Human Motion Analysis: A Review

J. K. Aggarwal and Q. Cai
Computer and Vision Research Center
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712
Tel: (512)471-1369
Fax: (512)471-5532

Abstract

Human motion analysis is receiving increasing attention from computer vision researchers. This interest is motivated by a wide spectrum of applications, such as athletic performance analysis, surveillance, man-machine interfaces, content-based image storage and retrieval, and video conferencing. This paper gives an overview of the various tasks involved in motion analysis of the human body. We focus on three major areas related to interpreting human motion: 1) motion analysis involving human body parts, 2) tracking of human motion using single or multiple cameras, and 3) recognizing human activities from image sequences. Motion analysis of human body parts involves the low-level segmentation of the human body into segments connected by joints, and recovers the 3D structure of the human body using its 2D projections over a sequence of images. Tracking human motion using a single or multiple cameras focuses on higher-level processing, in which moving humans are tracked without identifying specific parts of the body structure. After successfully matching the moving human image from one frame to another in image sequences, understanding the human movements or activities comes naturally, which leads to our discussion of recognizing human activities. The review is illustrated by examples.

1 Introduction

Human motion analysis is receiving increasing attention from computer vision researchers. This interest is motivated by applications over a wide spectrum of topics. For example, segmenting the parts of the human body in an image, tracking the movement of joints over an image sequence, and recovering the underlying 3D body structure is particularly useful for analysis of athletic performance as well as medical diagnostics. The capability to automatically monitor human activities using computers in security-sensitive areas such as airports, borders, and building lobbies is of great interest to the police and military. With the development of digital libraries, the ability to automatically interpret video sequences will save tremendous human effort in sorting and retrieving images or video sequences using content-based queries. Other applications include building man-machine user interfaces, video conferencing, etc. This paper gives an overview of recent approaches to the various levels of tasks needed to accomplish the analysis of human motion from image sequences.

In contrast to our previous review of motion estimation of a rigid body [3], this survey concentrates on motion analysis of the human body, which is a non-rigid form. Our discussion covers three areas: 1) motion analysis of the human body structure, 2) tracking of human motion using a single or multiple cameras, and 3) recognizing human activities from image sequences. The relationship among these three areas is depicted in Figure 1. Our review follows a bottom-up approach in describing the general tasks for each area. Motion analysis of the human body usually involves the extraction of the low-level feature, such as body part segmentation, joint detection and identification, and the recovery of 3D structure from the 2D projections in an image sequence. Tracking moving individuals using a single or multiple cameras involves applying visual features to detect the presence of humans directly, i.e., without considering the geometric structure of the body parts. Motion information, such as position and velocity incorporated with intensity values, is employed to establish matching between consecutive frames. After feature correspondence between successive frames is solved, the next step is to

0-8186-8040-7/97 $10.00

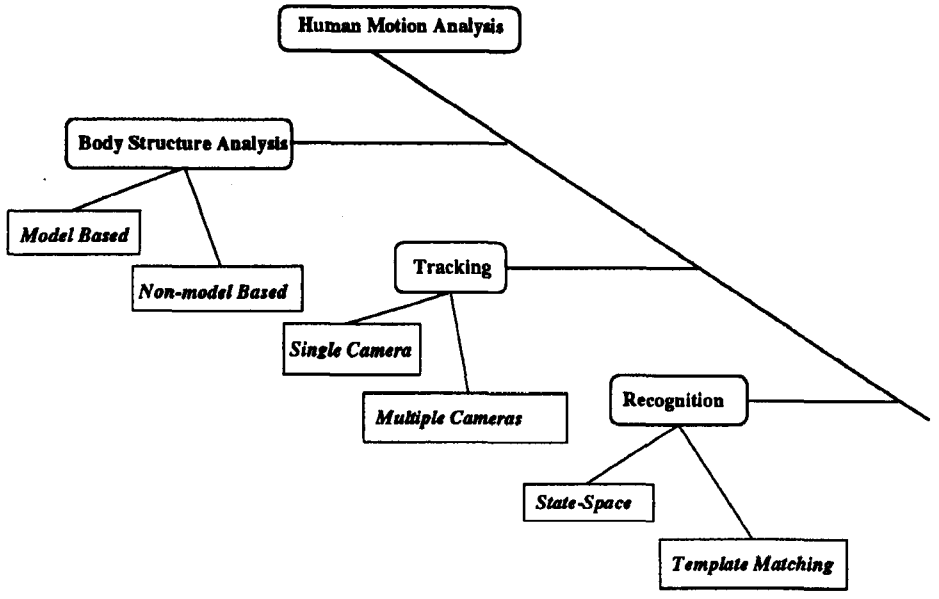

Figure 1: Relationship among the three areas of human motion analysis addressed in the paper.

understand the behavior of these features throughout the image sequence. Therefore, our discussion turns to a review of methods for recognition of human movements and activities.

There are two typical approaches to the motion analysis of human body parts, depending on whether *a priori* shape models are used. Figure 2 lists a number of the publications in this area over the past several years. The work denoted with an asterisk was developed in the Computer and Vision Research Center at The University of Texas at Austin. In each type of approach, the representation of the human body evolves from stick figures to 2D contours to 3D volumes as the complexity of the model increases. The stick figure representation is based on the observation that human motion is essentially the movement of the supporting bones. The use of 2D contours to represent the human body is directly associated with the projection of the human figure in images. Volumetric models, such as generalized cones, elliptical cylinders, and spheres, attempt to describe the details of a human body in 3D and thus require more parameters for computation.

With regard to the tracking of human motion without the use of body parts, we differentiate the work based on whether the subject is imaged at one time instant by a single camera or from multiple perspectives using different cameras. In both configurations, the features to be tracked vary from points to 2D blobs to 3D volumes. There is always a trade-off between feature complexity and tracking efficiency. Lower-level features, such as points, are easier to extract but relatively more difficult to track than higher-level features such as blobs and 3D volumes. Most of the work in this area is listed in Figure 3.

To recognize human activities from an image sequence, researchers typically use one of two types of approaches: approaches based on a state-space model or ones which use a template matching technique. In the first case, the features used for recognition have been points, lines, and 2D blobs. Methods using template matching usually apply meshes of a subject image to identify a particular movement. Figure 4 gives an overview of past research in this area. In some of the publications, recognition is conducted using only parts of the human images. Since these methods can be naturally extended to recognition of a whole body movement, we also include them in our discussion.

