IEEE CVPR-94, Seattle, Washington, June 20-24, 1994, pages 214-221.

A Survey of Motion Analysis from Moving Light Displays

Claudette Cédras and Mubarak Shah
Computer Science Dept.
University of Central Florida
Orlando, FL 32816

Abstract

Motion-based recognition deals with the recognition of objects or motions directly from the motion information extracted from a sequence of images. There are two main steps in this approach. The first consists of finding an appropriate representation for the objects or motions, from the motion cues of the sequence, and then organize them into useful representations. The second step consists of the matching of some unknown input with a model. This paper provides a review of recent developments in motion-based recognition.

1 Introduction

Motion perception plays an important role in the human visual system. It helps us recognize different objects and their motion in a scene, infer their relative depth, rigidity, etc. We tend to focus our attention on moving objects, while motionless objects are not as easily detectable. Our sensitivity and ease of perception and interpretation of motion suggests that our visual system is well adapted to temporal information.

Motion perception has been studied extensively using Johansson's moving light displays (MLDs) [15]. MLDs consist of bright spots attached to an actor dressed in black, and moving in front of a dark background. The collection of spots carry only 2D and no structural information, since they are not connected. A set of static spots remained meaningless to observers, while their relative movement created a vivid impression of a person walking, running, dancing, etc. The gender of a person, even the gait of a friend can be recognized based solely on the motion of those spots [4].

There are two theories about the interpretation of MLD type stimuli. In the first, people use motion information in the MLD to recover the 3D structure (structure from motion) and subsequently use the structure for recognition. The moving object would be identified first, then the motion it performs in the sequence would be sought. According to the second theory, motion information is directly used to identify a motion, without any structure recovery.

There has been significant interest over the last decade, in the computer vision community, in the structure from motion theory (e.g. [13, 29, 30]). In that approach, 3D coordinates of points on the moving objects and their 3D motion is recovered from a sequence of frames. This problem is formulated in terms of systems of nonlinear or linear equations given 2D location of moving points among a few frames. In these approaches, it is assumed that the recovered 3D structure will subsequently be used for recognition. However, 3D structure is not sufficient alone for robust and accurate recognition, and the reconstruction is sensitive to noise. Multiple cues like motion, specularities, textures, etc. , are needed.

The second theory of motion analysis deals with the direct use of motion information for recognition. In this approach, the emphasis is not on the static structure, and motion information is not extracted one frame at a time. Instead, a large number of frames is used to extract motion information in its continuum. The use of a longer sequence leads to recognition of higher level movements like walking, which consist of a complex and coordinated series of events that cannot be understood by looking at only a few frames. Therefore, more complex motions can be examined at a more appropriate level.

Motion-based recognition consists of the recognition of objects or motions directly from the motion information extracted from the sequence of images. Knowledge about the object or motion is used to construct models that will serve in the recognition process. There are two main steps in this approach. The first consists of finding an appropriate representation for the objects or motions, from the motion cues of the sequence, and then organize them into useful representations. Models are then created and extended as necessary. The second step consists of the matching of some unknown input with a model. Methods here often consist of pattern classification techniques.

Another way to use motion for recognition is to explicitly use shape and motion models to predict and recover the motion performed by an object. Motion here is defined as a sequence of the specification of parameters defining the shape of an object in time. This approach is often designated as tracking. Most of the work in this area pertains to the walking motion of a person.

2 Extraction of Motion Information and Matching

The first important step in motion-based recognition is the extraction of motion information from a sequence of images. There are generally two methods for extracting 2D motion: motion correspondence and optical flow. Optical flow is an approximation of the 2D flow field from image intensities, and methods for its computation are well documented. Motion correspondence is concerned with the matching of characteristic tokens through time. This correspondence results in what is called a motion trajectory, i.e. a sequence of locations (x, y) through time.

Most features used in motion and object representations are derived from motion trajectories and optical flow. For instance, motion trajectories can be parametrized in several ways: by finding speed and direction, velocity v_x and v_y, or spatiotemporal curvature. Parametrized representations can be analyzed to identify important motion events, i.e. particular occurrences in the motion, like a sudden change of direction or a stop, or they can be compared to other curves to determine their relative motion. Extraction of motion information over a region or over a whole image can also be used, as opposed to motion trajectories that carry information about a single point on an object. Features derived from an extended region are called region-based features.

