and their local orientations. The idea behind this is to make use of rich descriptions to interpret the images. Researchers have proposed many approaches based on these ideas and made some progress in recoving 3-D information from 2-D images. However, human body movement interpretation still remains a rather challenging and difficult problem. Although researchers have been working on this problem for more than ten years, they have not made a great deal of progress due to the following difficulties:

(1) The physical world is 3-D, but an image contains only a 2-D projection of

this reality. Not only will the projective image change with changes of position, orientation, and distance of the human body, but it is possible that different 3-D human body positions have the same 2-D projection.

(2) The human body is an extremely complex object, being highly articulated and capable of a bewildering variety of motions. Rotations and twists of the body parts occur in nearly every movement, and various parts of the body continually move in and out of occlusion.

The difficulties described above make it nearly impossible to infer 3-D structures directly from their 2-D projections. During this research, we simplify the domain by only considering a single human in an environment devoid of others objects. We propose a model-based interpretation approach to recover the 3-D information from its single view 2-D projections. The approach employs a human body model to match the image features and to predict the position of the body.

3.1 Model-Based Approach

The human visual system has no difficulty interpreting human motion. The motion of a human body can be correctly inferred from just a few lights placed on the body[Marr81]. One experience [YoungFu86] shows that only about 200 ms are required for the motion to be identified as human. This indicates that the human visual system has a remarkable capacity for interpreting changing images caused by rigid motion of 3-D objects. [Rashid79] has developed a method which does not employ a model of the human body. They have achieved some success in interpreting images with lights attached to the human body, which demonstrates that the model may play a lesser role in object separation and tracking tasks. But they mainly concentrate on movement with no occlusion of hands, etc, and with minimal complicated movements. In order to handle occlusion, to detect twist in body segments, to infer the locations of anatomical landmarkers which have no visible counterpart, and ultimately, to generate a rich semantic descriptions of human motion from images, it seems clear that a human body model is necessary. In this research, we have a set of consecutive human body movement images which have arms twisting and body occlusion. We apply a human body model with some kinematic constraints to our task of prediction process and recovery process.

3.2 System Overview

The method we have chosen in this human motion interpretation research is to combine low-level image processing with high-level motion interpretation. From the low-level image processing, we extract maximal useful information from the images, including geometric features such as lines, circles, etc. Interpretation builds on the results of this low level processing. The information from the single view 2-D images is interpreted in terms of a human model with well defined structure and kinematic constraints. Another task of interpretation is to predict the future positions of those feature points—this directs the task of image feature finding by matching the model of geometric symbols to the image characters at the predicted position. So basically, we divide our process into three parts: image analysis; image interpretation; and 3-D structure of the human body recovery and display in a graphics window.

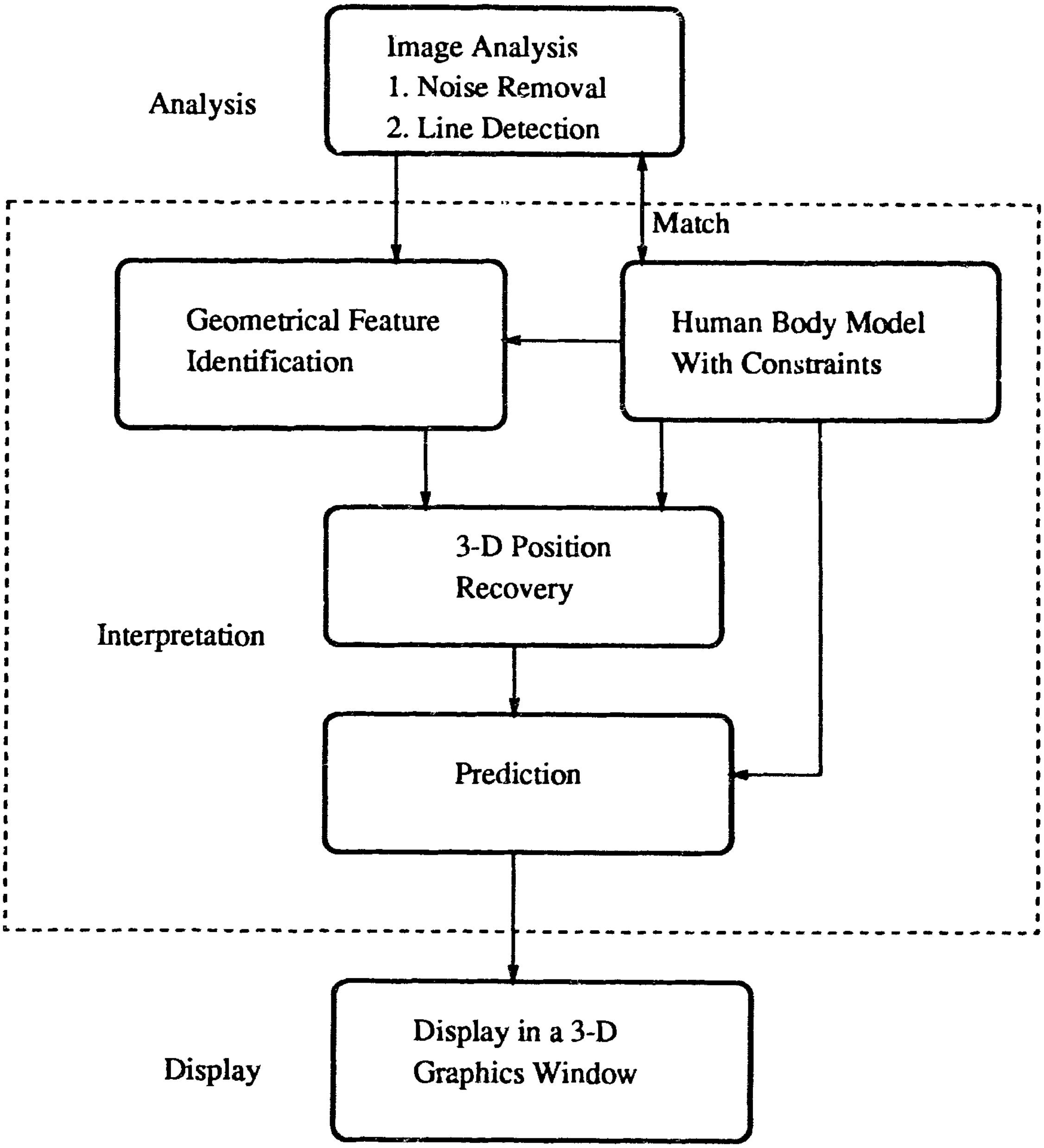


Figure 3.1: Configuration of the Model-Based System

3.2.1 Input to the System

A wooden model of a human body, which is about 20 centimeters tall, is used as the source of the pictures in this research. The physical structure of the simulated human body is known. This includes the length of body segments formed by angle joints and the ranges for each joint angle. Constraints on the body, including angle constraints, motion constraints and collision constraints are encoded into the model. Up to seventeen markers are attached to the feature positions of the wooden body model. These markers are geometric symbols, such as squares and circles with different colors. These are used to identify the positions of the head, the chest, the arm joints, the legs, and the feet.

The test images contain a set of consecutive samples of a simulated body motion, such as the arm swing movement. We execute each simulated body movement by moving the arms in a continuous, smooth function. The movement also embodies extensive knowledge and constraints of the human body, for instance, it will not move any limb beyond the limitations of its associated joint, nor will it move one body part through another. The simulated human body movements are based on gravity and balance.

The set of images is produced by a CCD video camera, digitized and stored in Image File Format(IFF). The machine we have chosen to analyze the images on is a Sun Sparcstation. All the images are stored in the workstation with the intensity value ranges from 0 for dark to 255 for light, thereby yielding a 512 × 512 × 8 resolution. The images were set up to give a rather strong contrast with respect to the background.

Two-dimensional images can be generated from a 3-D scene by parallel or perspective projection. Perspective projection is chosen here because human eyes and cameras have that function. Assuming that the focal length of the camera and its position relative to the simulated body is known, and since the viewpoint of the camera is invariant for a given set of body movement pictures, we can set up a relation between a camera-centered coordinate system and an object-centered coordinate system. The position of the human body in the object system is derived through multiplying the coordinates in the camera-centered system by the camera-object transformation matrix.

3.2.2 Image Analysis

The image analysis component is the only process which deals directly with the images. The input is a list of 2-D images with some flesh-colored areas which represents the attached feature areas. This process performs some basic image processing, such as impulsive noise removal and White Gaussian noise removal. After that, the image is processed to detect edges and do line linking. These line features are passed to the next stage of image interpretation to identify the attached geometric symbols to the human body model.

3.2.3 Image Interpretation

The image interpretation process is designed to recover 3-D position information using a model based search. The feature points are first identified by matching extracted line characters to the model features of circles or squares predefined to the human body model. The matching is under the constraint environment. Then, the information about the third dimension at these feature points is derived from the human model using the kinematic methods. The interpretation process also has the function of predicting the future position of the feature points based on distance constraints and motion constraints.

3.2.3.1 Description of the Model and the Constraints

The model of human body plays a important role in the interpretation process. It contains all of the 3-D knowledge about the human body and has the function of:

(1) inferring 3-D information for a matched 2-D feature points;

(2) being able to predict the future position of the feature points.

There are several representations of 3-D model: surface representations (points, polygons, or surface patches) and volume representations (polyhedra, ellipsoids, generalized cylinders, spheres). Considering that the stick figure has the advantage of propagating the geometric angle constraints and distance constraints throughout the model, we choose the stick figure [Calvert88] as our model representation. The nodes and links represent primitive topological and geometric constraints. The principal constraints we consider are:

(1) Angle constraints at each joint ;

(2) Distance constraints for a segment between two feature points;

(3) Motion constraints based on the cooperative movement of different parts of the body.

They are list in the appendix A.

3.2.3.2 Matching Methods Used in the Interpretation

The matching between the line patterns and the feature models is applied to identify the geometric symbols such as square, triangle, etc. (consisting of lines and curves), and to assign them different names corresponding to the human model. Also the matching has the task of predicting the future position of the feature points to help the first matching in future frames.

The image analysis process produces line and curve features. To group them and identify which pattern they represent, these line and curve features are matched to the predefined geometric models which attach to the human body model. Kinematic constraints are supplied to get an explicit identification. This high-level matching is actually a confirmation of the feature points identified from the image plane match.

A matched image description already has 2-D coordinate information; since the model has 3-D information about the simulated human body used as input, we can apply a kinematic method, starting from the chest point, then moving to the shoulder point, elbow point and wrist point, etc. to derive a complete 3-D description of all parts of the body.

3.2.4 Display of the Recovered 3-D Position of the Human Body

After the 3-D structure of the imaged human body is derived, we present the results by displaying them in a 3-D graphics window on a SUN Sparcstation. The display is programmed in C and makes use of the Hoops software package. It has a graphics interface that allows the user to rotate the simulated human body through 360^0 freedom and the user can also see the human body including the occluded parts from three directions: front, right, top. The display of the results uses a stick figure [Calvert82].

3.3 Expected Results

People have done a lot work on recovering 3-D information from images. But most of the successful work is based on two view images. We expect our system, given a sequence of single view images of a human in motion, should be capable of tracking the motion in 3-D space and "understanding" or describing the motion in some form.

Starting from a sequence of single view 2-D images showing a human body movement, we want the system to extract the attached feature information from the images, merge it with the human body model and its constraints, and generate a set of 3-D body structures representing possible body movements. The output is the 3-D position information recovered for each feature point. A 3-D stick figure representing the human body is built and displayed on a 3-D graphics workstation.

Occlusion occurs when one body parts interferes with the view of another. We would also like to investigate whether the system could detect and analyze the occlusion, find the positions of occluding parts, predict their reappearance, and estimate their 3-D coordinates.

Chapter 4

Image Analysis

In computer analysis of time-varying images, high-level understanding is achieved by building on the results of low-level processing. Low-level image processing mainly refers to feature extraction. The approaches described below show that researchers have been focusing on better approaches to detect the markers attached to the human body.

Huffman[Huffman71] reported a video-based method for tracking four markers, which were fixed to the specimen surface, and calculated the 2-D information from the measured marker positions and displacements. The markers were first identified using a threshold search, and the marker centers were found using a centroid formula. This method did not include on-line experimental control, and the data acquisition and analysis rate were really slow. In addition to this method, Humphrey et al.[Humphrey83] proposed a tracking algorithm utilizing a threshold search followed by a method of rows and columns(MRC) marker identification whereby the center of a marker is defined to be that pixel location corresponding to the maximum (light marker) or minimum (dark marker) sum of a row and a column of pixel intensities. An advantage of this approach is its ability to independently track four very small markers, thereby allowing estimation of shearing strains and the homogeneity of the strain field. However, due to hardware and software limitations, the data acquisition and experimental control were very slow. Another approach is dynamic scene segmentation. Dynamic scene segmentation consists of dividing images into changing parts and constant parts, and locating the significant moving feature parts in each image sequences.

The approaches introduced above did not use a model of the object. They all have the disadvantage of low processing speed due to searching the whole image area. The model based feature detection method which we propose here aims to provide a faster feature localization. The input to our method is a sequence of images of a simulated human body which has some geometric symbols to identify the joints. We begin our method by smoothing the images, extracting edge features, and identifying the geometric symbols through matching to a catalogue of stored 2-D geometric model descriptions. In this section, we introduce the image noise removal and edge detection algorithms.

4.1 Image Processing

The function of image processing is to get as much information as possible about the feature in the images. Although we provide a strongly contrasting background for the figure in the video image, nevertheless, when the pictures are digitized, there is a certain amount of noise due to various factors. The factors considered here include:

(a) Random white noise;

(b) CCD video camera device instabilities;

(c) Varying lighting condition.

These factors can cause severe distortions in the digital image and hence ambiguous features and poor recognition results. Mathematically, they are roughly divided into: (a) Impulsive noise which appears as random white spots of high positive grey level value in an image. It is usually caused by errors during the image acquisition or transmission through communication channels. Median filters have good impulsive noise filtering capabilities; (b) White Gaussian noise whose grey level value is in a normal distribution with a covariance σ . The noise energy is represented as σ^2. In the image, it results in some points with low grey level since usually the noise energy has a relative low value. Thresholding is efficient and simple to remove this noise.

4.1.1 Impulsive Noise Removal

A median filter has good performance in rejecting the sharp details due to the addition of impulsive noise; in this approach we replace the grey level of each pixel by the median of the grey levels in a neighborhood of that pixel, instead of by the average.

The principal function of median filtering is to force points with very distinct intensities to be more like their neighbors, thus actually eliminating intensity spikes that appear isolated in the area of the filter mask. In our approach, we use a 3×3 mask to implement the median filter. It was defined as follows: sort the 9 pixel level values of the mask, determine the median which is the fifth largest value among 9 values here, and assign that value to the central pixel.

We found that this nonlinear signal processing method is particularly effective when the noise pattern consists of strong, spike-like components, and where the characteristic to be preserved is edge sharpness which is utilized later.

4.1.2 White Gaussian Noise Removal

During thresholding, the brightness value of each pixel is compared to a threshold value, and the pixel is assigned to one of two categories depending on whether the pixel value is exceeded or not. Thus those low grey level points will belong to the background after thresholding.

For a 512 × 512 image, consider a pixel with value f(x, y) at the point (x, y); this is compared with the thresholding value t, then the points with brightness value greater than t remain at their original grey value; while those less than t are set to the black background value. i.e., new pixel value $f_{new}(x,y)$ will be given as:

$$f_{new}(x,y) = \begin{cases} 0 & \text{if } f(x,y) < t \\ f(x,y) & \text{if } f(x,y) \geq t \end{cases} \tag{4.1}$$

The main point in this method is the selection of the threshold value. Although we can select threshold values which depend on the grey level value for a region of the image, it is probably unnecessary and it is a very time consuming. Usually we select a fixed threshold value for the whole image area, provided the character of the image does not change abruptly. A histogram of grey-level content provides a global description of the appearance of an image and provides a good way to select the threshold value. It counts how often each brightness value occurs: the points belonging to the background give rise to a peak in this graph since there are many of them, and similarly for the points belonging to the wooden human body; whereas the noise points give rise to a valley since the points rare. We choose a fixed threshold value to place these noise points in the background.

4.2 Feature Extraction

This is the process of localizing those areas in the image where a potential feature is likely to be present. Most of the techniques can be adapted to detect either light or dark targets. Burton[14] used a double window filter. This filter is based on the contrast between the target and its immediate background. It consists of two rectangular windows, in which the inner window surrounds the target, and the outer window contains background. Range is used to control the window sizes. Sklansky[15] used a spoke filter(an eight-spoke digital mask) which is an extension of the Hough circle detector. It examines the local edge magnitude and direction. It needs preprocessing which includes intensity normalization, and an edge detector. Rubin[16] used a linear discriminant function of local features of the image to obtain points of interest, while the target is assumed to be in the neighborhood of these points.

Our approach starts by detecting edges and then performs edge thinning. *A priori* knowledge of the objects being viewed together with the appropriate feature model are used to identify and locate the feature points in the image.

4.2.1 Edge Detection

An edge is a piece of the boundary between two regions with relatively distinct grey-level properties. Basically, the idea underlying most edge-detection techniques is the computation of a local derivative operator.

The gradient of an image f (x, y) at location (x, y) is defined as the two-dimensional *vector*:

$$\mathbf{G}[f(x,y)] = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{4.2}$$

For edge detection, we are interested in the magnitude of G[f(x,y)] which is given by:

$$|G[f(x,y)]| = mag[G] = [(\partial f/\partial x)^2 + (\partial f/\partial y)^2)]^{1/2} \tag{4.3}$$

The gradient gives the maximum rate of increase of f(x,y) per unit distance in the direction of G. The *direction* of the gradient vector is represented by α(x,y) with respect to the x axis.

$$\alpha(x,y) = \arctan(G_y/G_x) \tag{4.4}$$

In digitized images, the magnitude of the first derivative $\partial f/\partial x$ and $\partial f/\partial y$ at every pixel can be used to detect the edge of an image in a number of ways. One approach is to use the first-order differences in a 3 × 3 neighborhood about point (x, y) to estimate the gradient [Niblack86](see figure 4.1). We define the component of the gradient vector in the x direction as:

$$G_x = (x_3 + 2x_6 + x_9) - (x_1 + 2x_4 + x_7) \tag{4.5}$$

x_1	x_2	x_3
x_4	x_5	x_6
x_7	x_8	x_9

Figure 4.1: 3 × 3 Mask for Edge Detection

and in the y direction as:

$$G_y = (x_7 + 2x_8 + x_9) - (x_1 + 2x_2 + x_3) \tag{4.6}$$

G_x and G_y are combined together using the equation 4.2 to get G[f(x, y)] to obtain an estimate of the gradient at that point. Processing on the whole image yields estimates of the gradient at all points in the image. The magnitude of G[f(x, y)] is compared to a threshold value t. If it exceeds t, an edge point is produced by setting a new pixel value 0 representing black. Otherwise, the point has the new pixel value 255 representing white.