The organization of the paper is as follows: Section 2 reviews work on motion analysis of the human body structure. Section 3 covers the research on the higher-level tasks of tracking human motion without identifying the human body parts. Section 4 extends the discussion to recognition of human activity in image sequences based upon successfully tracking the features between consecutive frames. Finally, section 5 concludes the paper by giving general comments on previous work in the area of human motion analysis and discusses possible future directions of research in this area.

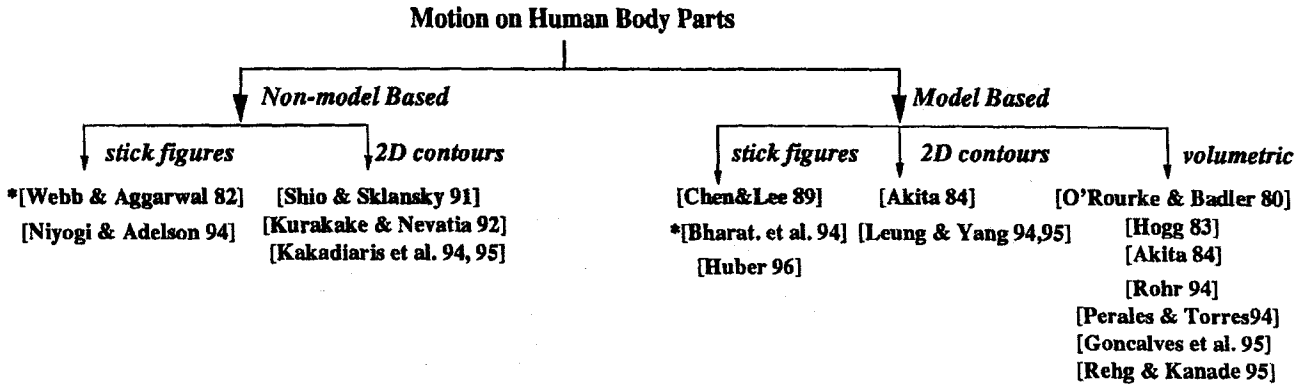

Figure 2: Past research on motion analysis of human body parts.

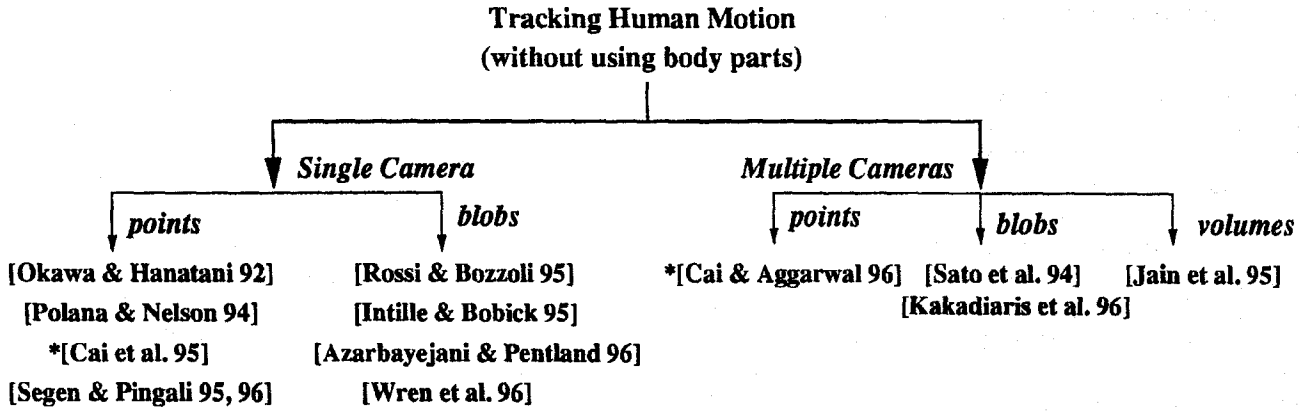

Figure 3: Past research on tracking of human motion without using body parts.

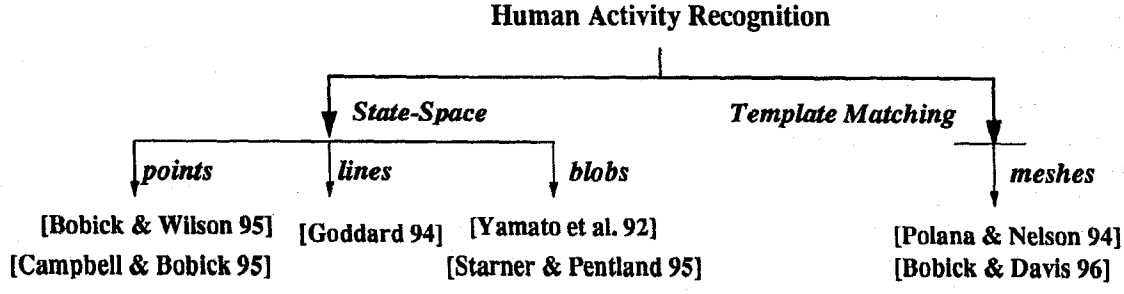

Figure 4: Past work on human activity recognition.

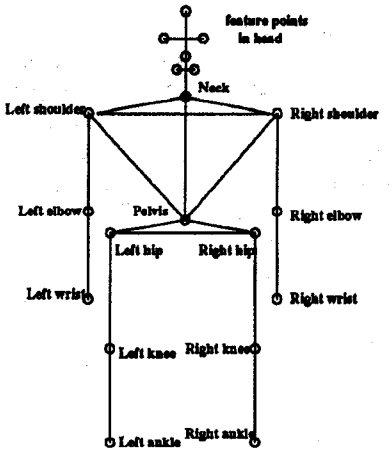

Figure 5: A stick-figure human model (based on Chen and Lee's work [11]).