In this section, we will examine the kind of information that can be extracted from a sequence of images, and how recognition or classification takes place.

2.1 Trajectory Parametrization

Trajectories are basically vector valued functions, since at each frame i we have two locations, x and y. Parametrization into single valued functions is very useful. One representation is the velocity v_x and v_y, i.e. the velocity in x and in y relative to time. Speed and direction are other useful parametrizations, and are respectively defined as $s_i = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$ and $d_i = \arctan\left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i}\right)$. Velocity, speed and direction are fairly easy to compute, and generate curves that are easy to interpret.

The spatiotemporal curvature κ of a trajectory is another common representation. It is determined as $\kappa = \frac{\sqrt{A^2+B^2+C^2}}{((x')^2+(y')^2+(t')^2)^{3/2}}$, where $A = \begin{vmatrix} y' & t' \\ y'' & t'' \end{vmatrix}$, $B = \begin{vmatrix} t' & x' \\ t'' & x'' \end{vmatrix}$, $C = \begin{vmatrix} x' & y' \\ x'' & y'' \end{vmatrix}$, and $|\cdot|$ denotes the determinant.

2.2 Relative Motion and Motion Events

In the trajectory parametrizations, absolute values of velocity, speed, direction and curvature were used. However, absolute values might sometimes be inadequate. Cutting and Proffitt [6] showed that relative motion is an important aspect in human visual perception. This kind of information should therefore be very helpful in computer vision systems. Multiple trajectories can be used to compute relative motion. For example, relative angles can be found as the joint angle between three points, in each frame. Angular velocities could then be determined from frame to frame.

Motion events are defined as significant changes or discontinuities in motion. A sudden change of direction or velocity, for example, can provide important clues to the type of object and/or its motion. Motion events are usually detected by the presence of discontinuities that can be found by looking at derivatives of the velocity, for example.

Gould and Shah's Trajectory Primal Sketch (TPS) is a representation for the significant changes in motion [12]. Changes are identified at various scales by computing the scale-space of the velocity curves v_x and v_y extracted from the trajectory of a point. This results in a set of TPS contours, each contour corresponding to a change in motion. The representation has been shown to distinguish basic motions like translation, rotation, projectile and cycloid.

Based on psychophysical considerations, Engel and Rubin [9] described the significant changes in motion as motion boundaries, i.e. motion events that partition a global motion into its psychological parts, of which they found five types: smooth starts, smooth stops, pauses, impulse starts and impulse stops. They used polar velocity representation (s, ϕ), and the features extracted for the detection of the perceptual boundaries are first and second derivatives of the speed s and the second derivative of the direction ϕ.

Goddard [11] used, as motion events, changes in rotational velocity of body segments along with changes in their direction. The angular velocity range was partitioned into six, and the four quadrants were used for the orientation. Any change in the orientation or an-

gular velocity constituted a motion event, which triggered some action in his connectionist system.

2.3 Region-Based Features

For certain types of objects or motions, the extraction of precise motion information for each single point is neither desirable nor necessary. Instead, the ability to have a more general idea about the content of a frame might be sufficient. Features generated from the use of information over a relatively large region or over whole images are referred to here as region-based features. For instance, Nelson and Polana [24] gathered a set of four features based on the computation of the normal flow, i.e. the flow component parallel to the gradient, over regions of interest. The features are the *mean flow magnitude divided by its standard deviation, the positive and negative curl and divergence estimates, the non-uniformity of flow direction* and the *directional difference statistics in four directions*. Once all those values are computed, they are put into vector form for classification. In another paper [26], the same authors use a similar vector representation in order to recognize periodic motions like walking and running. It is also based on the computation of the normal flow, from which statistics are gathered over a selected portion of the sequence of images.