The edge detection process may produce some disconnected border points and broad edges because of the image noise and selection of threshold value t. To delete those points and thin the edge, the output from the gradient operator is checked at each edge point. Only those gradient points that are maximum among several directions are kept. Referring to Figure 4.2, four directions labeled 1, 2, 3, and 4 are considered since direction 6 is the same as direction 4, 7 the same as 3, and so on. The two neighbors in the direction that is closest to the direction of the gradient g_p of the center pixel p are checked, and if g_p is the largest, the other two points are set to background and eliminated. The detailed algorithm is introduced in [Niblack86]. The result is an image with only the one best point across the border at any point.

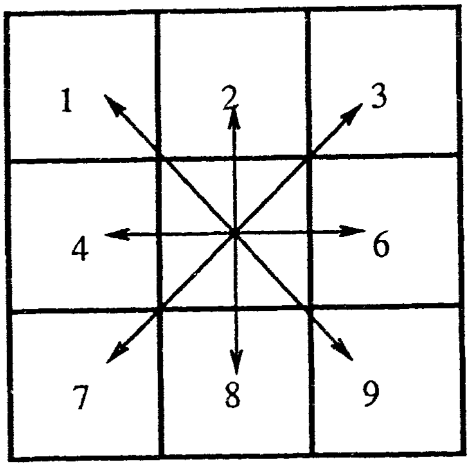


Figure 4.2: The Numbering Convention For Direction Around A Pixel

Chapter 5

Image Interpretation

In this chapter, we deal with the feature point identification and the 3-D recovery function of the system. Given a preprocessed image, the system should be able to apply the kinematic constraints to the prediction process and the matching process, to identify the image feature points along with their 2-D locations, and to infer the third dimension of the human body.

A model-based method is employed in our approach to interpret the single view 2-D images. It is basically divided into two steps: (a) identifying and labeling the line patterns corresponding to the human body model; and (b) interpreting the feature points and recovering the 3-D information. These steps are briefly described as follows:

(a) The image features are composed of line patterns. To identify the feature points, the first step is to group and label line patterns using *a priori* knowledge. The *a priori* knowledge includes the shape and position of geometric feature models, i.e., the circles, squares and triangles attached to the body model, and the initial posture of a sequence of human body movement. The feature models are applied to the matching process using instructions from the prediction process to find correspondences between components of the *image description* and the *feature model representations*. Kinematic constraints, like the distances between the elbow and wrist joints, along with motion rules, predict the future position where the feature point should appear and instruct the low-level feature model matching process. The output is a list of the image features with semantic names, together with their 2-D positions.

(b) The interpretation of the image feature points is implemented using the model knowledge together with kinematic methods. The human body model includes a complete 3-D data structure describing the human body. Each feature point follows the motion rule from the last n frames. The motion rule is applied to the recovery process by considering the direction and speed for the feature points in the last n frames. From the initial position, each feature point is tracked and its third dimension is inferred by calculating the difference between the projection length of each segment and its true length in the human body model. Kinematic constraints are provided to reject the infeasible positions of each feature point. For instance, to infer the 3-D position of the wrist points, its depth is recovered by calculating the squareroot of the difference between the square of the projection length of the segment $l'_{elbow-wrist}$ in the image plane and the square of the true length $l_{elbow-wrist}$ from the body model. The squareroot reflects the depth information for the wrist point. The recovery process is implemented under a tree structure: from the chest to shoulder, elbow, and wrist. This stage recovers the complete 3-D structure of the human body.

Finally, we also wish to know how well the recovery process performs when the feature points are not known at all positions, in particular, when the body parts of interest are occluded by another part.

In this chapter, we will first introduce the detailed human body model and its kinematic constraints, then describe the constraint based prediction process and matching process. Finally we focus on recovering the 3-D position of each feature point.

5.1 Human Model Description

A model is an organized representation of features which provides descriptions and information for image analysis and understanding. The model description can be roughly divided into two categories: 2-D model description and 3-D model description. In this thesis project, we are concerned about the 3-D model construction and representation. The construction of a 3-D model requires the model's coordinate system, its component axes in an image and the arrangement of the component axes in the model's coordinate system. This construction is appropriate when the viewpoint is fixed.

Our intention is to develop an appropriate 3-D model which would produce the 2-D appearance of the given body taken from a fixed point of view. The model is made up of structural parts which incorporate all the known information about the body. A description of it is given in the form of a relational structure in which the nodes correspond to features: geometric shapes and their 3-D positions. For example, the elbow point is expressed by a triangle at (12, 28, 140). We employ kinematic constraints to provide some properties of the human body to the model description.

5.1.1 Two kinds of Human Body Model

At the present time, volume models, surface models, and stick figures are the most widely utilized representations for the human body model. In the volume models, the body is decomposed into instances of one or more primitive volumes, such as cylinders, ellipsoids, or spheres. A few ellipsoids or spheres can capture the surface and longitudinal axis properties of many body parts. Among the volume models is the "BubbleMan" developed by Korein and Badler[Badler79] where the body is build up from the superposition of about 300 spheres. Another approach proposed by Herbison-Evans is a "Sausage Woman", which is built up from a smaller number of ellipsoids[MarrNishihara78]. The volume models are relatively efficient and robust, however, there are difficulties in refining them to give a truly realistic look.

Surface modeling is another approach. The body surface may be modeled by partitioning it into planar or curved patches. Representation based on a planar decomposition of the body surface can be implemented by vertices or polygons[Badler79]. The former has from 300 to 3000 vertices, which has the advantage of simple display primitives, but sacrifices the solid appearance of an actual body. The polygon representation has the advantage of solid rendering but has high display cost. Furthermore, polygon models of a jointed shape may yield unnatural results when the shape is changed at a joint.

5.1.2 Stick Figure Model Representation

A stick figure model is composed of joints and segments which contain kinematic constraints. Its representation for the human body is based strictly on the connectivity and flexibility of the body. We choose it as our model description in this project as it is easily implemented.

A 3-D stick figure model does not effectively portray complete human body movements, especially rotatory movements, twists of certain body parts, and contacts between body surfaces. To solve this problem, we employ three orthogonal graphics windows to display the recovered three dimensional structure of the human body. Then the twist of body parts, can be clearly viewed from three directions. In addition, the display is programmed to have a rotate function so that the viewer can understand the recovered body structure.

5.1.2.1 Composition of the Human Body Model

In our project, the stick figure model is defined to have 14 joints and 15 segments, as shown in Figure 5.1. The human body model is defined in a chest-centered coordinate system: with the x axis pointing to the right, y axis pointing up, and z axis pointing to away from the viewer. The model is facing towards the viewer and is in a left-handed coordinate system.

The model contains all of the system's "world knowledge" for the human body. It has a total of five degrees of freedom. The seg, ents and joints are linked together into a tree-structured skeleton. Each joint is a unique point connecting two segments. A segment is an abstract rigid body with an associated embedded coordinate system.

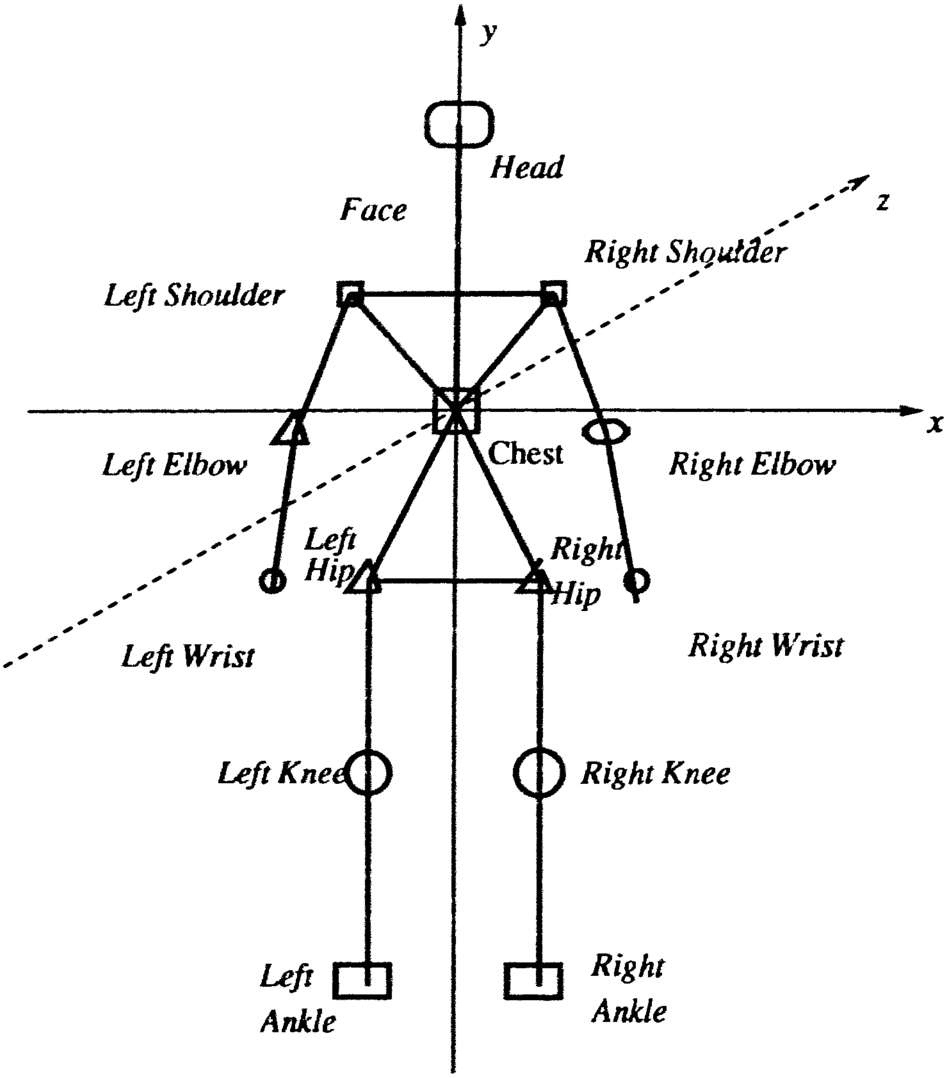


Figure 5.1: Stick Figure Representation of the Human Body Model

It has two joints located within its coordinate system. The segments, such as the torso formed by the chest, left shoulder and right shoulder, move rigidly; the only articulation permitted is at the joints. The angle constraints at the 14 joints and the lengths of all 15 segments are known. They are described in the next section as kinematic constraints applied to the model.

The geometric features attached to the human body are the components of the model. Image features such as edge, corner, line, curve, hole, and boundary curvature define individual feature components of an image. These features and their spatial relations are then combined to generate model descriptions. From the geometric symbols attached to the wooden human body, we build up our model by setting these geometric features, such as the circles and rectangles, at the corresponding locations of the model. A total of 14 unique geometric symbols are attached to the 14 joints of our stick figure model.

5.1.2.2 Object-Centered Coordinate System

The stick figure model is defined in an object-centered coordinate system. There are two kinds of object-centered coordinate systems that the 3-D model representation might use. In one, all the component axes of a description, from neck to wrist, are specified in a common frame. The other uses a distributed coordinate system, in which each component has its own local coordinate system. For the latter one, the spatial relations specified in a 3-D model description are always local to one of its components; for example, since the elbow point is specified in a local shoulder coordinate system, recovery is relatively simple to implement. We choose the local coordinate system to specify the relative arrangement of a 3-D model's component axes. In this tree structure of the human body model, the positions of each feature point are derived from the local coordinate system and its relative position to the chest axis.

5.2 Kinematic Constraints of the Human Body

Much of the research on vision relies on constraint analysis in matching and feature finding. For example, Marr[81] proposed compatibility constraints, a uniqueness constraint and a continuity constraint in his research work. Constraint analysis is especially useful when imposed on human body movement since kinematic constraints limit the scope of feasible movements. In a particular situation, such as when each joint of the body has movement limitations in angles and two body parts cannot occupy the same place at the same time, the specification of such constraints can be are very effective in recognizing each body part and justifying the result.

In this research work, we apply angle limitations, distances constraints and motion constraints to the human body model. These are the most general kinematic constraints for human body movement analysis. We check the violation of the constraints for any particular orientation of the body in the recovery process. Kinematic constraints arise from the structure of the simulated human body model and its 3-D properties projected onto the image plane. In the prediction and matching process, the constraints are actually the 2-D projection of constraints from the 3-D world. Each segment is permitted to have a certain orientation, which is expressed as 2-D angles limits projected from the 3-D world. The distance constraints between the body parts are also 3-D projections on the 2-D plane. However, in inferring the third dimension position of the feature points, 3-D constraints are directly applied.

5.2.1 Angle Constraints

Each joint of the human body has limits to its free angular movement, which give rise to constraints on the relative positions of body features on opposite sides of the joint. Angular constraints are used to reject particular illegal body positions during the prediction process.

Angle constraints are usually divided into four different categories associated with the human body joints. They are

(a) flexion/extension,

(b) abduction/adduction,

(c) rotation,

(d)bending.

Data on these angles can be easily found from references in the kinesiology. Some of the constraints are listed in appendix A. For the human body model employed in this project, the allowable ranges of the angles are slightly different from those in the references and we introduce them as follows:

(A) At the shoulder joint. Project the human body onto the x-y plane along the -z axis and the y-z plane along the -x axis respectively (Figure 5.2). Let the upper arm, either right or left, be specified by a vector **u** from the shoulder to the elbow. Assume that the projection vector of **u** onto the x-y plane and y-z plane is $\mathbf{u}_{(x,y)}$ and $\mathbf{u}_{(y,z)}$. Define:

θ = the angle from the **x** axis vector to the projected vector $\mathbf{u}_{(x,y)}$ (Figure 5.3).

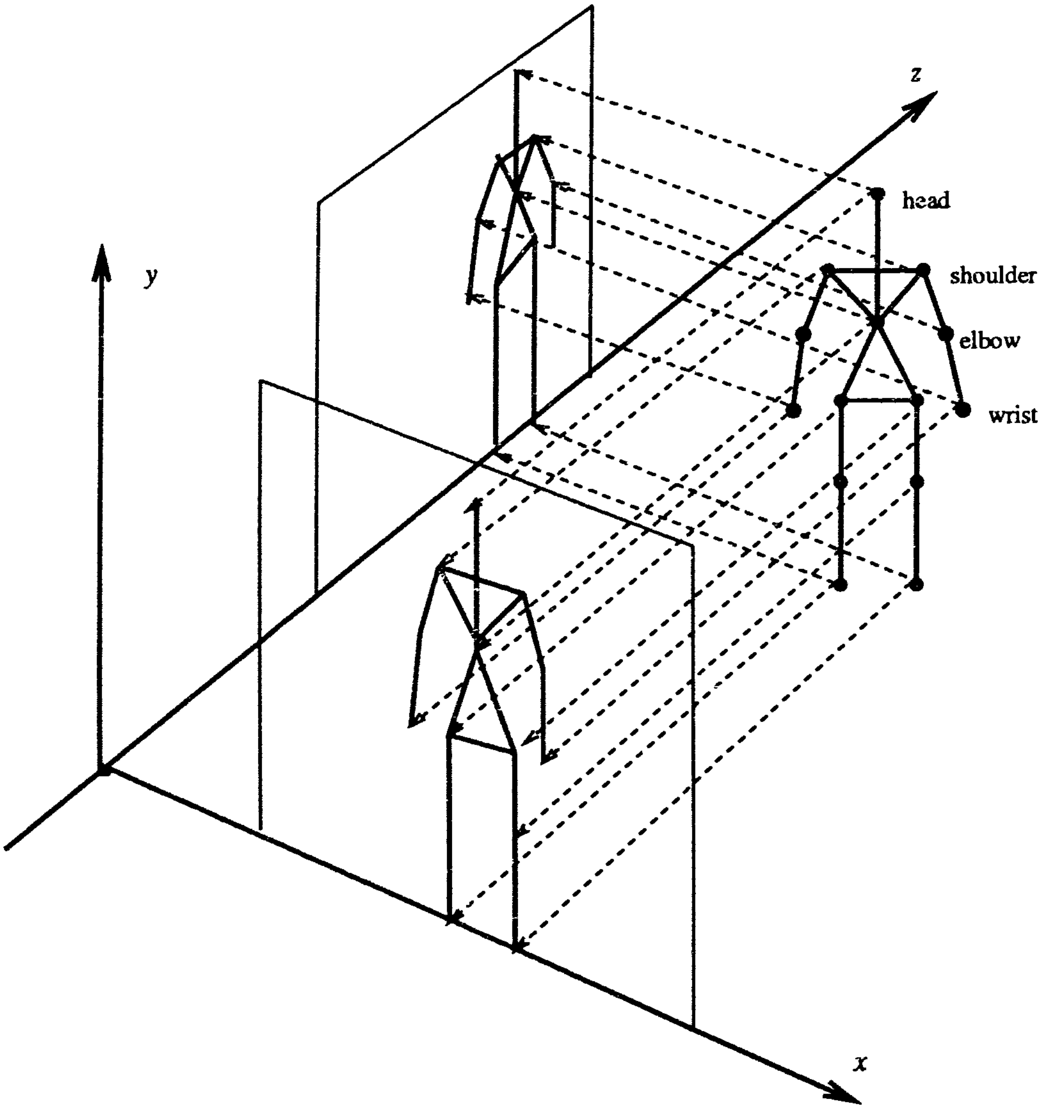


Figure 5.2: Human Body Model and Its Projection

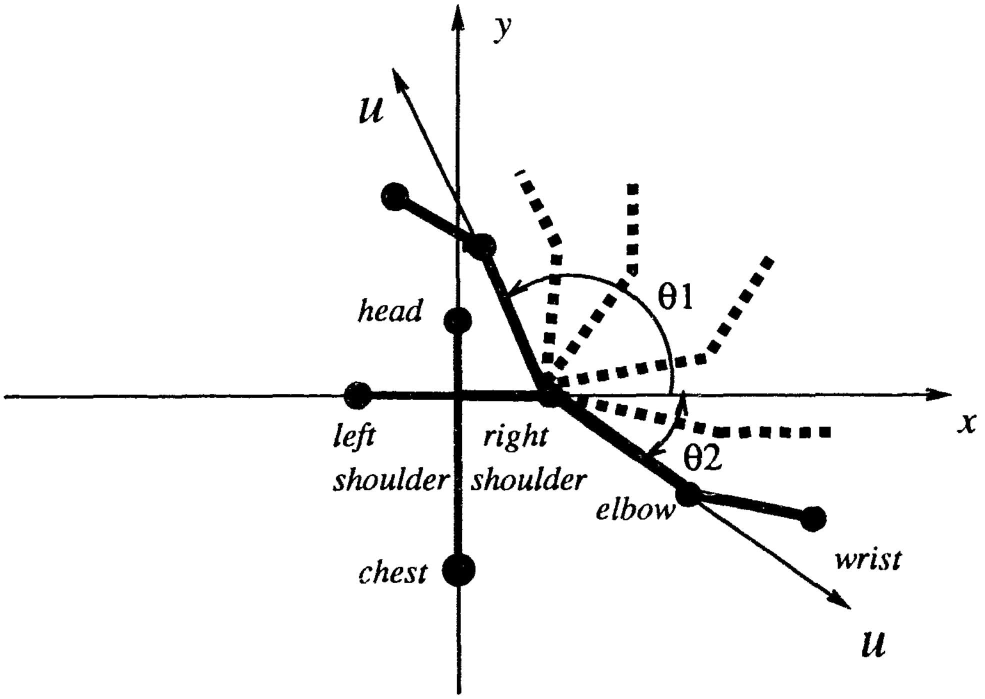


Figure 5.3: Right Shoulder Angle Constraint in x-y Plane Seen From the Front

ϕ = the angle from the -z axis vector to the projected vector $\mathbf{u}_{(y,z)}$ (Figure 5.4).