2 Motion Analysis of Human Body Parts

This section focuses on motion analysis of human body parts, i.e., approaches which involve 2D or 3D analysis of the human body structure through image sequences. Conventionally, human bodies are represented as stick figures, 2D contours, or volumetric models [1]. Thus, body segments can be approximated as lines, 2D ribbons, and elliptical cylinders, accordingly. Figures 5, 6, and and 7 show examples of the stick figure, 2D contour, and volumetric representations of the human body, respectively. In our later discussion, human body motion is addressed by the movement of the limbs and hands [50, 28, 6, 33], such as the velocities of the hand or limb segments, or the angular velocity of various body parts.

Two general strategies are used, depending upon whether information about the object shape is employed in the motion analysis, namely, model-based approaches and methods which do not use *a priori* shape models. Both methodologies follow the general framework of: 1) feature extraction, 2) feature correspondence, and 3) high-level processing. The major difference between the two methodologies is in the process of establishing feature correspondence between consecutive frames. Methods which assume *a priori* shape models match the 2D image sequences to the model data. Feature correspondence is automatically achieved once matching between the images and the model data is established. When no *a priori* shape models are available, however, correspondence between successive frames is based upon prediction

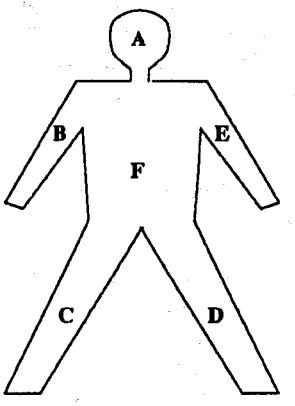

Figure 6: A 2D contour human model (similar to Leung and Yang's model [26]).

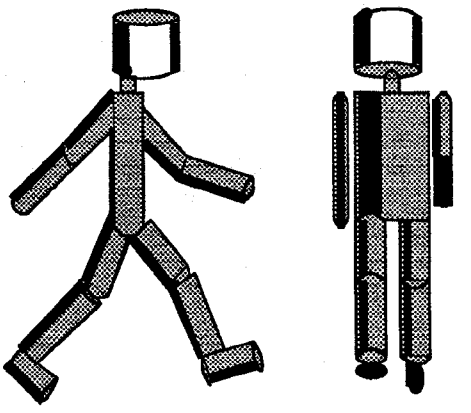

Figure 7: A volumetric human model (derived from Hogg's work [15]).

or estimation of features related to position, velocity, shape, texture, and color. These two methodologies can also be combined to complete processing at various levels, verify the matching between consecutive frames, and finally, accomplish complex, high-level tasks. Since we have addressed a certain amount of work in this area in [1], we will focus on the discussion of the very recent developments in this area, and describe others briefly.

2.1 Motion Analysis without a priori Shape Models

Most approaches for 2D or 3D interpretation of human body structure focus on motion estimation of the joints of body segments between consecutive frames. When no *a priori* shape models are assumed, heuristic assumptions are usually imposed to establish the correspondence of joints in an image sequence. These assumptions define the constraints on feature correspondence, decrease the search space, and, eventually, result in a unique match.

The simplest representation of a human body is the stick figure, which consists of line segments linked by joints. The motion of joints provides the key to motion estimation and recognition of the whole figure. This concept was initially considered by Johansson [23], who marked joints as *moving light displays* (MLD). Along this vein, Rashid [39] attempted to recover a connected human structure with projected MLD by assuming that points belonging to the same object have higher correlations in projected positions and velocities. Later, Webb and Aggarwal [47, 48] recovered the 3D structures of Johansson-type figures in motion. Their algorithm was based on the *fixed axis assumption*, which assumes that each rigid object (or parts of an articulated object) motion is constrained so that its axis of rotation remains fixed in direction. Therefore, the depth of the joints can be estimated from their 2D projections. Niyogi and Adelson [33], however, pursued another route to estimate the joint motion of human body segments. They first examined the spatial-temporal (XYT) braided pattern produced by the lower limb trajectories of a walking human and conducted gait analysis for coarse human recognition. Then the projection of head movements in the spatial-temporal domain was located, followed by the identification of other joint trajectories. These joint trajectories were then utilized to outline the contour of a walking human based on the observation that the human body is spatially contiguous. Finally, a more accurate gait analysis was performed using the outlined 2D contour, which led to a fine-level recognition of specific humans. Most of Bobick's studies [8, 11] concentrated on the trajectories of the MLDs of joints. However, they obtained the 3D positions of the joints directly from range data, and thus avoided the problem of feature correspondence. This strategy helped them to focus on higher-level processing, such as activity recognition and understanding, which will be reviewed in later sections.

Another way to describe the human body is by using 2D contours. Under such descriptions, the human body segments are analogous to 2D ribbons or blobs. For example, Shio and Sklansky [45] focused their work on 2D translational motion of human blobs. The blobs were grouped based on the magnitude and direction of the pixel velocity, which was obtained using techniques similar to the optical flow method [18]. The velocity of each part was considered to converge to a global average value over several frames. This average velocity corresponded to the motion of the whole human body. This observation led to identification of the whole subject via region grouping of blobs with a similar smoothed velocity. Kurakake and Nevatia [30] attempted to obtain the joint locations in images of walking humans by establishing correspondence between extracted ribbons. Their work assumed small motion between two consecutive frames, and feature correspondence was conducted using various geometric constraints. Joints were finally identified as the center of the area where two ribbons overlaps. Recent work by Kakadiaris *et al.* [26, 25] focused on body part decomposition and joint location from image sequences of the moving subject using a physics-based framework. In this work, the subject image is assumed to be one blob. As the subject moves and new postures occur, multiple new blobs are produced to replace the old ones, with each of them representing an emerging sub-part. Joints are determined based on relative motion and shape of two moving sub-parts.

2.2 Model-based Approaches

In the above subsection, we examined several approaches to motion analysis that do not use *a priori* shape models. Although this type of approach is necessary when no *a priori* shape models are available, it is usually difficult to establish feature correspondence between consecutive frames. On the other hand, daily observation teaches us that the human eye usually interprets moving figures using *a priori* shape models learned from previous experience. Thus most methods for the motion analysis of human body parts use models to fit to the given image, and, therefore, to accomplish high-level tasks at the final stage. As men-

tioned before, the models may be represented as stick figures, 2D contours, or volumetric models. We will discuss each of these below.