Eigen images extracted from a set of graylevel images of an object provide enough information to directly represent new images of that object. Kirby et al. [16] used that method with mouth images. The eigen images are the eigenvectors of the ensemble averaged covariance matrix $C = \frac{1}{P}\sum_{j=1}^{P} u^{(j)}u^{(j)T}$, where $u^{(j)}$ is the vector formed from the concatenation of the columns of the j'th image and u^T is the transpose of u. Each image in a sequence can then be expressed as a linear combination of the eigen images.

Darrell and Pentland used a set of model views for hand gestures [7]. Their method automatically stores the appropriate number of views necessary to represent the object using correlation. If optical flow can be reliably extracted, flow correlation might be more appropriate as compared to correlation using plain graylevels. Martin and Shah [21] used a sequence of dense optical flow around the mouth of a speaker, which are then correlated for matching.

Using binarized images, Yamato et al. [33] extracted a mesh feature from each frame of a sequence. An image is divided into a grid, and the proportion of black to white pixels in each grid element is computed; the ordered set of ratios for an image is called the mesh feature. Petajan et al. [25], also with binarized images, created a codebook of mouth opening images, i.e. a set of images of the different shapes of the mouth.

2.4 Matching

Once the representation has been defined and the features from both models and unknown sequences properly organized, a comparison must be made so that classification or recognition can take place. Since several methods use a vector type of representation, the matching can be performed in a fairly efficient manner with clustering techniques, which are very common and well documented. Other matching methods will obviously depend on the type of representation, and will thus vary accordingly.

3 Motion Recognition

3.1 Cyclic Motion Detection

The presence of cyclic motion in a sequence of images can reveal a lot about the object performing that type of motion. A rigid object can perform a cyclic movement, for example a ball in a pendulum motion, while an articulated object can perform much more complex motions. Furthermore, different cyclic motions could occur concurrently with the same or with different frequencies and phase relative to each other. Based on studies of the human visual system, Allmen and Dyer [2] argue that cyclic motion detection (1) does not depend on prior recognition of the moving object; (2) does not depend on the absolute position of the object; (3) needs sequences containing at least two complete cycles; (4) is sensitive to different scales, i.e. cycles at different levels of a moving object can be detected. To study cyclic motion, the authors used curvature as a low level description of motion, and its scale-space as a representation. A modified version of a uniform cost algorithm was used for cycle detection. An advantage of curvature scale-space is that it is possible to detect cycles at different scales.

Polana and Nelson [27] first computed a *reference curve*, which is the approximate trajectory of a moving object's centroid, and aligned the frames with respect to it. If the object presented some periodic motion, for example a person walking, the motion of the arms and legs would create periodic graylevel signals around the centroid. The periodic motion is extracted from the graylevel signals using a Fourier transform and an overall periodicity measure is computed. In an upcoming paper [26], the same authors describe a method for recognizing different periodic motions.

Tsai et al. [31] use spatiotemporal curvature as a representation. Some preprocessing is performed,

in particular an autocorrelation to emphasized self-similarity within the curvature function. A Fourier transform is finally applied to that signal to detect the presence of cycles and their period.

Tsai et al. and Polana and Nelson used the Fourier transform to directly detect periodicity, which seems a reasonable and sensible thing to do. Furthermore this method is more robust to uncorrelated noise. Recognition using cyclic motion has been reported in [31, 26].

3.2 Lipreading

Lipreading is a very difficult task, especially since certain phonemes can appear visually identical (phonemes are minimal meaningful units of sound from which two words can be distinguished). For instance, the phonemes "b", "p" and "m" sound different but look the same when spoken [10].

In Petajan et al. [25], the lipreading task is performed with the help of a codebook. The codebook is a set of mouth images containing all possible aspects of the mouth's appearance. In particular, the codebook contains images of the mouth opening area. A spoken word consists of a sequence of codebook images. Words are compared by computing a distance value between an unknown word and the models.

Finn and Montgomery use a combination of distances between different points around the mouth [10]. Twelve dots were placed around the mouth of speaker and tracked during the experiments; a total of fourteen distances were measured, and used as a feature vector. Recognition consisted of computing a total rms value between two utterances.

