UMENTATION PAGE

Form Approved
OMB No. 0704-0188

2

ion is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, ileting and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this ducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

2. REPORT DATE

3. REPORT TYPE AND DATES COVERED
FINAL/15 JUN 91 TO 14 OCT 94

4. TITLE AND SUBTITLE
DETECTION, STABILIZATION, AND IDENTIFICATION
OF MOVING OBJECTS BY A MOVING OBSERVER (U)

5. FUNDING NUMBERS
2304/A7
AFOSR-91-0288

6. AUTHOR(S)
Professor Randal Nelson

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Computer Science Department
Univ of Rochester
734 Computer Studies Building
Rochester, NY 14627

8. PERFORMING ORGANIZATION REPORT NUMBER
AEOSR-TR- 94 0250

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
110 DUNCAN AVE, SUITE B115
BOLLING AFB DC 20332-0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFOSR-91-0288

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE: DISTRIBUTION IS UNLIMITED

12b. DISTRIBUTION CODE
UL

13.

The stated goal of the research was to demonstrate that robustly computable motion features can be used directly as a means of detecting and recognizing moving objects. Specifically, the goal was to design, implement, and test a general framework for detecting movement from a moving platform, and recognizing both distributed motion activity on the basis of temporal texture, and complexly moving, compact objects on the basis of their action. This recognition approach contrasts with the reconstructive approach that has typified most prior work on motion. The underlying motivation is the observation that, for objects that typically move, it is frequently easier to identify them when they are moving than when they are stationary. Specifically, in the case of temporal texture, the researchers proposed to extract statistical spatial and temporal features from approximations to the motion field and use techiques analogous to those developed for grayscale texture analysis to classify regional activities such as windblown trees, ripples on water, or chaotic fluid flow, that are characterized by complex, non-rigid motion. For action identification, they proposed to use the spatial and temporal arrangement of motion features in conjunction with simple geometric image analysis to identify complexly moving objects such as machinery and locomoting people and animals. The proposed work has practical applications in monitoring and surveillance, and as a component of a sophisticated visual system.

14. SUBJECT TERMS

15. NUMBER OF PAGES
35

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR(SAME AS REPORT)

Final Technical Report
AFOSR Grant Number 91-0288
Detection, Stabilization, and Identification of Moving Objects by a Moving Observer

Randal C. Nelson
Department of Computer Science
University of Rochester
Rochester, New York 14627

Project Summary

The stated goal of the research was to demonstrate that robustly computable motion features can be used directly as a means of detecting and recognizing moving objects. Specifically, the goal was to design, implement, and test a general framework for detecting movement from a moving platform, and recognizing both distributed motion activity on the basis of *temporal texture*, and complexly moving, compact objects on the basis of their *action*. This *recognition* approach contrasts with the *reconstructive* approach that has typified most prior work on motion. The underlying motivation is the observation that, for objects that typically move, it is frequently easier to identify them when they are moving than when they are stationary. Specifically, in the case of temporal texture, we proposed to extract statistical spatial and temporal features from approximations to the motion field and use techniques analogous to those developed for gray-scale te:.ture analysis to classify regional activities such as windblown trees, ripples on water, or chaotic fluid flow, that are characterized by complex, non-rigid motion. For action identification, we proposed to use the spatial and temporal arrangement of motion features in conjunction with simple geometric image analysis to identify complexly moving objects such as machinery and locomoting people and animals. The proposed work has practical applications in monitoring and surveillance, and as a component of a sophisticated visual system.

By and large, the goal of the project were accomplished. A number of papers describing the work have appeared in technical journals and conferences, and prototype code implementing the algorithms as well as test data, is available by request. A detailed technical description of the work is contained in three papers that are attached to this report.

The first phase of the project addressed the classification of temporal textures via statistical characteristics of the associated motion fields. We developed a group of statistical measures involving first and second order characteristics of the motion field. These measures included differential quantities such as curl and divergence, and spatial statistics such as directional co-occurrence features. When incorporated into simple nearest-neighbor classifiers, these measure proved successful in distinguishing a number of natural temporal textures. Principle component analysis, carried out in the motion feature space was used to evaluate the relative effectiveness of the various measures. This work is described in the paper "Qualitative Recognition of Motion Using Temporal Texture" attached to this report.

The second phase of the work involved the detection, isolation, and tracking of periodically moving objects. This group includes objects such as walking and running people, running, flying, or swimming animals, and some sorts of machinery. To human observers, many of these objects can be more readily identified by their motion signatures than by their shape - particularly in low-resolution or high-clutter regimes. Identification of objects in this group is also important in many practical applications. We developed a technique based on the Fourier transform that allowed us to flag and isolate periodically moving objects in real scenes. The method is general, and applies to a wide variety of situations, including those with an actor is translating against a varying background, which cannot be characterized by a simple cyclical image. This work is described in the attached paper "Detecting Activities".

The final phase of the work involved the identification of periodic activities once they had been isolated in an image, e.g. whether the motion is produced by a walking or a running person, or something else entirely. Previous approaches to this problem have relied upon analysis of joint trajectories, often obtained by attaching lights to the limbs of an actor. The problem with this approach is that it is not clear how to obtain the required trajectories from a raw image sequence - the joints must be identified and tracked. It also is hard to generalize to other motions. We developed a method for classifying periodic movement based on low-level motion features. Basically, the detection and isolation procedures developed in the previous phases of the research allowed us to define a canonical form for arbitrary periodically moving objects. With the data in

this normalized form, a representation consisting of a spatiotemporal template of local motion features could be effectively used to classify a wide variety of moving objects. Since no prior models of the objects are required, the technique is more general than those based on joint trajectories. The method was demonstrated on a database of real-world image sequences containing a variety of movements including running and walking people, people on swings, and mechanical animals. The technique seems to have sufficient resolution to distinguish, for example, walking from running, as well as from less similar motions, across multiple actors. It does not have the resolution to reliably distinguish individuals on the basis of their gait. This work is described in the attached paper "Recognizing Activities".

Publications

Publications in pear reviewed journals and referenced book chapters resulting from grant sponsored research.

Ramprasad Polana and Randal C. Nelson, 1992, "Recognition of Motion from Temporal Texture", *Proc. IEEE Conference on Computer Vision and Pattern Recognition,* Champaign, Illinois, June 1992, 129-134.

Randal C. Nelson and Ramprasad Polana, 1992, "Qualitative Recognition of Motion Using Temporal Texture", *CVGIP Image Understanding,* Vol 56, 1, July 1992, 78-89. A short version appears in Proc. DARPA Image Understanding Workshop, San Diego, CA, Jan 1992.

Nelson, Randal C., Finding line segments by stick growing, to appear, IEEE Trans. PAMI.

Polana, Ramprasad and Nelson, Randal C., Detecting activities, Proc., Computer Vision and Pattern Recognition, 2-7, New York, NY, June 1993.

Polana, Ramprasad, and Nelson, Randal C. Detection of Activities, Journal of Visual Computing and Image Representation, To appear.

Polana, Ramprasad, and Nelson, Randal C., Recognizing activities, Under review.

Sponsored Personnel

Total researchers working with the Principal Investigator on this project

Faculty: Randal C. Nelson

Postdocs: none

Graduate Students: Ramprasad Polana

Other: none

CVGIP: IMAGE UNDERSTANDING
Vol. 56, No. 1, July, pp. 78–89, 1992

Qualitative Recognition of Motion Using Temporal Texture

RANDAL C. NELSON AND RAMPRASAD POLANA

Department of Computer Science, University of Rochester, Rochester, New York 14627

Received November 8, 1991; accepted March 18, 1992

We describe a method of visual motion recognition applicable to a range of naturally occurring motions that are characterized by spatial and temporal uniformity. The underlying motivation is the observation that, for objects that typically move, it is frequently easier to identify them when they are moving than when they are stationary. Specifically, we show that certain statistical spatial and temporal features that can be derived from approximations to the motion field have invariant properties, and can be used to classify regional activities such as windblown trees, ripples on water, or chaotic fluid flow, that are characterized by complex, nonrigid motion. We refer to the technique as *temporal texture analysis* in analogy to the techniques developed to classify gray-scale textures. This *recognition* approach contrasts with the *reconstructive* approach that has typified most prior work on motion. We demonstrate the technique on a number of real-world image sequences containing complex movement. The work has practical application in monitoring and surveillance, and as a component of a sophisticated visual system. © 1992 Academic Press, Inc.

1. INTRODUCTION

Who has not watched ripples spread across a pool and known water thereby? Or seen leaves shimmer their silver backs in a summer breeze and known a tree? Who has not known the butterfly by her fluttering? Or seen a distant figure walking and known there goes a man? In order to successfully interact with a dynamic world, an agent must interpret the activity around it. In the vision system, this requires the interpretation of visual motion. The everyday experience of visual motion incorporates a considerable element of recognition; this may even be its dominant attribute. Yet surprisingly, this aspect of motion has been neglected in the literature on computational motion analysis, which has emphasized instead, a reconstructive approach. We show here that robustly computable motion features can be used directly as a means of recognition. In particular, we argue that there exists a class of image motions, common in scenes of the natural environment, that are characterized by structural or statistical self-similarity in space and time. Typical examples might include ripples on a pool, a flock of birds, windblown grass or trees, and turbulent weather patterns in the atmosphere. Such motions, referred to as *temporal textures*, can be efficiently identified using statistical pattern recognition techniques based on invariant features of the motion field.

Visual motion has, of course, long been considered an important source of information in natural vision systems. Many of the (comparatively) unsophisticated systems, such as those possessed by insects and lower vertebrates, are essentially blind to anything that is not moving. Even in the more sophisticated systems possessed by higher vertebrates, including man, motion in the visual field retains an important role. Moving objects in a scene are typically the first attended to, and a wide variety of (semi)quantitative information relating to object segmentation, depth, three dimensional shape, and object and observer motion, seems to be derived from the visual motion field.

The potential wealth of derivable information inspired a large body of work on the computation of *exact* geometric quantities such as the 3-D shape of objects, their location, and the motion of the observer. This reconstruction problem is sometimes referred to as the *structure-from-motion problem*. Research has been typically divided into two main areas: finding 3-D information from 2-D projected motion assuming it is available, and determining projected motion from raw image sequences. Results have been obtained in both areas; however, the high-level shapes from motion algorithms tend to be very sensitive to the accuracy of the underlying motion information, and the accuracy of the computed motion information has typically been low. Consequently, only moderate success has been achieved in this area.

The emphasis on visual motion as a means of quantitative reconstruction of world geometry has tended to obscure the fact that motion can also be used for recognition. In fact, in biological systems, the use of motion information for recognition is often more evident than its use in reconstruction. A simple example occurs in the case of the common toad *Bufo bufo* for which any elongated object within a certain size range that exhibits motion along the long axis is identified as a potential food item, and elicits an orienting response [Ewar87]. Birds

1049-9660/92 $5.00

ignore the natural movement of trees in the wind, but respond immediately to the approach of a predator. More generally, stylized movements seem to be a universal form of communication between animals with eyes, from the aggressive posturing of various fiddler crabs (*Uca* species), to the mating dance of the blue footed booby (*Sula nebouxi*), to the expressive facial movements of baboons.

Humans have a remarkable ability to recognize different kinds of motion, both of discrete objects, such as animals or people, and in distributed patterns as in windblown leaves, or waves on a pond. A classic illustration of motion recognition by humans is provided by Moving Light Display experiments where the sole source of information about a moving actor is provided by lighted points attached to a few joints [Joha73]. People shown these images dismiss single frames as meaningless dot patterns but can recognize characteristic gaits such as running or walking, and even gender and familiar individuals from the sequential presentation.

Such abilities suggest that, in the case of machine vision, it might be possible to use motion as a means of recognition directly, rather than indirectly through a geometric reconstruction. In addition to the biological motivations, there are computational reasons for considering motion as a recognition modality. One advantage is that the motion field, insomuch as it can be extracted at all, is robust with respect to lighting changes, and much more simply related to shape than is image luminance. Furthermore, if the task is to find an object that is known to be moving, motion can be used to efficiently presegment the scene into regions of high and low interest. This can frequently be done even if the observer is itself moving [Nels90].

Recognition can thus be viewed as an alternative, more qualitative approach to utilizing visual motion. Structure-from-motion can be viewed as a general transformation of information in one form (time varying images) into a (presumably) more useful form (e.g., depth maps). Recognition, on the other hand, serves to identify a specific situation of interest to the system, for instance, the approach of a fly if you are a frog, or a bird if you are a fly. A reconstructed world model contains a lot of information, possibly enough to find a fly if you are a frog, but it also contains a lot of information that a frog has no interest in, and that was expensive to obtain.

The above illustrates a central point of the active/behavioral approach to vision, namely, that in any practical system, both the information extracted and its representation must take into account the function of the system. The primary reason is that the total quantity of information contained in a visual signal is far greater than any system needs or can handle. In most proposed applications of vision, all but a tiny fraction of this information is irrelevant. The fundamental problem of vision is determining what image information can be used and extracting an efficient representation for it. Despite this fact, the goal of machine vision has often been portrayed as the problem of devising information transforms that preserve as much of the original information as possible, albeit in a purportedly more convenient form, on the grounds that one never knows what one might need, and that information once thrown out cannot be recovered. Reconstructionist approaches in which the goal is to determine "intrinsic images" representing for example the distance, surface normal, relative velocity, reflectivity, and illumination for every point in the image, are of this sort. We believe that such a least commitment strategy is exactly the wrong approach. Assuming ignorance about a situation in which considerable structure exists is generally poor policy, and in the case of vision, it can be disastrous. The strategy that should be followed is to throw out as much information as quickly as possible on the grounds that what is thrown out does not have to be processed and does not tie up limited computational resources. This might be termed a strategy of most commitment. The behavioral approach provides a mechanism for deciding what can be thrown out via the use of a prior knowledge about the functionality of the system. Knowing exactly what information is needed and what it will be used for also permits the system to alter its interaction with the world dynamically, in order to make that information more readily obtainable.

We define qualitative vision as the computation of iconic image properties (qualities) having a stable relationship to functional primitives. These functional relations are the building blocks for visual behavior. The iconic nature of qualitative primitives provides the necessary information reduction; only a minute fraction of the original information is present, but it is directly relevant to the task at hand. Recognition is a qualitative statement in this sense. It classifies a situation in terms relevant to some functionality. In the case of motion, there are a variety of applications in which robustly computable motion information can be used for identification directly, and much more efficiently, than via traditional 3-D reconstruction.

An example of a directly useful motion feature is the regional divergence of the motion field, which can be used to detect approaching objects. In houseflies (*Musca domestica*) divergent flow activates a landing reflex when approaching a surface. We have implemented a collision avoidance system based on the divergence cue [Nels89]. The basic idea is that a region on a collision course with the observer will be expanding and thus display positive divergence. We utilized a set of features termed *directional divergences* $D_\phi \mathbf{f}$ parameterized by a polar angle ϕ and given by

$$D_\phi \mathbf{f} = \frac{\partial f_\phi}{\partial r_\phi} = \cos^2\phi \frac{\partial f_x}{\partial x} + \sin^2\phi \frac{\partial f_y}{\partial y} + \sin\phi\cos\phi\left[\frac{\partial f_x}{\partial y} + \frac{\partial f_y}{\partial x}\right],$$

where f_ϕ is the component of the motion field $\mathbf{f}$ in the ϕ direction. These are equivalent to 1-D divergences along various axes, and can be robustly computed from projected flow information, which is easier to obtain than the full motion field. The system was used to guide a robot vehicle between obstacles. Figure 1 shows the divergence produced by a pair of obstacles toward which the vehicle is moving.

The divergence is a simple example of a temporal texture, that is, a regional property that identifies an area as a certain sort of "stuff" (here stuff that might collide with you), somewhat as gray-level textures can identify regions in a static image. Examples of more complex temporal textures, which would require a combination of several motion features for classification, include the fluttering of leaves on a tree; the glitter of sunlight from distant water and wave motion in nearby water; the motion within a flock of birds, on top of an anthill, or in a crowd at a football game; the effect produced by moving near a fractal object such as a bush; a snowfall, a waterfall; the turbulent curl of smoke; and the swirl of clouds in a weather system.