From the kinematic point of view, the value of $\theta 1$ is restricted lie between 0° and 130° in the x, y coordinate system when the arm is above the shoulder; or to a value $\theta 2$ between 0° and 80° when the arm is below the shoulder (Figure 5.3). Similarly, ϕ is the angle at the right(left) shoulder point with a value $\phi 1$ between 0° and 70° in the y, z coordinate system when the arm is above the shoulder; or it is a value $\phi 2$ between 0° and 160° when the arm is below the shoulder (Figure 5.4).

These angle limitations are considered in our experiment. They define a specific motion area for the elbow joint. Checking these angles would reject the illegal body positions during the positioning process.

(B) At the elbow joint. We define the angle at the right and left elbow joint as

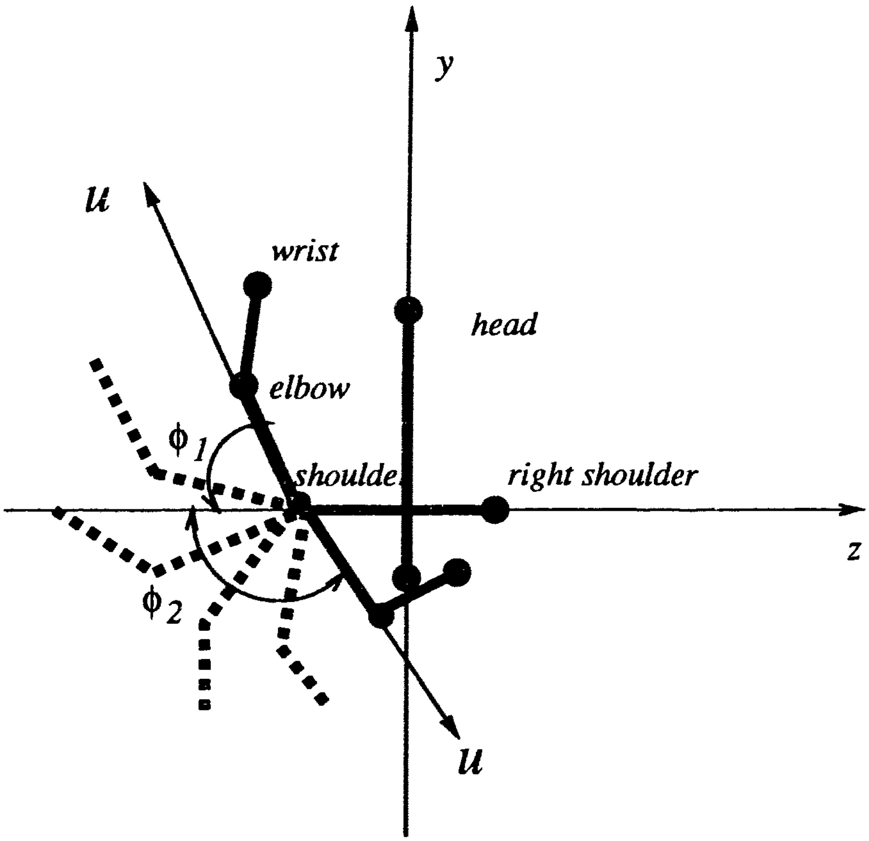


Figure 5.4: Right Shoulder Angle Constraint in y-z Plane Seen From the Side

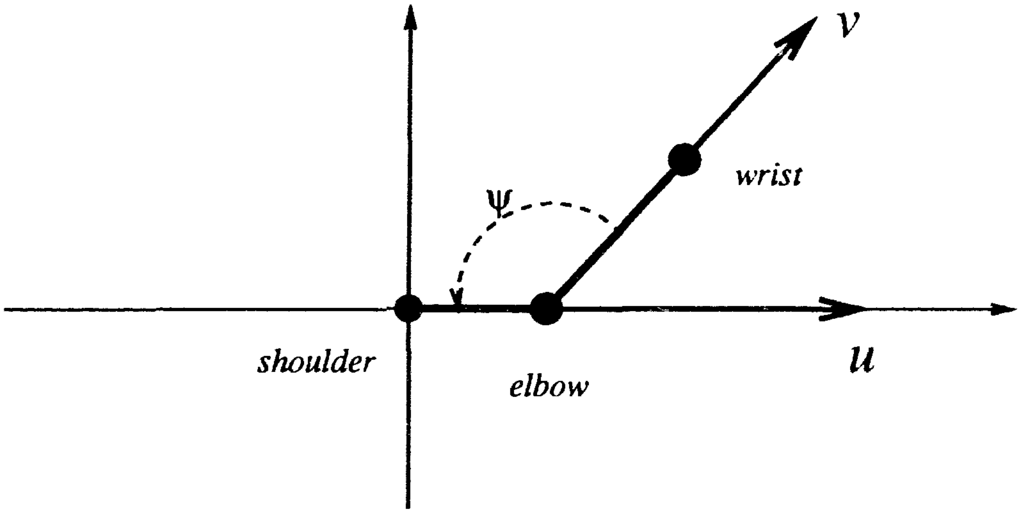


Figure 5.5: Right Elbow Angle Constraint

follows:

u = the vector from the shoulder to elbow, as before

v = the vector from the elbow to the wrist,

w = the cross product of v and u, and

ψ = the angle measured from v to -u.

Then the angle ψ has a value from 0° to 170° (Figure 5.5). By considering this angle constraint, we predict the position where the wrist joint might be and reject the impossible positions for the wrist joint.

(C) At the head joint. The angles associated with the head joint are defined by the head point and the chest point when the body moves its head forward or backward. Let the vector from the chest to head point have a projected vector $n_{(y,z)}$ on the y-z plane, also let the angles measured from the y axis vector to the vector $n_{(y,z)}$ be $\delta 1$ or $\delta 2$. Then the angle $\delta 1$ has a value from 0° to 70° for the forward angle and from 0° to 40° for the backward angle $\delta 2$ (Figure 5.6).

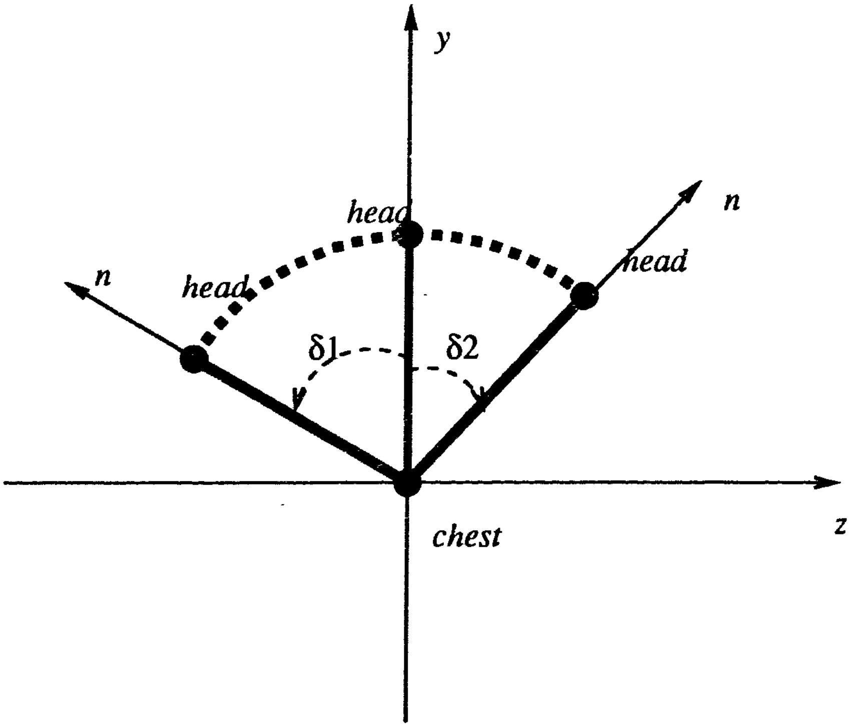


Figure 5.6: Head Angle Constraint

5.2.2 Distance Constraints

Distance constraints can make it easier to check the the violations for any particular orientation of the body and understand meaningful body postures. For example, when both arms are swinging, usually one is in front of the body and the other is behind the body. The distance between the two wrist point projections on the image plane is in an increasing function until they reach the very end positions. Also the wrist points should be in a circular area centered on the elbow feature points. The distance constraints between the two feature points used in this project are listed in appendix A. They are used in predicting the position of the feature appearance in the next frame. In addition, they can be used to reduce the combinations of possible solutions of recovered joints.

5.2.3 Motion Constraints

Another factor we consider in the recovery process is the motion constraints. Among a remarkably large number of different body postures, we shall not only apply the physical constraints mentioned above, but also make use of motion knowledge to obtain meaningful human body postures. In other words, we try to use *a priori* knowledge about human motion to set up a "Motion Model":

(1) Generally speaking, when the body is in motion, the two arms are not both in front of or behind the torso simultaneously. The same restriction also applies for the two legs. This give us information about one arm's position relative to the other.

(2) The elbow movement is cooperative with the shoulder. When these two joints both move, they will swing forward or backward at the same time. This rule also holds for the hip joint and the knee joint. Usually they swing in the same direction at the same time.

(3) Arms and legs move smoothly, most of the time. From this, we get an idea of how fast the arm is moving; the arm should be in an area which corresponds to its speed of the movement and observe "motion coherence". This helps us to find the two dimensional positions of the feature points.

Motion constraints are especially useful in feature point prediction. Most human movement is smooth and thus we predict the future position based on the motion model.

5.3 Image Feature Point Identification

The first task of the recovery process involves identifying the feature points from which the recovered 3-D structure of the human body is displayed. The identification process consists of finding correspondences between components of the image description and the human body representations, i.e., to get corresponding matches between the image geometric labels extracted by the low-level image processing and the predefined model. After this matching, the features are registered in a list of symbolic representations of image features. The whole identification process is divided into three steps as follows:

(1) Get the edge features of the geometric feature from the images by using edge

detection;

(2) Using *a priori* knowledge about the human body model, match the edge feature to the feature model in the predicted area;

(3) Get a symbolic representation of the geometric symbols along with their 2-D information in the image plane.

5.3.1 Image Feature Representations

Image feature points are the points necessary to construct a human body. The selection of image feature representation has a deep effect on feature point identification. The selection should follow the policy of being easy to implement and least sensitive with respect to the noise. Generally, the feature can be geometric characteristics of the object, such as corner, line, curve, etc; or an intensity function of the part at all locations(x, y); or even trichromatic luminance measurement in the color image.

In this project, we attach some geometric symbols with different colors to the wooden body as the image feature. The symbols are circles, squares, triangles, which have different red, yellow or brown colors. These features are the basis from which to generate the models in the modeling phase and are again used to match with the extracted line patterns from the images.

5.3.2 Matching Between Models and Image Characters

We have introduced the three step feature point identification process. The first step is to get the line patterns in a given input image by using edge detection. This is described in the previous chapter: image analysis. The second step is to find a set of patterns in the given image that approximately matches the model's features. In this section, we deal with this problem by providing a model-based feature matching method: the line patterns extracted from the image plane and those predefined by the model are compared.

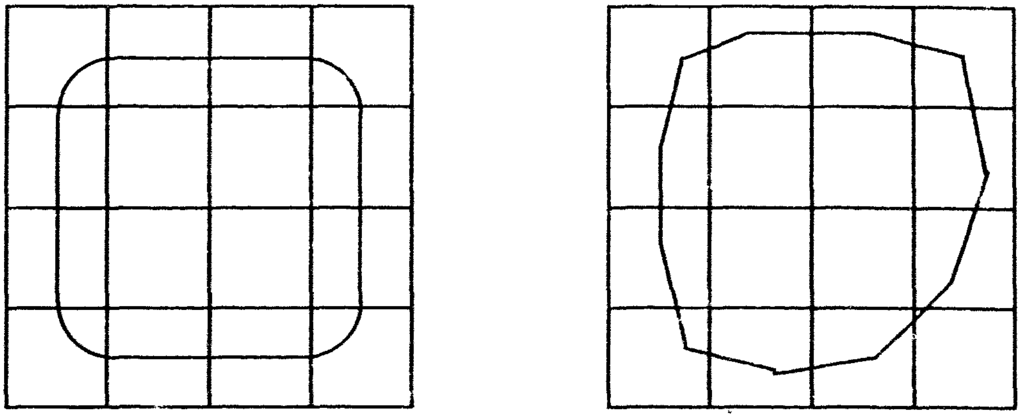
Let f(x,y) be the image field containing the part to be inspected, and let t(x,y) be the template of a defective pattern, like a circle with given radius or a triangle with given angles. The matching is commonly done by computing a similarity measure C(x, y) between f(x,y) and t(x, y) at all (x, y) in f(x, y), and then using a similarity measure for detection. The similarity measure used is cross correlation, defined as C(x, y) = f(i,j)*t(i-x,j-y). To be practical, t(x, y) is usually zero outside a small window.

In our project, the low-level image processing gives the labeled line features. The model used for matching is a 2-D geometric pattern on the input wooden human body. The matching method is used to find geometric symbols represented by these line features.

For example, to identify a circle pattern, a 4 × 4 mask is used as shown in Figure 5.7(a). Inside is a circle model whose curves are white on a black background: $f_1(x,y)$ is the grey value of the model mask. Value 1 represents white and value 0 represents black. An image character is shown in Figure 5.7(b), where $f_2(x,y)$ is the image grey value. The matching equation is given as:

$$T = f_1(x,y) \oplus f_2(x,y) \tag{5.1}$$

This equation produces the difference between two masks by implementing the ex-



(a)Feature Model

(b)Image Character

Figure 5.7: Matching Between Feature Models and Image Characters

clusive or logical function of $f_1(x, y)$ and $f_2(x, y)$ at the corresponding pixel. If the difference is less than a given threshold **t**, the image curve forms a circle; otherwise, it does not. To identify the pattern from these line features, we can use other models to check whether it belongs to a triangle, or a square, etc.

This matching uses the information gained at the low-level image processing level and the feature information included in the model description process. It assigns names to the different image characters which corresponds to the feature points.

5.3.3 Constraints in Prediction and Matching

When the matching between the image characters and feature model, it is not easy to determine the right ones when there is a large set of model descriptions. Ti assist the process, the kinematic constraints of the model are incorporated in a prediction-identification process to make predictions about the appearance of the feature in the next image by checking the constraints. The predicted feature is verified by matching.

Constraint analysis methods have been used in image interpretation. Badler [O'RourkeBadler80] provided a constraint propagation method in his human motion analysis research. He even proposed a constraint network based on the structure of the human model. Lee[LeeChen85] determined the 3-D human body postures by deleting any explanation which did not satisfy the kinematic constraints.

In this thesis project, we apply the kinematic constraints in the prediction process to assist the use of the corresponding image feature model to fulfill the low-level matching. Also, the constraints are employed in the recovery process to infer the third dimension of the human body. The physical kinematic constraints of the human body considered here include: angular constraints at the joints which limit bending and twisting, distance constraints between the rigid skeletal structures of the body, and motion constraints which limit the movement. These constraints are treated as relations between the positions of the features on the human body.

5.3.4 Initial Frame Image Feature Identification

The initial posture of the human body is assumed to be known, as are the general positions of the feature points in the first frame. Correspondences between components of the *image description* and the *model representations* are found through the matching method described in the last section.

At this stage, identification is implemented based on the prior knowledge about the position. Each feature point is located at the starting position of the entire body movement. The image preprocessing part produces line features of the geometric patterns which represent different image features of the human body. Then different feature models are applied to match these geometric patterns to identify the feature points. For example, a triangle model is matched to the geometric patterns at the head area to identify the head feature point.

The resulting symbolic description of the feature points is passed to the prediction-verification process for analysis of the next frame.