Mase and Pentland [22] observed that the most important features that affect mouth shape relate to the elongation and the opening of the mouth, affecting upper and lower lips. Computing average optical flow in four windows around the mouth of a speaker, and performing a principal component analysis on the flow components, two functions, mouth opening $O(t)$ and elongation $E(t)$ were created. $O(t)$ and $E(t)$ are computed at each frame, and time warping is performed, to normalize the time to speak each word. A sampling of the functions for each model was used for matching with a similar sampling of an unknown.

Martin and Shah [21] use a sequence of dense optical flow fields around the mouth of a speaker, which are spatially warped, temporally warped, then correlated, for matching with a sequence of optical flow frames. Spatial warping is used to locate the window containing the lips of each model frame with each input frame, while temporal warping addresses the problem of sequences of different length. Each model optical flow frame is compared to each optical flow frame of the unknown sequence using correlation.

Kirby et al. [16] chose to express mouth images as a linear combination of the fixed set of the eigenvectors of the ensemble averaged covariance matrix C (section 2.3). A spoken word made up of a sequence of P images can then be expressed as a $Q \times P$ matrix of coefficients computed with respect to the set of Q eigen images. Identification of particular words in a sequence using spatial eigenfunctions was performed using a template matching technique.

Methods for lipreading should attempt to achieve the following goals. They should be able to perform in unconstrained or little constrained environments, they should achieve speaker independence, address continuous speech and ultimately perform in real-time. None of the methods described above meet all these goals. Finn and Montgomery [10] place markers around the mouth of a speaker; Martin and Shah's work [21] is computationally expensive; Mase and Pentland [22] address continuous speech but the difficult part, according to the authors, is to detect the beginning of the first word. Petajan et al. [25] have an interesting method, since when the codebook is complete, a distance table is computed which makes image comparison fast. Kirby et al. [16] provide a nice basis for lipreading, however more extensive work is required.

3.3 Gesture Interpretation

Humans have the capability or can develop the ability to interpret gestures, and gestural languages have been developed to allow hearing impaired people to communicate more easily. Two studies are described below.

Darrell and Pentland [7] present an automatic view-based approach to build the set of models from which gesture models will be created. Model views of an object are built using normalized correlation. The object is tracked, and when the correlation score drops below a threshold, a new model view is added. This process is repeated until no more views are necessary. Gesture models are then created. A gesture is a sequence of views over time. Each frame of a gesture is correlated with each stored view of the object, and its score plotted, for each view, with respect to time. Dynamic time warping is performed to adjust all gestures to the same length. An unknown gesture is similarly correlated and its score plotted, for all view models. The matching is then done by comparing the correlation scores between the unknown and the model gestures.

Davis and Shah [8] report a simple method for hand gesture recognition by tracking the trajectory followed by each finger and using their motion as a basis for recognition. The direction of motion and displacement

of each finger are stored, along with the name of the gesture, in a simple data structure. A motion code is then derived, which consists of a five-bit number, each bit associated with a finger. A bit is set if motion of its corresponding finger is detected, and is reset otherwise. Such model gestures are stored in an array of linked lists, and the motion code serves as an index for all gestures sharing it. For an unknown gesture, the direction and displacement of each finger is computed, and its motion code derived. The matching will be performed by comparing the unknown only to those models sharing the same motion code as the unknown.

In automatic interpretation of hand gestures, similar goals as lipreading are to be attained: independence from the person performing the gesture, address continuous gesturing, use unconstrained environment and perform in real-time. Both works above do perform in real-time; as for the rest, much work remains.

3.4 Motion Verb Recognition

Motion verb recognition deals with the association of natural language verbs with the motion performed by a moving object.

Koller et al. [17] devised a method which automatically characterizes the trajectory of moving vehicles in an intersection. The motion verbs were divided into four different categories. Attributes, computed from the sequence, help describe more precisely the trajectory segments, the position of a vehicle with respect to the street or other objects, its orientation, velocity, etc. In addition, predicates are used, whose truth value is determined at each time instant. Using this information, an interpretation is sought, along with the time period for which it is true.

