Synergizing Spatial and Temporal Texture

Chin-Hwee Peh and Loong-Fah Cheong

***Abstract*—Temporal texture accounts for a large proportion of motion commonly experienced in the visual world. Current temporal texture techniques extract primarily motion-based features for recognition. We propose in this paper a representation where both the spatial and the temporal aspects of texture are coupled together. Such a representation has the advantages of improving efficiency as well as retaining both spatial and temporal semantics. Flow measurements form the basis of our representation. The magnitudes and directions of the normal flow are mapped as spatiotemporal textures. These textures are then aggregated over time and are subsequently analyzed by classical texture analysis tools. Such aggregation traces the history of a motion which can be useful in the understanding of motion types. By providing a spatiotemporal analysis, our approach gains several advantages over previous implementations. The strength of our approach was demonstrated in a series of experiments, including classification and comparisons with other algorithms.**

***Index Terms*—Normal flow, spatiotemporal texture, temporal texture.**

I. Introduction

In THE biological world, the most basic capabilities found in animals are based on motion [23]—they are critical for the animals' navigation and basic survival. However, in the field of computational vision, robotic systems still face huge challenge in fully realizing the potential of the rich motion cues. Complete recovery of the egomotion parameters is a notoriously ill-posed problem [2], [7], [11], [14], [37], [41], [44]. This, coupled with the view that the egomotion parameters must be computed before any of the motion-based functionalities can be accomplished, has seriously obscured the potential for using motion directly for many applications.

Recent researches in computational vision [16] emphasized the fact that, often, appropriate spatiotemporal representations, that are directly relevant to the tasks at hand, can be computed from the imagery without going through the ill-posed process of egomotion computation. Thus instead of one strict hierarchy, we can have a variety of visual processes, computed in parallel and using motion features of varying amount of complexity. This view is partially motivated by the results from brain sciences [45]: different parts of the visual cortex seem to perform functionally specialized operations. One example of such modules is the computation of the rate of approach of surrounding scene points [8], [15], [30], [35]. This can be used to identify situations where something is rapidly approaching the camera, indicating a threatening or aggressive situation. Another module is the recognition of objects through motion patterns, including both highly structured ones such as those produced by walking and running, and more statistical ones such as those due to fluttering leaves and flowing water. Our research efforts subscribe to this philosophy. Through a variety of such motion competences, a multifaceted usage of motion information for content representation can be effected, thus better representing the temporal semantics of an image sequence, without suffering from the instabilities typically associated with egomotion estimation. One of the motion competences exhibited by human is that of motion recognition capability, which is the focus of the current work. This paper builds on the works of [6], [31]; its contribution lies in the novel synergizing of the spatial and temporal aspects of motion field, resulting in a representation which we termed as spatiotemporal texture.

It is not uncommon for us to find object that exhibits characteristic motion with indeterminate spatial extent. The ensemble motion by a flock of birds taking flight, flowing water, fluttering leaves and waving flags are just some of the more common examples that serve to illustrate such motion. These commonly observed phenomena, together with the vast domain in which they exist, has prompted several researchers to formulate techniques that are able to discriminate, recognize and synthesize the distinctive motion patterns exhibited by these objects [3], [6], [31], [36]. In a pioneering work, Nelson and Polana [31] have shown the possibility of using these motion patterns to classify on a small set of image sequences. They coined the term temporal texture to define collectively motion patterns that exhibit statistical regularity but have indeterminate spatial and temporal extent. The key concern of their work is to establish computationally the viability of using temporal texture as a means for object recognition. Thus the features extracted are designed in such a way to isolate the temporal essence, independent of the spatial characteristics. This replicates computationally the classical Johansson's experiment [25] where human subjects were able to identify different activities from bright spots attached to an actor dressed in dark and moving in front of a dark background, without the aid of any structural information.

In the real world, of course, the phenomenon of temporal texture is always presented together with some spatial features, be it the more structured kind like a human hand, or the more statistical kind like fluttering leaves. In fact, the spatial structure in a way constrains the types of movement possible; there is no reason why these two intimately linked aspects of the same phenomenon should not be combined in a synergistic way, especially in areas of application where efficiency can be one of the key concerns. That is, from the computational viewpoint, if the spatial and temporal features can be represented in some

Manuscript received August 6, 2000; revised June 6, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Eric L. Miller.

The authors are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119260 (e-mail: elepehch@nus.edu.sg; eleclf@nus.edu.sg).

Digital Object Identifier 10.1109/TIP.2002.804265.

integral way, instead of being handled separately as is the case nowadays, redundancy would be removed, resulting in less computational load and more efficient indexing. For example, if the transient shape of a hand can be simultaneously extracted alongside with its motion, we can quickly further narrow down our search of a representation and recognize the motion of the hand at a faster rate.

We present a scheme whereby the spatial and temporal aspects of texture are coupled together. It can be viewed as an extension of [31] to encode more spatial and temporal information. This strategy has the advantages of better efficiency as well as possessing stronger discriminatory power. By mapping the magnitudes and directions of the normal flow (component of the optical flow projected along the intensity gradient direction), we obtained spatiotemporal textures. A set of sequential spatiotemporal textures is then superimposed to obtain a time-integral version of such textures. We then carried out a classification experiment to demonstrate the viability of using such textures as content identifiers. A comparative study of our algorithm with two existing algorithms was also carried out to assess the relative strengths and weaknesses of the algorithms.

The remaining sections are organized as follows. Section II discusses works related to this study. Section III outlines the motion information extracted for our analysis and builds appropriate representation used for subsequent matching. Sections IV and V describe the experiments conducted and discuss the results of the experiments. The paper closes with Section VI, which presents a summary of the present research study, some possible applications, as well as some thoughts on future research directions.

II. Related Work

The direct use of motion for object recognition has been realized computationally by Nelson and Polana in their qualitative analysis of temporal texture [31]. In their study, statistical features based on the magnitudes and directions of flow vectors were calculated to recognize different types of temporal textures. Their study highlighted the computational possibility of using low level motion features for recognition. However, as this is a pioneering work, the intriguing spatiotemporal relationships of moving parts of the objects were not fully addressed. The distinctive spatial features of the moving objects were also insufficiently exploited. Following [31], Bouthemy and Fablet [6] analyzed statistically on the temporal distribution of appropriate local motion-based measures to perform global motion characterization of video shots. Their method enhanced the temporal descriptive power by analyzing motion over an extended sequence, rather than just over two frames. However, the spatial characteristics of the moving object were again ignored. Otsuka *et al.* [32] extracted features of temporal texture based on tangent plane representation of motion trajectory. However, in their study, spatial and temporal features were separately determined. Furthermore, the extracting of the motion trajectory from temporal textures is by no means an easy task and renders its accuracy questionable.

The relevance of combining spatial features with motion is evidenced in the work of Davis and Bobick in which they represented and recognized action using view-based temporal templates [12]. They constructed a binary motion-energy image (MEI) which represents where motion has occurred in an image sequence, as well as a motion-history image (MHI) which is a scalar-valued image where intensity is a function of recency of motion. By matching the aggregate shape induced by an action with the stored models of views of known actions, different actions are identified. In this case, the cumulative transient shape of the body, when coupled with the bodily movement, forms the temporal templates that enable recognition. Our study adopts a similar approach in recognizing moving objects. However, instead of using shape and motion, texture and motion are used. This allows us to recognize objects whose shapes are more indeterminate in spatial extent. The way in which motion information is extracted from the image sequence also differs from [12].

There have been various works concerned with characterizing local estimates of spatiotemporal orientation [1], [18], [21]. The objective is either to recover optical flow or to characterize the dimensionality of the local orientation. Reference [43] has attempted to abstract from these spatiotemporal data a number of qualitative structural descriptions such as “unstructured,” “flicker,” “coherent motion,” “incoherent motion,” etc. However, due to the local nature of the input, only a small set of primitive classes could be distinguished. Our spatiotemporal representation utilizes a novel and richer input. It allows us to tap on the rich varieties of texture classification techniques which in turn are capable of yielding semantically more complex categories.