Chen and Lee [12] recovered the 3D configuration of a moving subject according to its projected 2D image. Their model used 17 line segments and 14 joints to represent the features of the head, torso, hip, arms, and legs (shown in Figure 5). Various constraints were imposed for the basic analysis of the gait. The method was computationally expensive, as it searched through all possible combinations of 3D configurations, given the known 2D projection, and required accurate extraction of 2D stick figures. Bharatkumar *et al.* [6] also used stick figures to model the lower limbs of the human body, where joints such as the hip, knee, and ankle were considered. They aimed at constructing a general model for gait analysis in human walking. Medial-axis transformations were applied to extract 2D stick figures of the lower limbs. The body segment angle and joint displacement were measured and smoothed from real image sequences, and then a common kinematic pattern was detected for each walking cycle. A high correlation (> 0.95) was found between the subject in real image sequences and the model, showing that the kinematic model is a good measure for detecting walking humans. Huber's human model [19] is a refined version of the stick figure representation. Joints are connected by line segments with a certain degree of constraint that can be relaxed by "virtual springs". Thus, this designed articulated kinematic model behaves analogously to a mass-spring-damper system. Motion and stereo measurements of joints are confined to a three-dimensional space called *Proximity Space* (PS). The human head serves as the starting point for tracking all PS locations. In the end, particular gestures were recognized based on the PS states of the joints associated with the head, torso, and arms.

Akita [4] focused on model-based motion analysis for real image sequences. Both stick figures and cone approximations were integrated and processed in a coarse-to-fine fashion. A key frame sequence of stick figures indicates the approximate order of the motion and spatial relationships between the body parts. The stick figure contains six segments: head, torso, arms, and legs. A cone model is included to provide knowledge of the rough shape of the body parts, which consists of six segments corresponding to the counterparts of the stick figure model. Perales and Torres also made use of both stick figure and volumetric representations in their work [36]. They introduced a predefined library with two levels of biomechanical graphical models. Level One is a stick figure tree with nodes for body segments and arcs for joints. Level Two is composed of descriptions of surface and body segments constructed with various 3D primitives used in computer graphics. Both levels of the model are applied in different matching stages.

Leung and Yang [31] applied a 2D ribbon model to recognize poses of a human performing gymnastic movements. The emphasis of their work is to estimate motion solely from the outline of a moving human subject. The system consists of two major processes: extraction of human outlines and interpretation of human motion. The 2D ribbon model is comprised of two components, the "basic" body model and the "extended" body model. The basic body model outlines the structural and shape relationships between the body parts, shown in Figure 6. The extended model consists of three patterns: the support posture model, the side view kneeling model, and side horse motion model. A modified edge detection technique was developed based on the work in Jain and Nagel [21] to generate a complete outline of the moving object images. A spatial-temporal relaxation process was proposed to determine which side of the moving edge belongs to the moving object. Two sets of 2D ribbons on each side of the moving edge, either a part of the body or that of the background, are identified according to their shape changes over time. The body parts are labeled according to the human body model. Then, a description of the body parts and the appropriate body joints is obtained.

Elliptical cylinders are one of the commonly used volumetric models for modeling human forms. Hogg [17] and Rohr [41] used the cylinder model originated by Marr and Nishihara [32], in which the human body is represented by 14 elliptical cylinders. Each cylinder is described by three parameters: the length of the axis and the major and minor axes of the ellipse cross section. The origin of the coordinate system is located at the center of the torso. Both Hogg and Rohr attempted to generate 3D descriptions of a human walking by modeling. Hogg [17] presented a computer program (WALKER) which attempted to recover the 3D structure of a walking person. Rohr applied eigenvector line fitting to outline the human image, and then fitted the 2D projections into the 3D human model using a distance measure similar to Hogg [17].

O'Rourke and Badler [35] conducted 3D human motion analysis by mapping the input images to an elaborate volumetric model. The model is a well-defined structure that consists of 24 rigid segments and 25 joints. The surface of each segment is defined by a col-

lection of overlapping sphere primitives. A coordinate system is embedded in the segments. Their model also includes the constraints of human motion, such as restrictions on joint angles, and a method to detect collisions between non-adjacent segments. Along the same vein, Rehg *et al.* [40] rendered two occluded fingers with several cylinders, and the center axes of the cylinders are projected into the center line segments of the 2D finger images. To track the motion of the self-occluded fingers, they assumed an invariant visibility order of 2D templates of these two occluded fingers to the viewing camera. Three possible occlusions were considered. These assumptions simplified the motion prediction and matching process between the 3D shape model and its projection in the image plane. Recent work by Goncalves *et al.* [16] addressed the problem of motion estimation of a human arm in 3D using a calibrated camera. Both the upper and lower arm were modeled as truncated circular cones, and the shoulder and elbow joints were assumed to be spherical joints. They used perspective projection of a 3D arm model to fit the blurred image of a real arm. Matching was conducted by recursively minimizing the error between the model projection and the real image through dynamically adapting the size and orientation of the model.

All of these approaches must match each real image frame to the corresponding model, which represents the human body structure at a abstract level. This procedure is itself non-trivial. The complexity of the matching process is governed by the number of model parameters and the efficiency of human body segmentation. When fewer model parameters are used, it is easier to match the feature to the model, but more difficult to extract the feature. For example, the stick figure is the simplest way to represent a human body, and thus it is relatively easier to fit the extracted lines into the corresponding body segments. However, extracting a stick figure from real images needs more care than searching for 2D blobs or 3D volumes.

3 Tracking Human Motion without Using Body Parts

In the previous section, we discussed human motion analysis requiring the geometric identification of joint connected body parts. The task of pre-recognition and locating features such as joints and body segments is a difficult one. It is computationally more efficient to track or recognize moving humans by directly using uninterpreted low-level visual features. We will discuss a number of methods that adopt such a strategy.

The objective of tracking is to establish correspondence of the image structure between consecutive frames based on features related to position, velocity, shape, texture, and color. Typically, the tracking process involves matching between images using pixels, points, lines, and blobs, based on their motion, shape, and other visual information [2]. There are two general classes of correspondence models, namely "iconic models", which use correlation templates, and "structural models", which use image features [2]. Iconic models are generally suitable for any objects, but only when the motion between two consecutive frames is small enough so that the object images in these frames are highly correlated. Because our interest is in tracking moving humans, which retain a certain degree of non-rigidity, our literature review is limited to the use of structure models.