The goal of Tsotsos' work [32] was to build an artificial intelligence system, called ALVEN, capable of using motion information to recognize normal and abnormal behavior of a heart's left ventricular motion. Natural language semantic components were developed to describe motion concepts using English motion verbs.

3.5 Temporal Textures Classification

In their paper, Nelson and Polana describe how the movement of the ripples on water, the wind in the leaves of trees, can be classified [24]. Those motions, referred to as temporal textures, show complex and non-rigid motions. The term temporal texture is used to emphasize that the motion patterns are of indeterminate spatial and temporal extent. Different statistical features based on optical flow fields (see section 2.3) were put into vector form, and were classified using a nearest centroid classifier.

4 Human Tracking and Recognition

This section will be concentrating on methods designed to recognize human motion. There are several ways to view this task. The first one is to recognize the action performed by a person in a scene, among a database of human action models, in a way similar to that of the previous section. The second way is to recognize the different body parts like arms, legs, head, throughout a sequence. This approach is referred to here as labeling. The third way is to define motion as a sequence of the specification of the parameters defining the configuration of an object in time, so that recognition amounts to the determination of the most plausible configuration in time. This approach has been used mostly with humans, and is called here tracking of human motion. The modeling of the human shape and of human motion plays an important role, especially in tracking.

4.1 Modeling of the Human Body

To properly study human motion, good body models must be defined, and several have been developed. The stick figure model consists of segments, usually connected at their endpoints and representing the body. This model can be seen as a skeleton (in the computer vision sense) and can be as detailed as necessary. Volumetric models are intended to better represent the complexity of the human body. Generalized cylinders, i.e. cylinders with an elliptical cross-section of constant size and shape, are most commonly used. The model can be as refined as necessary, by using a collection of component cylinders representing the different body segments, giving more detailed information about the spatial organization of the human shape.

4.2 Modeling of Human Motion

Human motion can be modeled using joint angles. Joint angles have been extensively studied in physical medicine [23]. They are more formally expressed as flexion/extension, abduction/adduction and rotation angles. Studies have also shown that the forward motion is almost constant within a walking cycle, while the vertical displacement of the head is relatively small considering the global motion. Normal locomotion is characterized by a smooth forward translation of the trunk and rhythmicity in the length of successive steps as well as in the duration of successive temporal components of the walking cycle.

In computer vision, joint angles plotted in time (joint curves) for one walking cycle have been used as a walking motion model for humans. They provide

sufficient information for the determination of the posture of a person, i.e. the relative position of each body segment, throughout the cycle. Other type of knowledge can also be extracted to provide additional information, for example constraints on possible angles for each joint, along with constraints on angular velocity. This kind of information can reduce the search space during tracking by constraining the possible angle variation between frames.

Another kind of approach for modeling motion is to use a sequence of stick figures, called key frame sequence, to model rough movements of the body. This key frame sequence consists of an ordered set of stick figures, each differing from its predecessor and successor, for instance, when a body segment has crossed or uncrossed another body segment.

4.3 Recognizing Body Parts

The main goals of the methods described here is to track and label each part of a body performing some action. The tracking consists of determining the location and shape of body parts from frame to frame, while labeling involves identifying them.

In Akita's work [1], the recognition of the parts is done in the following order: legs, head, arms and trunk. Correspondence between frames is established using one of two methods: when the position change of a segment is small enough, its position can be predicted from the previous frame using window code distances, which are defined in the paper. If window codes cannot find a correspondence, then a key frame sequence is used to find the current posture.

Leung and Yang [19] also tackle the problem of body labeling in a sequence. The labeling is made up of two steps. The region description process abstracts the segmented image (see [20]) to extract the antiparallel lines (*apars*) that will be used for labeling. The most appropriate apars are then chosen for body part identification according to heuristics related to width ratios and likelihood.

Akita's work, to our opinion, is simpler yet more complete than Leung and Yang's, which is more dependent on predefined thresholds. However Akita's work needs more a priori information: the motion must explicitly be described with the key frame sequence.