III. Spatiotemporal Representation

A. Normal Flow as Motion Information

Most motion-based analysis techniques rely heavily on the ability to estimate motion to a certain degree of accuracy (e.g., in determining the dominant direction and recovering structure from motion). Optical flow technique is usually used to estimate motion field from an image sequence. Unfortunately, optical flow cannot be estimated based on image intensities alone unless additional constraint is imposed (e.g., smoothness [22]). Such constraints are either difficult to implement in practice or are not true over the entire image.

Apart from the above said difficulty, the estimation of motion using optical flow usually involves iterations that require long processing time. This may generate a large amount of overheads rendering a recognition task inefficient. One of the solutions to reduce the processing time in video processing is the utilization of motion compensation component of the MPEG video encoder as a coarse-grained representation of the optical flow [13]. Nevertheless, motion vectors obtained in this method are highly inaccurate as spatial resolution is lost. Another solution is to utilize partial motion information whose computation does not require iteration (e.g., normal flow).

In this paper, normal flow is used as it can be more accurately computed and involves much less computation. Hardware implementation is also simpler (e.g., [29]). These are significant practical advantages that cannot be said for the use of optical flow. Although the full displacement is not recoverable, we

argue that the partial flow does provide sufficient information for the purpose of motion-based recognition.

If we assume that the image intensity I for a given scene point (x, y) remains unchanged over time t and $t+\delta t$, we may write [22]

$$I(x, y, t) = I(x+\delta x, y+\delta y, t+\delta t). \tag{1}$$

By expanding this equation and ignoring the higher order terms, we get

$$\delta x\frac{\partial I}{\partial x} + \delta y\frac{\partial I}{\partial y}t + \delta t\frac{\partial I}{\partial t} = 0. \tag{2}$$

Dividing the equation by δt, and denoting the partial derivatives of I by I_x, I_y, and I_t, we obtain

$$I_x\frac{\partial x}{\partial t} + I_y\frac{\partial y}{\partial t} + I_t = 0. \tag{3}$$

The above constraint provides only one linear equation with two unknowns [i.e., the local velocity vector $(\partial x/\partial t, \partial y/\partial t)$]. Additional constraints are needed to allow the velocity to be determine locally. This is formally known as the aperture problem. Equation (3), however, enables us to determine the normal velocity directly which is given by

$$u_n = \frac{-I_t}{\sqrt{(I_x^2 + I_y^2)}}. \tag{4}$$

Normal flow, being the amount of pixel movement along the image intensity gradient, naturally describes the temporal aspect of an object in dynamic motion. Apart from obtaining motion information, the spatial distribution of these flow vectors also captures the spatial aspect of the object undergoing motion. A spatial collection of the normal flow magnitudes reflects the physical constraints of the moving parts and their inertia to move. While a normal flow measurement at a particular pixel does not fully constrain its two-dimensional (2-D) motion flow, a spatial collection of the normal flow directions does anticipate the intended directions of move by the edges at a particular instant. Thus, it unveils the potential to move that is guided by the physical orientation of the edge. Our hypothesis is that the spatial patterns generated by the magnitudes and the directions of the normal flow are distinctive and unique to the object initiating that move.

A spatial collection of normal flows is in many ways similar[1] to a spatial distribution of edge gradient vectors. Thus spatial features of an object are inherently encoded in the normal flow plots. By exploiting the spatial distribution of the normal flows, the spatial texture of the moving body is integrated with the flow information. This synergy reveals powerful characteristics that are unique and innate to the moving body initiating the move. Our work can be regarded as a generalized approach in the analysis of temporal texture, synergizing both the spatial and the temporal aspect of object motion. Features such as divergence and curl described in [31], while not explicitly measured, were subsumed under our representations since we can always derive these parameters from the spatial collection of normal flow vectors.

[1]There are, however, important differences. First, the edge gradient vector measures the direction of the change in gray-level at a point. However, the direction of normal flow gives the direction of move of an edge. The two vectors either point in the same direction or point in opposite directions. Second, the magnitude of the gradient vector measures the strength of an edge whereas the magnitude of the normal flow measures the strength of a move by an edge. Hence, an object has to move before normal flow is present. On the contrary, an edge's presence is independent of motion.

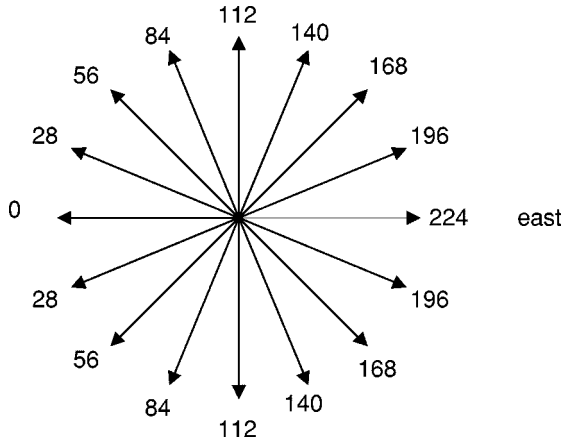

Fig. 1. Quantization of normal flow direction into 16 gray levels.

B. Spatiotemporal Texture Mappings

Aimed at providing a spatiotemporal analysis on the motion of the objects, the magnitudes and directions of normal flow are mapped into intensity levels for subsequent analysis. The resulting intensity images encode both the spatial and temporal aspects of the moving objects. Textures generated in this way are henceforth referred to as magnitude plots and directional plots for magnitudes and directions of the normal flow respectively.

In magnitude plot, the magnitudes of normal flow vectors are linearly mapped into the 256 gray scale levels. Only flow vectors with magnitudes ranging from zero to four are mapped. Any values larger than four are regarded as noise. This mapping is then further quantized into nine gray levels (at an interval of 32, with gray level 0 and 255 inclusive). Hence, the pixel values $f(i, j)$ in the magnitude plot has the following formulation: $f(i, j) \in \{0, 31, 63, \ldots, 255\}$. Visually, regions of darker gray would correspond to regions with larger normal flow magnitude. The magnitude plot, in effect, encodes the magnitude strength of the normal flow in terms of intensity at the respective pixel locations.

The direction of each normal flow vector is quantized into one of the eight directions as shown in Fig. 1 with an angle of 22.5° subtended between each pair of adjacent directions. The reference direction is arbitrary taken to be the one that points to the "east." Any flow vector with quantized direction similar to the reference will take up the gray level of the reference direction (i.e., level 224). The others will take up the gray level depending on the deviations of their quantized directions from the reference direction. The pixel values in the directional plot is represented by $g(i, j) \in \{0, 28, 56, \ldots, 255\}$. The largest deviation would correspond to the greatest jump in gray level from the reference direction (i.e., level 0). Stationary points appear as white (i.e., level 255). The resultant texture exhibits tones of gray, with the darkest region corresponding to region with flow vectors having the largest deviation from the reference direction.