5.3.5 Image Feature Identification in Consecutive Frames

For a sequence of motion images, it is not obvious how to find the feature points at each frame, especially in an image where the body parts are occluded. At this stage, the kinematic constraints play a critical role in predicting the feature point positions of the human body for the next frame. The kinematic constraints can be analysised individually. And at some places, several constraints are combined together to implement the prediction function.

The distance constraints and angle constraints are usually put in the form of inequalities between expressions, along with the possibility of including max and min. Equality can be encoded as two inequalities. For instance suppose a distance between the shoulder point and the elbow point is represented as a stick figure distance by the quantifier L_StoE, the distance between the elbow point and the wrist point is defined by the quantifier L_EtoW, and the angle between the upper arm and lower arm is defined by the quantifier A_ARMtoARM. Then the whole area

A which may be covered by the wrist point can be expressed as:

$$A \leq \sqrt{(L\_StoE)^2 + (L\_EtoW)^2 - 2 * (L\_StoE) * (L\_EtoW) * cos(A\_ARMtoARM)} \tag{5.2}$$

The subparts of the body are predicted and their 3-D structure are recovered using a flowchart. The area where the wrist point will appear is predicted based on the identification of the elbow point. Basically, the wrist point is located within a circle area centered at the elbow with a radius equal to the segment length between the elbow and the wrist. But the area could be shrunk because of the small step of the movement between each picture. For instance, the arm moves only a small angle θ during each step, and θ can be as small as 15° in our experiment. So, the wrist point exists only in a part area of the circle (Figure 5.8). It can be predicted as:

$$\begin{aligned} X\_Elbow &\leq X\_Wrist \leq X\_Elbow + (L\_EtoW) * cos(\theta) \\ Y\_Elbow &\leq Y\_Wrist \leq Y\_Elbow + (L\_EtoW) * sin(\theta) \end{aligned}$$

These local predictions are combined together to produce a global prediction which provides strong clues for feature points discrimination. The position of the feature in the next frame is predicted by simply continuing the body movement without change. For example, if the arm is in a swing motion, it will be predicted to keep swinging until it reaches the angular constraint(angular constraints are described above). From that point, the arm will swing backwards.

For some points, the movement is relatively small, almost nil during some body movements. For instance, during an arm swing motion, the position of the chest feature point does not change a lot. Instead, it remains in a fixed small rectangular area. At this stage, the feature point is identified relatively fast by searching the geometric model in this area and will not be confused with other geometric features.

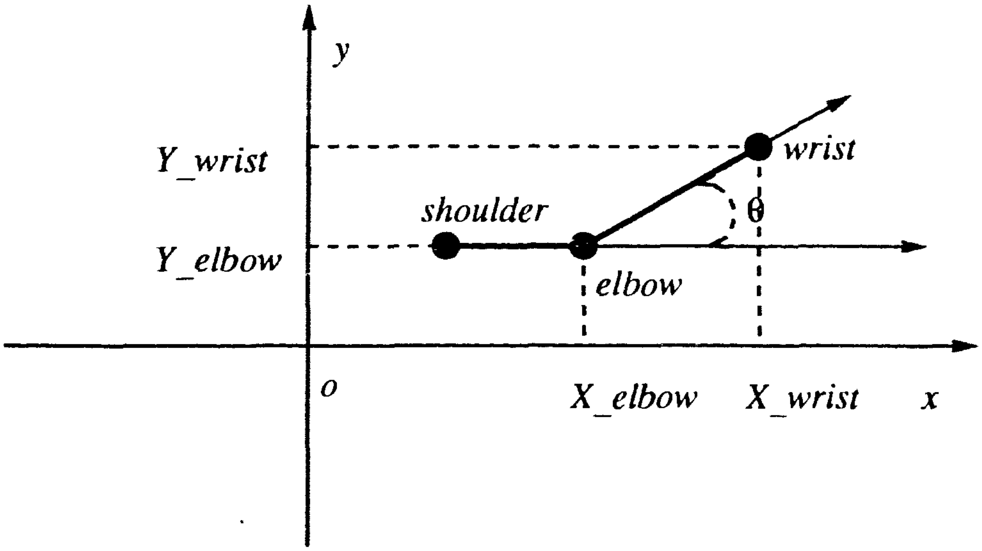


Figure 5.8: Prediction of the Wrist Point

5.4 Recovery of 3-D Structure of the Human Body

It is straightforward mathematically to transform from a 3-D scene model to its projected 2-D image, but the inverse problem is considerably more difficult. However, the combination of image data and *a priori* knowledge of the 3-D human body model can result in a constrained environment in which the human body position in 3-D can be estimated. The clues available *a priori* include:

(a) world knowledge about the human body model,

(b) compositions of the chest coordinate system,

(c) the relationship between the camera-centered and the object-centered coordinate systems, and

(d) the 2-D coordinates of the body in the image plane.

Given an image with identified feature points, our task is to transform the feature point positions in the 2-D focal plane coordinates to their location and orientation in the 3-D object-centered coordinate system; then starting from the feature point on the chest, the 3-D parameters of the objects are determined using a tree structure search. The third dimension of each feature point is calculated through the model knowledge and the kinematic constraints.

5.4.1 Relationship Between Object Structure and Camera Model

The locations of the body feature points in the image are specified relative to the viewer in a camera-centered coordinate system. The camera views the world along the negative z-axis of its coordinate system, with the y-axis pointing up, and the x-axis to the right. It is specified in a left-handed camera-centered coordinate system. The 3-D positions of the human body which we seek should be in an object-centered system. The object-centered coordinate system is defined on the chest of the human body, so it is this right-handed chest-centered coordinate system that is used to define the body in this project. A 3-D structure (x,y,z) in the chest-centered coordinate frame has a translational kinematic relationship which produces a 3-D structure (X,Y,Z) in the camera-centered coordinate frame. The relationship can be described by the following equation:

$$(x, y, z) * [T] = (X, Y, Z), \tag{5.3}$$

where [T] denotes a transformation matrix. The matrix [T] can be readily shown to be the product of a translation matrix, a rotation matrix, and a conversion matrix from the right-handed coordinate system to the left-handed coordinate system. The details for the rotation and translation motions can be found in [Roger74].

In this project, the human body is defined in the chest-centered coordinate system which is in the same direction as in the camera-centered coordinate system. Thus, only a translation and a conversion matrix are needed to calculate the transformation. Here all of these transformations are smooth and well behaved and the transform matrix [T] is known from the *a priori* knowledge about system. So it is straight forward to transfer each feature point coordinates from the chest-centered coordinate system to the camera-centered system, and the recovery process is implemented in the camera-centered coordinate system. The parameters we use later are all referred to that system. The recovered 3-D structure of the human body is finally converted by an inverse transformation to the chest-centered coordinate system.

5.4.2 Recovery Strategy – Tree Structure

The human body is represented by a hierarchical tree where each node represents one body feature point. The chest point is used as the root because it is typically nearest to the body's center of mass. Based on the stick-figure human model, we group all joints into five classes:

Class (a) = chest, left shoulder, left elbow, left wrist

Class (b) = chest, right shoulder, right elbow, right wrist

Class (c) = chest, head

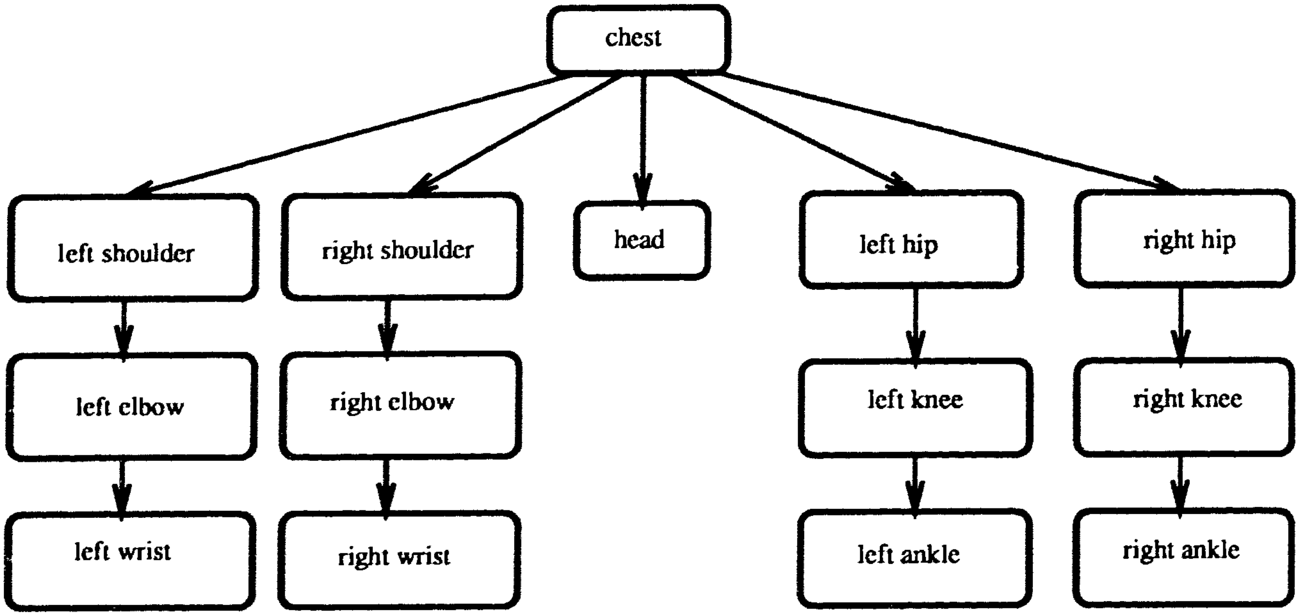


Figure 5.9: Tree Structure of the Recovery Process

Class (d) = chest, left hip, left knee, left ankle

Class (e) = chest, right hip, right knee, right ankle

The human body is modeled in a world coordinate system where the chest point is the root. The feature point at each node is connected by a serial kinematic chain which consists of n segments from the root point by n nodes. The Cartesian position of the end node can be expressed as a multiplicative function of matrices from the parent node coordinate frame to the current node frame. This consists of a translation to the origin of the joint followed by a rotation in the displaced frame to achieve the position in the world coordinate system.

Starting from the chest, we can find the coordinates of those joints in each class. For instance, the elbow point is modeled in a local coordinate system related to its parent node - the shoulder point. The position of the shoulder point is derived first, following by the position of the elbow point, which is found by multiplying the matrix relating the elbow to shoulder point and the matrix relating the shoulder point and chest point. Thus its position in the world frame is found. The position of the wrist point is also derived in this method under the tree structure.

5.4.3 Relation Between the Camera-Centered Coordinate System and Its Projection

A perspective projection is obtained by concatenation of a perspective transformation followed by a projection onto some 2-D "viewing" plane. The coordinates of the feature joint (x_c, y_c, z_c) in the camera-centered coordinate system and its projection (x'_c, y'_c) on the z'_c =0 image plane are related by a perspective projection matrix:

$$(x_c, y_c, z_c, 1)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/f \\ 0 & 0 & 1 \end{pmatrix} = (x'_c k, y'_c k, k), \tag{5.4}$$

where the parameter f is proportional to the camera focal length and the number k is a scalar factor. Solving this equation, we have:

$$x_c = x'_c k, \tag{5.5}$$

$$y_c = y'_c k, \tag{5.6}$$

$$z_c = f * (k-1). \tag{5.7}$$

If r is the focal ratio of a camera, and an object is parallel to the image plane of the camera at distances d from the center of the camera, then the image of the object will measure r/d in image plane coordinates.

The length of a segment, l_{se}, between the starting point (x_s, y_s, z_s) and ending point (x_e; y_e, z_e) can be expressed as

$$l_{se}^2 = (x_s - x_e)^2 + (y_s - y_e)^2 + (z_s - z_e)^2. \tag{5.8}$$

The length of the segment l_{se} is known from knowledge of the human body model. The equation can be expressed in terms of k and the solution for (x_c, y_c, z_c) is obtained after getting a value for k, and the structure of the human body in the camera-centered coordinate system is recovered. However, the equation 5.8 expressed in terms of k^2 shows that there exist two possible solutions for k. This produces feature joint positional ambiguity—Figure 5.10 illustrates these two possible solutions: A segment ab in the image plane may be projected from segment AB_1 or AB_2 which have the same segment length. Point B_1 and B_2 have the same projection point b.

5.4.4 Determination of Joint Structure

In this research, the human body is considered to be restricted by physical constraints. We apply kinematic constraints to reject unmeaningful configurations of the human body and to select those which fit into the body motion.

In our recovery process, the first step is to calculate the angle between a feature point and its parent joint. To calculate the angle: suppose the projection of the

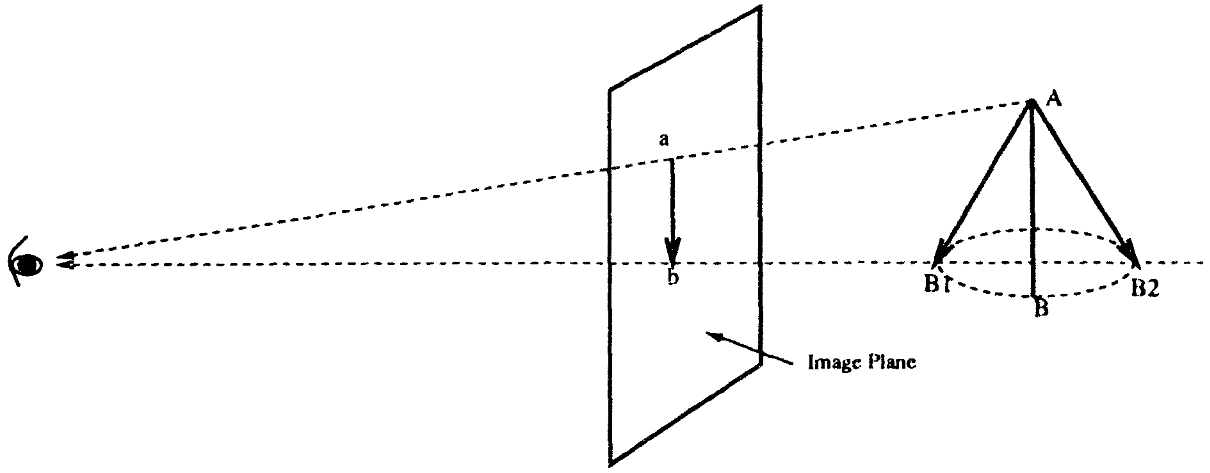


Figure 5.10: Two Possible Solutions For the Projected Feature Point b

distance anticipated between the two connected points is l_{se}. The *prior* model knowledge gives the actual length of this segment l_{SE}. The angle between the projected segment and the actual one is :

$$\theta = cos^{-1}[l_{se}/l_{SE}] \tag{5.9}$$

This angle is compared to the angle constraints applied to this point. If it exceeds its limitation, then this configuration is rejected. Otherwise, it is passed to the next step for further classification. Angle rejection is especially beneficial to those points which have two different angle limitations, such as for the head, which has a fairly large difference between the forward angle constraint and the backward one. If the angle between the head point and the chest point is over 25°, then it must belong to the forward configuration, since no human being could bend backwards through such a big angle. The angle constraints also apply to other angle joints, whose limitations were explained before in the Human Body Kinematic Constraints section.

The next step involves distance constraints and is implemented by calculating

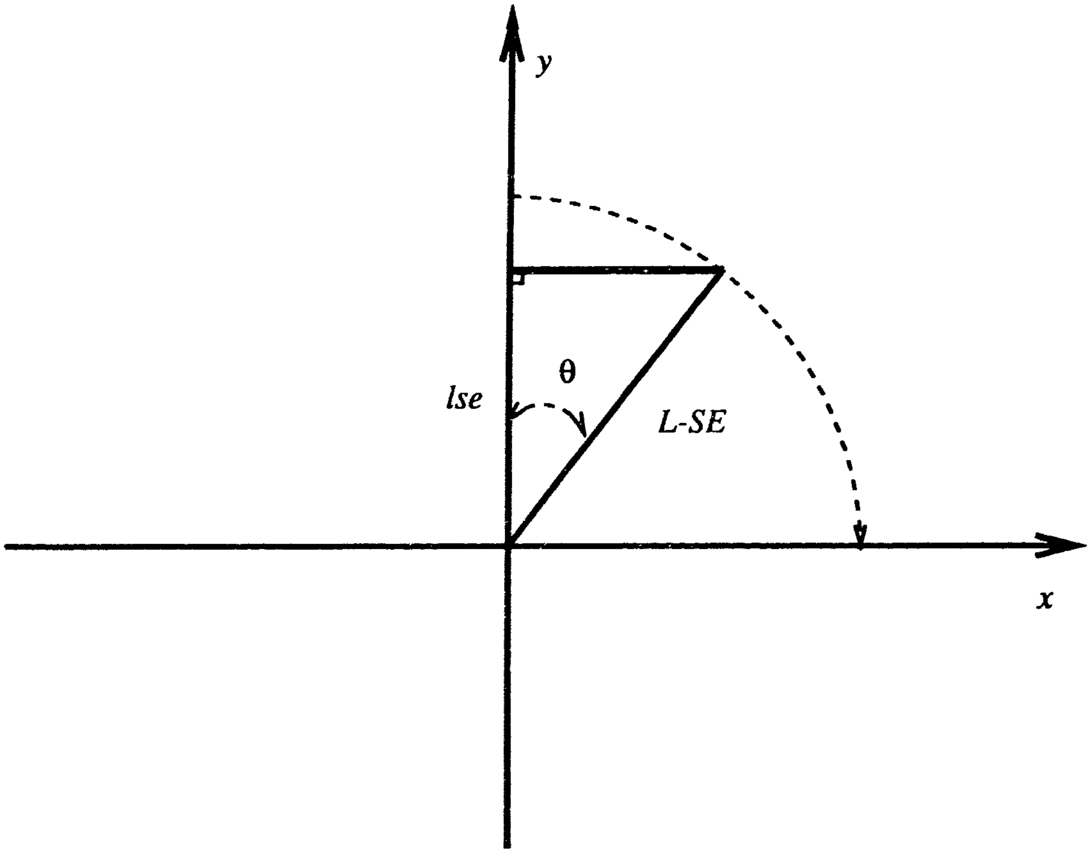


Figure 5.11: Angle Constraints Applied to the Feature Point

the distances between pairs of feature points. For instance, to identify the two wrist points, we may find the distance L_{lr} between the left hand and the right hand. For a hand-swing movement, it is obvious that the distance L_{lr} increases as the hands move away from a posture with arms close to the legs (in the initial frame) until they reach fully extented positions. From there, the distance L_{lr} decreases. The motion rules applied to the human body are used to reject those configurations which do not follow the distance function: decreasing, increasing, decreasing, $\cdots$. These distance constraints are applied to the feature points on the two arms.