There are a number of potential applications for motion recognition. One area in which it would be useful is in automated surveillance. Motion detection via image differencing can be used for intruder detection; however, such systems are subject to false alarms, especially in outdoor environments, since the system is triggered by anything that moves, whether it be a human, a dog, or a tree blown by the wind. Motion recognition techniques, both of the discrete and textural variety have the potential to disambiguate the motions of different origin. Another application is in industrial monitoring. Many manufacturing operations involve a long sequence of simple operations, each performed repeatedly and at high speed by a specialized mechanism at a particular location. It should be possible to set up one or more fixed cameras that cover the area of interest, and to characterize the allowed motions in each region of the image(s). Abnormal activity would violate the prior constraints and allow the location of a problem to be identified quickly. This sort of analysis would be particularly valuable for the fast detection and neutralization of catastrophic failures that traditional quality control systems might not identify in time to prevent major damage. A similar situation arises in security surveillance of a compound, where certain types of motion may be expected in certain areas and in certain situations (e.g., the opening of a gate after an approval signal has been sent) but not in others (e.g., a man climbing over a wall). Other possibilities include monitoring satellite imagery for developing storm systems and crowds for incipient disturbances. General motion recognition techniques could also be applied to areas such as gesture recognition [Rhyn86] and handwriting analysis.

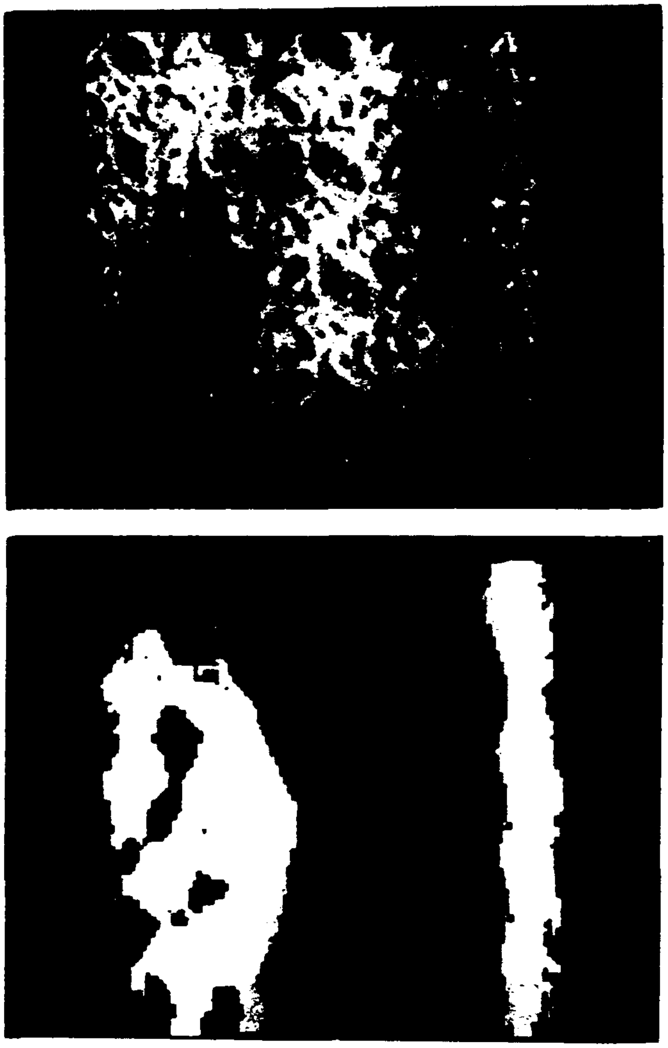

FIG. 1. Obstacle detection via flow field divergence.

2. BACKGROUND AND RELATED WORK

Motion recognition in general has received relatively little attention in the literature. Most computational motion work, as mentioned previously, has been concerned with various aspects of the structure-from-motion problem. There is a large body of psychophysical literature addressing the perception of motion, most of it concerned with primitive percepts. A modest amount of this work addresses more complicated motion recognition issues [Joha73, Cutt81, Hoff82, Hild87], but the models and descriptions have typically not been implemented. Various computational models of temporal structure have been proposed (e.g., [Chun86, Feld88]) but much of this work is at a fairly high level of abstraction and has not actually been applied to visual motion recognition except in rather artificial tests. Some of the best work in temporal pattern recognition has actually been done in the context of speech processing [Juan85, Tank87, Elma88].

A few studies have considered highly specific aspects of motion recognition computationally. Pentland [Pent89] considered lip reading, and implemented a system that could recognize spoken digits with 70–90% accuracy over five speakers. The system required the location of the lips to be entered by hand, and depended on an explicitly constructed lip model. Rashid [Rash80, Godd89] considered the computational interpretation of moving light displays, particularly in the context of gait determination. This work emphasized rather high-level symbolic models of temporal sequences, an approach made possible by the discrete nature of the moving light displays. The results were quite sensitive to discrete errors and thus highly dependent on the ability to solve the correspondence problem and accurately track joint and limb positions. This severely limits the general applicability of the method. Anderson *et al.* [Ande85] describe a method of change detection for surveillance applications based on the spectral energy in a temporal difference image. This has the flavor of the temporal texture analysis described here, but was not generalized to other motion features or more sophisticated recognition.

Some of the work done in the context of the structure from motion problem, particularly the methods that have been developed to obtain local motion information from image sequences, is relevant to temporal texture. Although we make somewhat different use of the information, this work has motivated and provided foundations for our approach, and it is thus appropriate to review the field.

A camera moving within a three-dimensional environment produces a time-varying image that can be characterized at any time t by a two-dimensional vector-valued function f known as the *motion field*. The motion field describes the two-dimensional projection of the three-dimensional motion of scene points relative to the camera. Mathematically, the motion field is defined as follows. For any point (x, y) in the image, there corresponds at time t a three-dimensional scene point (x', y', z') whose projection it is. At time $t + \Delta t$, the world point (x', y', z') projects to the image point $(x + \Delta x, y + \Delta y)$. The flow field at (x, y) at time t is given by

$$f(x, y, t) = \lim_{\Delta t \to 0} \left[\frac{\Delta x}{\Delta t}, \frac{\Delta y}{\Delta t} \right].$$

The motion field depends on the motion of the camera, the three-dimensional structure of the environment, and the three-dimensional motion (if any) of objects in the environment. If all these components are known, then it is relatively straightforward to calculate the motion field. In the traditional approach to motion analysis, the question has been whether the process can be inverted to obtain information about camera motion and structure of the environment. This is the basis of the structure-from-motion problem. The solution is not easy, and if arbitrary shapes and motions are permitted in the environment, there may not be a unique solution. However, it can be mathematically demonstrated that, in many situations, a unique solution exists.

The existence of such solutions has inspired a large body of work on the mathematical theory of extracting shape and/or motion information from the motion field. There have been two basic approaches to the problem. The first utilizes point correspondences in one or more images, generally under the assumption of environmental rigidity [Ullm79, Tsai81]. This is equivalent to knowing the motion field at isolated points of the image. Several authors have obtained closed form solutions to the shape from motion problem in this formulation, obtaining a set of linearized equations (Long81, Tsai84]. The second approach uses information about the flow and its derivatives in a local neighborhood under some assumption about the structure of environmental surfaces (e.g., they are planar) [Praz81, Boll87, Waxm87]. In this case, the end result is a set of equations relating the flow field derivatives to the camera motion and the three-dimensional structure of the environment. Most of these studies, however, have started with the assumption that detailed and accurate information, either in the form of point correspondences or dense motion fields, is available. Unfortunately, the solutions to the equations are frequently inordinately sensitive to small errors in the motion field. In the case of point correspondences, Tsai and Huang [Tsai84] report 60% error for a 1% perturbation in input for some instances using their method. This error sensitivity is due both to inherent ambiguities in the motion fields produced by certain camera motions, at least over restricted fields of view, and (in the second approach) to the reliance on differentiation of the flow field, which amplifies the effect of any error present in the data.

The two approaches to obtaining shape from motion utilize somewhat different methods for extracting motion information from image sequences. The methods using point correspondences rely on matching techniques similar to those employed in stereo vision [Moro79, Marr79, Barn80]. This process is well known to be difficult since features may change from one image to the next, and even appear and disappear completely.

Techniques for computing dense motion fields have relied heavily on differential methods, which attempt to determine the motion field from local computations of the spatial and temporal derivatives of the gray-scale image. The first derivative methods originally proposed by Horn and Schunk [Horn81] must deal with what is known as the *aperture problem*, which refers to the fact that only the component of optical flow parallel to the local image gradient can be recovered from first-order differential information. Intuitively, the aperture problem corresponds to the fact that for a moving edge, only the component of motion perpendicular to the edge can be determined. This effect is responsible for the illusion of upward motion produced by the rotating spirals of a barber pole where either vertical or horizontal motion could produce the local motion of the edges, and the eye chooses the wrong one. In order to determine both components of the flow field vector, information must be combined over regions large enough to encompass significant variations in the gradient direction. The most common method of doing this involves some form of regularization [Horn81, Anan95, Nage86]; however, such methods often result in blurring of motion discontinuities. A nonblurring method known as *constraint line clustering* has been proposed by Schunck [Schu84]. Techniques using higher order derivatives to avoid the aperture problem have been proposed [Nage83, Uras88]; however, these suffer from stability problems due to multiple differentiation and typically require extensive smoothing to produce clean results. Other methods include spatiotemporal energy methods [Heeg87], Fourier methods based on phase correlation [Burt89], and direct correlation of image patches [Barn80, Litt88]. Recent work by Anandan [Anan89] provides a common framework into which many of these methods can be incorporated.

A potential problem with most of the above approaches is the assumption that the motion field manifests itself locally as a rigid 2-D motion of an image pat :h. Unfortunately, the local apparent motion of the image, known as the *optical flow*, does not necessarily correspond to the 2-D motion field. The most obvious demonstrations are pathological examples. For instance, a spinning, featureless sphere under constant illumination has zero optical flow, but a nonzero motion field. Conversely, a stationary sphere under changing illumination has nonzero optical flow, but zero motion field. Image patches also undergo various nonrigid deformations such as expansion and skewing. Verri and Poggio [Verr87] have shown that only under special conditions of lighting and movement do the motion field and the optical flow correspond exactly. They also show, however, that for sufficiently high gradient magnitude, the agreement can be made arbitrarily close. This corresponds to the intuition that for strongly textured images the motion field and the optical flow are approximately equal. A few authors have attempted to explicitly include some of these effects (e.g., [Burt89]), but it is not clear that any great advantage has been obtained thereby.

On the whole, despite a great deal of effort expanded in devising flow invariants, regularization methods, and matching techniques, neither correspondence nor flow field methods have yielded data sufficiently accurate to allow the theoretical structure-from-motion results to be reliably applied. Adiv [Adiv85] argues that inherent near ambiguities in the 3-D structure-from-motion problem may make the goal of extracting information sufficiently precise to allow uniform application of the theoretical solutions unattainable in practice. Verri and Poggio [Verr87] make essentially the same point, arguing that the disagreement between the motion field and the optical flow makes the computation of sufficiently accurate quantitative values impractical.

An alternative is to devise qualitative applications that can make use of inaccurate flow information [Thom86, Nels88, Nels89]. The motion recognition strategies proposed here represent one such application. Many of the motion features proposed in the next section are qualitative in the sense that their detection does not rely on highly accurate measurements of the motion field. In fact, useful motion features can be obtained from partial information such as the projected flow computed in the first step of the Horn and Schunck procedure, for example, the directional divergence used for obstacle avoidance in [Nels89]. To reiterate, our idea is to use motion information for identification directly, rather than proceeding indirectly, through the reconstruction of an analog 3-D world model.

3. TEMPORAL TEXTURE

Classical gray-level texture analysis is concerned with the identification of spatial invariances in the gray-level patterns in an image region. These invariances may be either structurally or statistically defined. The basic idea is to characterize different sorts of "stuff" of indeterminate spatial extent in terms of such invariances. In this article we extend this basic idea into the temporal dimension with the idea of recognizing similar stuff in dynamic scenes. This is motivated in part by the existence of a large class of natural phenomena that seem to have characteristic motions, but indeterminate spatial extent. Examples include windblown trees or grass, turbulent flow in cloud patterns, ripples on water, falling snow, and the motion of a flock of birds or a crowd of people. The motion in a temporal texture is distinct from that in patterns such as walking and cycling, which involve structure at a single location.

Temporal texture could be analyzed directly as a three-dimensional signal using generalizations of the techniques applied to two-dimensional fields. However, since most changes along the time dimension are due to motion in the image, it makes sense to preprocess the time-varying image to obtain motion information, as it is in object motion that the physical invariances lie. In this case, a natural choice is the optic flow field. The basic source of information is thus a time-varying vector field representing an approximation to the two-dimensional motion field induced by movement in the world. Such a field contains considerably more information than the scaler valued field associated with gray-level texture analysis. In addition, the direction and magnitude of motion have a more direct relationship to typically salient events in the world than the gray level of a single pixel. Consequently, certain types of recognition might be expected to be easier. For example, in the right context, fast downward motion could be taken as evidence of a falling object. It is difficult to envisage making any similar statement about (say) gray level 147. A problem with using optic flow is that it is difficult compute accurately. One solution is to devise measures that are insensitive to inaccuracy. Another is to utilize partial information. An example is the gradient parallel component of the optic flow, which is simpler to compute locally from an image sequence than the full motion field.

Despite the differences in domain, some techniques of spatial texture analysis are applicable to temporal textures. Spatial texture analysis is traditionally performed using either statistical or syntactic methods. Statistical methods utilize measures of local features that are expected to be similar within patches of the same texture. Examples of measurements that have been used include gray-level cooccurrence matrices [Hara73, Conn80], Fourier power spectra [Bajc76, Chen82], and average magnitude response of filter masks [Laws80, Mali89]. There are also several methods based on estimation parameters for a description of a region in terms of some texture model. Examples include autoregressive models [Kash82] and Markov random fields [Kane82]. Syntactic approaches are most appropriate for highly regular textures and involve analyzing the geometric arrangement of primitive structural elements. In the case of natural temporal textures, techniques similar to the statistical gray-level methods seem most appropriate, and most of the features described in this article are of this type. As with spatial textures, the main criteria for selecting features are that they change little within a given texture (i.e., an area of the same stuff), and that they vary significantly between different textures.

The dimensionality of the vector-valued flow field and the fact that measures can be made in both space and time allow considerable latitude in designing features. Since textures are characterized by statistical regularities in the occurrence of local structure, extraction of features useful for classification generally involves at least two tiers of processing: A local feature extraction stage, and (at least one) spatially or temporally extended integration stage. Local features can be any useful quantity that can be associated with a point in the image. Examples include flow magnitude and direction, differential measures such as divergence and curl, and local uniformity measures. The spatiotemporal motion energy filters introduced by Heeger [Heeg87] could also provide useful measures in this context. Typically these are expected to vary within a texture, thus necessitating the integration phase. Extended measures are most frequently based on quantities such as means or variances, but other extended measures, such as Fourier coefficients and cooccurrence statistics, can be used. The most typical structure for a temporal texture feature involves extended spatial or temporal (or both) measures of spatiotemporal microfeatures. Features can also be derived from extended spatial measure of extended temporal features and vice versa.

In order to simplify the motion preprocessing, we considered features based on the gradient parallel component of the motion field, also referred to as the normal flow. The simplest local motion measures are the magnitude and direction of the normal flow. We examine several statistical features based on the distribution of these first-order quantities. The direction and magnitude can be combined locally, both spatially and temporally, to obtain second-order local motion measures. We also examine features based on the distribution of some second order measures. All these are described below.

A useful statistic based on the distribution of the normal flow magnitude is the average flow magnitude divided by its standard deviation. The scaling by the standard deviation has the effect of making the measure robust under scaling changes. One way to think of this statistic is as a measure of "peakiness" in the velocity distribution. It is invariant under translation, rotation, and temporal and spatial scaling.

We also considered statistics of second-order flow magnitude features, namely, estimates of the divergence and curl of the motion field obtained from the normal flow. Positive and negative divergence and positive and negative curl were taken as separate features to give four different second-order features. The features used are the mean values of these quantities over the region of interest. They are invariant with respect to rotation and translation, but not scaling. If scale invariant features are desired, ratios of the differential measures can be used.

A useful first-order statistic can be derived from the distribution of flow directions. Intuitively, what is being measured is the nonuniformity in direction of motion. Our non-uniformity statistic was computed by discretizing the direction into eight possible values, computing a histogram over the relevant neighborhood of the image, and summing the absolute deviation from a uniform distribution. It should be noted that the normal flow direction at a pixel is parallel (or antiparallel) to the gradient direction. Thus measures based on the normal flow direction alone depend on the underlying intensity texture. To reduce this dependence, the normal flow directions in the histogram are normalized by the four-way histogram of gradient directions. This feature is invariant under translation, rotation, and temporal and spatial scaling.