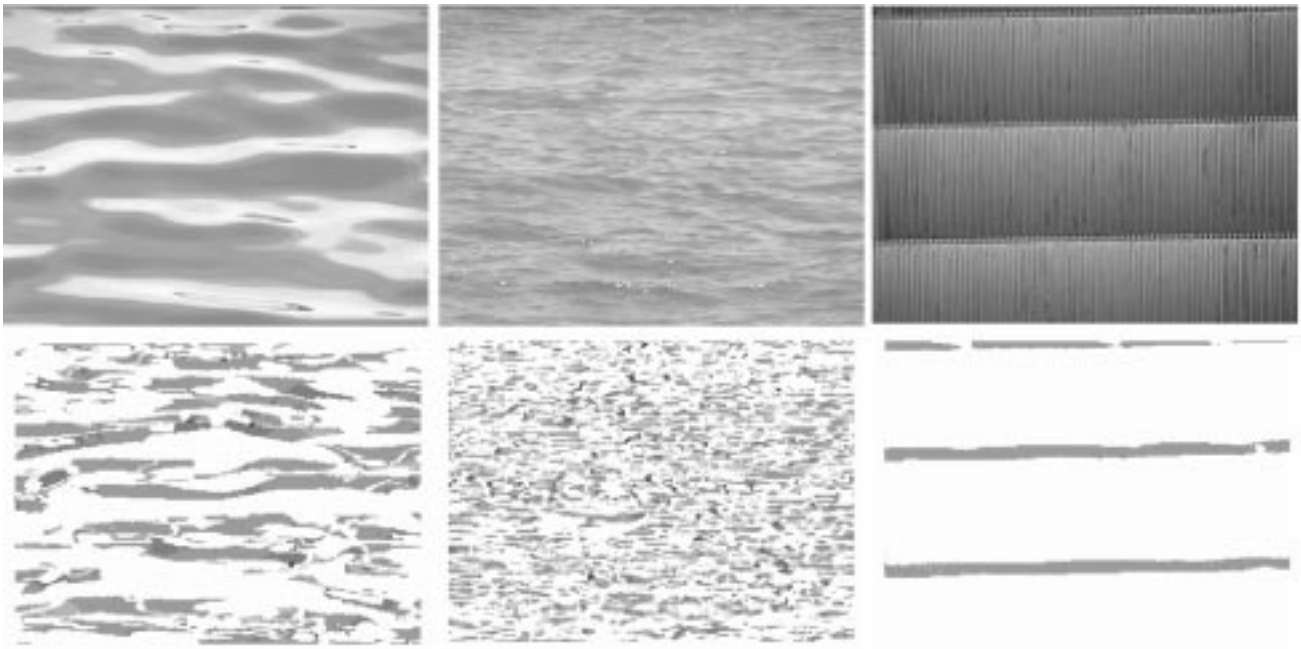

Fig. 2. Images of waves (different scales) and ascending escalator with their respective directional plots at bottom.

C. Selected Characteristics of Spatiotemporal Texture

We used three aspects of visual texture (i.e., coarseness, directionality, and contrast) to elucidate the spatiotemporal texture formation process and to justify the inclusion of the spatial aspect in our representation. In the actual experiment, a more comprehensive set of texture features was extracted. Fig. 2 shows some examples of directional plots.

- In general, coarseness is dependent on the size, the spatial structure and the level of coherence in the motion of the moving objects or parts. For instance, the escalator motion is more rigid and oriented than that generated by the waves. The moving elements in the image are also larger in this case. Due to the discrete tessellation of the directional plots, neighboring directions are mapped to the same gray level and grouped together. Hence, the texture generated is very much coarser. The same argument applies for the magnitude plots.
- Directionality in directional plot is dependent on both spatial and temporal factors. One key factor is the number of "layered" or occluding types of motion found. Flows that are highly laminar or stream-like also tend to form texture with strong directionality. Besides, directionality also relies strongly on the characteristics of the movement itself. The magnitude plot gives an overview on the global distribution of the magnitude strength. Due in part to the physical nature of the moving objects, some motion has strength that is uniformly distributed along a particular direction. It differentiates oriented motion from nonoriented one.
- In the directional plot, contrast indicates the amount of directional uniformity of normal flow, as the gray levels represent the amount of deviation from the reference direction. With a more directionally diversified type of motion, the dynamic range of gray levels in the texture will be broader, thereby resulting in a texture of higher contrast (e.g., fluttering leaves). Objects with many moving parts also tend to have more edges and thus, will result in higher contrast. Instead of measuring directional uniformity of normal flow, the contrast in magnitude plot reflects the dynamic range of magnitude strength spatially. We get a texture of high contrast if a given scene has wide-ranging normal flow magnitudes, and if these magnitudes were evenly distributed throughout the scene.

From the above discussion, we see that temporal information can be highly constrained by the structural primitives of the object. However, these temporal primitives are neither exclusive nor unique to any single object. Thus, the mere use of temporal information without any spatial information retention can be insufficient to fully describe the moving object. In the physical world, there exist objects with very similar, though not exactly alike, temporal characteristics (e.g., banner and grass; waves and wind-blown grassland). In this case, the presence of spatial information can serve as an important cue to discriminate these objects.

D. Extended Spatiotemporal Texture

Notwithstanding the fact that a certain degree of the spatiotemporal properties innate to a moving object has already been encapsulated by the spatiotemporal texture, the rich temporal attributes displayed by a moving object over an extended sequence may still be under-represented. In view of this, we further the concept of spatiotemporal texture by separately superimposing the magnitude and directional plots in time. Such superimposition traces motion history and forms images similar to a series of transitory projection of an object onto our retina when flashed before our eyes. Spatiotemporal textures extended this way are referred to as extended magnitude plots (EMPs) and extended directional plots (EDPs) for magnitudes and directions of the normal flow respectively. Fig. 3 illustrates some examples of EMPs and EDPs.

The mathematical expression of the extended spatiotemporal plots is defined as follows.

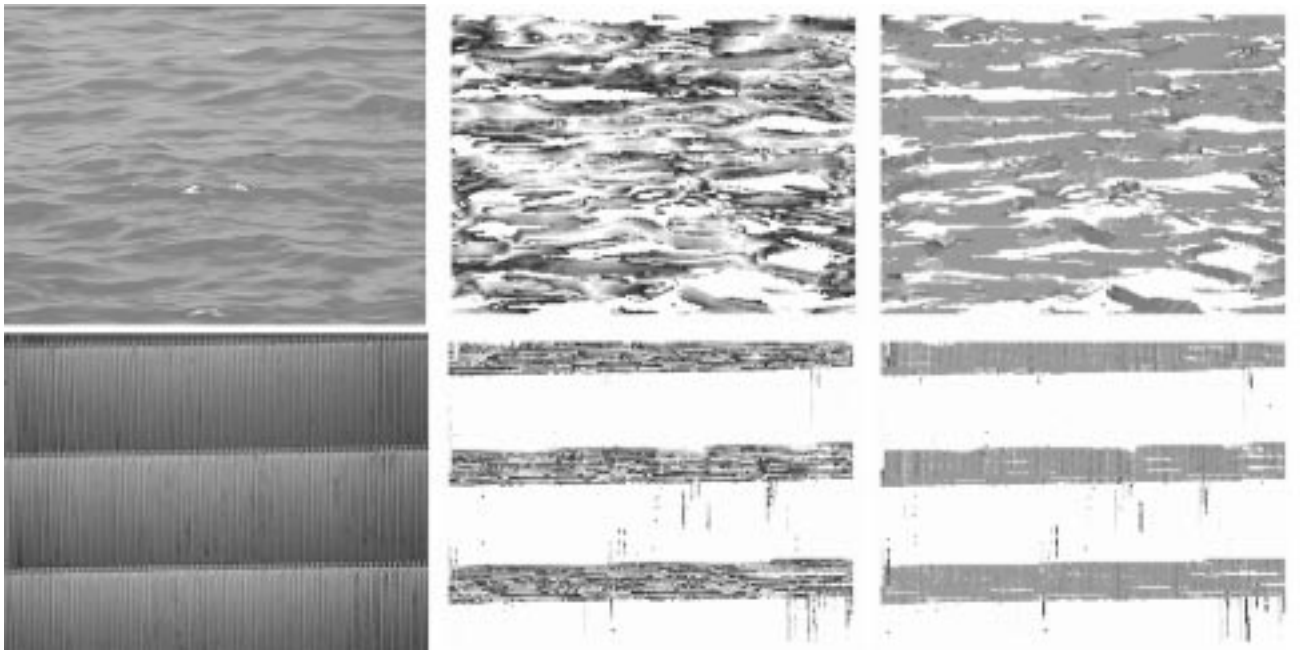

Fig. 3. Some examples of images with their extended magnitude and directional plots (col 1: images; col 2: extended magnitude plots; and last col: extended directional plots).