The above two steps can eliminate many infeasible body configurations. Although the mathematical calculation is quite complicated, the results are very effective in clearing the ambiguities.

5.4.5 Occluding Parts Interpretation

A feature is occluded when the feature detection process can not find a particular body feature after examining the image plane. Thus, it is assumed that the feature region must be hidden behind another part of the body. To derive the position of these occluded feature points, a major strategy is to use the body's visible components to predict the occluded feature. When this is done, the representation is slightly weakened in terms of the uniqueness criterion, but this is better than no estimate.

Our approach to finding the position of the occluded feature points is to depend on the n last frames and the neighboring feature points to get the track of these points. In a smooth movement, the body produces approximately a constant moving speed. From the feature points in the last n frames, we calculate the moving speed and direction. The speed may not be exactly the same from frame to frame. So, we choose its average moving speed over the last three frames as that for the current frame and use it to estimate the position.

The estimated position can be checked by the visible body parts. For instance, to check the occluded position of one wrist point, we first calculate the distance between the wrist and elbow points. The distance must satisfy the distance constraints introduced in the human body model. If the constraints are not satisfied, the occluded parts need to be estimated again. Second, the distance between the left and right wrist points is also calculated. It should follow the change of distance function for this movement. The estimated position is only tentative. The results are displayed in the 3-D graphics window to be checked.

Another pair of the visible parts which are related to the occluded part are used to predict the reappearance of the occluded part. For cooperative movement of the body, when the visible wrist point is moving backwards, the occluded wrist point will likely be moving in the contrary direction with the same speed as the visible wrist point. The data is the same as the movement towards occlusion. It will appear when it moves through that distance behind the body.

The interpretation of results from a sequence of single view images is presented in the next chapter: Display and Discussion of Results.

Chapter 6

Display and Discussions of Results

In this chapter, we describe how the results are displayed using the HOOPS 3-D software package on the Sun Sparcstation. Some of the experimental results which were obtained using the system are discussed.

6.1 Display of the Recovered 3-D Structure of the Human Body

To evaluate the recovery system, it is necessary to view the physical realization of the final results. In this project, the evaluation of the recovery of the 3-D structure of the human body is fulfilled by displaying the stick figure representation in a 3-D graphics window and checking whether the body posture matches that in the input images. We require that the display function in the graphics window:

(a) be able to rotate the bodies for view from different directions;

(b) have interactive controls.

HOOPS, which provides the tools for modifying, querying, and displaying graphics in three dimensions on the Sun Sparcstation, is employed in our system.

6.1.1 HOOPS Graphics Software Package

HOOPS is a database oriented software package which stores information about geometric primitives, cameras, lights, rendering and modeling attributes. The database is organized as a tree-shaped hierarchy, where related elements are grouped together in segments which are the units of organization within the database. Each segment may also contain other segments, hence the resulting tree structure.

In the display implementation, segments consisting of two connecting feature points are transformed as the tree-elements while the chest feature point acts as the root of the tree structure. The data is easy to organize because of the hierarchical structure of the bodies.

6.1.2 Display of Results

The display consists of three parts. On the right-hand side of the display area there are a number of buttons, which can be selected by using the mouse to control and to select various body posture. On the left is the stage window where the human body is displayed in space. The stage view can be rotated. On the top of the

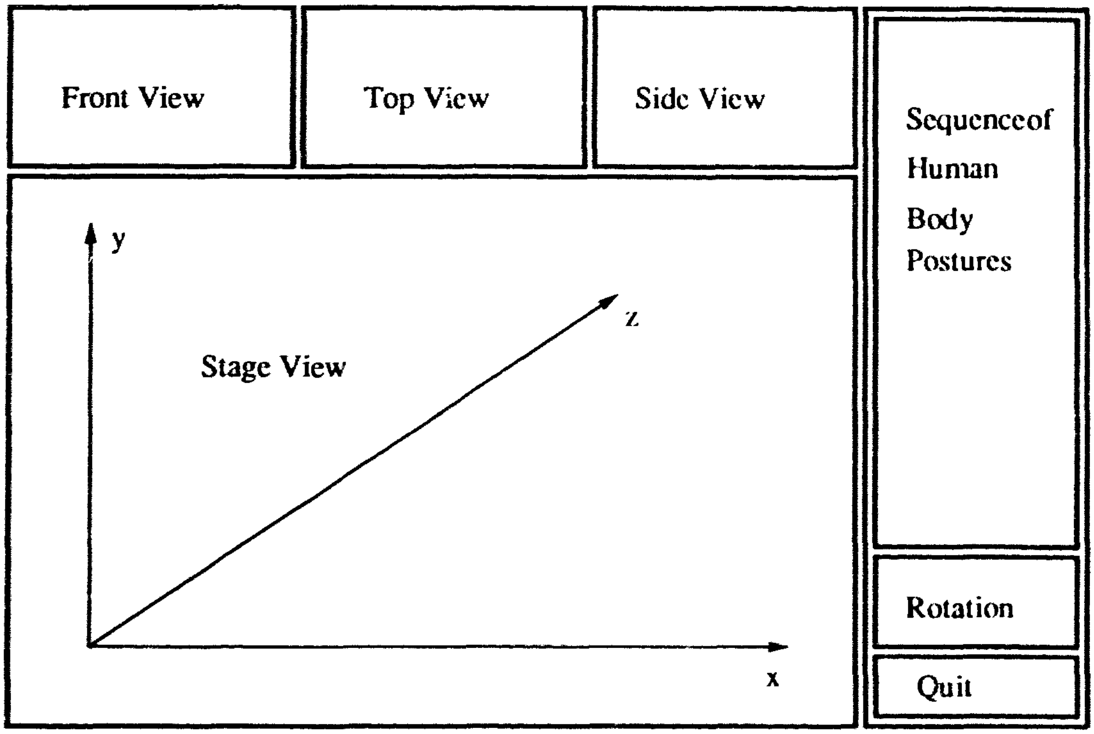


Figure 6.1: Configuration of the Display Window

window, there are three other windows which display the body posture from three fixed orthogonal directions.

A 3-D human body is first displayed in a natural neutral standing position. The interactive 3-D workstation allows the viewer to select the postures of the body corresponding to the different frames and to rotate it arbitrarily. The user can move back and forth in the sequence of human body movements.

6.2 Experimental Results

In this section, we introduce two of the experiment results obtained in the project. The images and their corresponding displays are shown in appendices B and C.

Example 1: Gestural Motion

The first example sequence consists of 6 frames of symmetric arm motions, roughly in the form of a "applauding arm swing". There is no vertical body motion in this sequence. Appendix B shows the input and output of the system for each frame in the sequence. The left column shows the actual input images. Geometric symbols are attached to the feature point positions of the body model. There are no overlapping regions for the feature positions in this image sequence.

From *a priori* knowledge of the human body movement, it is not difficult to trace the feature points and the ambiguous positions are cleared by checking the angular limitations. The angle constraints from the motion rule forbid the arms to swing backwards. The recovered structure of the human body for each frame is shown in the right column. The display window illustrates the recovery system functions.

Example 2: Arm Swing Movement

The second test contains a set of motion pictures where the right hand moves in front of the body from the initial posture, while the left hand moves behind the body bending the arm at the elbow. When the arms reach the top position, they swing backwards. The depth function of the feature points on the right arm increases when the arm moves towards the camera. The depth function of the left arm decreases since the arm moves in the contrary direction. This motion rule and its corresponding distance constraints are applied to the inference process. The system correctly tracks the visible feature points on the arm from the knowledge of the arm movement. The position of the occluded elbow point on the right arm is estimated from the last 3 frames. For instance, to get the position of the wrist point in frame 6, we calculate the movement speed and direction of that feature point from frame 3 to 5. The tentative explanation of the occluded point comes from this information and the motion rule of this arm. Although this data is hard to verify since they are not recorded during taking the experiment, the display in the 3-D graphics window shows that the results are reasonable. The system also predicts the appearance of the occluded points when they swing back from the behind of the body.

6.3 Discussion of Results

The goal of this model-based vision system is to demonstrate the ability to recover the 3-D structure of the human body from a sequence of single view images. The recovery process includes techniques of image preprocessing, feature point prediction and identification, and the derivation of the third dimension based on the model and kinematic constraints.

Example 1 described above is a simple test case, where motions exhibited are very limited and straightforward. The swinging movement of the arms is in a plane with little motion in depth. The results given show that the system performed quite well. The system tracked the geometrical symbols attached to the feature points and sent them along with their 2-D positions to the recovery process. The kinematic constraints restricted the 3-D position of each feature points and helped to recover the depth.

Example 2 has more complex motion: the arms are hidden behind the human body in some frames. But it as handled by the same basic paradigm: the visible feature points were matched with corresponding model features, while the hidden joints were estimated from the last n frames. The depth information was acquired by rejecting the infeasible body postures. The appearance of the occluded feature points on the hidden arm was predicted according to the motion rule applied to the other arm.

The examples illustrate the ability of the system to match the feature model and the extracted image features. However, the task is simplified because the image features are geometric shapes placed on the human model and there is *prior* knowledge of the movement. Even for occluded body features, a less precise identification is possible by using motion rules calculated over a number of frames. Although these testing sequences are simple, they do illustrate that the motion can be tracked without completely examining each image. Note that in the analysis we do not require the difference of two consecutive input images, we do not produce a picture of the model and subtract it from an image frame, and we do not perform any others expensive image processing techniques to get the parameters of movement. The results can be obtained by looking at only a fraction of the pixels in each image frame. In this system implementation, some *prior* factors are considered. These include:

(1) A human body model and its kinematic constraints: this is the basis for this implementation of the system.

(2) A high image sampling rate so that there is only small motion between consecutive frames. This is of especial benefit to the prediction and identification of the feature points.

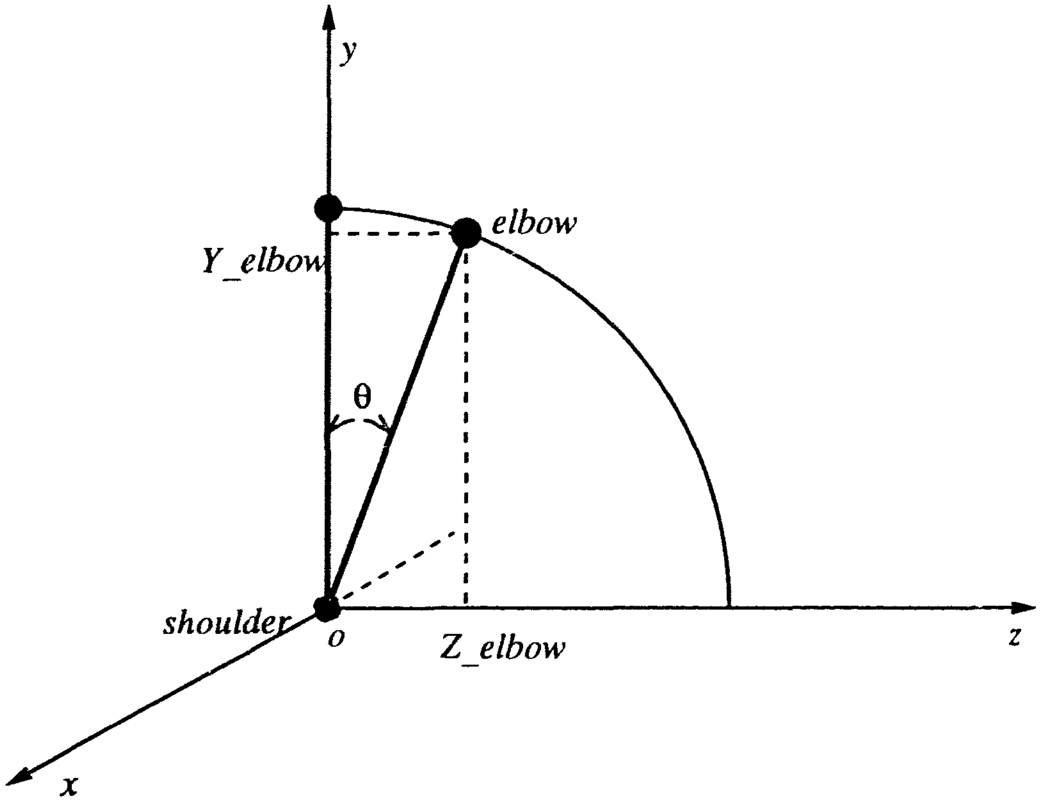


Figure 6.2: Relation between the Depth Information and Swing Angle

(3) Good initial guesses for feature point position: the initial guesses have a big impact on the feature joint search and iterative solutions. A good initial guess is often required for convergence of this method.

(4) Efficient use of motion information: motion parameters are estimated from more than two frames.

These factors are very important in implementing the recovery process. In assessing the results, two more factors, accuracy and speed are considered.

(1) Although the limited amount of data makes it difficult to verify the accuracy, it still can be checked by viewing the results in the 3-D graphics window.

Figure 6.2 illustrates the relation between the swing angle of the upper arm and its depth information. Assume the elbow is swinging backwards and the angle θ is between the shoulder-elbow segment and the y axis. The length of the segment is l_{se}. The depth Z_{elbow} can be expressed by:

$$Z_{elbow} = l_{se} \times sin(\theta) \tag{6.1}$$

$$Y_{elbow} = l_{se} \times cos(\theta) \tag{6.2}$$

and therefore

$$\frac{\partial Z_{elbow}}{\partial \theta} = l_{se} \times cos(\theta) \tag{6.3}$$

$$\frac{\partial Y_{elbow}}{\partial \theta} = l_{se} \times sin(\theta) \tag{6.4}$$

The equation above shows that when θ is small, the projection on the image plane Y_{elbow} changes slowly, while the depth Z_{elbow} changes faster. At this time, the segment is almost parallel to the image plane x-y. A totally contrary situation exists when the θ is large, as the segment is nearly perpendicular to the image plane. This situation requires highly accurate identification of the feature points on the image plane. For example, an one pixel error when $\theta = 0°$ might produce an error $\sqrt{2 * l_{se} - 1} \approx 1.41 \times l_{se}$ to depth Z_{elbow}, whereas the same one pixel error when $\theta = 90°$ produces an error $\sqrt{l_{se}^2 - 1} \approx l_{se}$. This shows that the error in identifying the feature points produces larger error when θ is small than that when θ is large.

The image acquisition process is a possible source of error which gives image noise and blurs image features. Thus the system produces mismatches between the image features and feature models and generates errors in locating the feature points. This effects the recovery of the third dimension of the human body.

(2) Speed is another factor to be considered in the system. Image processing, which includes feature point prediction, searching and matching, and selecting from a set of feature descriptions, is the most costly to compute. The graphics display also requires some processing, but 80% of the processing for each frame is taken up by the low level image processing tasks.

To speed up the processing, parallel techniques could be used for the the image processing. Note that if more feature points are added to the body model, more geometrical features are needed to attach to the model, and feature matching will be more difficult and will require proportionately more time. The tradeoff, of course, is that large images take longer to process.

Chapter 7

Conclusions and Directions For Future Work

The principal objective of this thesis was to implement a vision system to recover the 3-D structure of the human body from a sequence of single view 2-D images. The experiments illustrated that the system, which uses a high-level prediction process and a low-level feature point identification driven by a well-defined human body model, was able to track the feature points attached to the human body in each image frame and give a 3-D explanation to the body movement.

Both low-level and high-level processes are important in vision. Good data gathered from the low-level is a critical prerequisite to reasonable performance at the higher levels. Image noise removal and geometric line feature model matching are introduced to identify the feature points attached to the human body. However, without the human model and prediction results from the high-level, this bottom-up searching would not be able to get an efficient matching.

One of the major bottlenecks for model-based vision is in the acquisition and representation of the models. In this project a simple stick figure is used as a model for the 3-D human body. Knowledge about the human body, its kinematic constraints and attached geometric features are encoded in the stick figure model representation. The analysis of the kinematic constraints predicts an area where the feature points should appear in the next frame. It was also used to reject the infeasible body positions during the recovery process.

Our system required that the images of body motion be restricted to situations where the feature points appear in the images. This gives rise to recovery confusion when the human body is involved in the very complicated movements including both rotational and translational motion. The feature points which were used to represent the structure of the human body are based on their geometrical shapes. This could be extended by using color and texture which are two other important cues for human recognition of objects. These two factors have been brought in a 3-D vision system which is used to automatically acquire 3-D kinematic data[Geurtz91]. This is a promising direction for future development which could be extended to such fields as the analysis of athlete training, patrol robots in unknown environments and the capture of movement for animation.

At this stage, the system has relied on feature point identification to distinguish body parts from one another. A possible different approach would employ a more robust and reliable human body model based on the 3-D shape of the body and using all the motion knowledge. The matching between the images and the model would be based on the shape of the body rather than some selected points. The system would use reasoning to explain the body position with the aid of motion knowledge and other kinematic constraints.

Appendix A

List of Kinematic Constraints

Table A-1 illustrates the general angle limitations at each angle joints. Table A-2 lists the lengths of rigid segments of the human body model employed in this project.