Second-order measures of the normal flow direction distribution can be derived from the difference statistics, which give the number of pixel pairs at a given offset that differ in their values by a given amount. These difference statistics can be represented by a cooccurrence matrix of the normal flow direction surrounding a pixel. Cooccurrence matrices are computed for four directions (horizontal, vertical, positive diagonal, and negative diagonal) at a distance proportional to the average flow magnitude. This yields invariance with respect to scaling. In each direction the ratio of the number of pixel pairs differing in direction by at most one to the number of pixel pairs differing by more than one is computed. This ratio is the sum of the first two difference statistics to the sum of the last three difference statistics. Logarithms of the resulting ratios are used as a feature in each of the four directions, and represent a measure of the spatial homogeneity of the flow. These features are invariant under translation, rotation, and scaling.

4. EXPERIMENTAL RESULTS

A set of image sequences representing both oriented temporal textures such as flowing water and nonoriented textures such as leaves fluttering in the wind was digitized. In addition, sequences representing uniform expansion and rotation of a textured scene were obtained. These were used in classification experiments utilizing the features described above. Seven different texture samples, listed below, were used for the experiments:

A. fluttering crepe paper bands

B. cloth waving in the wind

C. motion of tree in the wind

D. flow of water in a river

E. turbulent motion of water

F. uniformly expanding image produced by forward observer motion

G. uniformly rotating image produced by observer roll.

Representative examples of scenes and derived flow are illustrated in Figs. 2A–2E. Figure 3 illustrates the temporal dimension for two of the cases, showing a horizontal slice through the spatiotemporal solid. The temporal axis runs vertically.

For each sample texture, two image sequences consisting of 16 256 × 256 pixel frames taken at 30 Hz were split into quadrants to obtain eight independent sample image sequences of 128 × 128 pixels. The normal flow field was computed between each consecutive pair of image frames using a multiresolution flow computation, with the direction of normal flow quantized to one of eight directions. The end result of the processing was a sample of eight normal flow sequences of 15 frames each for each texture.

Classification experiments were run using a nearest centroid classifier. More elaborate classifiers could be used, but the nearest centroid method gives a fairly direct indication of the utility of the features. The features used were those described in the previous section, namely

a. mean flow magnitude divided by standard deviation

b. positive and negative curl and divergence estimates

c. nonuniformity of flow direction

d. directional difference statistics in four directions.

Normalization constants were computed so that the ensemble mean for each feature was 1. No more sophisticated normalization procedure was found necessary.

The first four samples of each texture are used as a training set to compute the centroid of the cluster corresponding to that texture in the feature space. The different feature values are converted into common units by mapping the average of the resulting centroids to a unit vector. Table 1 contains the values of these features for each flow sample. It can be seen that, overall, the within sample variation is smaller than the between sample variation as desired. No single feature is sufficient to distinguish all the textures, but for each texture, there is at least one feature that clearly separates it from the others. For example, as would be expected, texture A, containing an approaching object, is distinguished by high divergence. For texture B, containing moving vertical bands, the second-order difference feature in the vertical direction clearly separates it from the rest.

The remaining four samples are tested using a nearest centroid classification scheme. The results of classification are summarized in Table 2. Note that none of the features alone is sufficient to separate all the textures, but the combination gives 100% success in the classification

TABLE 1
Sample Features

	a	b				c	d			
Texture	Mag	Pos div	Neg div	Pos curl	Neg curl	Dir	Hor	Vert	Pos diag	Neg diag
A: Bands	1.083	0.591	−0.698	0.275	−0.199	0.488	5.013	9.084	5.192	5.178
	1.009	9.640	−0.570	0.240	−0.223	0.663	4.679	8.356	4.962	5.055
	1.081	0.467	−0.625	0.212	−0.209	0.837	4.358	7.878	4.518	4.409
	1.221	0.544	−0.548	0.188	−0.203	0.694	5.319	8.954	5.540	5.452
B: Cloth	1.417	0.648	−0.620	0.314	−0.322	0.928	4.265	6.170	4.806	4.289
	1.529	0.530	−0.557	0.323	−0.335	0.917	4.939	5.681	6.149	4.826
	1.282	0.610	−0.597	0.317	−0.308	0.942	3.390	4.972	3.882	3.338
	1.393	0.610	−0.647	0.337	−0.342	0.902	3.596	4.732	4.276	3.563
C: Plant	0.964	0.708	−0.216	0.196	−0.297	0.947	1.481	2.103	2.276	1.466
	1.064	0.306	−0.434	0.263	−0.176	0.952	1.556	2.287	2.353	1.574
	0.882	0.527	−0.436	0.258	−0.279	0.968	1.262	1.868	2.055	1.239
	0.951	0.386	−0.392	0.294	−0.264	0.970	1.300	1.871	2.053	1.243
D: Water	1.293	0.446	−0.550	0.191	−0.161	0.864	4.637	5.148	7.154	4.792
	1.494	0.486	−0.382	0.171	−0.187	0.814	5.025	5.617	7.038	5.110
	1.258	0.517	−0.585	0.206	−0.186	0.885	4.297	4.777	6.218	4.505
	1.512	0.448	−0.528	0.222	−0.225	0.887	3.869	4.176	6.073	3.876
E: Turbulence	1.123	0.728	−0.637	0.400	−0.399	0.946	2.454	2.972	3.962	2.521
	1.206	0.811	−0.587	0.376	−0.408	0.929	2.616	3.052	4.303	2.699
	1.106	0.595	−0.769	0.422	−0.397	0.945	2.186	2.733	3.671	2.250
	1.062	0.799	−0.526	0.430	−0.427	0.945	2.164	2.611	3.677	2.152
F: Approach	1.099	0.462	−1.001	0.268	−0.231	0.947	2.175	3.167	2.661	2.241
	1.076	0.397	−0.954	0.266	−0.206	0.922	2.668	3.327	3.791	2.785
	1.028	0.336	−0.942	0.248	−0.186	0.922	2.366	3.173	3.272	2.490
	1.018	0.422	−1.018	0.331	−0.257	0.918	2.597	3.458	3.375	2.683
G: Roll	1.182	0.437	−0.395	0.095	−0.584	0.929	2.952	4.076	3.523	3.025
	1.204	0.621	−0.420	0.083	−0.663	0.942	3.257	4.185	4.077	3.394
	1.032	0.382	−0.353	0.053	−0.660	0.935	2.923	3.970	3.627	3.076
	1.087	0.528	−0.337	0.110	−0.725	0.943	2.788	3.782	3.597	2.906

TABLE 2
Classification Results

Feature combination	Correct classification	Percentage success
All	28	100
b. d	28	100
a. d	24	85
b. c	21	75
d	21	75
b	20	71

of the test cases. In fact, the second-order features alone are sufficient for successful classification in all cases.

We also performed a principal component analysis of these features to gauge the relative importance of different features in producing the variation in the sample values. The first three principal components of the entire data set are shown in Table 3. Note that the first principle component has a high eigenvalue and relatively high proportions of the second-order features, particularly positive and negative divergence. This is consistent with the finding that the second-order features alone are sufficient for classification in this case. The principal components

TABLE 3
Principle Components

Comp	a	b				c	d				Eigenvalue
1	0.29	0.51	0.95	0.02	−0.29	−0.20	0.15	0.20	0.18	0.16	54.95
2	−0.25	0.42	0.38	0.78	0.07	0.10	−0.82	−0.37	−0.89	−0.69	5.82
3	−0.09	0.78	−0.50	0.12	−0.16	0.01	0.16	0.06	−0.00	0.13	2.54

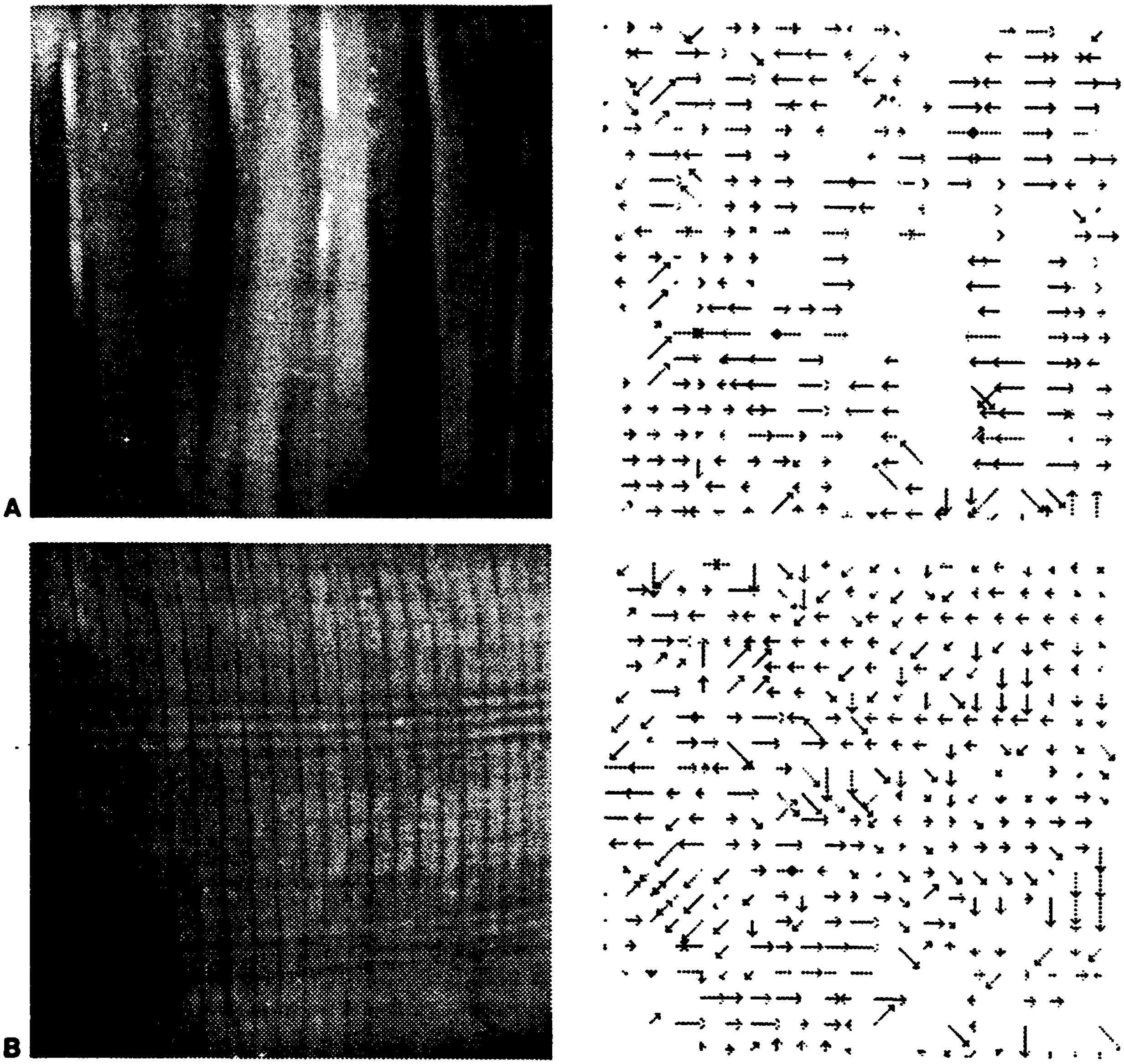

FIG. 2. (A) Image and flow for paper bands. (B) Image and flow for cloth. (C) Image and flow for plant leaves. (D) Image and flow for water. (E) Image and flow for turbulent motion.

within each sample contain small absolute coefficients for the same second-order features, showing that these features are most useful in classification.

5. CONCLUSION AND FUTURE WORK

We have described a method of motion recognition using temporal textures. This technique uses statistical measures of local motion features as components of a feature vector that can be used in standard classification methods. We identified several motion features that appear to have desirable properties for recognition, and illustrated their utility in classifying a sample of real-world temporal textures. Future work includes the analysis of other feature classes, including purely temporal features of the flow as well as Fourier techniques.

We also plan to extend the technique to the recognition of compact, possibly nonrigid, objects. This differs from textural recognition in that it is typically the detailed arrangement of features (in space and time), which we term the object's *action*, rather than regional statistics, that constitute the basis for identification. Though such an extension will not provide a general solution to the object recognition problem, we think that there are a number of

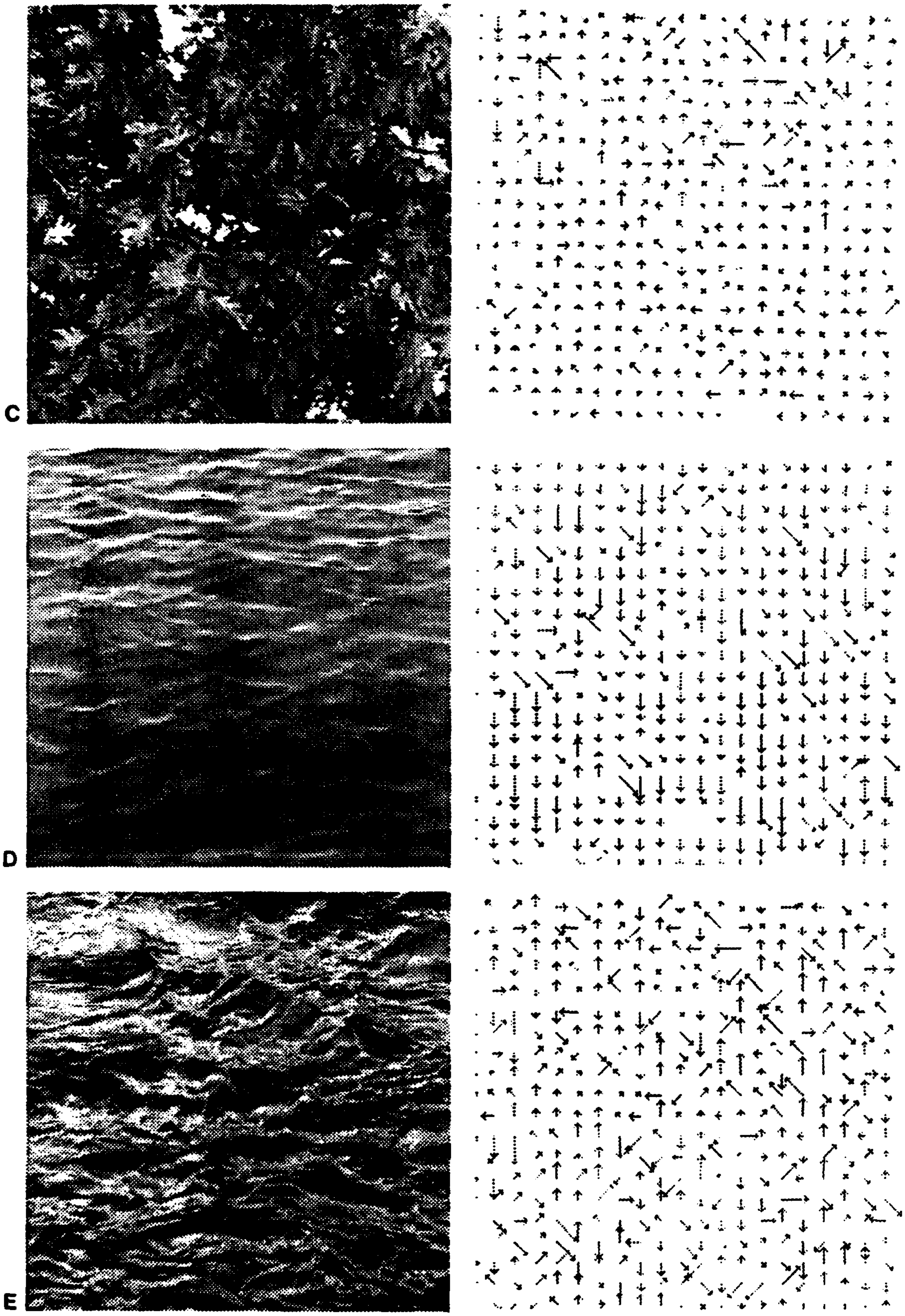

FIG. 2.—*continued*

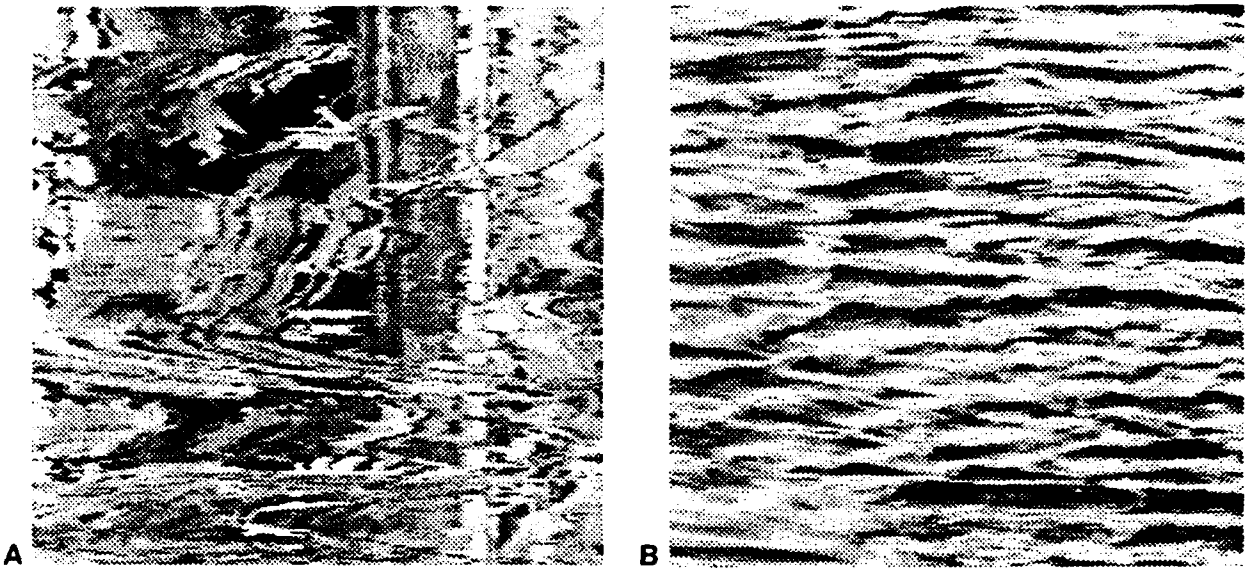

FIG. 3. (a) Temporal slice for leaves. (b) Temporal slice for turbulence.

situations in which motion recognition techniques can make identification far easier than it would be using static image processing alone. For example, something big and oblong moving smoothly and horizontally in the vicinity of a road is probably a car. This conjunction of features is fairly simple to compute, particularly compared to the requirements of static analysis, which must be able to tell cars from boulders, architectural clutter, and shadows on the road. Similarly, the toad in Section 1 assumes that anything small (it has a notion of distance and hence of size from crude stereo), oblong, and moving in the direction of its long axis is good to eat (or at least is worth a taste).