4.4 Three-Dimensional Tracking

Three-dimensional tracking (3D tracking) consists of determining the 3D position of a body, along with its posture, i.e. the relative position of its parts, from the frame by frame analysis of a sequence. The human body is modeled using stick figures or generalized cones. The analysis of a frame provides the information necessary to update the posture and position of the body model in space. The validity of the updated model is verified by comparing its projection on the image plane with the edges extracted from the sequence.

Tracking with Stick-Figure Models. The work done by Chen and Lee [5] is divided in two parts. The first consists of finding all *allowed* 3D configurations of the body for each frame. This process is described in [18], where the possible location of all joints is determined in space, and knowledge of physical and motion constraints is used to eliminate invalid configurations. The second part is to find the sequence of configurations that would best represent the walking motion. The normal movement during walking is a smooth and continuous motion. In this study, this movement is considered as a collection of smooth and continuous angular motions of all body segments, expressed as a nearly constant angular velocity, or equivalently, a close to null angular acceleration. An overall angular acceleration function over all frames was defined such that minimizing the function results in finding the sequence of configurations leading to a smooth motion.

Tracking with Volumetric Models. Tracking with volumetric models is more complex because of the larger number of parameters required to represent the model itself. Two studies will be reported below. The body model, for both, consists of a cylinder for each hand, arm, forearm, foot, leg, thigh, trunk and head, and both use joint angle curves for the same joints, as motion model. The posture is parametrized by a value ranging from 0 to 1, i.e. from this value, all joint angles for that posture can be found. In both, this parameter is determined and plotted as a function of the frame number, along with the body position with respect to the world coordinates.

In Hogg's work [14], a frame by frame analysis provides an estimate of the person's 3D position and its posture. Important parameters are the posture parameter $PSTR$, speed and direction of motion. The purpose is thus to find the sequence of assignments of the parameters that satisfies the constraints that defines the walking motion. The tracking is done through a function called TRACK. TRACK uses a function called SEARCH that seeks the optimal parameter assignments for the current image, which are constrained by the parameters chosen for the previous frame along with the model constraints. The search space for each parameter is partitioned into a set of closed intervals. An evaluation function will be given

representatives of those intervals, and the results used as part of a plausibility function. The plausibility is a weighted function of the plausibility of each part of the model, and is computed with the help of the projection of the model contours and the actual image edge points. The process is then repeated for each frame.

Rohr's method comprises two phases [28]. The first phase, called the initialization phase, provides an estimate for the posture parameter *pose* and 3D position of the body using a linear regression method. The second phase, using the estimate of the first phase, uses a Kalman filter approach to incrementally estimate the model parameters. The 3D position and *pose* estimates are determined by matching projected model contours with the image edge, and serve as a current measurement vector for the filter. An overall similarity between the model edges and the graylevel edges is computed. The parameters are chosen such that the overall similarity is maximized. Knowledge of *pose* and its time derivatives reduces the search space. The state vector for the Kalman filter is $p_k = (X_k, X'_k, Y_k, Y'_k, Z_k, Z'_k, pose_k, pose'_k)^T$, where (X_k, Y_k, Z_k) is the 3D position in frame k, x' is the first time derivative. At each time, the 3D position and *pose* parameter are fed to the Kalman filter which will then provide an estimate for the new position and *pose* in the next frame.

The two approaches described above seem very similar. Both use the same type of 3D model, $PSTR$ and *pose* carry the same information, they both use joint angle curves for the same joints of the body. They however differ in several ways. In particular, Rohr uses medical data as a basis for the joint angle curves, while Hogg uses data from only one person. The Kalman filter also provides more robust and smoother results.

4.5 Human Motion Recognition

Human motion presents a special challenge because of the amount of possible configurations of the body. The different motions of the parts need to be determined with respect to each other. The scope of Goddard's thesis [11] is the recognition of 400ms moving light displays generated from actors using a connectionist approach. Given the set of trajectories of points in a sequence, line segments are extracted. The goal is to combine line segments together to form legs or arms, then to combine pairs of arms and pairs of legs to form upper and lower body limbs, and finally to combine upper and lower limbs for the description of the complete motion. However, not only do the arms and legs pairs need to be properly linked in space, their relative motion in time must also correspond to proper body motion. To achieve this level of complexity, a hierarchical system is described, that combines segments into components, components into assemblies, and so on, up to scenarios. Scenarios represent temporal series of events, with information on sequence and duration. Goddard demonstrated that the changes in angular velocities are sufficient for recognition.