Joints	Flexion	Extension	Abduction	Adduction	Rotation	Bending
Shoulder	0-180	0-50	0-180	0-50		
Hip	0-90	0-40	0-45	0-30		
Elbow	0-160	0				
Knee	0-130	0				
Pelvis	0-75	0-30			0-30(L/R)[1]	0-35(L/R)
Neck	0-45	0-55			0-70(L/R)	0-35(L/R)

Table A.1: Angle Constraints at Joints

L/R: Left/Right

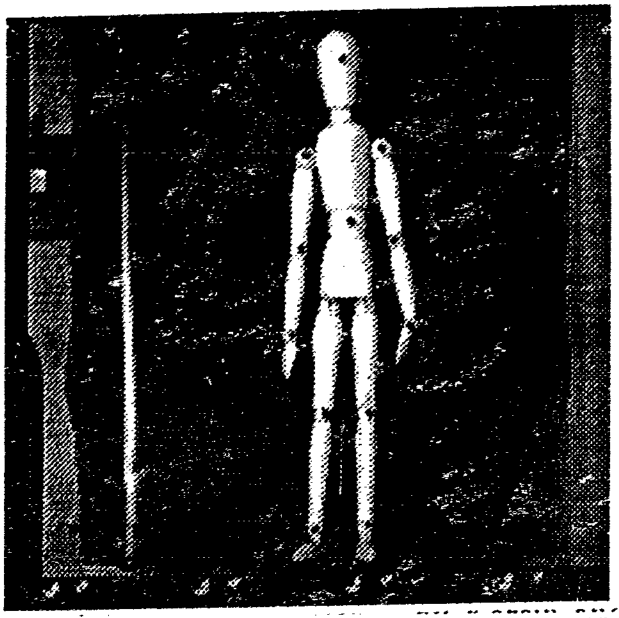
Rigid Segment	Length (mm)
Chest-to-Head	142.0
Chest-to-Left Shoulder	24.0
Chest-to-Right Shoulder	23.0
Chest-to-Left Hip	33.0
Chest-to-Right Hip	35.0
Left Elbow-to-Left Shoulder	82.0
Left Elbow-to-Left Wrist	74.0
Left Shoulder-to-Right Shoulder	62.0
Right Elbow-to-Right Shoulder	81.0
Right Elbow-to-Right Wrist	75.0
Left Knee-to-Left Hip	94.0
Left Knee-to-Left Ankle	98.0
Left Hip-to-Right Hip	29.0
Right Knee-to Right Hip	95.0
Right Knee-to-Right Ankle	95.0

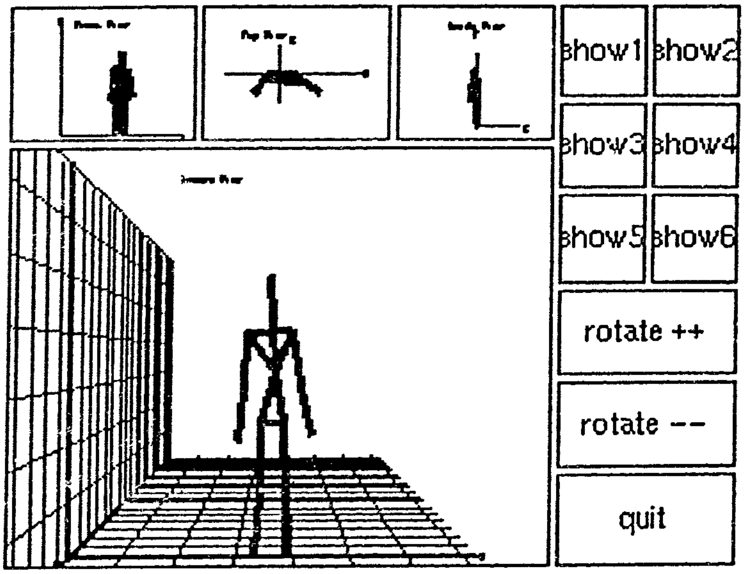
Table A.2: Lengths of rigid Segments of a Human Body

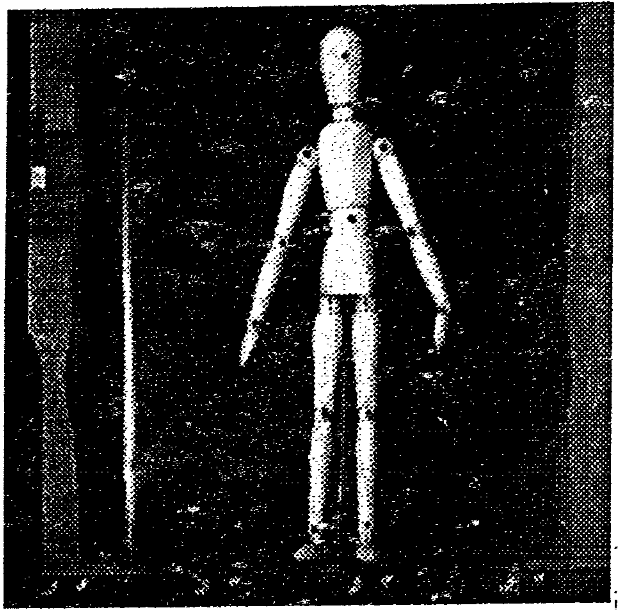
Appendix B

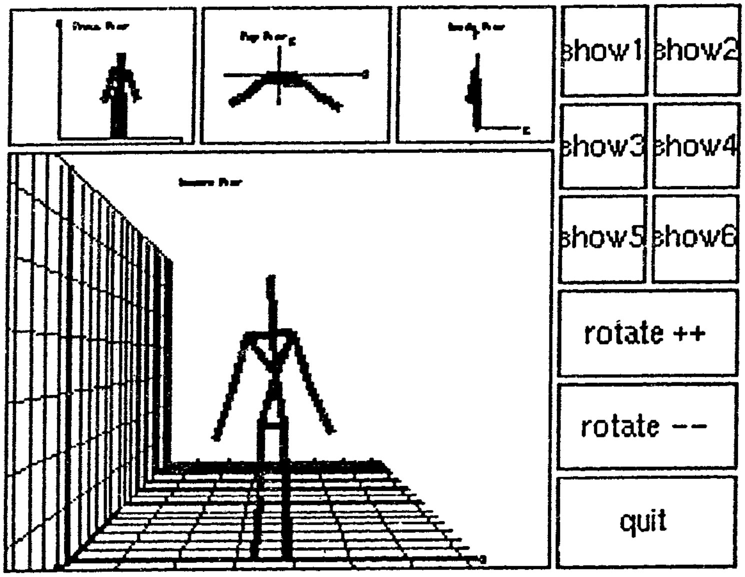
Experiment 1: Gestural Motion

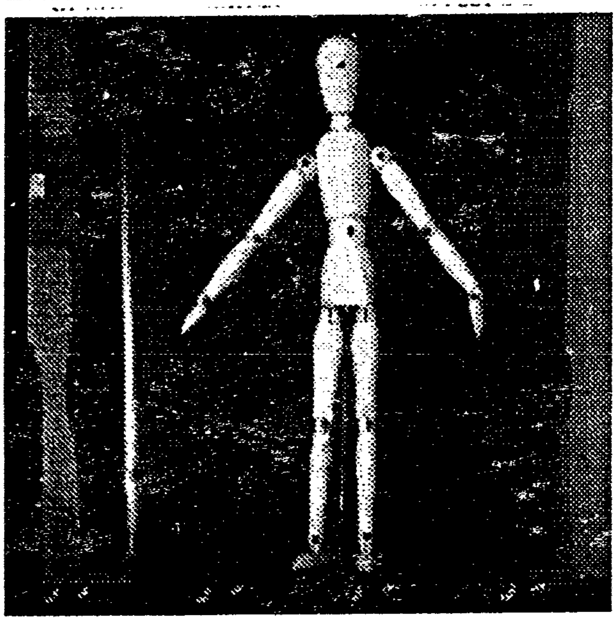
The following contains a sequence of 6 human body motion pictures. The left column is the original input images. The right ones are the output of the recovered human body structure corresponding to the left one. Both of the hands are always in front of the body.

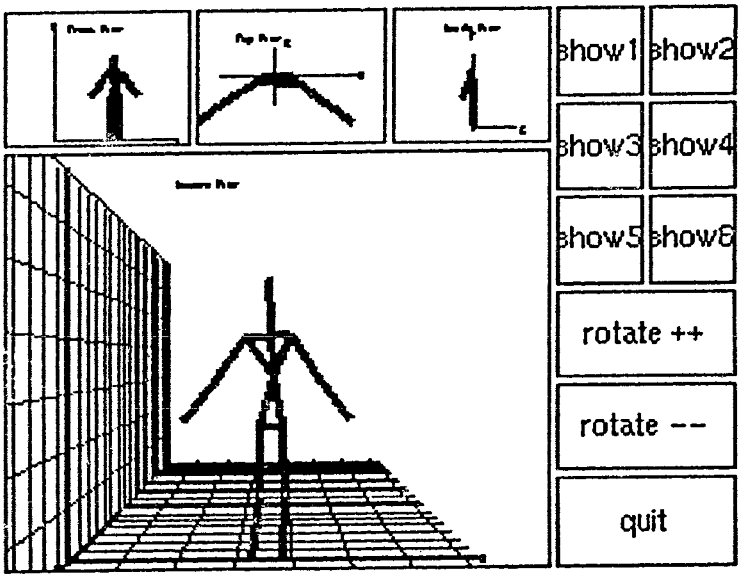


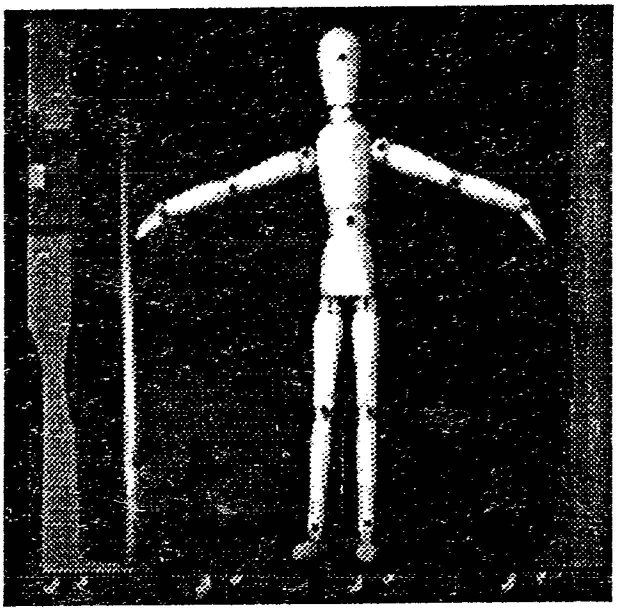

show1
show2
show3
show4
show5
show6
rotate ++
rotate --
quit

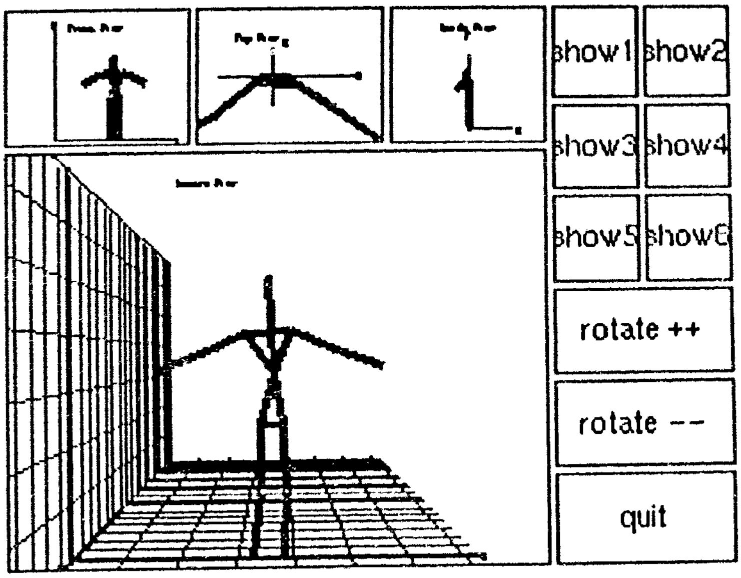



show1
show2
show3
show4
show5
show6
rotate ++
rotate --
quit

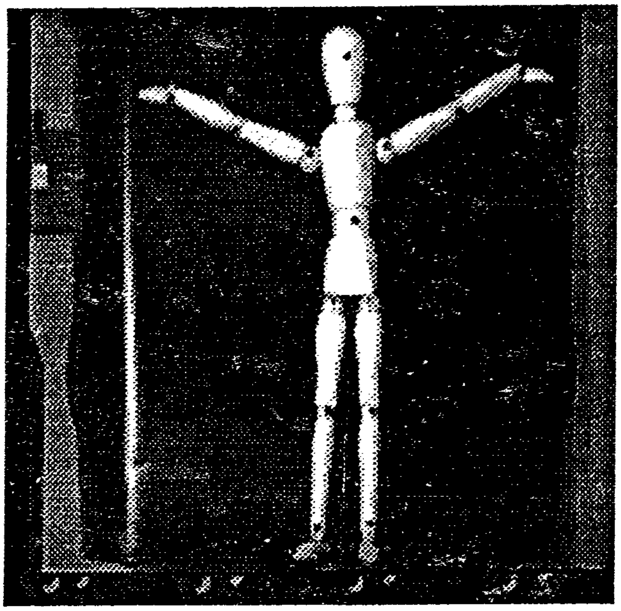


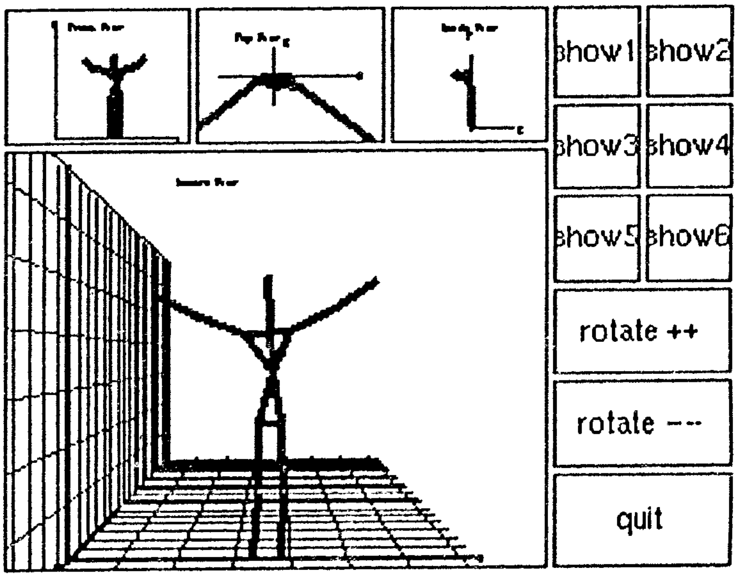

show1
show2
show3
show4
show5
show6
rotate ++
rotate --
quit



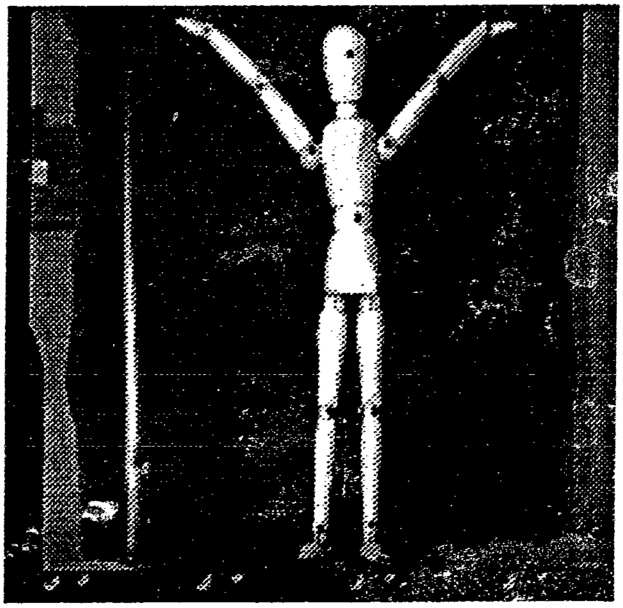


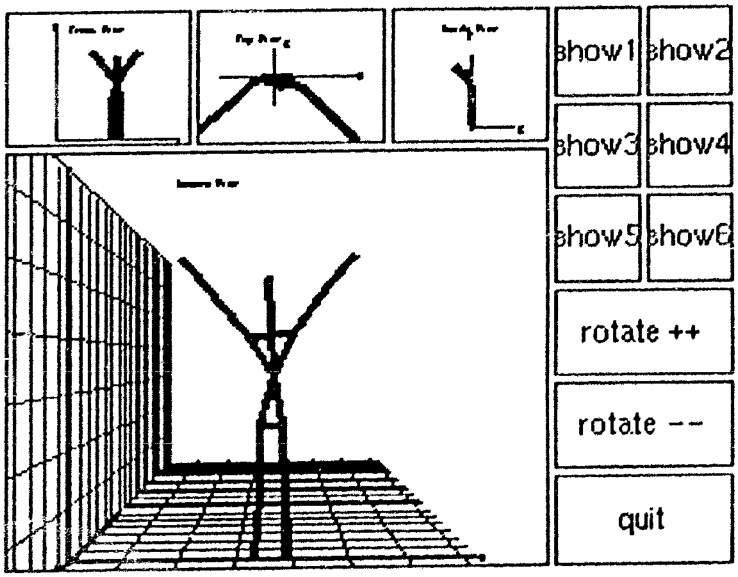
show1
show2
show3
show4
show5
show6
rotate ++
rotate --
quit





show1
show2
show3
show4
show5
show6
rotate ++
rotate --
quit



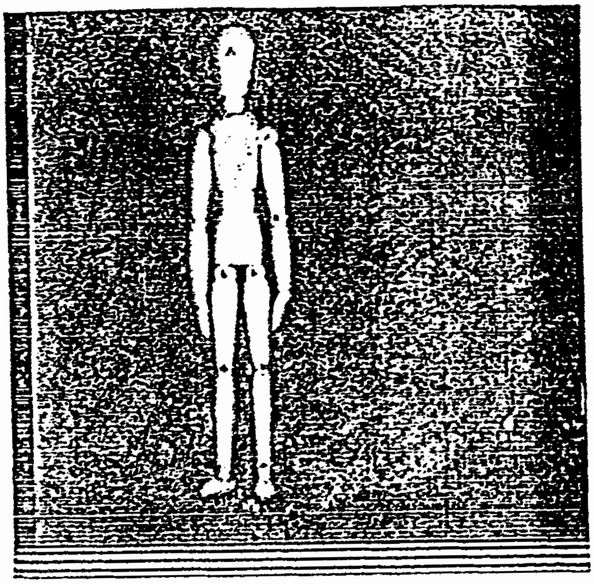


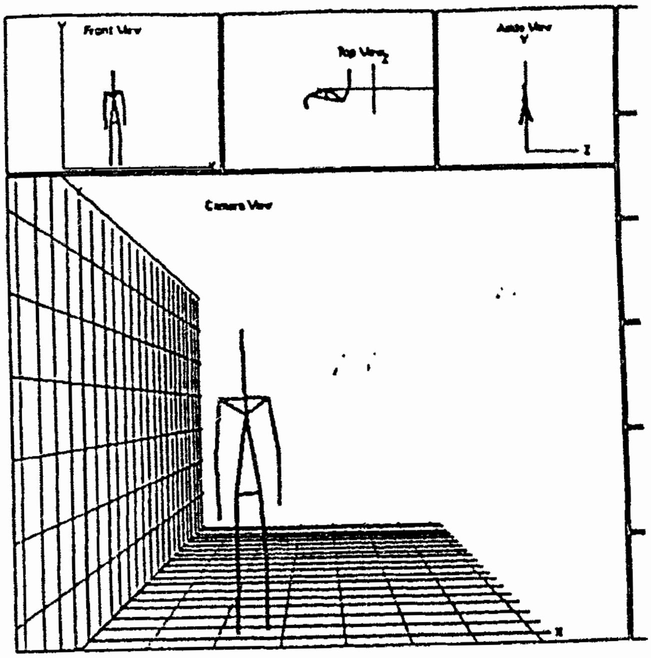
show1
show2
show3
show4
show5
show6
rotate ++
rotate --
quit