The simplest technique is to use conjunctions of motion and geometric features, and more generally, spatio-temporal templates specifying the rough spatial arrangement of motion (and geometric) features. More sophisticated pattern recognition techniques include the generalized Hough transform [Ball81] and hypothesize-and-test schemes [Grim86]. A candidate for handling time sequences for which a fixed template is insufficiently flexible is the formalism of hidden Markov models [Baum70, Jeli76, Juan85]. These have been used primarily for speech recognition, but the technique is valid for a wide variety describable by sequence of discrete symbols having an underlying probabilistic relationship.

REFERENCES

[Adiv85] G. Adiv. Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field, in *Proceedings, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1985*, pp. 70–77.

[Anan85] P. Anandan and R. Weiss. Introducing a smoothness constraint in a matching approach for the computation of optical flow fields, in *Proceedings, Third Workshop on Computer Vision: Representation and Control, 1985*, pp. 186–194.

[Anan89] P. Anandan. A computational framework and an algorithm for the measurement of visual motion, *Int. J. Comput. Vision* **2**, 1989, 283–310.

[Ande85] C. H. Anderson, P. J. Burt, and G. S. van der Wal. Change detection and tracking using pyramid transform techniques, in *Proceedings, SPIE Conference on Intelligent Robots and Computer Vision, Boston, MA, 1985*, pp. 300–305.

[Bajc76] R. Bajcsy and L. Lieberman. Texture gradient as a depth cue, *Comput. Graphics Image Process.* **5**, 1976, 52–67.

[Ball81] D. H. Ballard. Generalizing the Hough transform to detect arbitrary shapes, *Pattern Recognit.* **13**(2), 1981, 111–122.

[Barn80] S. T. Barnard and W. B. Thompson. Disparity Analysis of Images, *IEEE Trans. PAMI* **2**(4), 1980, 330–340.

[Baum70] L. E. Baum and J. Eagon. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Math. Stat.* **41**, 1970, 164–171.

[Boll87] R. C. Bolles Epipolar Plane Analysis: An Approach to Determining Structure from Motion, in *Proceedings, International Joint Conference on Artificial Intelligence, 1987*, pp. 7–15.

[Burt89] P. J. Burt, J. R. Bergen, R. Hingorani, R. Kolczynski, W. A. Lee, A. Leung, J. Lubin, and H. Shvayster. Object tracking with a moving camera, *Proceedings, IEEE Workshop on Motion, Irvine, CA, 1989*.

[Chen82] C. H. Chen. A study of texture classifications using spectral features, in *Proceedings, 6th International, Conference on Pattern Recognition, Munich, 1982*, pp. 1064–1067.

[Chun86] H. W. Chun. a representation for temporal sequence and duration in massively parallel networks: Exploiting link connections, in *Proceedings of AAAI-86, August, 1986*.

[Conn80] R. W. Conners and C. A. Harlow. A theoretical comparison of texture algorithms, *IEEE Trans. PAMI* 2(3), 1980, 204–222.

[Cutt77] J. E. Cutting and L. T. Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues, *Bull. Psychon. Soc.* **9**, 1977, 353–356.

[Cutt81] J. E. Cutting, Six tenets for event perception, *Cognition* **10**, 1981, 71–78.

[Dins88] I. Dinstein, A new technique for visual motion alarm. *Pattern Recogn. Lett.* **8**(5), 1988, 347.

[Elma88] J. E. Elaman, *Finding Structure in Time*, Technical Report 8801, Center for Research in Language, University of California, San Diego, 1988.

[Ewar87] J. P. Ewart, Neuroethology of releasing mechanisms: Prey-catching in toads, *Behav. Brain Sci.* **10**, 1987, 337–405.

[Feld88] J. E. Feldman, *Time, Space and Form in Vision*, Technical Report 244, University of Rochester Department of Computer Science, 1988.

[Godd89] N. H. Goddard, Representing and recognizing event sequences, *Proceedings, AAAI Workshop on Neural Architectures for Computer Vision, Minneapolis, August, 1988.*

[Grim86] W. E. L. Grimson, The combinatorics of local constraints in model-based recognition and localization from sparse data, *J. ACM* **33**(4), 1986, 658–686.

[Hara73] R. M. Haralick, K. Shanmugam, and I. Dinstein, Textural features for image classification, *IEEE Trans. Systems Man Cybernet.* **3**(6), Nov. 1973, 610–621.

[Heeg87] D. Heeger, Optical flow from spatio-temporal filters, *Proceedings 1st International Conference on Computer Vision, 1987*, pp. 181–190.

[Hild87] E. C. Hildreth and C. Koch, The analysis of visual motion from computational theory to neural mechanisms, *Ann. Rev. Neurosci.* **10**, 1987.

[Hoff82] D. D. Hoffman and B. E. Flinchbaugh, The interpretation of biological motion, *Biol. Cybernet.* **42**, 1982, 195–204.

[Horn81] B. K. P. Horn and B. G. Schunk, Determining optical flow, *Artif. Intell.* **17**, 1981, 185–204.

[Jeli76] F. Jelinek, Continuous speech recognition by statistical methods, *Proc. IEEE* **64**, 1976, 532–556.

[Joha73] G. Johansson, Visual perception of biological motion and a model for its analysis, *Perception Psychophy.* **14**, 1973, 201–211.

[Joha76] G. Johansson, Spatio-temporal differentiation and integration in visual motion perception, *Psychol. Res.* **38**, 1976, 379–393.

[Juan85] B. H. Juang and L. R. Rabiner, Mixture autoregressive hidden Markov models for speech signals, *IEEE Trans. Acoustics Speech Signal Processing* **33**(6), Dec. 1985, 1404–1413.

[Kane82] H. Kaneko and E. Yodogawa, A Markov random field application to texture classification, in *Proceedings, Pattern Recognition and Image Processing, Las Vegas, June, 1982*, pp. 221–225.

[Kash82] R. L. Kashyap, R. Chellappa, and A. Khotanzad, Texture classification using features derived from random field models, *Pattern Recognit. Lett.* **1**, 1982, 43–50.

[Laws80] K. I. Laws, *Textured Image Segmentation*, Ph. D. dissertation, Department of Engineering, University of Southern California, USCIPI Report No. 940, Jan. 1980.

[Litt88] J. J. Little, H. H. Bulthoff, and T. Poggio, Parallel optical flow using local vote counting, in *2nd International Conference on Computer Vision*, 1988, 454–459.

[Long81] H. C. Longuet-Higgins, A computer algorithm for reconstructing a scene from two projections, *Nature* **293**, 1981.

[Mali89] J. Malik and P. Perona, A computational model of texture segmentation, in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition, 1989*, pp. 326–332.

[Marr79] D. Marr and T. Poggio, A Computational Theory of Human Stereo Vision, *Proc. R. Soc. London B* **204**, 1979, 301–328.

[Moro79] H. P. Morovec, Visual Mapping by a Robot Rover, *Proc. IJCAI, 1979*, pp. 598–600.

[Nage83] H. H. Nagel, Displacement vectors derived from second order intensity variations in image sequences, *Comput. Vision Pattern Recognit. Image Process.* **21**, 1983, 85–117.

[Nage86] H. H. Nagel and W. Enkelmann, An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE Trans, PAMI* **85**, Sept. 1986, 565–593.

[Nels88] R. C. Nelson and J. Aloimonos, Finding motion parameters from spherical flow fields (or the advantages of having eyes in the back of your head) *Biol. Cybernet.* **58**, 1988, 261–273.

[Nels89] R. C. Nelson and J. Aloimonos, Using flow field divergence for obstacle avoidance in visual navigation, *IEEE Trans. PAMI* **11**(10), Oct 1989, 1102–1106.

[Nels90] R. C. Nelson, Moving object detection by a moving observer: Two qualitative methods, in preparation.

[Pent89] A. Pentland and K. Mase, *Lip Reading: Automatic Visual Recognition of Spoken Words*, M.I.T. Media Lab Vision Science Technical Report 117, Jan. 1989.

[Praz81] K. Prazdny, Determining the instantaneous direction of motion from optical flow generated by a curvilinear moving observer. *Comput. Vision Graphics Image Process.* **22**, 1981, 238–248.

[Rash80] R. Rashid, *Lights: A System for the Interpretation of Moving Light Displays*, PhD thesis, Department of Computer Science, University of Rochester, 1980.

[Reic88] W. Reichard and M. Engelhaaf, Movement detectors provide sufficient information for local computation of 2-D velocity field, *Naturwissenschaften* **74**, 1988, 313–315.

[Rime90] R. Rimey and C. M. Brown, Sequential behavior as a selective attention mechanism: Modeling eye movements with hidden Markov models, in preparation.

[Schu84] B. G. Schunck, Motion segmentation by constraint line clustering in, *IEEE Workshop on Computer Vision: Representation and Control, 1984*, pp. 58–62.

[Tank87] D. W. Tank and J. J. Hopfield, Concentrating information in time: Analog neural networks with applications to speech recognition problems, in *Proceedings, First International Conference on Neural Networks, 1987*, pp. 455–468.

[Thom86] W. B. Thompson and J. K. Kearney, Inexact vision. in *Workshop on Motion, Representation, and Analysis, May, 1986*, pp. 15–22.

[Tsai81] R. Y. Tsai and T. S. Huang, Estimating 3-D motion parameters of a rigid planar patch I, *IEEE ASSP* **30**, 1981, 525–534.

[Tsai84] R. Y. Tsai and T. S. Huang, Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces, *IEEE Trans. PAMI* **6**, 1984, 13–27.

[Ullm79] S. Ullman, The interpretation of structure from motion, *Proc. R. Soc. London B*, **203**, 1979, 405–426.

[Uras88] S. Uras, F. Girosi, and V. Torre, A computational approach to motion perception, *Biol. Cybernet.* **60**, 1988, 79–87.

[Verr87] A. Verri and T. Poggio, Against quantitative optical flow, in *International Conference on Computer Vision, June 1987*, pp. 171–180.

[Waxm87] A. Waxman, Image flow theory: A framework for 3-D inference from time varying imagery, in *Advances in Computer Vision* (C. Brown, Ed.), Erlbaum, Hillsdale, NJ. 1987.

Detecting Activities

Ramprasad Polana and Randal Nelson

Department of Computer Science
University of Rochester
Rochester, New York 14627
Email: polana@cs.rochester.edu and nelson@cs.rochester.edu

Abstract

The recognition of repetitive movements characteristic of walking people, galloping horses, or flying birds is a routine function of the human visual system. It has been demonstrated that humans can recognise such activity solely on the basis of motion information. We present a novel computational approach for detecting such activities in real image sequences on the basis of the periodic nature of their signatures. The approach suggests a low-level feature based activity recognition mechanism. This contrasts with earlier model-based approaches for recognizing such activities.

1 Introduction

The motion recognition ability of the human visual system is remarkable. People are able to distinguish both highly structured motion, such as that produced by walking, running, swimming or flying birds, and more statistical patterns such as that due to blowing snow, flowing water or fluttering leaves. The classic demonstration of pure motion recognition by humans is provided by Moving Light Display experiments [Johansson, 1973]. More subtle movement characteristics can be distinguished as well. For example, human observers can identify the actor's gender and even identify the actor if known to them by his or her gait. Similar discrimination abilities using motion alone have been observed in non-human animals as well [Ewart, 1987]. This biological use of motion probably reflects the fact that for certain tasks, visual motion provides more effective cues than other modes of visual perception. Motion is a particularly useful cue for certain types of recognition due to the fact that it is relatively easy to extract the motion field independent of illumination and shading of the image.

As a first step towards motion recognition by a machine, we define three classes of motion according to the spatial and temporal uniformity exhibited, so that different motions can be recognized using different techniques appropriate to their inherently different characteristics. We define *temporal textures* to be the motion patterns of indeterminate spatial and temporal extent, *activities* to be motion patterns which are temporally periodic but are limited in spatial extent, and *motion events* to be isolated simple motions that do not exhibit any temporal or spatial repetition. Examples of temporal textures include wind blown trees or grass, turbulent flow in cloud patterns, ripples on water, the motion of a flock of birds etc. Examples of activities are walking, running, rotating or reciprocating machinery, etc. Examples of motion events are isolated instances of opening a door, starting of a car, throwing a ball etc.

It turns out that temporal textures can be effectively treated with statistical techniques analogous to those used in gray-level texture discrimination. A previous paper [Polana and Nelson, 1992] describes this. Activities and motion events, on the other hand, are more discretely structured, and techniques similar to those used in static object recognition would be expected to be useful in their classification. Since different sorts of techniques must be used to distinguish the different sorts of motion, it would be useful to have a method for making a preliminary classification of the motions present in an image. In this paper, we describe a robust method for detecting and localizing periodic activities, including ones, such as walking or flying, that involve simultaneous translation of the actor. The method is based on frequency domain analysis of an image in which low-level motion information has been used to isolate and track likely locations of activity. The method also suggests a way of using low-level structural features to classify activities once they have been detected.

Motion recognition techniques, both of the discrete and textural variety have the potential to disambiguate the motions of different origin. The motions of many natural objects can be classified as periodic activities, including human walking. Duplication of the recognition ability of these motions in machine systems would be useful in a number of applications, such as automated surveillance. Motion detection via image differencing can be used for intruder detection; however such systems are subject to false alarms, especially in outdoor environments, since the system is triggered by anything that moves, whether it be a human, a dog, or a tree blown by the wind. Another application is in industrial monitoring. Many manufacturing operations involve a long sequence of simple operations each performed repeatedly and at high speed by a specialized mechanism at a particular location. It should be possible to set up one or more fixed cameras that cover the area of interest, and to characterize the allowed motions in each region of the image(s).

2 Related Work

Although motion plays an important role in biological recognition tasks, motion recognition in general, has received little attention in the literature compared to the volume of work on static object recognition. Most computational motion work in motion in fact, has been concerned with various aspects of the structure-from-motion problem. There is a large body of psychophysical literature addressing the perception of motion, most of it concerned with primitive percepts. A modest amount of this work addresses more complicated motion recognition issues [Johansson, 1973, Cutting, 1981, Hoffman and Flinchbuagh, 1982, Hildreth and Koch, 1987], but the models and descriptions have typically not been implemented. Various computational models of temporal structure, have been proposed (e.g. [Chun, 1986, Feldman, 1988]) but much of this work is at a fairly high level of abstraction, and has not actually been applied to visual motion recognition except in rather artificial tests.