Discrimination between different tennis strokes was investigated by Yamato et al. [33] using Hidden Markov Models (HMM). They can be seen simply as symbol generating machines. An image sequence is processed in three steps. In the first, a mesh feature is extracted, then associated to a symbol by a clustering technique. From this process, a sequence of output symbols is derived, one symbol for each frame of the sequence. In the second step, sequences are used to train the HMMs. There are as many HMMs as there are motions to be recognized. During this phase, the parameters describing an HMM are optimized such that it has a high probability of generating the sequence of output symbols derived from a particular motion. Finally, given a sequence of output symbols from an unknown motion, we want to find the HMM that is most likely to generate the same sequence as the unknown. The likelihood is computed using a probabilistic approach.

There are several advantages to that technique, namely its probabilistic nature, along with its versatility, since it can be readily generalized to probably any type of motion. Goddard's work is very impressive, but the connectionist part of the work seems complex.

5 Conclusion and Future Directions

Motion-based recognition consists of the recognition of objects or motions directly from motion information extracted from the sequence of images. Knowledge about the object or motion is used to construct models that will serve in the recognition process.

There are still several problems to be addressed. In the case of multiple moving objects in a scene, proper segmentation remains a difficult task. This is why experiments are usually performed in constrained environments or with special apparatus. If motion-based recognition methods are to be used more widely, feature extraction will have to be performed in noisy environments and without any particular enhancements.

Perceptual organization of trajectories or spatiotemporal curves is an emerging theme. It has been shown that spatial invariants exist, which permit us to infer 3D information from a 2D image projection. Those types of invariants applied to motion could be very insightful. For example, 2D elliptical trajectories imply a rotation motion in 3D; a set of elliptical trajectories with parallel major and minor axes corresponds

to the motion of points of a single rotating object in 3D. The determination of those types of motion invariants that are reliable and stable provides a new avenue for this type of research. Similarly, the clustering of spatiotemporal flow curves can provide a representation for coherent motions like a translation or rotation [3]. A hierarchical clustering of these curves can lead to the detection of different objects, their particular motion, their occlusion/disocclusion, and even relative and common motion could be inferred. Dynamic perceptual organization can be a very useful research direction that could lead to interesting approaches and results.

A significant part of future research will remain application oriented. Applications will furthermore preferably run in real-time, and hardware solutions will be necessary.

Acknowledgement: The research reported in this paper was supported by NSF grants CDA-9122006 and IRI-922076. An extended version of this paper under more appropriate title *Motion Based Recognition: A Survey* is available by anonymous ftp from eustis.cs.ucf.edu (132.170.108.42) under /pub/tech-paper/survey.ps.Z.

References

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

[2] M. C. Allmen and C. R. Dyer. Cyclic Motion Detection Using Spatiotemporal Surfaces and Curves. In *ICPR-90*, pages 365–370.

[3] M. C. Allmen and C. R. Dyer. Computing Spatiotemporal Relations for Dynamic Perceptual Organization. *CVGIP:IU*, 58(3):338–351, 1993.

[4] C. D. Barclay, J. E. Cutting, and L. T. Kozlowski. Temporal and Spatial Factors in Gait Perception that Influence Gender Recognition. *Perception and Psychophysics*, 23(2):145–152, 1978.

[5] Z. Chen and H.-J. Lee. Knowledge-Guided Visual Perception of 3-D Human Gait from a Single Image Sequence. *SMC*, 22(2):336–342, 1992.

[6] J. E. Cutting and D. R. Proffitt. The Minimum Principle and the Perception of Absolute, Common, and Relative Motions. *Cognitive Psychology*, 14:211–246, 1982.

[7] T. J. Darrell and A. P. Pentland. Recognition of Space-Time Gestures Using a Distributed Representation. Technical Report TR-197, M.I.T. Media Laboratory Vision and Modeling Group, 1992.