Feature-based tracking typically starts with feature extraction, followed by feature matching over a sequence of images. The criteria for selecting a good feature are its robustness to noise, brightness, contrast, and size. To establish feature correspondence between successive frames, well-defined constraints are usually imposed to eliminate invalid matches and distinguish a unique correspondence. There is a trade-off between feature complexity and tracking efficiency. Lower-level features, such as points, are easier to extract but relatively more difficult to track than higher-level features such as lines, blobs, and polygons.

We discuss the problem of tracking human motion in two scenarios – with a single camera setup and with a distributed-camera configuration. Under the first setup, images are taken from the view of a single camera, whereas the a distributed-camera tracking system uses several cameras fixed at various locations of the monitored area to capture images simultaneously. In both cases, various levels of features could be used to establish matching in successive frames. The major difference is that the features used for matching using images taken from multiple perspectives must project to the same spatial reference, while tracking using a single camera does not have this requirement.

3.1 Single Camera Tracking

Most methods for tracking moving humans use image sequences taken from a single camera. Features used for tracking are usually points and motion blobs. Polana and Nelson [37] observed that the movements of arms and legs converge to that of the torso. In their work [37], each walking subject image was bounded by a rectangular box, and the centroid of the bounding

box was used as the feature to track. Positions of the center point in the previous frames were used to estimate the current position. Therefore, correct tracking was resolved even when the two subjects were occluded to each other in the middle of the image sequence. Cai *et al.* [10] also focused on tracking the movements of the whole human body using a viewing system with 2D translational movement. They focused on dynamic recovery of still or changing background images. The image motion of the viewing camera was estimated by matching the line segments of the background image. Then, motion-compensated frames were constructed to adjust three consecutive frames into the same spatial reference. In the final stage, subjects were tracked using the center of the bounding boxes and estimated motion information. Segen and Pingali's [44] people tracking system utilized the corner points of moving contours as the features for correspondence. These feature points were matched in forward and backward orders between two successive frames using a distance measure related to position and curvature values of the points. The matching process implies that a certain degree of rigidity of the moving human body and small motion between consecutive frames was assumed. Finally, short-lived or partially overlapped trajectories were merged into long-lived paths.

Another commonly used feature for tracking is 2D blobs or meshes. In Okawa and Hanatani's work [34], background pixels were voted as the most frequent value during the image sequence [45]. The meshes belong to a moving foot were then detected by incorporating low-pass filtering and masking. Finally, the distance between the human foot and the viewing camera was computed by comparing the relative foot position to the pre-calibrated CAD floor model. Rossi and Bozzoli [42] also used moving blobs to track and count people crossing the field of view of a vertically mounted camera. Occlusion of multiple subjects was avoided due to the viewing angle of the camera, and tracking was performed using position estimation during the period when the subject enters the top and the bottom of the image. Intille and Bobick [20] solved their tracking problem by taking advantage of the knowledge of a so called "closed-world". A "closed-world" is a space-time domain where the knowledge of all possible objects present in the image sequences are available. They illustrated their tracking algorithm using the example of a football game, where the background and the rules for play are known *a priori*. Camera motion was removed by establishing homographic transforms between the football field model and its model using landmarks in the field. The players were detected as moving blobs, and tracking was performed by template matching the neighbor region of the player image between consecutive frames. Pentland *et al.* [49, 5] explored the blob feature thoroughly. In their work, blobs were not restricted to regions due to motion, and could be any homogeneous areas, such as color, texture, brightness, motion, shading, or a combination of these. Statistics such as mean and covariance were used to model the blob features in both 2D and 3D. In [49], the feature vector of a blob is formulated as (x, y, Y, U, V), consisting of a spatial (x, y) and color (Y, U, V) information. A human body is constructed by blobs representing various body parts such as head, torso, hands, and feet. Meanwhile, the surrounding scene is modeled as a texture surface. Gaussian distributions were assumed for both models of human body and background scene. Finally, pixels belong to the human body were assigned to different body part blobs using the log-likelihood measure. Later, Azarbayejani and Pentland [5] recovered the 3D geometry of the blobs from a pair of 2D blob features via nonlinear modeling and recursive estimation. The 3D geometry included the shape and the orientation of the 3D blobs along with the relative rotation and translation to the binocular camera. Tracking of the 2D blobs was inherent in the recovery process, which iteratively searched the equilibria of the nonlinear state space model across image sequences.

3.2 Multiple Camera Tracking

As stated above, most previous tracking methodologies have been limited to a single camera configuration. The disadvantage of using only one camera is that the area captured by the camera is relatively narrow due to the limited field of view of one single camera. To enlarge the monitored area, one strategy is to mount multiple cameras at various locations of the area of interest, so that if the subject disappears from the field of view of one camera, it will appear in the view of another camera in the system. A multiple camera setup also helps to solve the ambiguity of matching when subject images are occluded to each other. Only in very recent years has work on tracking of human motion from multiple perspectives emerged [43, 22, 27, 9]. Compared to the problem of tracking moving humans from a single camera, establishing feature correspondence between images captured at different locations is more challenging because the features are recorded in different spatial coordinates. All features to be tracked must be adjusted to the same spatial reference before matching is performed. Recent work by Cai and Aggarwal [9] uses multiple

points belonging to the medial axis of the human upper body as the feature to track. These points are sparsely sampled and assumed to be independent of each other, which preserves a certain degree of non-rigidity of the human body. Location and average intensity of the feature points are integrated to find the most likely match between two consecutive frames imaged from different viewing angles. One feature point is assumed to match to its corresponding epipolar line via motion estimation and pre-calibration of the cameras. Multivariate Gaussian distributions are assumed for the class-conditional probability density function of features of candidate subject images. Experimental results of tracking moving humans in indoor environments using a prototype system equipped with three cameras indicated robust performance and a potential for real time implementation. Sato *et al.* [43], on the other hand, treated a moving human as a combination of various blobs of its body parts. All distributed cameras are calibrated in the world coordinate system, which correspond to a CAD model of the indoor environment. The blobs of body parts were matched over image sequences using their area, average brightness and rough 3D position in the world coordinates. The 3D position of a 2D blob was estimated based on height information by measuring the distance between the center of gravity of the blob and the floor. These small blobs were then merged into an entire region of the human image using the spatial and motion parameters from several frames. Kakadiaris and Metaxas [24] consider the problem of tracking humans using as inputs from three cameras to estimate 3D human motion. Kelly *et al.* [27, 22] adopt a similar strategy [43] to construct a 3D environmental model. They introduce a feature called voxels, which are sets of cubic volume elements containing information such as which object belongs to this pixel and the history of this object. The depth information of the voxel is also obtained using height estimation. Moving humans are tracked as a group of these voxels from the "best" angle of the viewing system. A significant amount of effort is made in [22, 27] to construct a friendly graphical interface for browsing multiple images sequences of a same event.