Let τ be the number of previous spatiotemporal textures to be superimposed to form the extended plot. We use the subscript t to index time-dependent quantity. Let P_t be a nonempty, finite set of pixel locations in a $M \times N$ spatiotemporal texture formed by image frame t and $t+1$. Of this entire set of pixel location, we let R_t be a subset of P_t that contains moving points, and S_t, which is the complement of R_t, to be another subset which have members with points that are stationary or featureless (i.e., points with zero flow values). That is

$$P_t = S_t \bigcup R_t = \{(0, 0), (0, 1), \ldots, (M, N)\}$$

with the above representation, the value of EMP at time t and at pixel location (i, j) may be written as

$$E_t(i,j) = \begin{cases} f_t(i,j), & \text{if } (i,j) \in R_t \\ 255, & \text{if } (i,j) \in \bigcap_{t-\tau \le k \le t} S_k \\ f_{\tau_o}(i,j), & \text{if } (i,j) \in \bigcap_{\tau_o < k \le t} S_k \bigcap R_{\tau_o} \\ & \text{and } t > \tau \ge t - \tau \end{cases}$$

where f_t is the mapped magnitude of a moving point in R_t; and f_{τ_o} corresponds to the mapped magnitude of a moving point in R_{τ_o} and this spatial point does not move or is featureless subsequently from frame τ_o to t. A point is regarded as moving if its quantized normal flow magnitude is nonzero and is less than four (beyond which the flow is considered as noise).

The expression for the EDP is similar to that for EMP, except that $f_t(i, j)$ is replaced by $g_t(i, j)$, where

$$g_t(i, j) \in \{0, 28, 56, \ldots, 255\}.$$

Textural features in the extended plots unveil several characteristics (e.g., rigidity, coherence and size) of the moving object apart from those inherited from the spatiotemporal textures. By selecting a suitable value for τ, we can invariably adjust the level of spatial and temporal information to be encoded in the plots. Ideally, a different τ value should be given to different motion type to achieve an optimal spatio–temporal representation. Unfortunately, this can only be done empirically. A small value of τ would naturally encode more structural than temporal information. A large value of τ, though provide more temporal information, renders the extraction of spatiotemporal information difficult.

- A coarse texture in the EMP is caused by a moving object with moving parts that advance at a relatively constant rate. In the EDP, a coarse texture is due to an object that has edges moving in some prevalent directions. Besides, coarseness also reveals the size and coherence of the moving parts of an object. A highly constrained motion gives fine texture due to its inability to move over a large distance. In general, an object exhibiting periodic motion will have fine texture due to the oscillation of magnitudes and directions inherent in the move.
- Directionality of the texture is dependent on both spatial and temporal factors. Moving objects with highly laminar or stream-like motion will typically generate texture with strong directionality. Objects with most of the edges oriented to a particular direction will also yield textures of high directionality.
- A low contrast in the EDP is likely to be caused by moving objects with normal flow directions that do not deviate much from one another over time. On the other hand, a high-contrast EMP is obtained if the scene has wide-ranging normal flow magnitudes.

IV. Experiment I: Classification

The main objective of this experiment is to illustrate the ability of the extended spatiotemporal texture approach to discriminate various motion classes.

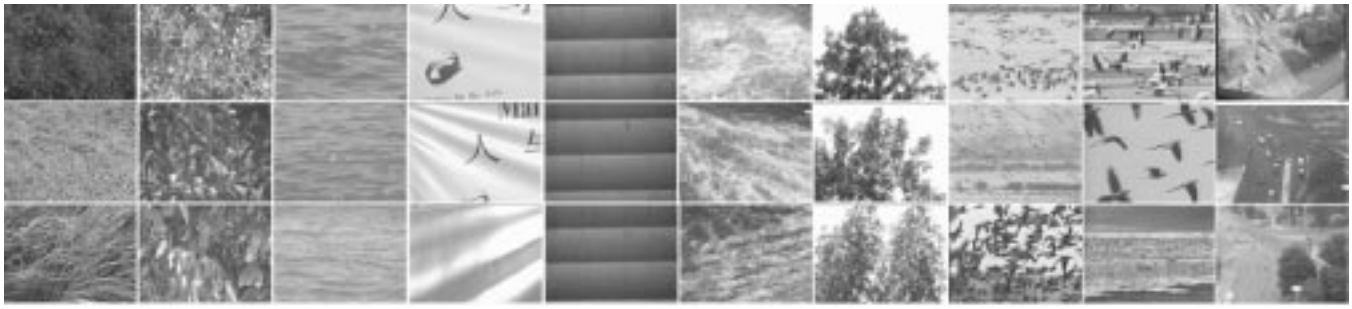

Fig. 4. Image sequences used in the experiments. (Cols. 1 to 10: Class A to class J; Row 1: Training/Test Sequence 0; Row 2: Test Sequence 1; Row 3: Test Sequence 2).

TABLE I
DIFFERENT CLASSES OF MOVING ENTITY

Class	Name	Description
A	Grass	Wind blown grass
B	Bush	Fluttering leaves
C	Waves	Gentle sea waves
D	Banner	Wind blown banner
E	Escalator	Moving escalator
F	Turbulence	Rough water turbulence
G	Tree	Wind blown tree
H	Birds I	Flock of birds in quartering flight
I	Birds II	Flock of birds in directional flight
J	Cars	Moving cars taken at road junction

A. Experimental Input

A set of image sequences containing a representative set of both oriented motion, such as flowing water, and nonoriented one, such as fluttering leaves was used as the input to the experiment. Motions of varying degrees of rigidity can also be found within the set of input. Besides, natural versus man-made, as well as deterministic versus nondeterministic type of motion are also represented in the image set. Altogether, we used ten classes of motion entities with varying motion characteristics (Table I).

For each class of motion entity, three image sequences consisting of 100 240 $\times$ 320 pixel frames each captured at 25 Hz were used. The first image of each sequence is as shown in Fig. 4. These image sequences were taken at different locations with different moving objects of the same class. For example, the motion class for leaves contains leaves of various shapes and sizes, and the motion class for escalator contains both ascending and descending escalators. The image sequences were either recorded using a tripod-mounted handheld digital camcorder or directly captured from some documentary videos. Sequences in Class J were downloaded from Image Sequence Server at KOGS/IAKS Universität Karlsruhe.

B. Procedure

For each image sequence in Fig. 4, the normal flow field was computed between each consecutive pair of image frames using the gradient-based approach described in [17]. In order to prevent possible contamination by any spurious or erroneous flow value computed, only flow vectors with magnitude ranging from zero to four were used. The selected normal flow vectors were then transformed into spatiotemporal texture plots as described in Section III-B. Five consecutive plots were then superimposed to form the EMP and EDP (i.e., $\tau = 5$). Hence, a total of 19 samples were obtainable from each image sequence. A set of texture features was then extracted from the EMPs and EDPs to form the feature set for that sequence.

The treatment of the feature samples is as follows. The first ten samples of the first sequence (Training Sequence) from each motion class were used as the training samples. These form clusters of feature vectors in the feature space which will be defined in Section IV-C. The remaining nine texture plots of the first sequence (Test Sequence 0) and the entire set of the other two sequences (Test Sequences 1 and 2) were used as verification set and test set respectively. These three sets were then classified using the k nearest-neighbor classification scheme based on standard Euclidean distance. We used the value of $k = 3$ throughout our experiment with the majority wins. All features were given equivalent weighting in arriving at the Euclidean distance. Any sample whose nearest Euclidean distance qualified as an statistical outlier was classified as unknown. This outlier's distance was computed based on Grubbs test [5] with 95% confidence level. The mean and variance for conducting the test was obtained from all the minimum Euclidean distance computed in our classification experiments.

C. Features Used

A subset of the features computed by the gray-level co-occurrence matrix (GLCM) [19], difference statistics [42], and Fourier spectrum [26] techniques were first shortlisted (Table II). Readers are referred to the original papers for full details on these classical texture features. A feature space consisting of a mix of features from the above three techniques was selected so that the features complement each other. In this way, we inherited a wide varying set of descriptors that could describe important textural properties of the plot. The different feature values extracted were normalized into common units by mapping the average to a unit vector. The dimension of the feature space sums up to 12 for both the magnitude and the directional plots. For simplicity, the distance d for the pixel-pair in the co-occurrence matrix and difference statistic computations was chosen to be 1 for all textural analysis in this experiment. In GLCM computation, the feature computed from each GLCM over four directions (i.e., 0°, 45°, 90°, and 135°) was averaged.