Appendix C

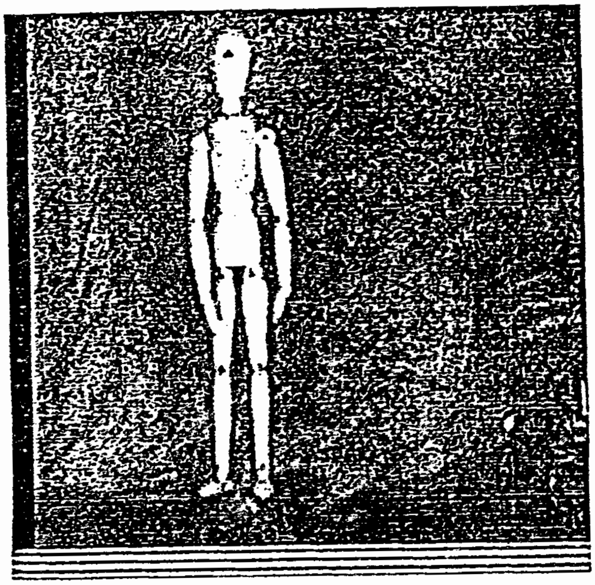
Experiment 2: Arm Swing Movement

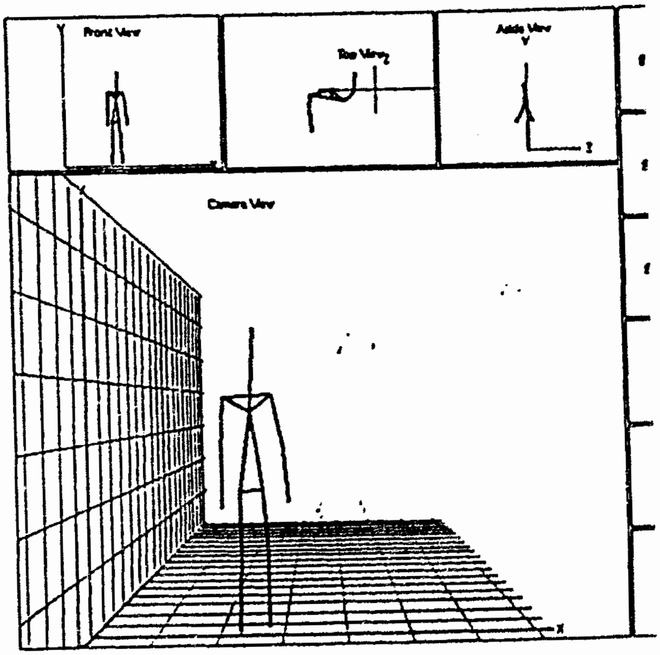
The following has 12 image frames about arms swinging movement. The left column is the input pictures. In the images, the right hand moves from the initial posture to the front of the body, then back to the initial position. The left hand moves to the back of the human body then reappear to the image plane. The right column displays the recovered 3-D structure of the human body. The occluded body parts are also displayed.