4 Human Activity Recognition

In this section, we review work on recognizing human activities from image sequences. Usually, human action recognition is based on successfully tracking the human through images sequences, and thus is considered to be a higher level task. A large body of literature is devoted to human facial motion recognition and emotion detection, which fall into the category of elastic non-rigid motion. In this paper, we are only interested in motion involving a human body, i.e., human body motion as a form of articulated motion. The difference between articulated motion and elastic motion is well addressed in [1]. There also exists a certain amount of work on human image recognition by applying geometric features, such as 2D clusters [45, 9], profile projects [29], texture and color information [14, 51]. Since motion information was not incorporated in the recognition process, we exclude them from further discussion.

For human activity or behavior recognition, most efforts have been concentrated on using state-space approaches [13] to understand the human motion sequence [50, 15, 8, 11, 46]. Another approach is to use the template matching technique [37, 7] to compare the feature extracted from the given image sequence to the pre-stored patterns during the recognition process. The advantage of using the template matching technique is its inexpensive computational cost; however, it is relatively sensitive to the variance of the movement duration. Approaches using state-space models, on the other hand, define each static posture as a state. These states are connected by certain probabilities. Any motion sequence as a composition of these static poses is considered a tour going through various states. Joint probabilities are computed through these tours, and the maximum value is selected as the criterion for classification of activities. Under such a scenario, duration of motion is no longer an issue because each state can repeatedly visit itself. However, approaches using these methods usually need intrinsic nonlinear models and do not have closed-form solutions. As we know, nonlinear modeling also requires searching for a global optimum in the training process, which requires complex computing iterations. Meanwhile, selecting the proper number of states and dimension of the feature vector to avoid "underfitting" or "overfitting" remains an issue. In the following subsections, template matching and state-space approaches [13] will be discussed.

4.1 Approaches using Template Matching

We start with approaches which use template matching. So far, the feature used for recognition in this category has been 2D meshes. Based on successfully tracking a moving human image from image sequences, Polana and Nelson [37] compute the optical flow fields [18] between consecutive frames and divide each flow frame into a spatial grid in both X

and Y directions. The motion magnitude in each cell is summed, forming a high dimensional feature vector used for recognition. To normalize the duration of the movement, they assume that human motion is periodic and divide the entire sequence into a number of cycles of the activity. Motion in a single cycle is averaged throughout the number of cycles and differentiated into a fixed number of temporal divisions. Finally, activity recognition is processed using the nearest neighbor algorithm. Recent work by Bobick and Davis [7] follows the same vein, but extracts the motion feature differently. They interpret human motion in an image sequence by using *motion-energy* images (MEI) and *motion-history* images (MHI). The motion images in a sequence are calculated via differencing between successive frames and then thresholded into binary values. These motion images are accumulated in time and form MEI, which are binary images containing motion blobs. The MEI are enhanced into MHI, where each pixel value is proportional to the duration of motion at that position. Each action consists of MEIs and MHIs obtained from images sequences captured from various viewing angles. Moment-based features are extracted from MEIs and MHIs and employed for recognition using template matching.

4.2 State-Space Approaches

State space models have been widely used to predict, estimate, and detect signals over a large variety of applications. One representative model is perhaps the Hidden Markov Model (HMM), which is a probabilistic technique for the study of discrete times series. HMM has been very popular in speech recognition, but only recently has it been adopted for recognition of human motion sequences in computer vision [50]. Its model structure could be summarized as a hidden Markov chain and a finite set of output probability distributions [38]. The basic structure of an HMM is shown in Figure 8, where S_i represents each state connected by probabilities to other states or its own, and $y(t)$ is the observation derived from each state. The main tool in HMM is the Baum-Welch (forward-backward) algorithm for maximum likelihood estimation of the model parameters. Features to be recognized in each state vary from points and lines to 2D blobs. We will address past developments in this area according to the complexity of the feature used for recognition.

Bobick [8] and Campbell [11] applied 2D or 3D Cartesian tracking data sensed by MLDs of body joints for activity recognition. To state more specifically, the trajectories of multiple part joints form a

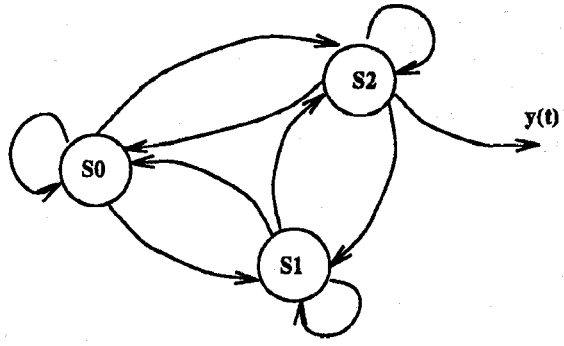

Figure 8: The basic structure of an Hidden Markov Model.

high dimensional phase space, and this phase space or its subspace are employed as the feature to recognize. Both of them are intended to transform the continuous motion into a set of discrete symbol representations. They apply the trajectories of the high dimensional phase space or its subspace as the feature for recognition. The feature vector in each frame in a motion sequence is portrayed as a point in the phase space, which belongs to a certain state. A typical gesture is defined as an order sequence of these states restricted by motion constraints. In Campbell [11], the learning/training process is conducted by fitting the unique curve of a gesture into the subspace of the full phase space into low-order polynomials. Gesture recognition is based on the maximum value of the correlation between the predictor and the current motion. Bobick [8], on the other hand, applied the *k-means* algorithm to find the center for each cluster (or state) using the sample points of the trajectories. Classifying motion trajectories to a gesture was accomplished through dynamic programming.