Goddard [1989] considers recognizing event sequences from Moving Light Display (MLD) images. His work addresses the representation of motion event sequences and their recognition assuming certain invariant image features. His input consists of the joint angles and angular velocities computed from the motion of the dots in the light displays. The joint angles and angular velocities are invariant to rotation in the image plane, scale and translation. A challenging part in computing these invariants is to recover the connectivity of the individual dots (by body parts) in the MLD images. A domain independent approach to this problem is given by Rashid. Rashid [Rashid, 1980, O'Rourke and Badler, 1980] considered the computational interpretation of moving light displays, particularly in the context of gait determination. This work emphasized rather high-level symbolic models of temporal sequences, an approach made possible by the discrete nature of the moving light displays. The results were quite sensitive to discrete errors and thus highly dependent on the ability to solve the correspondence problem and accurately track joint and limb positions. This severely limits the general applicability of the method.

A few studies have considered highly specific aspects of motion recognition computationally. Pentland [Pentland and Mase, 1989] considered lip reading, and implemented a system that could recognize spoken digits with 70%-90% accuracy over 5 speakers. The system required the location of the lips to be entered by hand, and depended on an explicitly constructed lip model. Some temporal pattern recognition work has been done in the context of speech processing [Juang and Rabiner, 1985, Tank and Hopfield, 1987, Elaman, 1988]. But the applicability of the techniques to motion recognition has not been considered.

Anderson et al. [Anderson *et al.*, 1985] describe a method of change detection for surveillance applications based on the spectral energy in a temporal difference image. This was not generalized to other motion features or more sophisticated recognition. Koller, Heinze and Nagel [1991] developed a system that tracks moving vehicles and characterizes their trajectory segments in terms of natural language concepts. Gould and Shah [1989] represent motion characteristics of moving objects by recording the important events in their trajectory. They propose the use of the resulting *trajectory primal sketch* in a motion recognition system. Allmen and Dyer have developed a method of extracting spatiotemporal curves corresponding to moving objects and applied the technique to detection of cyclic motions [Allmen and Dyer, 1990]. All the above require the difficult task of robustly computing the trajectories or spatiotemporal curves from image sequences before attempting recognition, and the demonstrations of their techniques involve only synthetic image sequences.

3 Activity Detection

Activities involve a regularly repeating sequence of motion events. If we consider an image sequence as a spatiotemporal solid with two spatial dimensions x, y and one time dimension t, then repeated activity tends to give rise to periodic or semi-periodic gray level signals along smooth curves in the image solid. We refer to these curves as *reference curves*. If these curves could be identified and samples extracted along them over several cycles, then frequency domain techniques could be used in order to judge the degree of periodicity.

Before defining the reference curves, first we shall formalize the concept of a periodic object. An object is defined as a set of points P. Associated with each $p \in P$ is a function $X_p(t)$ giving its location (in a fixed 3D coordinate system) as a function of time. A stationary periodic object (ie. a stationary object exhibiting periodic activity) has the property that $X_p(t) = X_p(t+\tau)$ for all $p \in P$, where τ is the time period for one cycle of the activity and is independent of p. We now define a translating periodic object. Such an object has the property that $X_p(t) = Y_p(t) + Z(t)$, where Y_p satisfies $Y_p(t) = Y_p(t+\tau)$ and $Z(t)$ is a path in 3D space independent of p. It can be assumed that $Z(0) = 0$ so that $X_p(0) = Y_p(0)$. Intuitively, a periodic object characterized by $Y_p(t)$ is translated along the path $Z(t)$ (we are assuming the object does not undergo any rotation and the viewing angle does not change). If we compensate for the translation of the object, we would be looking at a stationary periodic object as shown by the equation: $X_p(t) - Z(t) = Y_p(t) = Y_p(t+\tau) = X_p(t+\tau) - Z(t+\tau)$. Note that $Z(t)$ is not necessarily periodic. Note also that a stationary periodic object is a special case of translating periodic object with no translation, or in other words $Z(t) = 0$ for all t.

Corresponding to each point p of a translating periodic object, we define a 3D-reference curve $R_p(t)$ to be the path $X_p(0) + Z(t)$. We also define a 2D-reference curve $r_p(t)$ corresponding to a point p of the object, to be the projection of $R_p(t)$ onto the image plane over time (hence $r_p(t)$ is a curve in (x, y, t) space). The gray-level signal along the 2D-reference curve $r_p(t)$ is determined by the set of points of the object that appear along the 3D-reference curve $R_p(t)$. It can be shown that the same set of points of the object recur periodically along each reference curve $R_p(t)$. For example, the point p is on the reference curve $R_p(t)$ at time zero, and it coincides with the reference curve at regular intervals of τ (since $X_p(\tau) = Y_p(\tau) + Z(\tau) = Y_p(0) + Z(\tau) = X_p(0) + Z(\tau)$). Similarly, every other point of the object on the reference curve $R_p(t)$ recurs along $R_p(t)$ at intervals of τ.

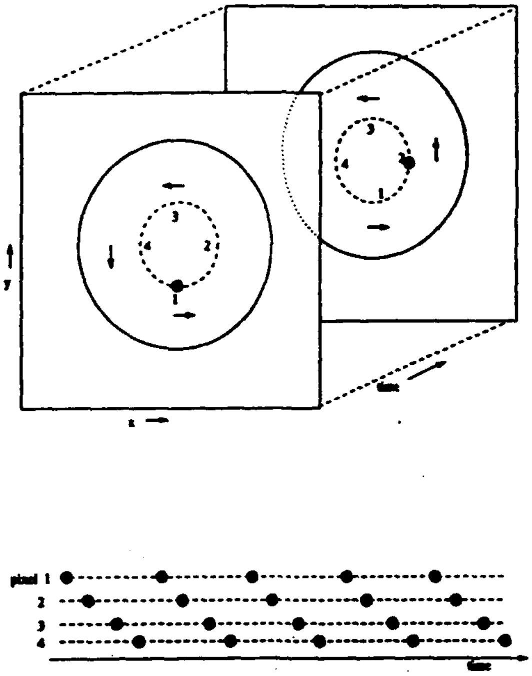

Figure 1: stationary circular rotation: temporal frequency and phase

We shall illustrate the concept with two examples, one stationary activity (one produced by a stationary periodic object) and the other involving a uniform translation of the actor, i.e. a locomotory activity. If the activity is stationary, the reference curves are lines parallel to the temporal dimension. For example, a circularly rotating ring gives rise to a temporally periodic signal at every pixel. This is illustrated in figure 1. In the case of uniform translation, the curves are straight lines at some angle that depends on the velocity. For general translation and perspective projection, the lines associated with a given actor approaching the camera, form a bundle with a common intersection, the vanishing point. For many practical situations, however, the vanishing point is far enough removed that the lines can be considered to be effectively parallel.

Consider the case of human walking. This is an example of a non-stationary activity; that is, if we attach a reference point to the person walking, that point does not remain at one location in the image. If the person is walking with constant velocity, however, and is not too close to the camera, then the reference point moves across the image on a path composed of a constant velocity component modulated by whatever periodic motion the reference point undergoes. Thus, if we know the average velocity of the person over several cycles, we can compute the spatiotemporal line of motion along which the periodicity can be observed. If the person moves with average velocity (u, v) the spatiotemporal line of motion will be determined by the equations $(x, y) = (u, v) * t + (x_0, y_0)$, where (x, y) is the position of the object in space at time t and (x_0, t_0) is the position at time zero. This applies to any object undergoing constant velocity locomotion.

3.1 Periodicity Detection

From Fourier theory we know that any periodic signal can be decomposed into a fundamental and harmonics. That is, we can consider the energy of a periodic signal to be concentrated at frequencies which are integral multiples of some fundamental frequency. This implies that if we compute the discrete Fourier transform of a sampled periodic signal, we will observe peaks at the fundamental frequency and its harmonics. Hence, in theory, the periodicity of a signal can be detected by obtaining its Fourier transform and checking whether all the energy in the spectrum is contained in a fundamental frequency and its integral multiples.

The real-world signals, however are seldom perfectly periodic. In the case of signals arising from activity in image sequences, disturbances can arise from errors in the uniform translation assumption, varying background and lighting behind a locomoting actor, and other sources. In addition, for computational purposes, we need to truncate the signal at some finite length which may not be an exact integral multiple of its period. Nevertheless, the frequency defined by the highest amplitude often represents the fundamental frequency of the signal. Hence we can get an idea of the periodicity in a signal by summing the energy at the highest amplitude frequency and its multiples, and comparing that quantity to the energy at the remaining frequencies. In practice, since peaks in a Fourier transform tend to be slightly broadened for a variety of reasons, including the finite length of the sample, we define the periodicity measure p_f of a signal f as a normalized difference of the sum of the power spectrum values at the highest amplitude frequency and its multiples, and the sum of the power spectrum values at the frequencies halfway between. That is,

$$p_f = (\sum_i F_{iw} - \sum_i F_{(iw+w/2)})/(\sum_i F_{iw} + \sum_i F_{iw+w/2})$$

where F is the energy spectrum of the signal f and w is the frequency corresponding to the highest amplitude in the energy spectrum.

The measure is normalized with respect to the total energy at the frequencies of interest so that it is one for a completely periodic signal and zero for a flat spectrum. In general, if a signal consists of frequencies other than one single fundamental and its multiples, its periodicity measure will be low.

Because the signal along any given reference curve in the image solid may be ambiguous, we need a way of combining periodicity measures of a number of signals from reference curves associated with the same actor. The simplest idea would be simply to sum the power spectra of the various signals, and apply the periodicity measure to the resultant curve. Unfortunately, this does not work, primarily because, although there is a fair amount of energy at the fundamental frequency, and quite a few signals in which high periodicity is present, there are also a lot of samples where the periodicity is not evident, or which appear periodic at some other frequency. The net affect, is that all this energy at other frequencies can swamp the main signal if they are combined additively. What does work, is a form of non-maximum suppression, where the periodicity measure is obtained for each power spectrum separately. Each frequency w is then assigned a value equal to the sum of the periodicity measures P_w from all the signals whose highest amplitude occurred at that frequency. The result is the same as suppressing all but the maximum frequency in each transform, weighting each by the periodicity measure of the signal, and summing them. The maximum value of this combined signal is taken as the fundamental frequency, and the associated periodicity measure is the average of the periodicity measures of the contributing signals.

Thus, the periodicity measure P for an entire image sequence is defined as

$$P = \max_w (P_w / n_w)$$

where n_w and P_w are the number of pixels at which the highest amplitude frequency is w and the sum of periodicity measures at those pixels respectively.

Finally, in order to apply the technique to real data, we need a way of extracting reference curves and the associated signals from an image sequence. In the following, we assumed that any activity that existed in the data would be either stationary, or locomotory in a manner that produced an overall translating motion. We also assumed that there was at most one actor in the scene, though a certain amount of background motion could be tolerated. A third assumption is that the viewing angle and the scene illumination does not change significantly so that the intensity along the reference curves remains periodic. The first assumption turns out not to be too restrictive – a large number of natural periodic activities fit into one of the two categories. The second can be relaxed with some additional preprocessing. Refer to the discussions section for how this can be achieved and how the other assumptions can be relaxed as well.

The first step of the algorithm is to identify locations in the scene where movement of any sort is occurring. This is done by computing the normal flow magnitude at each pixel between each successive pair of frames using a spatiotemporal differential method. Those pixels at which significant motion is present are marked, and the centroid of the marked pixels computed in each frame. The mean velocity (if any) of the actor is then computed by fitting a linear trajectory to the sequence of centroids. This is where the one-actor assumption comes into play. If several actors were present, simple clustering techniques could be used to isolate the regions in the scene corresponding to different activities. The reference curves were taken as the lines in the spatiotemporal solid parallel to that generated by the linear-fitted trajectory of the centroid. Signals were extracted along these curves, and those that displayed significant spread over a period of at least half as long as the signal were selected for processing. This had the effect of eliminating the need to process regions in which no motion occurred, as well as regions affected only by an occasional blip. The periodicity measures for all signals extracted is computed and are used in computing periodicity measure P for the entire image sequence as described above.

3.2 Experiments

We ran experiments on four different activities, and a number of non-periodic motions. The sequences were first recorded on video and then digitized later with suitable temporal sampling so that at least four cycles of the activity were captured in 128 frames. Following is a description of each activity and the conditions under which they were digitized.

- Walk: A person walking across a room viewed in profile. Six sequences of 128 frames of size 128x128 pixels were obtained. Half the sequences contained one person and the other half a second.
- Exercise: A person performing jumping jacks. Four sequences of 128 frames of 128x128 pixels, two each of two different people.
- Swing: A person swinging viewed from the side. Six sequences of 128 frames of 128x128 pixels, three each of two different people.
- Frog: A toy frog simulating swimming activity viewed from above. Four sequences of 128 frames of 64x256 pixels.
- Nonperiodic: Various sequences taken from television shows and live outdoor shots: splashing water, closeup of crowd at a political rally, a plane flying overhead, a robot hand picking up and manipulating objects (2 sequences), the input to an eye tracker (eyeball movements), leaves fluttering in the wind, turbulent flow in a stream. In all, 8 sequences of 128 frames of 128x128 pixels.

The swing and exercise activities were shot outdoors and contained background motion as well. Among the periodic activities, a single sequence of uniform rotation is included as well. Sample images of these activities are shown in figures 2 and 3.

The periodicity measures computed using the above algorithm are plotted for all 20 periodic and all 8 non-periodic sequences in figure 4. As is evident from the graphs and the projected scatter plot, the technique separates complex periodic from non-periodic motion nicely. The requirement that an empirically determined threshold be used is not a great drawback in this case, nor is it particularly surprising, since even the the intuitive notion of periodic activity falls on a continuum. Is the motion of a branch waving somewhat irregularly in the wind periodic or non-periodic? Here, we classified it as non-periodic, but it had one of the higher periodicity measures, as might be expected.

4 Discussion

Our periodic activity detection algorithm can be summarized as follows:

- *Input:* The input to the algorithm is a digitized image sequence consisting of 128 frames of resolution 128x128 pixels.
- *Output:* A periodicity measure indicating the amount of periodicity in observed in the image sequence. This is used to decide whether the image sequence contains a periodic activity and if so, to locate the region of the activity.
- *Step 1.* Compute normal flow magnitude at each pixel between each successive pair of frames using the differential method.
- *Step 2.* Mark pixels corresponding to significant motion in the scene by thresholding the normal flow magnitude. Compute centroid of the marked pixels in each frame. Compute the mean velocity (if any) of the actor by fitting a linear trajectory to the sequence of centroids. Take reference curves to be the lines in the spatiotemporal solid parallel to the linear trajectory of centroids of motion.
- *Step 3.* Extract gray-level signals along the reference curves. Compute the dominant frequency w and the periodicity measure P_w for each individual signal extracted.
- *Step 4.* Compute overall periodicity measure P for the image sequence using formula given in the last section.

We have assumed a number of things for the method to work correctly. First, we assumed that there is only one actor in the scene in approximately constant linear locomotion, and that the motion of the actor is significantly higher than that of the background motion. We also assumed that the viewing angle and scene illumination does not change significantly. Further, it is assumed that the entire image sequence consists of at least four cycles of the periodic activity if there is any. The following is a discussion of some of the merits of the algorithm and some approaches to deal cases where the assumptions are violated.

The method we described satisfies the several desirable invariances. It is invariant to image illumination, contrast, translation, rotation and scale. It is also invariant to the magnitude of locomotory motion and the speed of the activity. It is also fairly robust with respect to small changes in viewing angle. The periodicity measure does not depend on the number of pixels involved in the activity. If desired, a restriction on the minimum number of pixels can be imposed so that only activities of a minimum size can be recognized. The swing and exercise sequences were taken outdoors where there is a small amount of background motion. This comprises not only moving trees and plants, but also moving people and occasional crossing of a car. The thresholding stage on motion magnitude in step 2 of the algorithm (in our implementation one-half pixel per frame is used) eliminates small background motion, but it can not eliminate larger background motion such as produced by a car passing. That periodicity can be detected even in this case demonstrates that the technique is reasonably tolerant of background clutter and an occasional disturbance. The technique also provides a method for localizing activity in the scene by back-projecting the reference curves having high periodicity measures into the image solid.

So far we have assumed that the actors giving rise to the activity move with constant velocity along linear paths. The case of nonlinearly moving objects can be handled by tracking the object of interest given a coarse estimate of its initial location and velocity. This would generate reference curves that were not straight lines. We have already demonstrated the usefulness of the centroid of motion for computing the velocity of linearly moving objects. It could also be used for tracking the actors moving on more complex trajectories. Use of the motion centroid can be unreliable in estimating the centroid of the object if the shape of the object changes as it moves. In this case use of a prediction and correction mechanism using past values over a sufficiently long period can help.