We then selected a subset of these features using the χ^2 method described in [27] to reduce the feature space dimension

TABLE II
PRELIMINARY TEXTURAL FEATURES CHOSEN FOR THE EXPERIMENT

Features	Analytical Technique	Measure	Mathematical Formula
Feat1	Gray-Level	Inertia	$\sum(i-j)^2 P_d(i,j)$
Feat2	Co-occurrence	Shade	$\sum(i+j-\mu_x-\mu_y)^3 P_d(i,j)$
Feat3	Matrix	Correlation	$\frac{1}{\sigma_x\sigma_y}[\sum ij P_d(i,j) - \mu_x\mu_y]$
Feat4	Fourier Spectrum	Centered at 45^o	$\sum_{22.5^o \le tan^{-1}(v/u) < 67.5^o} \lvert F(u,v)\rvert^2$
Feat5	(Angular)	Centered at 135^o	$\sum_{112.5^o \le tan^{-1}(v/u) < 157.5^o} \lvert F(u,v)\rvert^2$
Feat6	Difference Statistics	Mean	$\sum kP_d(k)$ where $P_d(k) = \sum_{\lvert i-j\rvert=k} P_d(i,j)$

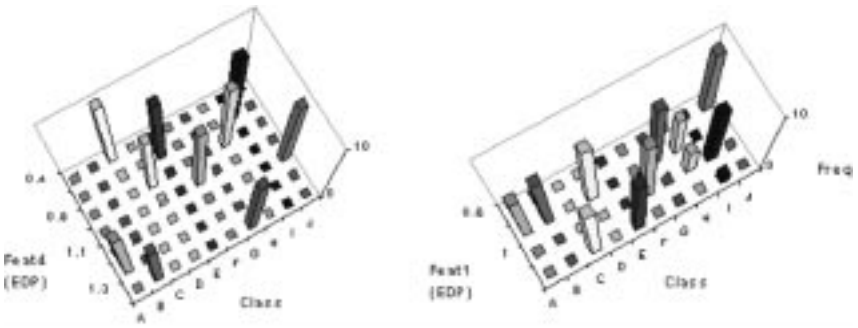

Fig. 5. Two examples of partitioned features. Left: a good feature and right: a bad feature.

and remove any redundant feature. In this method, a statistic significance test is conducted on the relationship between the values of a feature and the classes. Pairs of adjacent intervals with the lowest χ^2 values are merged until all pairs of intervals have χ^2 values exceeding a critical value. Fig. 5 shows both an example of a good and a bad feature partitioned at $\chi^2 = 0.95$. In this example, Feat4(EDP) could single out Class A, C, F, and J distinctively from the other classes. On the contrary, Feat1(EDP) was unlikely to separate any of the classes. We chose a subset of six features Table III that can best discriminate the nine motion classes based on the training set. Fig. 6 demonstrates the partitioning of the six selected features. Since CFeat1 measures the contrast of EDP and both Class D and E are likely to give low contrast EDP, it is difficult for one to separate the two classes based on CFeat1 alone. However, due to the difference in coarseness in the two EDP's, CFeat2 discriminates the two classes relatively well. Similarly, it is difficult for one to discriminate between Class G, H, and I using CFeat3 since the line structures found in these classes are quite similar. These line structures, however, differ significantly from that of Class E. Hence, Class E is well separated from these classes using CFeat3. Similar argument can be extended to all other feature measures.

D. Results and Discussion

Table IV tabulates the classification results while Table V traces the assignments of any misclassified samples. It can be concluded that the experiment carried out had successfully classified the majority of the classes. A close examination of the results reveals that image sequences which yield the lower success rates can be attributed to significant object size difference between the training and the test sets. One underlying reason for this poor performance is in the fixing of the parameter d at 1 for both the co-occurrence matrix and the difference statistics computations throughout the experiment. The classification results associated with these sequences could be readily improved with a proper range of d chosen to account for the individual scale difference. For instance, we could compute a number of GLCMs with several d values for each feature and use the maximum resultant statistical measure for classification. Alternatively, a χ^2 method could be used to select the d value that contains the most structure [46]. As expected, the classification rate for Test Sequence 1 of Class B increased appreciably from 8/19 to 18/19 when $d = 6$ and jumped to 19/19 when $d = 10$. Similarly, the rate for Test Sequence 2 of Class B increased from 11/19 to 14/19 for both $d = 5$ and $d = 10$. The classification rate for Test Sequence 2 of Class I climbed from 7/19 to 9/19 and 13/19 when we set $d = 8$ and $d = 4$, respectively.

The above argument on scale difference, however, may not be true for Test Sequence 2 of Class G (Tree) where the size of the object is of comparable scale with that of the training sequence. There was no apparent improvement in performance as we varied d. We attribute this poor performance to the extraction of excessive spatial information from the plots. A close observation in the image sequences of Class G shows that Test Sequence 2 differs from the other two sequences mainly in its structural makeup. The tree in the Test Sequence 2 is more leafy (denser) with minimal amount of exposed branches. As the Fourier Spectrum (Angular) analysis is highly sensitive to orientation of line structure in a texture, the algorithm fails to assign most of the samples in Test Sequence 2 to Class G. Instead, most of the samples were classified as Class A (Grass). This misclassification

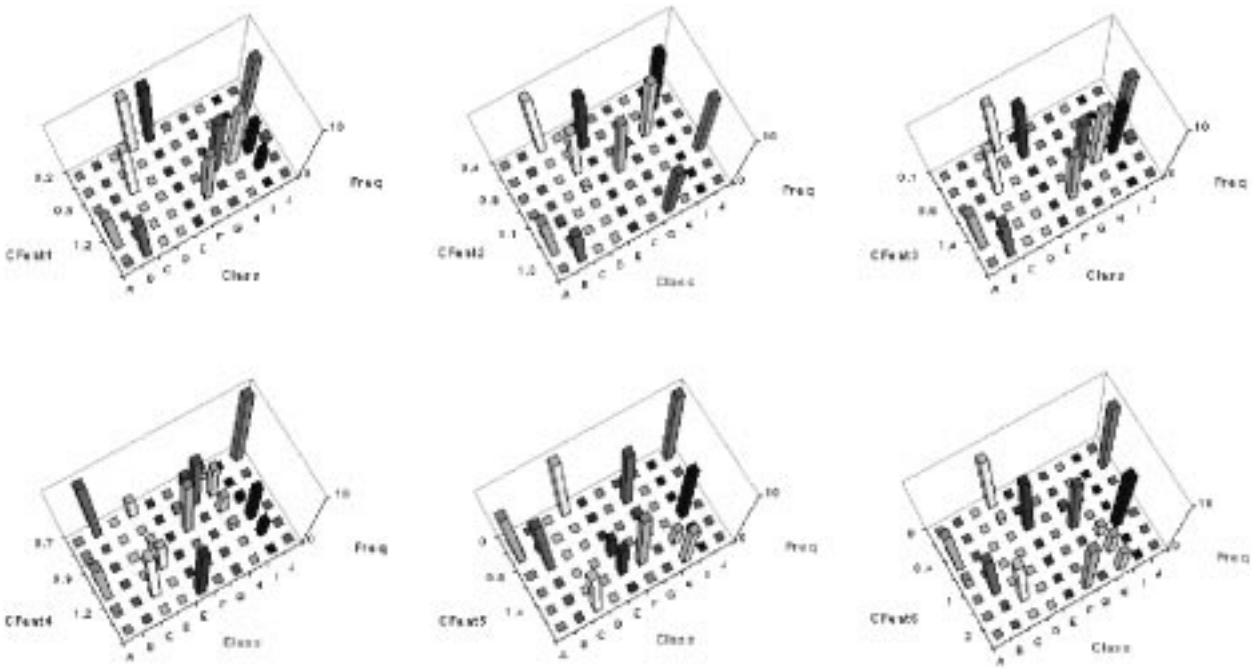

Fig. 6. Subset of six chosen features partitioned at $\chi^2 = 0.95$.