Goddard's human movement recognition [15] focused on the low limb segments of the human stick figure. 2D projections of the joints were directly used, without interpretation, as inputs, and features for recognition were encoded by coarse orientation and coarse angular speed of the line segments in the image plane. Although Goddard [15] did not directly apply HMM in his work for discrimination of human gaits, he also considered a movement as a composition of *events* linked by time *intervals*. Each *event* is independent of the movement. The links between these *events* are restrained by feature and temporal constraints. This type of structure is called a scenario, which is easily extended by assembling scenarios into composite scenarios. Learning is done by training the parallel network mapped by a feature hierarchy. Matching real scene

kinematics to modeled movements involves visual motion feature in the scene, *a priori* events matched to the scenario, and the temporal constraints on the sequence and time intervals.

Another commonly used feature for identifying human movements or activities is 2D blobs, obtained from any homogeneous regions based on motion, color, texture, etc. The work by Yamato *et al.* [50] is perhaps the first one on recognition of human action in this category. Mesh features of binary moving human blobs are used as the low-level feature for learning and recognition. Learning was implemented by training the HMMs to generate symbol patterns for each class. Optimization of the model parameters is achieved using the Baum-Welch algorithm. Finally, recognition is based on the output of the given image sequence using forward calculation. They tested sequences with six tennis strokes and achieved recognition rates ranging from 70% to 100%, depending on the number of training patterns. Recent work by Starner and Pentland [46] applied a similar method to recognition of American Sign Language. Instead of directly using the low-level mesh feature, they introduced the position of a hand blob, its angle of axis of least inertia, and eccentricity of its bounding ellipse in the feature vector.

5 Conclusion

We have given an overview of past developments in human motion analysis. Our discussion has focused on three major tasks: 1) motion analysis of the human body parts, 2) high-level tracking of human motion using a single or multiple cameras, and 3) recognition of human movements or activities based on successfully tracking features over an image sequence. Motion analysis of the human body parts is essentially the 2D or 3D interpretation of the human body structure using the motion of the body parts over image sequences, and involves low-level tasks such as body part segmentation, joint location and detection. Tracking human motion is a higher level task in which the parts of the human body are not explicitly identified, e.g., the human body is considered as a whole when establishing matches between consecutive frames. Tracking procedures depend on whether the subject is imaged at one time instant by a single camera or from multiple perspectives using different cameras. Tracking a subject from images taken from multiple perspectives at the same time instant requires that features be projected into a common spatial reference. The task of recognizing human activity over image sequences assumes that feature tracking for recognition has been accomplished. Two typical approaches are addressed: those based on a state-space model and those based on template matching the given images to a pre-stored pattern. Template matching is easy to implement, but sensitive to noise and the time interval of the movements. State-space approaches, on the other hand, overcome these drawbacks but usually involve complex iterative computation.

The key to successful execution of high-level tasks is to establish feature correspondence between consecutive frames, which still remains a bottle-neck in the whole processing. Typically, constraints on human motion or human models are assumed in order to decrease the ambiguity during the matching process. However, these assumptions may not closely fit the real situation or may introduce other tractable issues, such as the difficulty in estimating model parameters given the real image data. Recognition of human motion is just in its infancy, and there exists a trade-off between computational cost and motion duration accuracy for methodologies based on state-space models and template matching. New techniques are expected to improve the performance and, meanwhile, decrease the computational cost.

References

[1] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata. Non-rigid motion analysis: Articulated & elastic motion. accepted by *CVGIP: Image Understanding*.

[2] J. K. Aggarwal, L. S. Davis, and W. N. Martin. Correspondence process in dynamic scene analysis. *Proc. of the IEEE*, 69(5):562–572, 1981.

[3] J. K. Aggarwal and N. Nandhakumar. On the computation of motion of sequences of images - a review. *Proc. of the IEEE*, 76(8):917–934, 1988.

[4] K. Akita. Image sequence analysis of real world human motion. *Pattern Recognition*, 17(1):73–83, 1984.

[5] A. Azarbayejani and A. Pentland. Real-time self-calibrating stereo person tracking using 3-[d] shape estimation from blob features. In *Proc. of Intl. Conf. on Pattern Recognition*, pages 627–632, Vienna, Austria, August 1996.

[6] A. G. Bharatkumar, K. E. Daigle, M. G. Pandy, Q. Cai, and J. K. Aggarwal. Lower limb kinematics of human walking with the medial axis transformation. In *Proc. of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects*, pages 70–76, Austin, TX, 1994.

[7] A. F. Bobick and J. Davis. Real-time recognition of activity using temporal templates. In *Proc. of IEEE Computer Society Workshop Applications on Computer Vision*, pages 39–42, Sarasota, FL, 1996.

[8] A. F. Bobick and A. D. Wilson. A state-based technique for the summarization and recognition of gesture. In *Proc. of 5th Intl. Conf. on Computer Vision*, pages 382–388, 1995.

[9] Q. Cai and J. K. Aggarwal. Tracking human motion using multiple cameras. In *Proc. of Intl. Conf. on Pattern Recognition*, pages 68–72, Vienna, Austria, August 1996.

[10] Q. Cai, A. Mitiche, and J. K. Aggarwal. Tracking human motion in an indoor environment. In *Proc. of 2nd Intl. Conf. on Image Processing*, volume 1, pages 215–218, Washington, D.C., October 1995.

[11] L. W. Campbell and A. F. Bobick. Recognition of human body motion using phase space constraints. In *Proc. of 5th Intl. Conf. on Computer Vision*, pages 624–630, 1995.

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

[13] J. Farmer, M. Casdagli, S. Eubank, and J. Gibson. State-space reconstruction in the presence of noise. *Physics D*, 51D:52–98, 1991.

[14] D. A. Forsyth and M. M. Fleck. Identifying nude pictures. In *Proc. of IEEE Computer Society Workshop Applications on Computer Vision*, pages 103–108, Sarasota, FL, 1996.

[15] N. H. Goddard. Incremental model-based discriminiation of articulated movement from motion features. In *Proc. of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects*, pages 89–95, Austin, TX, 1994.