The detection scheme also assumes that there is only one activity in the scene except for some background clutter. If there are multiple activities in the scene, this detection technique can still be applied provided the activities can be spatially isolated so that they do not interfere with each other. In this case they can segmented using the motion information and later tracked separately. Even an occasional crossing of different activities can be tolerated as long as the regions can be separated again later. In our experiments, the periodic activity samples consist of at least four cycles of the activity. Minimum four cycles were used to detect the actual frequency given that there is a considerable amount element of non-repetitive structure from the background in the case of translating actors.

The complexity of detection is proportional to the number of pixels involved in the activity. About half the work is computing the fast Fourier transforms at each of the pixels. Most of the rest of the time is occupied by the motion detection process. The detection procedure currently runs on an SGI machine using four processors and it take approximately 15 seconds to process a 128 frame sequence of 128x128 images.

4.1 Recognition of Activities

The first stage in recognizing an activity is to detect that an activity exists, and localize it in the scene. This paper has described a technique for accomplishing this. Future work will utilize information computed in the detection stage for recognition and classification of specific activities. The detection scheme utilizes only the magnitude of the Fourier transform to obtain the periodicity measure. The phase of the Fourier transform is also computed at each location in the image and we propose to use this information along with other low-level information in the image, for recognition. For example, walking can be described as a sequence of motion events regularly occurring at each spatial location. The cycle of motion events at different spatial locations in the image have a fixed phase difference. These phase differences are valuable in characterizing the activities.

5 Conclusion

We have described a method of activity detection. This technique uses a periodicity measure on gray-level signals extracted along spatiotemporal reference curves. We have illustrated the technique using real-world examples of activities, and shown that it robustly detects complex periodic activities, while excluding non-periodic motion. We proposed a technique to recognize these activities using the detection scheme described here. It is not clear how much the periodicity alone is useful for recognition but we believe the phase information is valuable for activity recognition. Future work will concentrate on the development of robust phase features that can be used in conjunction with previously developed motion and gray-level features to classify activities.

References

[Allmen and Dyer, 1990] M. Allmen and C.R. Dyer. Cyclic motion detection using spatiotemporal surface and curves. In *Proc. Int. Conf. on Pattern Recognition*, pages 365–370, 1990.

[Anderson *et al.*, 1985] C. H. Anderson, P. J. Burt, and G. S. van der Wal. Change detection and tracking using pyramid transform techniques. In *Proc. SPIE Conference on Intelligent Robots and Computer Vision*, pages 300–305, 1985.

[Chun, 1986] H.W. Chun. A representation for temporal sequence and duration in massively parallel networks: Exploiting link connections. In *Proc. AAAI*, 1986.

[Cutting, 1981] J.E. Cutting. Six tenets for event perception. *Cognition*, pages 71–78, 1981.

[Elaman, 1988] J.E. Elaman. Finding structure in time. Technical Report 8801, Center for Research in Language, Univ. of California, San Diego, 1988.

[Ewart, 1987] J.P. Ewart. Neuroethology of releasing mechanisms: Prey-catching in toads. *Behavioral and Brian Sciences*, 10:337–405, 1987.

[Feldman, 1988] J.E. Feldman. Time, space and form in vision. Technical Report 244, University of Rochester, Computer Science Department, 1988.

[Goddard, 1989] N.H. Goddard. Representing and recognizing event sequences. In *Proc. AAAI Workshop on Neural Architectures for Computer Vision*, 1989.

[Gould and Shah, 1989] K. Gould and M. Shah. The trajectory primal sketch: A multi-scale scheme for representing motion characterestics. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 79–85, 1989.

[Hildreth and Koch, 1987] E.C. Hildreth and C. Koch. The analysis of visual motion from computational theory to neural mechanisms. *Annual Review of Neuroscience*, 1987.

[Hoffman and Flinchbuagh, 1982] D.D. Hoffman and B.E. Flinchbuagh. The interpretation of biological motion. *Biological Cybernatics*, pages 195–204, 1982.

[Johansson, 1973] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.

[Juang and Rabiner, 1985] B.H. Juang and L.R. Rabiner. Mixture autoregressive hidden markov models for speech signals. *IEEE Trans. Acoustics, Speech and Signal Processing*, 6:1404–1413, 1985.

[Koller *et al.*, 1991] D. Koller, N. Heinze, and H.-H. Nagel. Algorithmic characterization of vehicle trajectories from image sequences of motion verbs. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pages 90–95, 1991.

[O'Rourke and Badler, 1980] J. O'Rourke and N.I. Badler. Model-based image analysis of human motion using constraint propagation. *PAMI*, 3(4):522–537, 1980.

[Pentland and Mase, 1989] A. Pentland and K. Mase. Lip reading: Automatic visual recognition of spoken words. Technical Report 117, M.I.T. Media Lab Vision Science, 1989.

[Polana and Nelson, 1992] R. Polana and R.C. Nelson. Temporal texture recognition. In *Proc. of CVPR*, pages 129–134, 1992.

[Rashid, 1980] R.F. Rashid. *LIGHTS: A System for Interpretation of Moving Light Displays*. PhD thesis, Computer Science Dept, University of Rochester, 1980.

[Tank and Hopfield, 1987] D. W. Tank and J. J. Hopfield. Concentrating information in time: analog neural networks with applications to speech recognition problems. In *Proceedings of the First International Conference on Neural Networks*, pages 455–468, 1987.

Acknowledgements

This work is supported by contracts NSF IRI-9010692 and AFOSR 91-0288.

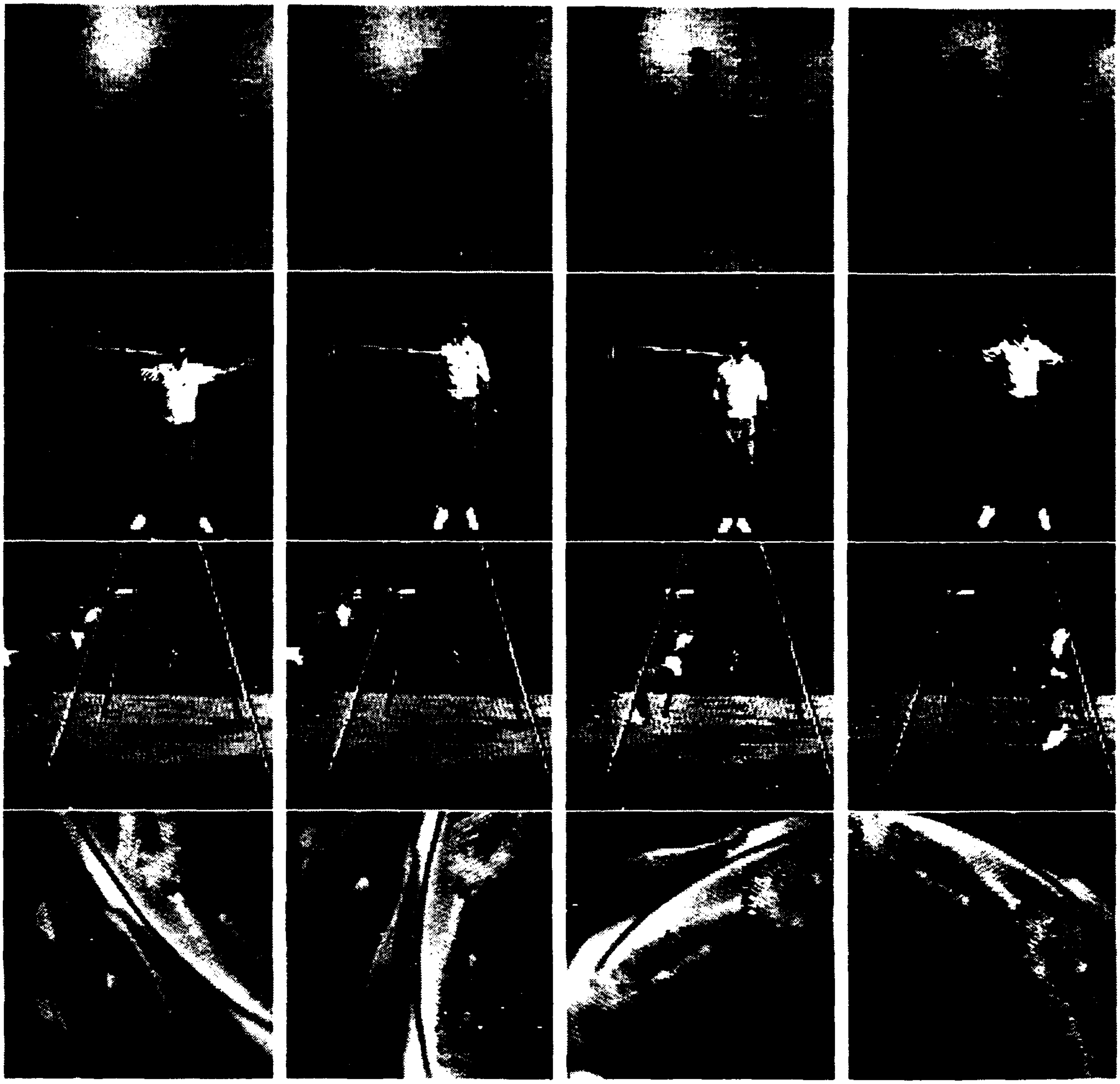

Figure 2: Sample images from periodic sequences: walk, exercise, swing and rotation

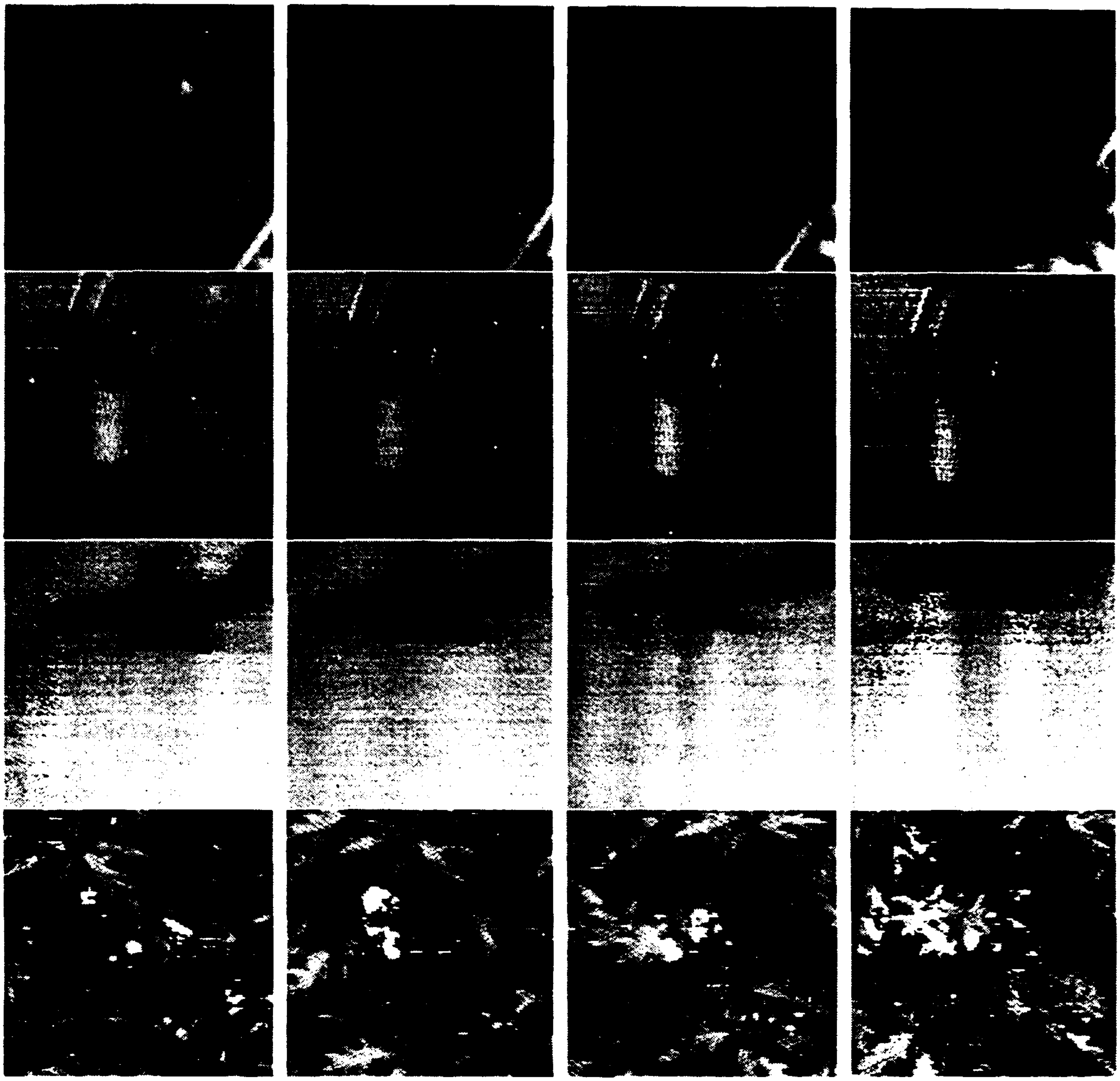

Figure 3: Sample images from nonperiodic sequences: people, robot hand, plane and leaves

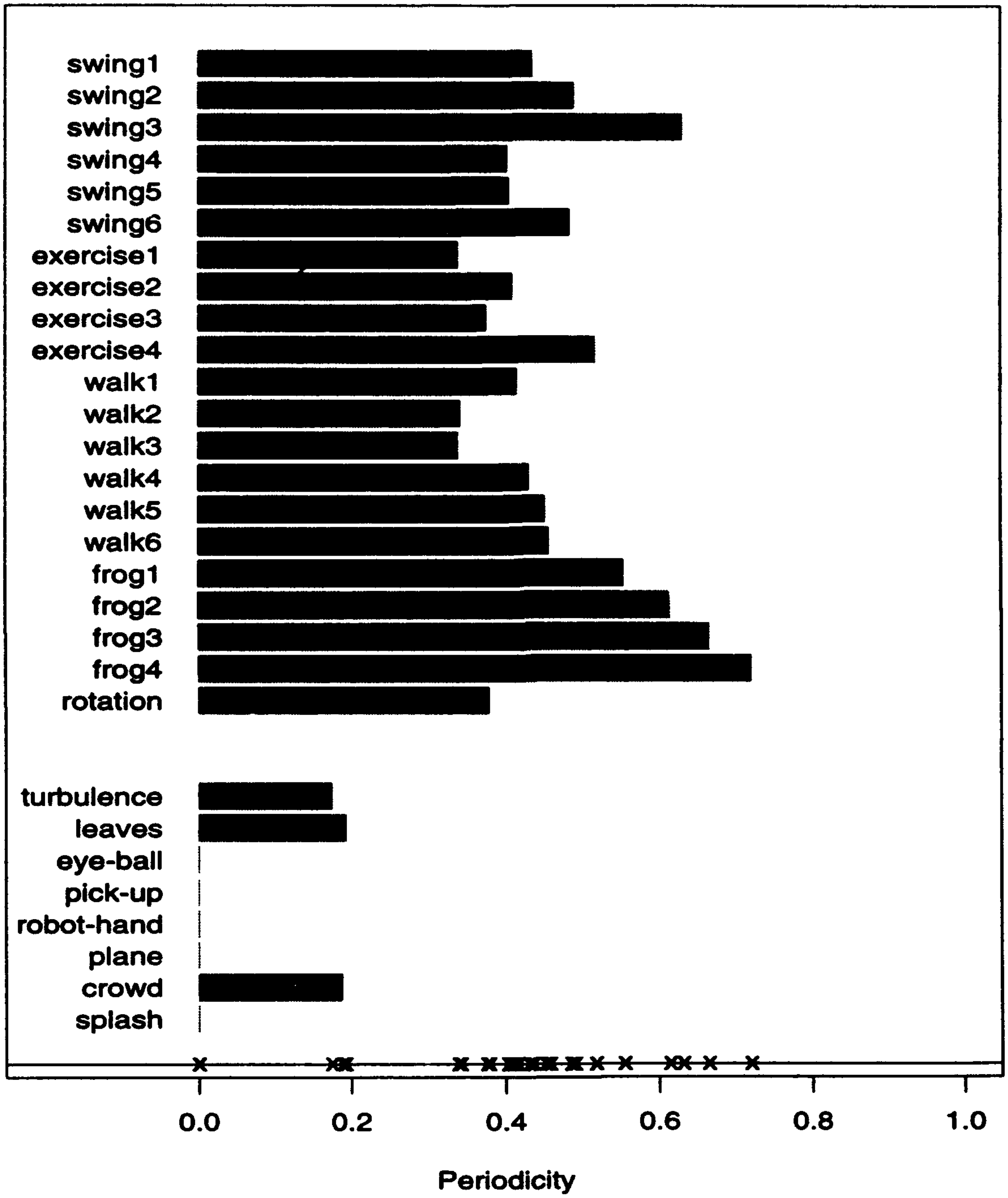

Figure 4: Periodicity measure for Periodic and Nonperiodic sequences

Recognizing Activities

Ramprasad Polana and Randal Nelson

Department of Computer Science
University of Rochester
Rochester, New York 14627
Email: polana@cs.rochester.edu and nelson@cs.rochester.edu

Abstract

The recognition of repetitive movements characteristic of walking people, galloping horses, or flying birds is a routine function of the human visual system. It has been demonstrated that humans can recognize such activity solely on the basis of motion information. We demonstrate a general computational method for recognizing such movements in real image sequences using what is essentially template matching in a motion feature space coupled with a technique for detecting and normalizing periodic activities. This contrasts with earlier model-based approaches for recognizing such activities.