TABLE III
FINAL TEXTURAL FEATURES USED FOR THE EXPERIMENT

Features	Source	Analytical Technique	Measure
CFeat1	EDP	GLCM	Inertia
CFeat2		Difference Statistics	Mean
CFeat3		Fourier (Angular)	At 45^o
CFeat4	EMP	GLCM	Correlation
CFeat5		GLCM	Shade
CFeat6		Fourier (Angular)	At 135^o

TABLE IV
CORRECT CLASSIFICATION WITH OUR ALGORITHM

Class	Test Seq. 0	Test Seq. 1	Test Seq. 2	Success Rate
A	9/9	19/19	17/19	95.74%
B	9/9	8/19	11/19	59.57%
C	9/9	19/19	12/19	85.11%
D	9/9	19/19	19/19	100.00%
E	9/9	19/19	19/19	100.00%
F	9/9	19/19	19/19	100.00%
G	9/9	19/19	8/19	76.60%
H	9/9	19/19	19/19	100.00%
I	9/9	7/19	17/19	70.21%
J	9/9	17/19	15/19	87.23%

is due to the sharing of some spatiotemporal resemblances between Test Sequence 2 of Class G and Training Sequence of Class A.

Fig. 7 shows the projection of the training samples onto its first two principal components using the Principal Component Analysis method. Distinct clusters, each corresponding to a motion class, were formed. This clear separation of the clusters naturally speaks for and is highly correlated to the high success rates obtained in our classification.

TABLE V
CLASSIFICATION AND MISCLASSIFICATION WITH OUR ALGORITHM (%)—E.G., FOR CLASS A, PERCENTAGE OF SAMPLES ASSIGNED TO CLASS A, B AND G WERE 95.74%, 2.13% AND 2.13% RESPECTIVELY. CLASS U REPRESENTS UNCLASSIFICATION DUE TO LARGE EUCLIDEAN DISTANCE COMPUTED OR TIES WITH OTHER CLASSES

	A	B	C	D	E	F	G	H	I	J	U
A	**95.74**	2.13					2.13				
B	4.26	**59.57**				23.40	6.38		6.38		
C			**85.11**					2.13	2.13	6.38	4.26
D				**100.00**							
E					**100.0**						
F						**100.0**					
G	23.40						**76.60**				
H								**100.0**			
I							21.27	8.51	**70.21**		
J	2.13			6.38			4.26			**87.23**	

We have explicitly shown the success of our approach in a supervised classification experiment. It is also interesting to access the performance under unsupervised classification. A k-means unsupervised clustering was conducted on the Training/Test Sequence 0 using the standard Euclidean distance criteria. The numbers of clusters were prespecified to be 10. The six features used were similar to those used in the original supervised experiment. The iteration stopped when no sample can be moved to a centroid closer than the one that it currently belonged. After three iterations, the clusters formed were as given in Table VI. We concluded that the algorithm had clustered the majority of the classes successfully with some confusion between Class F and H and I. This was due to the spread in the distribution of Class H as shown in Fig. 7.

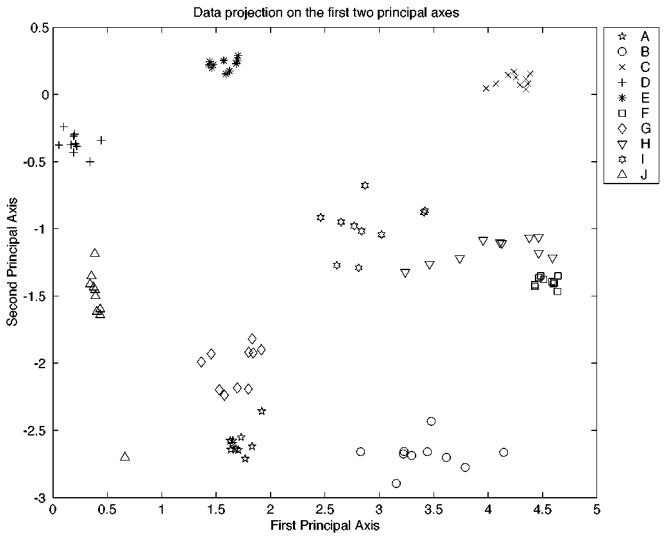

Fig. 7. Training samples projected onto the first two principal components.

TABLE VI
CLUSTER MEMBERSHIPS FORMED BY k-MEANS CLUSTERING

Cluster	1	2	3	4	5
Number(Class)	19(E)	19(B)	19(A)	18(J)	19(C)
Cluster	6	7	8	9	10
Number(Class)	19(F)+8(H)	19(D)	19(G)	1(J)	19(I)+11(H)

V. EXPERIMENT II: COMPARISON WITH TWO ALGORITHMS

We compare our algorithm with another two temporal texture analysis algorithms. We specifically chose the algorithms of Nelson and Polana [31], and Bouthemy and Fablet [6] for comparison due to the affinity of their works with ours although their works are mainly concerned with the recognition of motion class rather than particular moving entities. The objective of this experiment is to empirically benchmark the relative strengths and weaknesses in the three algorithms in identifying the moving entities.

In our algorithm, we have chosen the stacking factor to be five. This means that each single extended spatiotemporal texture had been derived from the superimposition of five consecutive spatiotemporal textures in a sequence. In Nelson and Polana's algorithm, the features obtained from five consecutive normal flow plots were averaged. Five consecutive normal flow plots were used to derive the temporal co-occurrence matrix in Bouthemy and Fablet's algorithm. Classification was performed with the k nearest-neighbor classifier $(k = 3)$ based on Euclidean distance with equal weighting for all features in the three algorithms. Similar to Experiment I, the first ten sets of feature samples from Sequence 0 were used as the training data. The remaining nine samples and the full set of the other two test sequences with 19 samples each were used for testing.

A. Nelson and Polana's Algorithm

Table VII summarizes the features extracted from the normal flow field in Nelson and Polana's algorithm. The classification results obtained were tabulated in Tables VIII and IX.

TABLE VII
FEATURE SET OF NELSON AND POLANA'S ALGORITHM

Feature	Measure
Feat1	Inverse coefficient of variation
Feat2	Positive divergence
Feat3	Negative divergence
Feat4	Positive curl
Feat5	Negative curl
Feat6	Non-uniformity of flow direction
Feat7	Directional difference statistics in the horizontal direction
Feat8	Directional difference statistics in the vertical direction
Feat9	Directional difference statistics in the +ve diagonal direction
Feat10	Directional difference statistics in the -ve diagonal direction

TABLE VIII
CORRECT CLASSIFICATION WITH NELSON AND POLANA'S ALGORITHM

Class	Test Seq. 0	Test Seq. 1	Test Seq. 2	Success Rate
A	9/9	0/19	0/19	19.15%
B	8/9	1/19	1/19	21.28%
C	9/9	18/19	0/19	57.47%
D	9/9	19/19	19/19	100.00%
E	8/9	17/19	17/19	91.49%
F	9/9	18/19	14/19	87.23%
G	9/9	19/19	19/19	100.00%
H	5/9	15/19	18/19	80.85%
I	9/9	0/19	0/19	19.15%
J	9/9	10/19	19/19	80.85%

TABLE IX
CLASSIFICATION AND MISCLASSIFICATION WITH NELSON AND POLANA'S ALGORITHM (%)—E.G., FOR CLASS A, PERCENTAGE OF SAMPLES ASSIGNED TO CLASS A, B AND G WERE 19.15%, 40.43% AND 40.43% RESPECTIVELY. CLASS U REPRESENTS UNCLASSIFICATION DUE TO LARGE EUCLIDEAN DISTANCE COMPUTED OR TIES WITH OTHER CLASSES

	A	B	C	D	E	F	G	H	I	J	U
A	**19.15**	40.43					40.43				
B	4.26	**21.28**				29.79				27.66	17.02
C			**57.47**			6.38		17.02	2.13	6.38	10.64
D				**100.00**							
E					**91.49**						8.51
F						**87.23**	2.13				10.64
G							**100.00**				
H						17.02		**80.85**			2.13
I	6.38	2.13			25.53		31.91		**19.15**		14.89
J									19.15	**80.85**	

In this algorithm, Classes D, E, F, G, H, and J scored the highest success rate with D and G achieving perfect score. Moderate performance is observed in Class C. Class C's moderate success rate can be attributed to the low success rate in its Test Sequence 2. While all of the three image sequences depict gentle waves, the waves in Test Sequence 2 differ significantly in scale from the training sequence. Despite Nelson and Polana's claim to scale invariance of their feature

TABLE X
FEATURE SET OF BOUTHEMY AND FABLET'S ALGORITHM

Features	Measure
Feat1	Average
Feat2	Variance
Feat3	Dirac
Feat4	Angular Second Moment
Feat5	Contrast

set, the overall success rate of Class C seems to suggest that either the feature set is not scale-invariant or that insufficient temporal information has been encoded into the features to recognize this similarity in motion. Despite the close similarity in the motion type exhibited by the three sequences, the low percentage of correctly matched samples from Test Sequence 2 appears to bear out the aforementioned hypotheses.