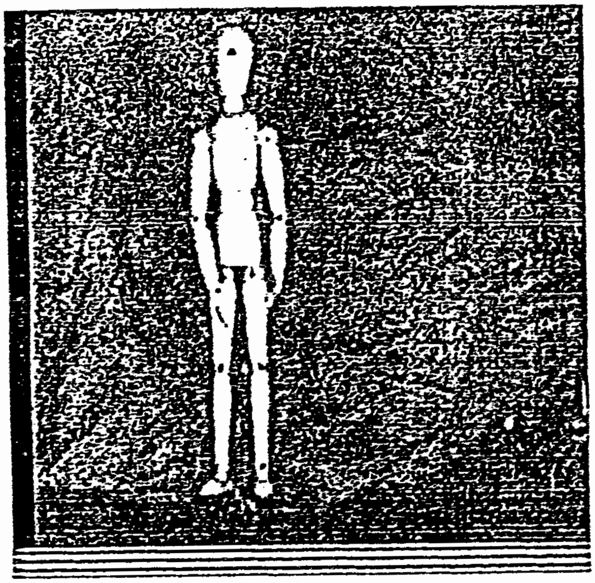


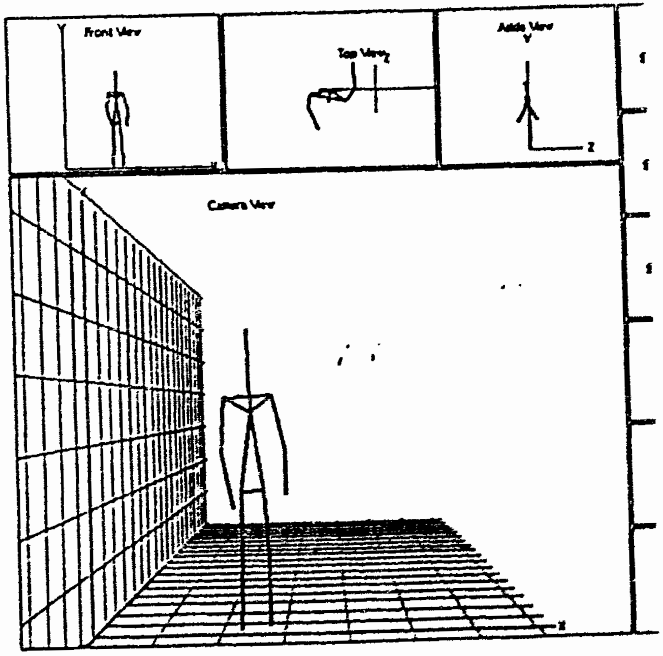
m1.cxap4



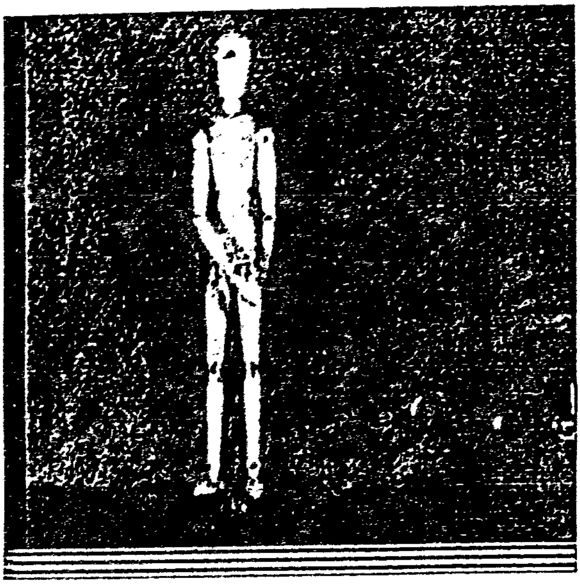


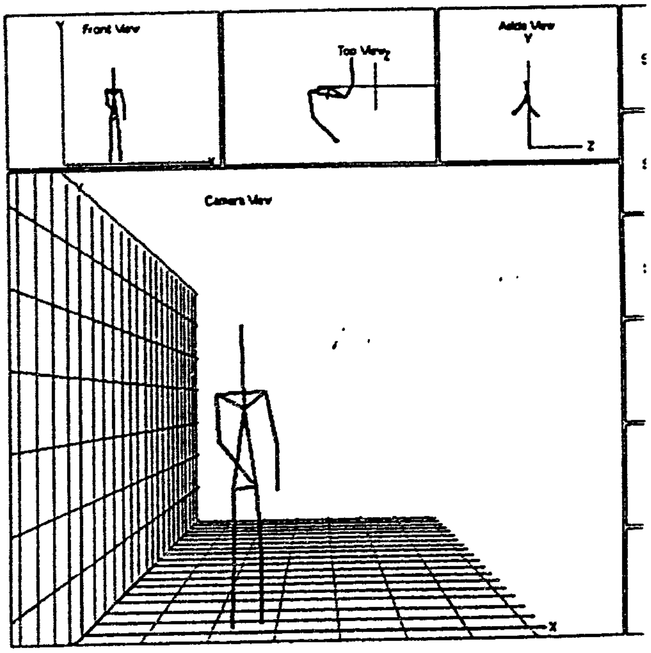
m2.cxap4



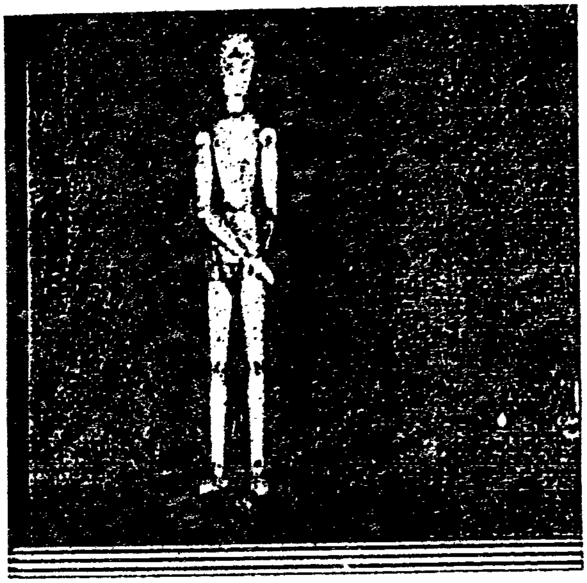


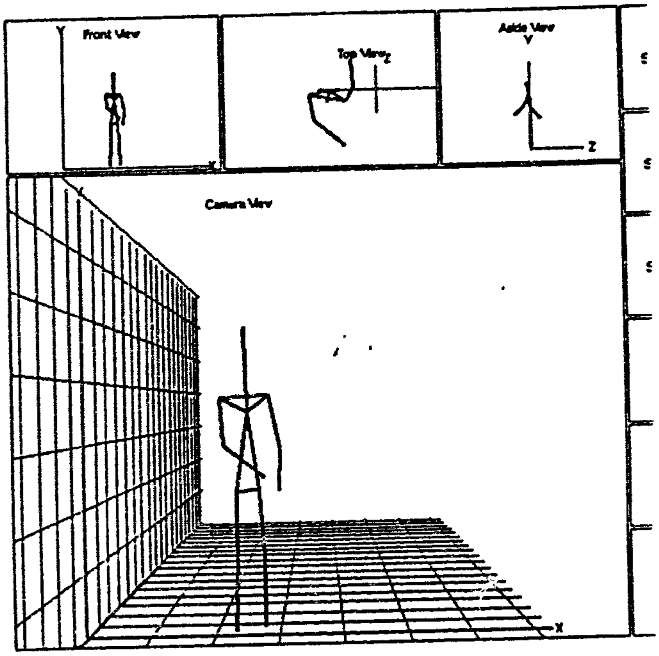
m3.cxap4



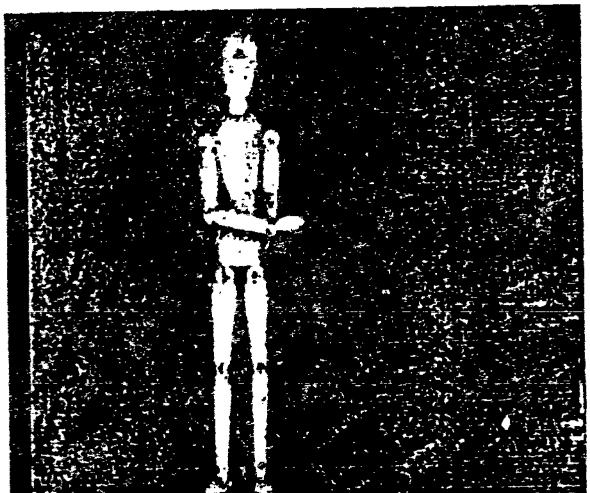


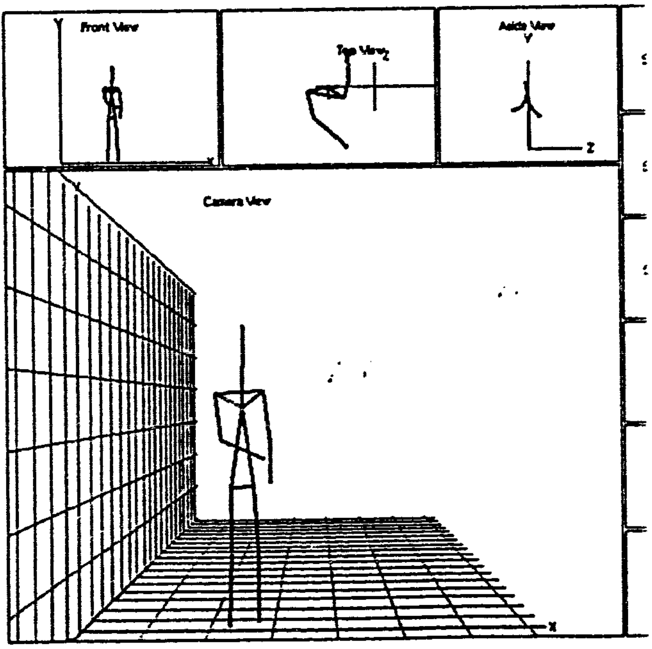
m4.cxap4



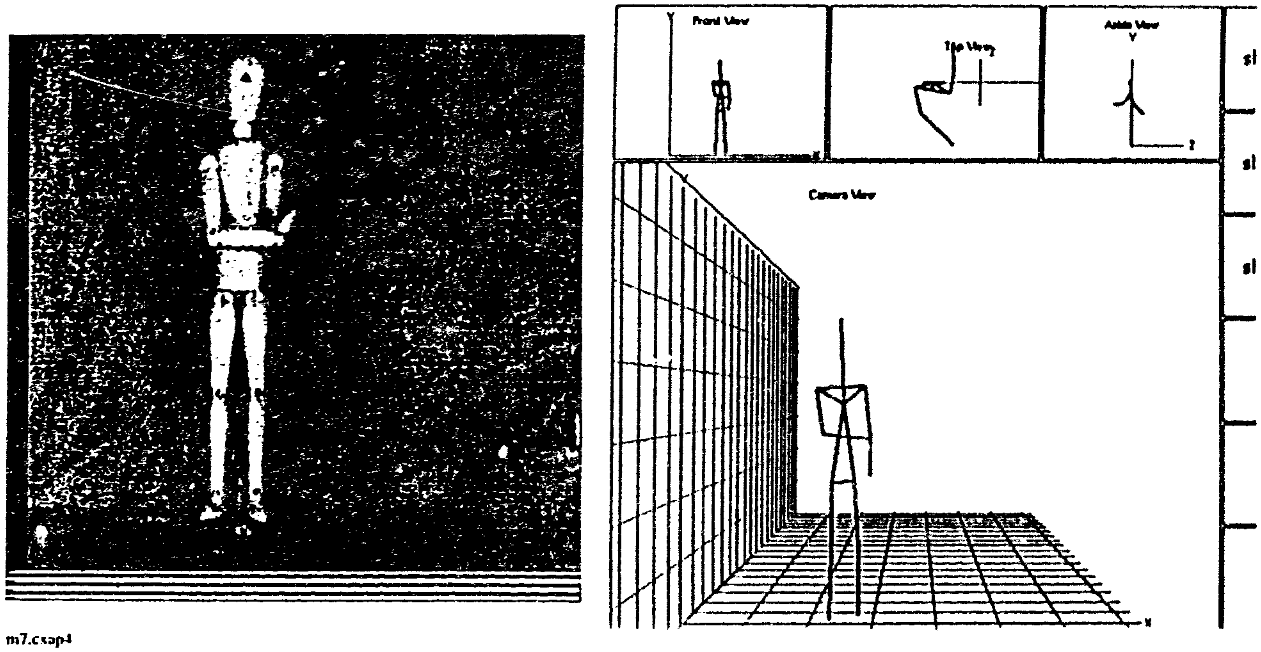


m5.cxap4

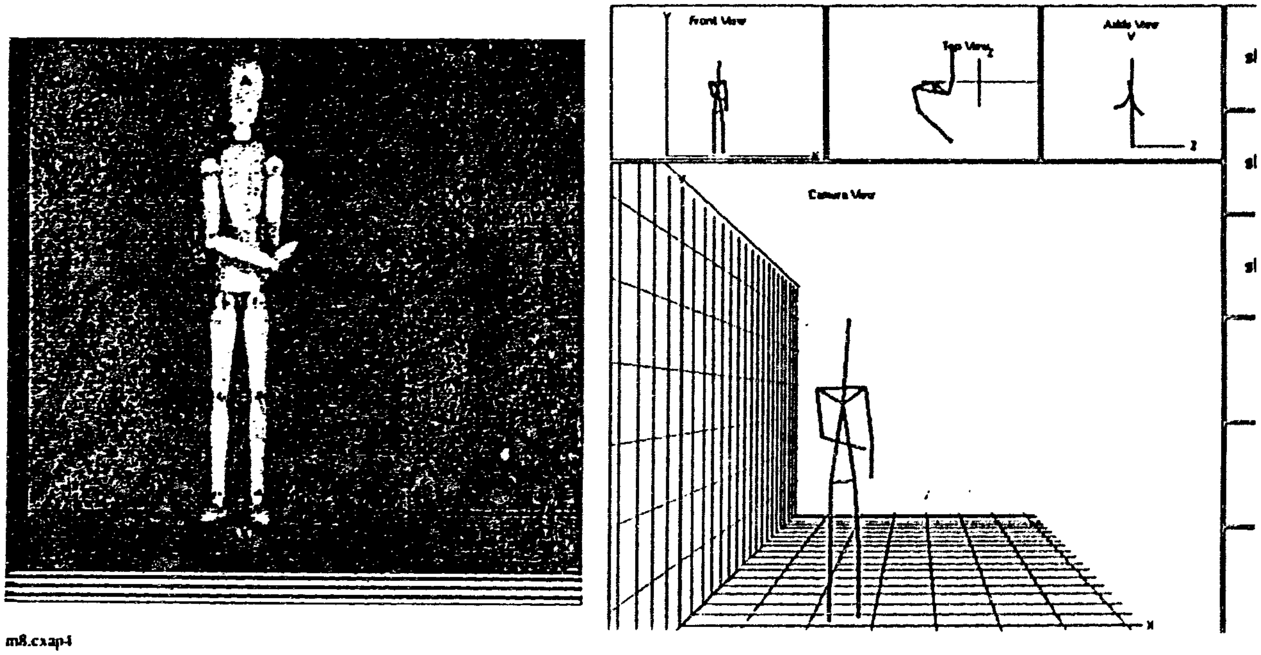




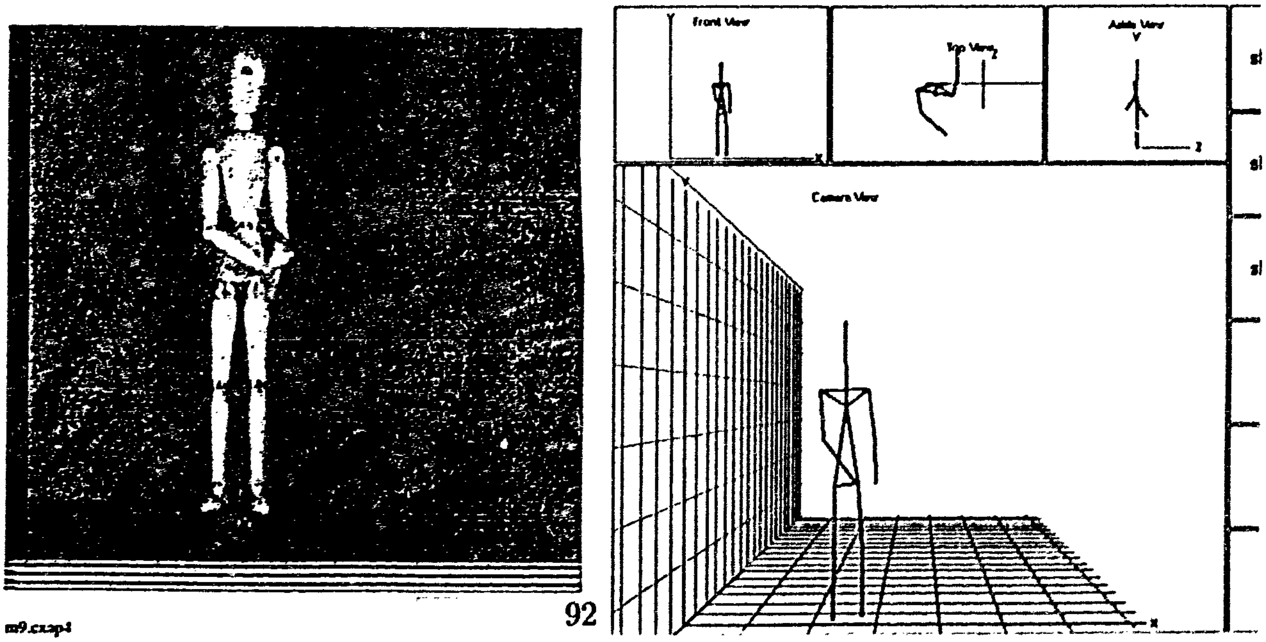
m6.cxap4



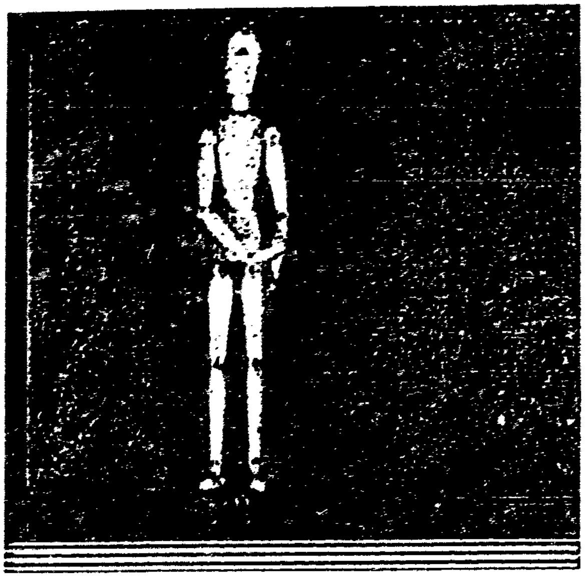
m7.cxap4



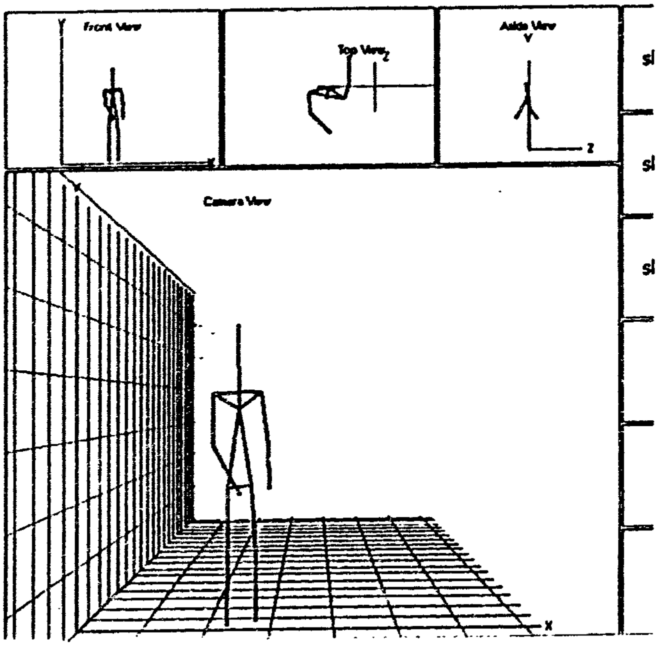
m8.cxap4

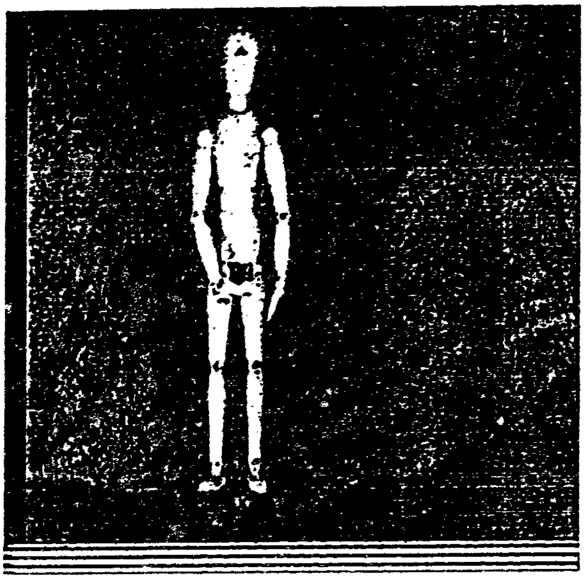


m9.cxap4

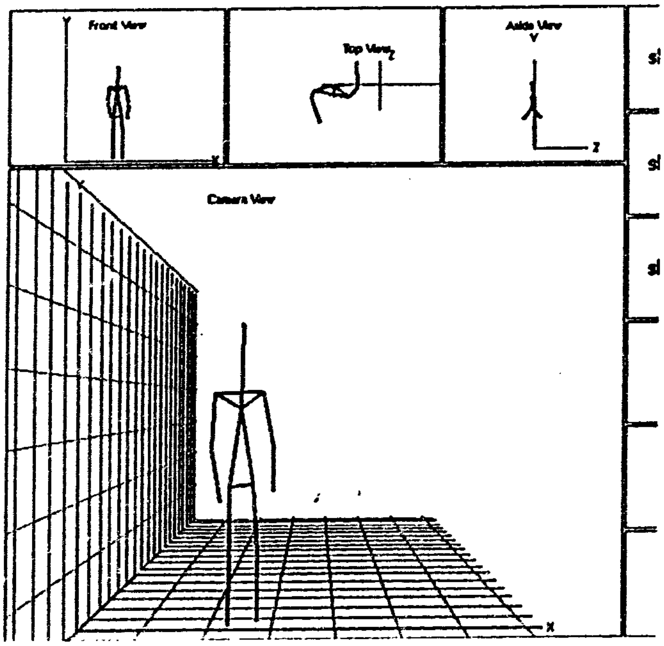


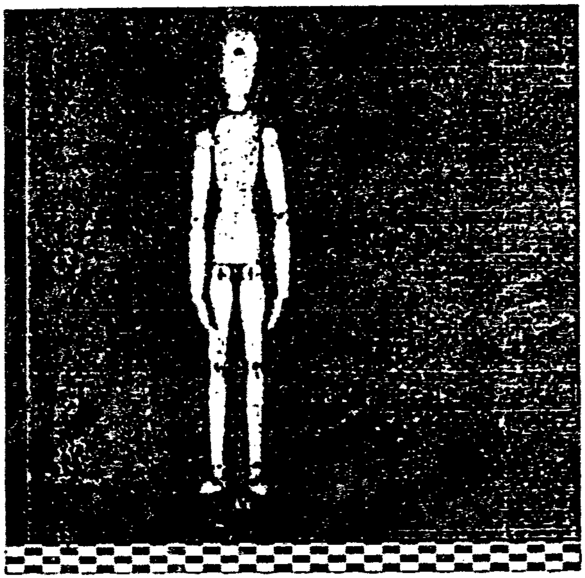
m10.cxap4



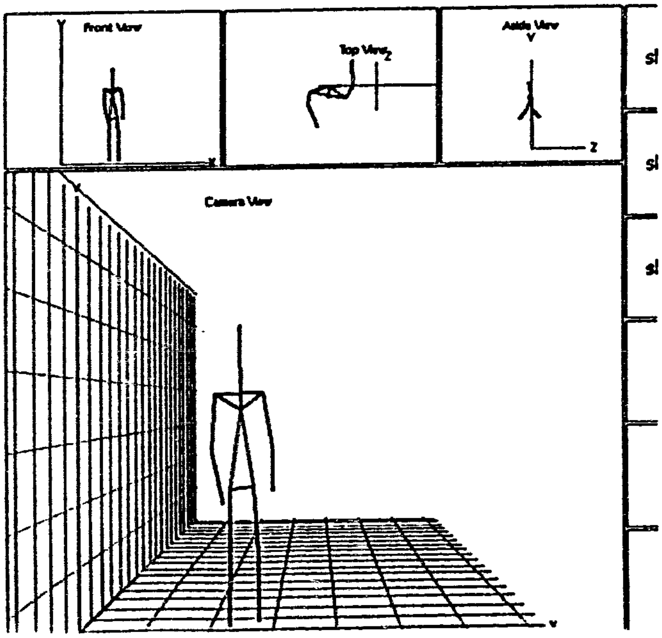


m11.cxap4





m12.cxap4



REFERENCES

[Badler79] Norman I. Badler and S. W. Smoliar, **"Digital Representations of Human Movement,"** *Ass. Comput. Mach. Comput. Surveys*, Vol. 11, pp. 19-38, March 1979.

[Badler87] Norman I. Badler, Kamran H. Manoochehri, and Graham Walters, **"Articulated Figure Positioning by Multiple Constraints,"** *IEEE Computer Graphics and Applications*, pp. 28-38, June 1987.

[Geurtz91] A. M. Geurtz, **"Three-Dimensional Human Motion Estimation: An Image Processing Approach,"** *Proc. of the International Symposium on 3-D Analysis of Human Movement*, pp. 19-21, July 1991.

[Brooks81] Rodney A. Brooks, **"Symbolic Reasoning Among 3-D Models and 2-D Images,"** *Artificial Intelligence*, 17, pp. 285-348, 1981.

[Brooks83] Rodney A. Brooks, **"Model-Based Three-Dimensional Interpretations of Two-Dimensional Images,"** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-5, No. 2, Mar. 1983.

[Calvert82] T. W. Calvert, J. Chapman, and A. Palta, **"Aspects of the Kinematic Simulation of Human Movement,"** *IEEE Computer Graphics and Applications*, pp. 41-50, Nov. 1982.

[Calvert88] T. W. Calvert, **"The Challenge of Human Figure Animation,"** *Graphics Interface'88, Proceedings*, pp. 203-210, 1988.

[Clowes71] M. B. Clowes, **"On Seeing Things,"** *Artificial Intelligence*, 2, pp. 79-112, 1971.

[ConzalezWintz84] Rafael C. Conzalez and Paul Wintz, **"Digital Image Processing,"** *Addison-Wesley Publishing Company*, 1984.

[FreemanLoutrel67] H. Freeman and P. P. Loutrel, **"An Algorithm for the Solution of the Two-Dimensional 'hidden-line' Problem,"** *IEEE trans. on Electronic Computer*, EC-16, PP. 784-790, 1967.

[Gibson66] J. J. Gibson, **"The Senses Considered As Perceptual Systems,"** *Houghton Mifflin, Boston*, 1966.

[Guzman68] A. Guzman, **"Decomposition of A Visual Scene Into Three-Dimensional Bodies,"** *Proc. AFIPS Fall Joint Comput. Conf.,* Vol. 33, pp. 291-304, 1968.

[Hannah80] Richard E. Hannah, **"Interpretation of Clinical Gait Analysis Data,"** *Proc. Canadian Society for Biomechanics Conf.,* pp. 80-81, October, 1980.

[Herman84] M. Herman, **"Matching Three-Dimensional Symbolic Description Obtained From Multiple Views of A Scene,"** *Proc. IEEE Conf. Computer Vision and Pattern Recognition, San Francisco. Ca.*, pp. 585-590, 1985.

[Horn77] B. K. P. Horn, **"Understanding Image Intensities,"** *Artificial Intelligence*, 8, pp. 201-231, 1977.

[Huffman71] D. A. Huffman, **"Impossible Objects As Nonsense Sentence,"** *Machine Intelligence*, Vol. 6, Edinburgh Univ. Press, Edinburgh, U.K., 1971.

[IkeuchiHorn80] K. Ikeuchi and B. K. P. Horn, **"Numerical Shape From Shading and Occluding Boundaries,"** *Artificial Intelligence*, 17 pp. 141-184, 1981.

[Kanade81] Takeo Kanade, **"Recovery of the Three-Dimensional Shape of an Object From A Single View,"** *Artificial Intelligence*, 17 pp. 409-460, 1981.

[Kender80] J. R. Kender, Ph.D. Thesis, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1980.

[LeeChen85] Hsi-Jian Lee and Zen Chen, **"Determination of 3D Human Body Postures From A Single View,"** *Computer Vision, Graphics, and Image Processing*, 30, pp. 148-168, 1985.

[Lowe89] David G. Lowe, **"Fitting Parameterized 3-D Models To Images,"** *Technical Report 89-26 UBC*, Dec. 1989.

[Marr81] D. Marr, **"Vision,"** *Freeman, San Francisco*, 1981.

[MarrNishihara78] D. Marr and H. K. Nishihara, **"Representation and Recognition of The Spatial Organization of Three-Dimensional Shapes,"** *Proc. R. Soc. Lond.*, B 200, pp. 269-293, 1978.

[MulliganMackworthLawrence89] I. Jane Mulligan, Alan K. Mackworth, and Peter D. Lawrence, **" A Model-Based Vision System For Manipulator Position Sensing,"** *Technical Report 89-13, UBC*, June 1989.

[Niblack86] Wayne Niblack, **"An Introduction to Digital Image Processing,"** *Prentice-Hall International Ltd.*, pp. 117-118, 1986.

[O'RourkeBadler80] Joseph O'Rourke and Norman I. Badler, **"Model-Based Image Analysis of Human Motion Using Constraint Propagation,"** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-2, No.6, Nov. 1980.

[Pavlidis80] Theo Pavlidis, **"Algorithms For Graphics and Image Processing,"** *Computer Science Press*, 1980.

[Perry90] Tekla S. Perry, **"Biomechanically Engineered Athletes,"** *IEEE Spectrum*, April, 1990.

[Potter77] J. L. Potter, **"Scene Segmentation Using Motion Information,"** *Comput. Graphics and Image Processing*, Vol.6, pp. 558-581, Dec. 1977.

[RoachAggarwal80] J. W. Roach and J. K. Aggarwal, **"Determining the Movement of Objects From A Sequence of Images,"** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-2, No. 6, pp. 554-562, Dec. 1980.

[Roberts65] L. G. Roberts, **"Machine Perception of Three Dimensional Solids,"** *Optical and Electrooptical Information Processing*, MIT Press, 1965.

[RogersAdams76] David F. Rogers and J. Alan Adams, **"Mathematical Elements For Computer Graphics,"** *McGraw-Hill Book Company*, 1976.

[Ullman76] S. Ullman, **"On Visual Detection of Light Sources,"**, *Biol. Cybernet.*, 21 pp. 205-212, 1976.

[Witkin81] Andrew P. Witkin, **"Recovering Surface Shape and Orientation From Texture,"** *Artificial Intelligence*, 17 pp. 17-45, 1981.

[YoungFu86] Tzay Y. Young and King-Sun Fu, **"Handbook of Pattern Recognition and Image Processing,"** *Academic Press Inc.*, 1986.