[8] J. W. Davis and M. Shah. Gesture Recognition. *ECCV-94*.

[9] S. A. Engel and J. M. Rubin. Detecting Visual Motion Boundaries. In *Proc. Workshop on Motion*, 107–111, 1986.

[10] K. E. Finn and A. A. Montgomery. Automatic Optically-Based Recognition of Speech. *PRL*, 8:159–164, 1988.

[11] N. H. Goddard. *The Perception of Articulated Motion: Recognizing Moving Light Displays*. PhD thesis, University of Rochester, 1992.

[12] K. Gould and M. A. Shah. The Trajectory Primal Sketch: a Multi-Scale Scheme for Representing Motion Characteristics. In *CVPR*, 1989.

[13] D. D. Hoffman and B. E. Flinchbaugh. The Interpretation of Biological Motion. *Biological Cybernetics*, 42:195–204, 1982.

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

[15] G. Johansson. Visual Perception of Biological Motion and a Model for its Analysis. *Perception and Psychophysics*, 14(2):210–211, 1973.

[16] M. Kirby, F. Weisser, and G. Dangelmayr. A Model Problem in the Representation of Digital Image Sequences. *PR*, 26(1):63–73, 1993.

[17] D. Koller, N. Heinze, and H.-H. Nagel. Algorithmic Characterization of Vehicle Trajectories from Image Sequences by Motion Verbs. In *CVPR*, pages 90–95, 1991. Extended version.

[18] H.-J. Lee and Z. Chen. Determination of 3D Human Body Postures from a Single View. *CVGIP*, 30:148–168, 1985.

[19] M. K. Leung and Y.-H. Yang. A Region Based Approach for Human Body Motion Analysis. *PR*, 20(3):321–329, 1987.

[20] M. K. Leung and Y.-H. Yang. Human Body Motion Segmentation in a Complex Scene. *PR*, 20(3):55–64, 1987.

[21] G. A. Martin and M. Shah. Lipreading Using Optical Flow. In *Proc. Nat. Conf. on Undergraduate Research*, 1992.

[22] K. Mase and A. Pentland. Lip Reading: Automatic Visual Recognition of Spoken Words. Technical Report 117, M.I.T. Media Lab Vision Science, 1989.

[23] M. P. Murray. Gait as a Total Pattern of Movement. *American Journal of Physical Medicine*, 46(1):290–333, 1967.

[24] R. C. Nelson and R. Polana. Qualitative Recognition of Motion Using Temporal Texture. *CVGIP:IU*, 56(1):78–89, 1992.

[25] E. D. Petajan, B. Bischoff, D. Bodoff, and N. M. Brooke. An Improved Automatic Lipreading System to Enhance Speech Recognition. In *SIGCHI '88: Human Factors in Computing Systems*, pages 19–25, 1988.

[26] R. Polana and R. C. Nelson. Recognizing Activities. Submitted to CVPR 1994.

[27] R. Polana and R. C. Nelson. Detecting Activities. In *CVPR*, pages 2–7, 1993.

[28] K. Rohr. Towards Model-Based Recognition of Human Movements in Image Sequences. *CVGIP: IU*, 59(1):94–115, 1994.

[29] H. Shariat and K. Price. How to Use More than Two Frames to Estimate Motion. In *Proc. Workshop on Motion*, pages 119–124, 1986.

[30] M. Subbarao. Interpretation of Image Motion Fields: A Spatiotemporal Approach. In *Proc. Workshop on Motion*, pages 157–165, 1986.

[31] P.-S. Tsai, M. Shah, K. Keiter, and T. Kasparis. Cyclic Motion Detection. Technical Report CS-TR-93-08, C.S. Dept., Univ. of Central Florida, 1993.

[32] J. K. Tsotsos, J. Mylopoulos, H. D. Covvey, and S. W. Zucker. A Framework for Visual Motion Understanding. *PAMI*, 2(6):563–573, 1980.

[33] J. Yamato, J. Ohya, and K. Ishii. Recognizing Human Action in Time-Sequential Images Using Hidden Mardov Model. In *CVPR*, pages 379–385, 1992.