[16] L. Goncalves, E. D. Bernardo, E. Ursella, and P. Perona. Monocular tracking of the human arm in 3d. In *Proc. of 5th Intl. Conf. on Computer Vision*, pages 764–770, Cambridge, MA, 1995.

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

[18] B. K. P. Horn and B. G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–204, 1981.

[19] E. Huber. 3D real-time gesture recognition using proximity space. In *Proc. of Intl. Conf. on Pattern Recognition*, pages 136–141, Vienna, Austria, August 1996.

[20] S. S. Intille and A. F. Bobick. Closed-world tracking. In *Proc. Intl. Conf. Comp. Vis.*, pages 672–678, 1995.

[21] R. Jain and H. H. Nagel. On the analysis of accumulative difference pictures from image sequences of real world scenes. *IEEE Trans. on PAMI*, 1(2):206–214, 1979.

[22] R. Jain and K. Wakimoto. Multiple perspective interactive video. In *Proc. of Intl. Conf. on Multimedia Computing and Systems*, pages 202–211, 1995.

[23] G. Johansson. Visual motion perception. *Sci. American*, 232(6):76–88, 1975.

[24] I. A. Kakadiaris and D. Metaxas. Model based estimation of 3D human motion with occlusion based on active multi-viewpoint selection. In *Proc. of IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, pages 81–87, San Francisco, CA, 1996.

[25] I. A. Kakadiaris and D. Metaxas. 3d human body model acquisition from multiple views. In *Proc. of 5th Intl. Conf. on Computer Vision*, pages 618–623, 1995.

[26] I. A. Kakadiaris, D. Metaxas, and R. Bajcsy. Active part-decomposition, shape and motion estimation of articulated objects: A physics-based approach. In *Proc. CVPR*, pages 980–984, Seattle, WA, 1994.

[27] P. H. Kelly, A. Katkere, D. Y. Kuramura, S. Moezzi, S. Chatterjee, and R. Jain. An architecture for multiple perspective interactive video. In *Proc. of ACM Conf. on Multimedia*, pages 201–212, 1995.

[28] W. Kinzel. Pedestrian recognition by modelling their shapes and movements, in progress in image analysis and processing III. In *Proc. of the*

7th Intl. Conf. on Image Analysis and Processing 1993, pages 547–554, Singapore, 1994.

[29] Y. Kuno and T. Watanabe. Automatic detection of human for visual surveillance system. In *Proc. of Intl. Conf. on Pattern Recognition*, pages 865–869, Vienna, Austria, August 1996.

[30] S. Kurakake and R. Nevatia. Description and tracking of moving articulated objects. In *11th Intl. Conf. on Pattern Recognition*, volume 1, pages 491–495, Hague, Netherlands, 1992.

[31] M. K. Leung and Y. H. Yang. First sight: A human body outline labeling system. *IEEE Trans. on PAMI*, 17(4):359–377, 1995.

[32] D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. In *Proc. R. Soc. London*, volume B, pages 269–294, 1978.

[33] S. A. Niyogi and E. H. Adelson. Analyzing and recognizing walking figures in XYT. In *Proc. CVPR*, pages 469–474, Seattle, WA, 1994.

[34] Y. Okawa and S. Hanatani. Recognition of human body motions by robots. In *Proc. IEEE/RSJ Intl. Conf. Intelligent Robots and Systems*, pages 2139–2146, Raleigh, NC, 1992.

[35] J. O'Rourke and N. I. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Trans. PAMI*, 2:522–536, 1980.

[36] F. J. Perales and J. Torres. A system for human motion matching between synthetic and real image based on a biomechanic graphical model. In *Proc. of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects*, pages 83–88, Austin, TX, 1994.

[37] R. Polana and R. Nelson. Low level recognition of human motion (or how to get your man without finding his body parts). In *Proc. of IEEE Computer Society Workshop on Motion of Non-Rigid and Articulated Objects*, pages 77–82, Austin, TX, 1994.

[38] A. B. Poritz. Hidden Markov Models: A guided tour. In *Proc. IEEE Intl. Conf. on Acoust. Speech and Signal Proc.*, pages 7–13, 1988.

[39] R. F. Rashid. Towards a system for the interpretation of moving light display. *IEEE Trans. on PAMI*, 2(6):574–581, November 1980.

[40] J. M. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *Proc. of 5th Intl. Conf. on Computer Vision*, pages 612–617, Cambridge, MA, 1995.

[41] K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP: Image Understanding*, 59(1):94–115, 1994.

[42] M. Rossi and A. Bozzoli. Tracking and counting people. In *1st Intl. Conf. on Image Processing*, pages 212–216, Austin, Texas, November 1994.

[43] K. Sato, T. Maeda, H. Kato, and S. Inokuchi. CAD-based object tracking with distributed monocular camera for security monitoring. In *Proc. 2nd CAD-Based Vision Workshop*, pages 291–297, Champion,PA, Feburary 1994.

[44] J. Segen and S. Pingali. A camera-based system for tracking people in real time. In *Proc. of Intl. Conf. on Pattern Recognition*, pages 63–67, Vienna, Austria, August 1996.

[45] A. Shio and J. Sklansky. Segmentation of people in motion. In *Proc. of IEEE Workshop on Visual Motion, IEEE Computer Society*, pages 325–332, October 1991.

[46] T. Starner and A. Pentland. Visual recognition of American Sign Language using Hidden Markov Models. In *Proc. Intl. Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, June 1995.

[47] J. A. Webb and J. K. Aggarwal. Visually interpreting the motion of objects in space. *IEEE Computer*, pages 40–46, August 1981.

[48] J. A. Webb and J. K. Aggarwal. Structure from motion of rigid and jointed objects. In *Artificial Intelligence*, volume 19, pages 107–130, 1982.

[49] C. Wren, A. Azarbayejani, T. Darrel, and A. Pentland. Pfinder: Real-time tracking of the human body. In *Proc. SPIE*, Bellingham,WA, 1995.

[50] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using Hidden Markov Model. In *Proc. IEEE Conf. CVPR*, pages 379–385, Champaign, IL, June 1992.

[51] J. Yang and A. Waibel. A real-time face tracker. In *Proc. of IEEE Computer Society Workshop Applications on Computer Vision*, pages 142–147, Sarasota, FL, 1996.