1 Introduction

The motion recognition ability of the human visual system is remarkable. People are able to distinguish both highly structured motion, such as those produced by walking, running, swimming or flying animals and birds, and more statistical patterns such as those due to blowing snow, flowing water or fluttering leaves. The classic demonstration of pure motion recognition by humans is provided by Moving Light Display experiments [Johansson, 1973], where human subjects were able to distinguish activities such as walking, running or stair climbing, from lights attached to the joints of an actor. More subtle movement characteristics can be distinguished as well. For example, human observers can identify the actor's gender, and even identify the actor if known to them, by his or her gait. Similar discrimination abilities using motion alone have been observed in non-human animals as well [Ewart, 1987]. This biological use of motion probably reflects the fact that for certain tasks, visual motion provides more effective cues than other modes of visual perception. Motion is a particularly useful cue for certain types of recognition due to the fact that it is relatively easy to extract the motion field independent of illumination and shading of the image.

As a first step towards motion recognition by a machine, we define three common classes of visual motion on the basis of the spatial and temporal regularity of the signal. Different recognition techniques apply to the different classes. We define the first class, *temporal textures* to be motion patterns that exhibit statistical regularity but have indeterminate spatial and temporal extent. Examples of temporal textures include wind blown trees or grass, turbulent flow in cloud patterns, ripples on water, the motion of a flock of birds etc. The second class, *activities*, consists of motion patterns that are temporally periodic and possess compact spatial structure. Examples of activities include walking, running, rotating or reciprocating machinery, etc. A third class *motion events* consists of isolated simple motions that do not exhibit any temporal or spatial repetition. Examples of motion events are isolated instances of opening a door, starting of a car, throwing a ball etc. On can imagine other combinations of attribute, e.g. spatially periodic and temporally limited, but these to not seem to occur broadly in natural visual environments.

It turns out that temporal textures can be effectively treated with statistical techniques analogous to those used in gray-level texture discrimination. A previous paper [Polana and Nelson, 1992] describes this. Activities and motion events, on the other hand, are more discretely structured, and techniques similar to those used in static object recognition would be expected to be useful in their classification.

In this paper, we describe a robust method for recognizing activities, including ones, such as walking, that involve simultaneous translation of the actor. In an earlier paper [Polana and Nelson, 1993], we described an algorithm to detect periodic activities in an image sequence making use of the periodic nature of the activity. The recognition algorithm utilizes the periodic activity detection algorithm as a first step in the computation of a normalized a feature vector which is then used to classify detected activity as one of several known activities.

Motion recognition algorithms. both for temporal texture and activity, have potential applications in several areas. One area is automated surveillancer. Motion detection via image differencing can be used for intruder detection: however such systems are subject to false alarms. especially in outdoor environments, since the system is triggered by anything that moves, whether it is a person, a dog, or a tree blown by the wind. Motion recognition techniques can be used disambiguate such situations. Another application is in industrial monitoring. Many manufacturing operations involve a long sequence of simple operations each performed repeatedly and at high speed by a specialized mechanism at a particular location. It should be possible to set up one or more fixed cameras that cover the area of interest, and to characterize the allowed motions in each region of the image(s).

2 Related Work

Although motion plays an important role in biological recognition tasks, motion recognition in general, has received little attention in the literature compared to the volume of work on static object recognition. Most computational motion work in motion in fact, has been concerned with various aspects of the structure-from-motion problem. There is a large body of psychophysical literature addressing the perception of motion, most of it concerned with primitive percepts. A modest amount of this work addresses more complicated motion recognition issues [Johansson, 1973, Cutting, 1981, Hoffman and Flinchbuagh, 1982, Hildreth and Koch, 1987], but the models and descriptions have typically not been implemented. Various computational models of temporal structure, have been proposed (e.g. [Chun, 1986, Feldman, 1988]) but much of this work is at a fairly high level of abstraction, and has not actually been applied to visual motion recognition except in rather artificial tests.

A specialized area that has seen some attention is the interpretation of moving light displays. Goddard [1989] considers recognizing event sequences from such images. His work addresses the representation of motion event sequences and their recognition assuming certain invariant image features. His input consists of the joint angles and angular velocities computed from the motion of the dots in the light displays. The joint angles and angular velocities are invariant to rotation in the image plane, scale and translation. A challenging part in computing these invariants is to recover the connectivity of the individual dots (by body parts) in the MLD images. A domain independent approach to this problem is given by Rashid [Rashid, 1980, O'Rourke and Badler, 1980]. This work considers the computational interpretation of moving light displays, particularly in the context of gait determination. This work emphasized rather high-level symbolic models of temporal sequences, an approach made possible by the discrete nature of the representation. The results were quite sensitive to discrete errors and thus highly dependent on the ability to solve the correspondence problem and accurately track joint and limb positions. This severely limits the general applicability of the method.

A few studies have considered highly specific aspects of motion recognition computationally. Anderson et al. [Anderson *et al.*, 1985] describe a method of change detection for surveillance applications based on the spectral energy in a temporal difference image. This was not generalized to other motion features or more sophisticated recognition. Pentland [Pentland and Mase, 1989] considered lip reading, and implemented a system that could recognize spoken digits with 70%-90% accuracy over 5 speakers. The system required the location of the lips to be entered by hand, and depended on an explicitly constructed lip model. Some temporal pattern recognition work has been done in the context of speech processing [Juang and Rabiner, 1985, Tank and Hopfield, 1987, Elaman, 1988], but the applicability of the techniques to motion recognition has not been considered.

Finally, there is a body of work based on the analysis of trajectories. Koller, Heinze and Nagel [1991] developed a system that tracks moving vehicles and characterizes their trajectory segments in terms of natural language concepts. Gould and Shah [1989] represent motion characteristics of moving objects by recording the important events in their trajectory. They propose the use of the resulting *trajectory primal sketch* in a motion recognition system. Allmen and Dyer have developed a method of extracting spatiotemporal curves corresponding to moving objects and applied the technique to detection of cyclic motions [Allmen and Dyer, 1990]. Tsai et al. [Tsai *et al.*, 1993] have also worked on cyclic motion detection using curvature trajectories to detect cycles by means of Fourier domain techniques. All the above require the difficult task of robustly computing the trajectories or spatiotemporal curves from image sequences before attempting recognition and the demonstrations of their techniques involve principally synthetic image sequences.

3 Detecting Activities

The first step in recognizing an activity is to determine that an activity exists, and localize it in the scene. In an earlier paper we have described a technique for accomplishing this [Polana and Nelson, 1993]. The present work will utilize the information computed in the detection stage for recognition and classification of specific activities.

Activities involve a regularly repeating sequence of motion events. If we consider an image sequence as a spatiotemporal solid with two spatial dimensions x, y and one time dimension t, then repeated activity tends to give rise to periodic or semi-periodic gray level signals along smooth curves in the image solid. We refer to these curves as *reference curves*. If these curves could be identified and samples extracted along them over several cycles, then frequency domain techniques could be used in order to judge the degree of periodicity.

To clearly define the reference curves, we need to formalize the concept of a periodic object. An object is defined as a set of points P. Associated with each $p \in P$ is a function $X_p(t)$ giving its location (in a fixed 3D coordinate system) as a function of time. A stationary periodic object (ie. a stationary object exhibiting periodic activity) has the property that $X_p(t) = X_p(t + \tau)$ for all $p \in P$, where τ is the time period for one cycle of the activity and is independent of p. A slight generalization gives us a definition for a translating periodic object. Such an object has the property that $X_p(t) = Y_p(t) + Z(t)$. where Y_p satisfies $Y_p(t) = Y_p(t + \tau)$ and $Z(t)$ is a path in 3D space independent of p. It can be assumed that $Z(0) = 0$ so that $X_p(0) = Y_p(0)$. Intuitively, a periodic object characterized by $Y_p(t)$ is translated along the path $Z(t)$ (we are assuming the object does not undergo any rotation and the viewing angle does not change).

If we can determine the translational path of the object by some sort of tracking procedure, then we need only consider stationary periodic objects as shown by the equation: $X_p(t) - Z(t) = Y_p(t) = Y_p(t + \tau) = X_p(t + \tau) - Z(t + \tau)$. More formally, corresponding to each point p of a translating periodic object, we define a 3D-reference curve $R_p(t)$ to be the path $X_p(0) + Z(t)$. We also define a 2D-reference curve $r_p(t)$ corresponding to a point p of the object, to be the projection of $R_p(t)$ onto the image plane over time (hence $r_p(t)$ is a curve in (x, y, t) space). The gray-level signal along the 2D-reference curve $r_p(t)$ is determined by the set of points of the object that appear along the 3D-reference curve $R_p(t)$. It can be shown that the same set of points of the object recur periodically along each reference curve $R_p(t)$. For example, the point p is on the reference curve $R_p(t)$ at time zero, and it coincides with the reference curve at regular intervals of τ (since $X_p(\tau) = Y_p(\tau) + Z(\tau) = Y_p(0) + Z(\tau) = X_p(0) + Z(\tau)$). Similarly, every other point of the object on the reference curve $R_p(t)$ recurs along $R_p(t)$ at intervals of τ.

Given an image sequence containing a moving object, the detection scheme works as follows: First, the object is tracked using a low-level process based on aggregation of moving pixels. The track is used to generate reference curves and sample motion signals are extracted along them. Each of the signals is processed using frequency domain techniques to compute a measure of periodicity. The periodicity measures of individual signals are combined to produce a periodicity measure for the entire tracked object, which is then thresholded to decide whether a periodic activity is present in the sequence.

The following is a step-by-step description of the periodic activity detection algorithm:

- *Input:* The input to the algorithm is a digitized 256-level gray-valued image sequence.
- *Output:* A periodicity measure indicating the amount of periodicity in observed in the image sequence. This is used to decide whether the image sequence contains a periodic activity and if so, to locate the region of the activity.
- *Step 1.* Compute normal flow magnitude at each pixel between each successive pair of frames using a differential method.
- *Step 2.* Mark pixels corresponding to significant motion in the scene by thresholding the normal flow magnitude. Compute centroid of the marked pixels in each frame. Compute the mean velocity (if

any) of the actor by fitting a linear trajectory to the sequence of centroids. Take reference curves to be the lines in the spatiotemporal solid parallel to the linear trajectory of centroids of motion. This simple tracking process is currently adapted to a single actor moving linearly, but is easily extended to multiple actors and other paths as long as the tracks are smooth. and the actors are separated most of the time.

- *Step 3.* Extract motion signals along the reference curves. Compute the dominant frequency and the periodicity measure for each individual signal extracted. We define the periodicity measure p_f of a signal f as a normalized difference of the sum of the power spectrum values at the highest amplitude frequency and its multiples, and the sum of the power spectrum values at the frequencies halfway between. That is,

  $$p_f = (\sum_i F_{iw} - \sum_i F_{(iw+w/2)})/(\sum_i F_i)$$

  where F is the energy spectrum of the signal f and w is the frequency corresponding to the highest amplitude in the energy spectrum.

- *Step 4.* For each frequency w assign a value equal to the sum of the periodicity measures P_w from all the signals whose highest amplitude occurred at that frequency. Compute overall periodicity measure P for the image sequence using formula $P = \max_w(P_w/n_w)$ where n_w and P_w are the number of pixels at which the highest amplitude frequency is w and the sum of periodicity measures at those pixels respectively.

A more complete discussion of the periodicity detection process and the assumptions made can be found in the previously cited paper.

4 Recognizing Activities

Once an activity has been identified and tracked in a scene, the next step is to recognize it. The tracking and periodicity detection algorithms provide spatial and temporal normalization that can be used to simplify the recognition procedure. In particular, recall that the periodicity detection procedure provides a periodicity measure for each active pixel in a tracked object. By backprojecting this measure, we can locate the pixels in each frame that display periodicity at the dominant frequency. Since these pixels are likely to belong to the actor of interest, we can use this backprojection to refine our initial segmentation, which was based solely on aggregate motion. By fitting a frame to this refined segmentation we compensate for variation in spatial scale and position. Currently this is done on the assumption that the distance of the actor does not change significantly over the sample (typically 4 cycles), but a simple change in the frame-fitting procedure can allow for smooth scale change as might result during an approaching motion. Similarly, the fundamental frequency allows us to frame the activity in time, and compensate for variation in temporal scale (i.e. frequency).

The end result of the normalization procedure is a spatio-temporal solid containing the activity of interest in a form that is invariant to spatial scale, spatial translation, and temporal scale. The next step is to compute a descriptor for this solid that can be used to classify the activity it represents. This sounds like a three dimensional template match, and in fact, with the appropriate motion features in the slots of the template, such an approach works well. Essentially, we capitalize on the fact that a periodic activity is characterized by regularly repeating motion events that have fixed spatial and temporal relationships to each other.

In more detail, the process is as follows. We divide one cycle of the spatio-temporal solid representing the activity into XxYxT cells by partitioning the two spatial dimensions into X, Y divisions respectively and the temporal dimension into T divisions. We then select a local motion statistic and compute the same statistic in each cell of the spatiotemporal grid. The feature vector in this case is composed of XYT elements each of which is the value of the statistic in a particular cell – essentially a three dimensional template.

One issue that affects the measures described above is the fact that so far, the normalized spatio-temporal solid, while corrected for temporal scale (frequency) is not corrected for temporal translation (phase). There are a couple of ways to handle this. One is to pick some robust temporal feature to define zero phase, and normalize all samples with respect to this feature. One feature that works fairly robustly is to take the time of maximum difference between total motion in the left and right half fields. Alternatively, since the pattern matching phase of the algorithm currently represents only a small fraction of the total computational effort, and the temporal resolution of the pattern is typically small (i.e., less than 10 samples per cycle), we can simply try a match at each possible phase and pick the best. We have found in our experiments that this method works better than the first. Hence, the results are reported using this kind of matching only.

We experimented with three different local statistics. The first was the dominant motion direction in each cell. This is approximated by computing the histogram of normal flow directions weighted by the corresponding normal flow magnitude and selecting the direction with highest histogram value. The second statistic represented the summed motion magnitude in the dominant motion direction. The third statistic is simply the summed normal flow magnitude in each cell. The directional information is ignored in this case. As it turned out, this last statistic, which is some ways the simplest, worked best.

4.1 Experiments

We ran experiments on seven different types of activities. The image sequences were first recorded on video and then digitized later with suitable temporal sampling so that at least four cycles of the activity were captured in 128 frames. Following is a description of each activity and the conditions under which they were digitized.

- Walk: A person walking on a treadmill.
- Exercise: A person exercising on a machine.
- Jump: A person performing jumping jacks.
- Swing: A person swinging viewed from the side.
- Run: A person running on a treadmill.
- Ski: A person skiing on a skiing machine.
- Frog: A toy frog simulating swimming activity viewed from above.

All samples were digitized at a spatial resolution of 128x128 pixels, except those for walk and run which were digitized at a resolution of 64x128 pixels. Pixels were 8 bit gray levels. The swing and exercise activities were shot outdoors and contained background motion.

We first digitized eight samples of each activity by the same person under the same conditions with respect to scene illumination, background, and camera position. We created the reference database taking half of the samples belonging to each activity. In other words, the reference database consists of four samples of each of the seven activities. Sample images of these activities are shown in figures 1. The remaining four samples of each activity are used to create the test database. In addition, we digitized four samples of walking by a different person and eight samples of the frog under different lighting conditions and different background and foreground gradients. These samples also differed from the reference database in frequency, speed of motion, and spatial scale. Examples of these samples are shown in Figure 2 These samples were added to the test database. The samples in the test database were classified by a nearest centroid classification technique using the samples in the reference database as training set.

We conducted experiments using the three local motion statistics described above. In each case the feature vector consists of the local statistic computed over each of a set of cells constituting a partition of the spatio-temporal solid. We divided each spatial dimension into four divisions and the temporal dimension into six divisions, so that we get a feature vector of length 96. To reiterate, the three local statistics were: direction of maximum motion (where the directions are quantized into eight sectors), the motion magnitude in maximum motion direction, and total motion magnitude in each cell. Sample features vectors are illustrated in 3 using the total motion magnitude statistic for a walk and a run sequence.