Indeed, the normal flow distribution is strongly linked to the spatial feature in the sense that it is located at the edges. Therefore it opens to question whether the feature set used by Nelson and Polana encodes more of the spatial or the temporal information. By the preceding reasoning, it is not surprising that Classes D, E, and G scored high success rates since the image sequences under those classes are similar in scale and spatial appearance within its own classes. It also accounts for why Classes F, H, and J, and especially Classes A, B, and I did not perform well. In particular, sequences used in Class I bear very little spatial resemblance to each other. Even the similarity in motion pattern is a more elaborate one (local cluster of fluttering wings) that requires a more sophisticated feature descriptor to make explicit. Using the Nelson and Polana's feature set these samples are apt to be grouped under Classes E and G.

B. Bouthemy and Fablet's Algorithm

To better handle the temporal evolution in temporal textures, Bouthemy and Fablet [6] proposed the use of temporal co-occurrence matrices to characterize motion types. Instead of normal flow, they used a locally weighted average of normal flow as described in [6]. The computed values are then quantized into 16 levels ranging from 0 to 4 pixel units. The setting of the co-occurrence matrix is similar to the spatial case, except that the pairwise pixels are chosen in the temporal sense (refer to the original paper for a detailed description). We prepared our experiment as per described in [6] with the temporal pixel pair distance d_t chosen to be 1. Table X summarizes the feature set used by Bouthemy and Fablet, while Tables XI and Table XII tabulate the classification results.

From Table XI, we see that the algorithm turned in rather good performance for Classes D, G, and H, and moderate performance for Classes E and J. Bouthemy and Fablet's algorithm differs significantly from the Nelson and Polana's algorithm in that it is more concerned with the evolution of the normal flow field over time rather than its characteristics at a particular instance. Thus, there is very little of the spatial information in its computed feature. This appears to account for the result of Class A. A patch of wind-blown grass generates motion pattern that is over any short time interval similar to that experienced by a moving banner though they are subject to somewhat different types of physical constraint. This partial similarity in temporal evolution resulted in a large number of samples from Class A being misclassified into Class D. In particular, Test Sequence 1 consists essentially of shorter grasses with less swaying in its motion. By ignoring the spatial and possibly more complex form of temporal information, the algorithm judged Test Sequence 1 of Class A to be more similar to Class D, whereas the human visual system seems capable of combining both the spatial and temporal information in its discrimination, which is what our algorithm attempts to emulate.

The classes with the worst performance were B, C, F, and I. In particular, Class B samples are equally likely to fall under Classes B, G, and H. Moreover, Class C samples tend to be classified under Class F. While Class C and Class F both belong to fluid type of motion, the motion inherent in Class F is typically more haphazard than Class C's. In addition, Class F samples are apt to be classified under Class H. This apparently points to the insensitivity of the Bouthemy and Fablet's feature set to the different types of stochastic motion. The feature set was also unable to characterize samples from Class I due to its elaborate motion. Finally, the poor performance for Test Sequence 1 of Class J was probably due to the presence of heavy snowfall in the foreground which overlapped with the actual vehicular motion.

TABLE XI
CORRECT CLASSIFICATION WITH BOUTHEMY AND FABLET'S ALGORITHM

Class	Test Seq. 0	Test Seq. 1	Test Seq. 2	Success Rate
A	7/9	1/19	10/19	38.30%
B	1/9	7/19	1/19	19.15%
C	8/9	2/19	3/19	27.66%
D	8/9	19/19	15/19	89.36%
E	5/9	13/19	8/19	55.32%
F	9/9	0/19	0/19	19.15%
G	9/9	14/19	18/19	87.23%
H	5/9	18/19	19/19	89.36%
I	0/9	2/19	0/19	4.26%
J	9/9	0/19	19/19	59.57%

TABLE XII
CLASSIFICATION AND MISCLASSIFICATION WITH BOUTHEMY AND FABLET'S ALGORITHM (%)—E.G. FOR CLASS A, PERCENTAGE OF SAMPLES ASSIGNED TO CLASS A, D, I, AND U WERE 38.30%, 57.47%, 2.13% AND 2.13% RESPECTIVELY. CLASS U REPRESENTS UNCLASSIFICATION DUE TO LARGE EUCLIDEAN DISTANCE COMPUTED TO TIES WITH OTHER CLASSES

	A	B	C	D	E	F	G	H	I	J	U
A	**38.30**			57.47					2.13		2.13
B		**19.15**		8.51	6.38		23.40	21.28	4.26		17.02
C			**27.66**	2.13	6.38	40.43		17.02	2.13	4.26	
D				**89.36**						6.38	4.26
E		2.13	2.13		**55.32**	4.26		23.40	2.13		10.64
F					4.26	**19.15**		74.47			2.13
G							**87.23**				12.77
H		2.13	2.13		2.13			**89.36**	2.13		2.13
I	4.26	14.89	4.26	10.64	6.38	10.64	19.15	8.51	**4.26**		17.02
J	27.66			12.77						**59.57**	

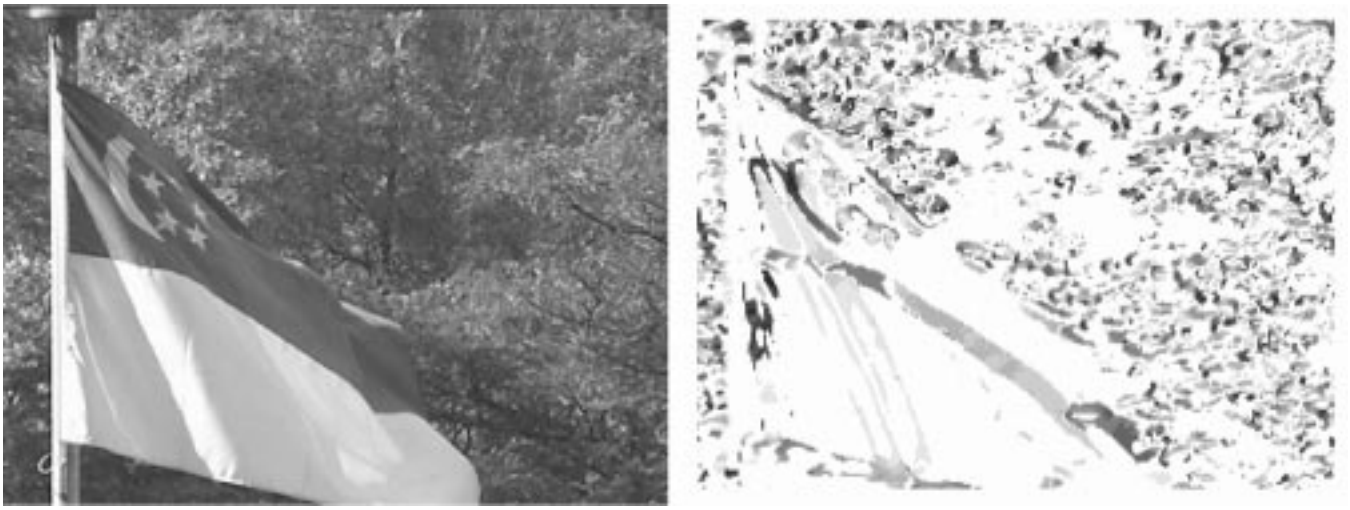

Fig. 8. Image and directional plot of a waving flag and fluttering leaves in the background.

VI. Success Factors in Our Algorithm

In contrast to the other two algorithms, our algorithm has distinctly better success rates for all the classes with Classes D, E, F, and H yielding perfect results. Slightly lower success rates were observed for Classes B, G, and I. The high success rate in Class D provides good empirical evidence of the ability to typify line-like temporal textures. It also demonstrated its ability to distinguish temporal textures with varying degrees of regularity, in view of its successful discrimination of samples from Class C and F. In comparison with the other two algorithms, our approach was evidently more effective in the characterization of the more complex temporal textures from Classes H and I.