We initially computed the feature vectors by finding a zero phase marker within a cycle using the method described previously. However, more reliable results were achieved by matching each test feature vector with the reference feature vector six times, corresponding to different temporal offsets, and choosing the best match obtained. The results reported below utilize the latter method. This classification resulted in correct classification of every sample in the test database, including the samples using a different actor and different backgrounds, which were not represented in the reference database.

The results of classification using different variations are shown in terms of percentage of test cases correctly classified in table 1. Somewhat to our surprise, the simplest statistic - the total motion magnitude gave better results than either of the statistics involving direction of motion. The reason for this turned out to be related to the resolution of our images. In order to digitize enough frames to test the technique, we had subsampled the images to 128 x 128 pixels. After filtering for periodicity, significant motion, and direction, it was often the case that few pixels with all these properties were left in any one cell, which made for a large amount of stochastic noise in the signal. Simply put, we didn't have high enough resolution data to appropriately utilize the more specific statistics.

The percentage of correct classification does not give a full indication for the quality of classification. Hence, we also illustrate the results by the confusion matrix which shows how closely test samples belonging to various classes match the reference samples of the different classes. The confusion matrix using the total motion magnitude statistic is shown in Figure 4. A large square indicates a good match. As can be seen from this table, some motions, for instance the swimming frog, do not resemble anything else in the database, while others, for instance running and skiing, are more likely to be confused. The results seem to correspond more or less to human intuition about how similar the motions are.

Feature vector	*Total Test Samples*	*Correct Classified*	*Percent Success*	*Failures*
direction of maximal motion	40	32	80	walk by different actor and frog under different gradients
magnitude in maximal direction	40	39	97.5	walk by different actor
total motion magnitude	40	40	100	None

Table 1: Classification results

5 Discussion

The following is a step-by-step description of the periodic activity recognition algorithm:

- *Input:* The input to the algorithm is a digitized 256-level gray-valued image sequence consisting of at least four cycles of a periodic activity.

- *Output:* A known class into which the activity is classified by the algorithm.

Added Clutter Percentage	*Total Test Samples*	*Successfully Detected*	*Correctly Classified*
25	4	3	3
50	4	3	3
75	4	2	0
100	4	2	0
150	4	1	0
200	4	0	0

Table 2: Classification results with motion clutter (samples are of walk)

- *Step 1.* Compute normal flow magnitude at each pixel between each successive pair of frames using a differential method.

- *Step 2.* Locate and track the activity in the image sequence using periodicity detection algorithm described in section 2.

- *Step 3.* Normalize the activity using pixels exhibiting periodic motion and compute a feature vector.

- *Step 4.* Classify the activity using nearest centroid algorithm.

The method we have described displays several desirable invariances. It is robust to varying image illumination and contrast because the method uses only motion information which is invariant to these. It is also invariant to spatial and temporal translation and scale due to the normalization of the feature vectors, and the multiple temporal matching. It is also fairly robust with respect to small changes in viewing angle. The swing and exercise sequences were taken outdoors where there is a small amount of background motion. This comprises not only moving trees and plants, but also moving people and an occasional crossing of a car. That the activities can be detected even in this case demonstrates that the technique is somewhat tolerant of background clutter and the occasional disturbance.

To understand how much background clutter can be tolerated by this technique, we have experimented with the walk samples by adding motion clutter produced by blowing leaves This structured motion clutter is added in a controlled fashion so that its mean magnitude represents a varying percentage of the mean magnitude of the signal, and the resulting samples are classified using the total motion magnitude statistic. The results are tabulated in 2. The results show that the recognition scheme can tolerate motion clutter whose magnitude is equal to one half that of the activity, and it displays degraded, but still useful performance for even higher clutter magnitudes.

We have assumed that the actors giving rise to the activity move with constant velocity along linear paths. The case of nonlinearly moving objects can be handled by tracking the object of interest given a coarse estimate of its initial location and velocity, (e.g. with a Kalman filter). This would generate reference curves that are not straight lines. We have already demonstrated the usefulness of the centroid of motion for computing the velocity of linearly moving objects, and providing a rough initial segmentation. It could also be used for tracking the actors moving on more complex trajectories. Use of the motion centroid can be unreliable in estimating the centroid of the object if the shape of the object changes as it moves. In this case use of a prediction and correction mechanism using past values over a sufficiently long period can help.

The detection scheme also assumes that there is only one activity in the scene except for some background clutter. If there are multiple activities in the scene, this detection technique can still be applied provided the activities can be spatially isolated so that they do not interfere with each other. In this case they can be segmented using the motion information and tracked separately. If a predictive tracker is used, an occasional crossing of different activities can be tolerated as long as the regions can be separated again later. In our experiments, the periodic activity samples consist of at least four cycles of the activity. Four cycles were needed to reliably detect the fundamental frequency given that there is a considerable amount of non-repetitive structure from the background in the case of translating actors.

The complexity of recognition is proportional to the number of pixels involved in the activity. More than half the work is computing the motion vectors at every pixel and then computing the fast Fourier transforms at each of moving pixels. The remaining time is spent computing the feature vector, the time for which depends on the local motion statistic computed. For a 128 image sequence, computation of the feature vector of motion magnitudes takes about 3 seconds. The classification algorithm currently runs on an SGI machine using four processors and it takes maximum 20 seconds to process a 128 frame sequence of 128x128 images.

6 Conclusion

We have described a general technique for periodic activity recognition. This technique uses a periodicity measure to detect the activity and then a feature vector based on motion information to classify the activity into one of several known classes. We have illustrated the technique using real-world examples of activities, and shown that it robustly recognizes complex periodic activities.

References

[Allmen and Dyer, 1990] M. Allmen and C.R. Dyer. Cyclic motion detection using spatiotemporal surface and curves. In *Proc. Int. Conf. on Pattern Recognition*, pages 365–370, 1990.

[Anderson *et al.*, 1985] C. H. Anderson, P. J. Burt, and G. S. van der Wal. Change detection and tracking using pyramid transform techniques. In *Proc. SPIE Conference on Intelligent Robots and Computer Vision*, pages 300–305, 1985.

[Chun, 1986] H.W. Chun. A representation for temporal sequence and duration in massively parallel networks: Exploiting link connections. In *Proc. AAAI*, 1986.

[Cutting, 1981] J.E. Cutting. Six tenets for event perception. *Cognition*, pages 71–78, 1981.

[Elaman. 1988] J.E. Elaman. Finding structure in time. Technical Report 8801. Center for Research in Language, Univ. of California, San Diego, 1988.

[Ewart, 1987] J.P. Ewart. Neuroethology of releasing mechanisms: Prey-catching in toads. *Behavioral and Brian Sciences*, 10:337–405, 1987.

[Feldman, 1988] J.E. Feldman. Time, space and form in vision. Technical Report 244, University of Rochester, Computer Science Department, 1988.

[Goddard, 1989] N.H. Goddard. Representing and recognizing event sequences. In *Proc. AAAI Workshop on Neural Architectures for Computer Vision*, 1989.

[Gould and Shah, 1989] K. Gould and M. Shah. The trajectory primal sketch: A multi-scale scheme for representing motion characterestics. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 79–85, 1989.

[Hildreth and Koch, 1987] E.C. Hildreth and C. Koch. The analysis of visual motion from computational theory to neural mechanisms. *Annual Review of Neuroscience*, 1987.

[Hoffman and Flinchbuagh, 1982] D.D. Hoffman and B.E. Flinchbuagh. The interpretation of biological motion. *Biological Cybernatics*, pages 195–204, 1982.

[Johansson, 1973] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.

[Juang and Rabiner, 1985] B.H. Juang and L.R. Rabiner. Mixture autoregressive hidden markov models for speech signals. *IEEE Trans. Acoustics, Speech and Signal Processing*, 6:1404–1413, 1985.

[Koller *et al.*, 1991] D. Koller, N. Heinze, and H.-H. Nagel. Algorithmic characterization of vehicle trajectories from image sequences of motion verbs. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pages 90–95, 1991.

[O'Rourke and Badler, 1980] J. O'Rourke and N.I. Badler. Model-based image analysis of human motion using constraint propagation. *PAMI*, 3(4):522–537, 1980.

[Pentland and Mase, 1989] A. Pentland and K. Mase. Lip reading: Automatic visual recognition of spoken words. Technical Report 117, M.I.T. Media Lab Vision Science, 1989.

[Polana and Nelson, 1992] R. Polana and R.C. Nelson. Temporal texture recognition. In *Proc. of CVPR*, pages 129–134, 1992.

[Polana and Nelson, 1993] R. Polana and R.C. Nelson. Detecting activities. In *Proc. of CVPR*, pages 1–6, 1993.

[Rashid, 1980] R.F. Rashid. *LIGHTS: A System for Interpretation of Moving Light Displays*. PhD thesis, Computer Science Dept, University of Rochester, 1980.

[Tank and Hopfield, 1987] D. W. Tank and J. J. Hopfield. Concentrating information in time: analog neural networks with applications to speech recognition problems. In *Proceedings of the First International Conference on Neural Networks*, pages 455–468, 1987.

[Tsai *et al.*, 1993] P.-S. Tsai, M. Shah, K. Keiter, and T. Kasparis. Cyclic motion detection. Technical Report CS-TR-93-08, Computer Science Dept, University of Central Florida, 1993.

Acknowledgements

This work is supported by contracts NSF IRI-9010692 and AFOSR 91-0288.

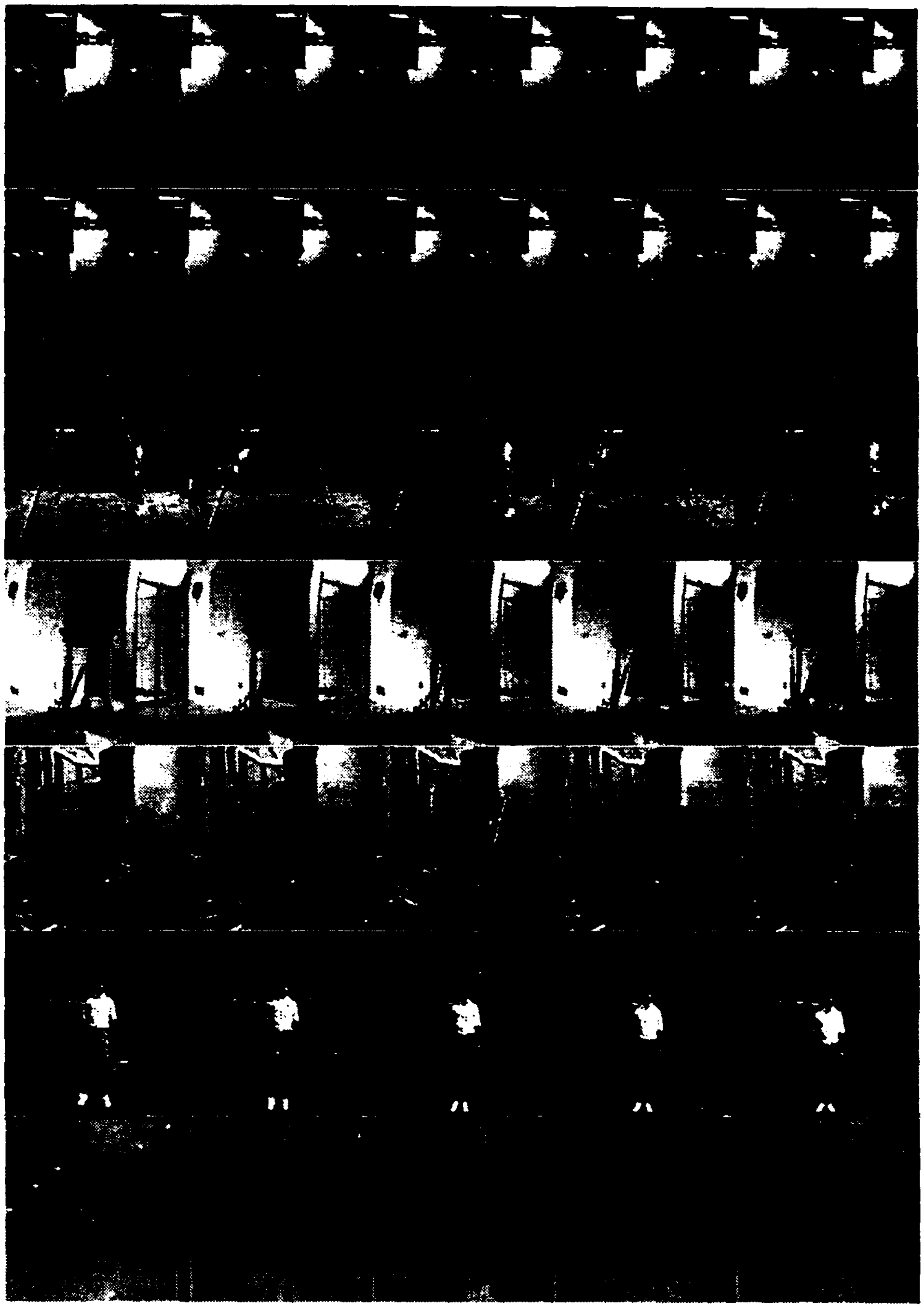

Figure 1: Sample images from periodic activities: walk, run, swing, jump, ski, exercise and toy frog

Figure 2: Sample images of walk by a different actor and toy frog under different background and frequency

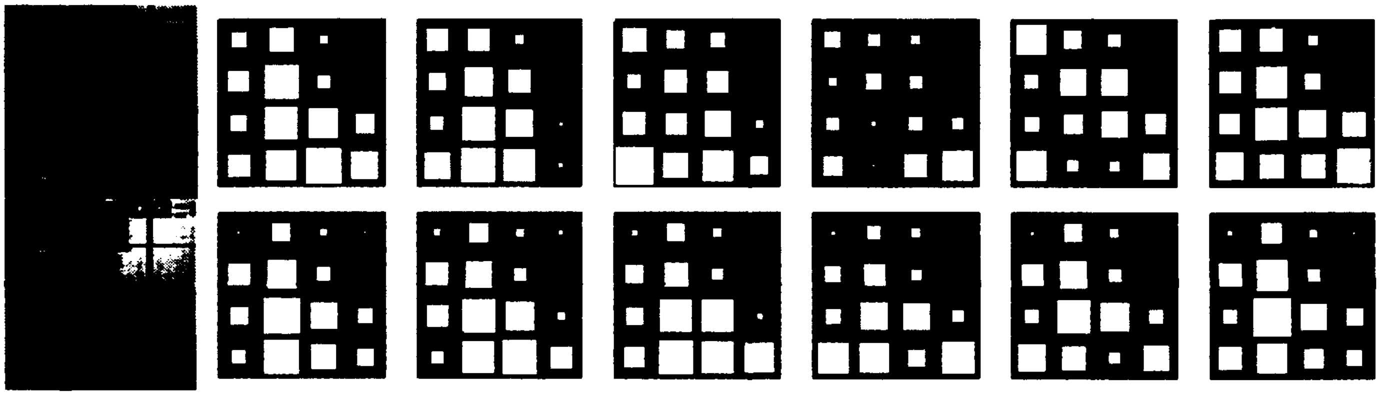

Figure 3: Sample total motion magnitude feature vector for a sample of walk (top) and a sample of run (bottom), one cycle of activity is divided into six time divisions shown horizontally, each frame shows spatial distribution of motion in a4x4 spatial grid (size of each square is proportional to the amount of motion in the neighborhood).

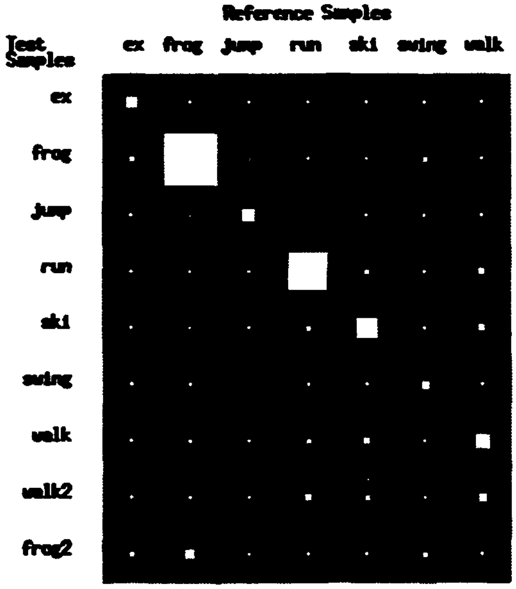

Figure 4: Confusion matrix for the feature vector using total motion magnitude