With exception to Class G, the overall success rates of the other classes were greater than those achieved in the other two algorithms. The better performance of our algorithm could be due to several factors: 1) The superiority of the feature set. For instance, features resulting from Fourier spectrum analysis are not used at all in the other two algorithms. 2) The richness of the input utilized by our algorithm—the extended spatiotemporal texture encodes both the spatial and the temporal information in an integral manner. Possibly both factors play a part too.

VII. Conclusions

The evidence gathered in the experiments demonstrated the superiority of our approach in distinguishing between different spatiotemporal textural qualities. There exists two underlying reasons for the success of our algorithm: the richness of the extended representations as well as the choice of feature extractors. From a theoretical perspective, the unsatisfactory performance of the Nelson and Polana's algorithm in the classification results could be attributed to the temporal deficiency of the inputs prior to feature extraction. Apart from the use of normal flow, no other explicit techniques are employed to enhance the encoding of temporal information into the inputs. There is also insufficient spatial information being extracted. On the contrary, the Bouthemy and Fablet's algorithm, while addressing the temporal deficiency found in the Nelson and Polana's algorithm did not aim to encode spatial information. The temporal co-occurrence matrices while characterizing the normal flow magnitude distribution across a fixed time frame did not consider the intra-relationship between normal flow magnitudes within a frame. Moreover, the distribution of normal flow directions, which represents a potential source of information, while readily available, was not exploited.

The main contribution of this paper lies in the introduction of the notion of spatiotemporal texture and the realization of its synergistic usage. Through the employment of the extended spatiotemporal texture plots, spatial textural qualities together with temporal information are inherently encoded as input to the feature extraction stage. In addition, we have used a wide range of analytical techniques for feature extraction which results in the robustness of our algorithm. While many have argued that basing features on normal flow alone suffers the disadvantage of discarding too much information, related works on psychophysics [9], [25], [28], [34] as well as the successful application of normal flow computation in the present study suggest that partial information alone would suffice for the purpose of recognition. As a result, simple recognition algorithms requiring minimal computational power can be implemented.

There are several areas in which the concept of spatiotemporal textures can be applied. Often in video indexing, we are particularly interested in indexing moving objects. Optic flow histogram [4], [10], [24] or variance of the flow [38], [39] is usually employed to characterize global motion or activity. However, it can be difficult for us to quantify or describe some motion types, especially those which are indeterminate in temporal extent. In such cases, there is neither dominant direction nor motion features that we can reliably depend on. With the present system's ability to characterize motion of such nature, we could have created categories like "flag-like motion," "fluttering leaves," and "water turbulence" to better relate the different motion experienced. Often, these categories are directly linked to high-level semantics, for instance, the waving flag motion of Fig. 8 is a better indication of the object identity than other features like "redness."

To facilitate object-based video coding, there is often a need to segment an image into different moving regions. Traditionally, this has been achieved using flow segmentation [40]. Such techniques would encounter difficulties when the motion flow field is of a statistical kind (e.g., Fig. 8 which contains both waving flag motion and the fluttering motion of leaves). By transforming the flow field into a texture map, we can tap on the rich varieties of texture segmentation techniques [20], [33] available to accomplish the segmentation task.

Future research efforts should gear toward devising a more wide-ranging methods to obtain other descriptive features from the extended spatiotemporal textures. Another important research direction lies in using learning techniques to adaptively and judiciously combine the various spatial and temporal features to form better discriminators. Lastly, in this paper, it has been assumed that the camera is stationary with minimal jitters. If the camera is moving, then the camera's motion needs to be compensated. This remains a challenging problem.

REFERENCES

[1] E. Adelson and J. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Amer. A*, vol. 2, no. 2, pp. 284–2992, 1985.

[2] G. Adiv, "Inherent ambiguities in recovering 3-D motion and structure from a noisy flow field," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 477–489, May 1989.

[3] E. Ardizzone, A. Capra, and M. L. Cascia, "Using temporal texture for content-based video retrieval," *J. Vis. Lang. Comput.*, vol. 11, pp. 241–252, 2000.

[4] E. Ardizzone and M. L. Cascia, "Video indexing using optical flow field," in *Proc. IEEE Int. Conf. Image Processing*, Lausanne, Switzerland, Sept. 1996, pp. 831–834.

[5] V. Barnett, T. Lewis, and V. Rothamsted, *Outliers in Statistical Data*, 3rd ed. New York: Wiley, 1994.

[6] P. Bouthemy and R. Fablet, "Motion characterization from temporal cooccurences of local motion-based measures for video indexing," in *Proc. Conf. Pattern Recognition*, Brisbane, Australia, Aug. 1998, pp. 905–908.

[7] L. Cheong, C. Fermüller, and Y. Aloimonos, "Effects of errors in the viewing geometry on shape estimation," *Comput. Vis. Image Understand.*, vol. 71, no. 3, pp. 356–372, Sept. 1998.

[8] R. Cipolla and A. Blake, "Surface orientation and time to contact from image divergence and deformation," in *Proc. 2nd Eur. Conf. Computer Vision*, Santa Margherita Ligue, Italy, 1992, pp. 187–202.

[9] J. Cutting and D. Proffitt, *Gait Perception as an Example of How We May Perceive Events in Intersensory Perception and Sensory Integration*, R. Walk and H. L. Pick, Eds. New York: Plenum, 1981, pp. 249–273.

[10] Y. Deng and B. S. Manjunath, "Content-based search of video using color, texture and motion," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Santa Barbara, CA, 1997, pp. 534–537.

[11] K. Daniilidis and M. E. Spetsakis, "Understanding noise sensitivity in structure from motion," in *Visual Navigation: From Biological Systems to Unmanned Ground Vehicles*, Y. Aloimonos, Ed. New York: Lawrence Erlbaum, 1995.

[12] J. Davis and A. Bobick, "The representation and recognition of action using temporal templates," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 1997, pp. 928–934.

[13] N. Dimitrova and F. Golshani, "R_x for semantic video database retrieval," in *Proc. ACM Multimedia 94*, San Francisco, CA, Oct. 1994, pp. 219–226.

[14] Z. Duric and Y. Aloimonos, "Estimating the heading direction using normal flow," *Int. J. Comput. Vis.*, vol. 13, no. 1, pp. 1–56, 1994.

[15] Z. Duric, A. Rosenfeld, and J. Duncan, "The applicability of Green's theorem to computation of rate of approach," in *Proc. ARPA Image Understanding Workshop*, Monterey, CA, Nov. 1994, pp. 1209–1217.

[16] C. Fermüller and Y. Aloimonos, "Vision and action," *Image Vis. Comput.*, vol. 13, no. 10, pp. 725–744, 1995.

[17] C. Fermüller, "Passive navigation as a pattern recognition problem," *Int. J. Comput. Vis.*, vol. 14, no. 2, pp. 147–158, 1995.

[18] D. Fleet and A. Jepson, "A cascaded approach to the construction of velocity selective mechanisms," Dept. of Comput. Sci., University of Toronto, RBCV-TR-84-6, Tech. Rep., Dec. 1984.

[19] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[20] R. M. Haralick and L. G. Shapiro, "Survey: Image segmentation," *Comput. Vis., Graph., Image Process.*, vol. 29, pp. 100–132, 1985.

[21] D. Heeger, "A model for the extraction of image flow," *J. Opt. Soc. Amer. A*, vol. 4, no. 8, pp. 1455–1471, 1997.

[22] B. K. P. Horn and B. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, no. 1–3, pp. 185–203, Aug. 1981.

[23] G. Horridge, "The evolution of visual processing and the construction of seeing systems," *Proc. R. Soc. Lond. B*, vol. 230, pp. 279–292, 1987.

[24] A. K. Jain, A. Vailaya, and W. Xiong, "Query by video clip," *Multimedia Syst.*, vol. 7, no. 5, pp. 369–384, 1999.

[25] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Percep. Psychophys.*, vol. 14, pp. 201–211, 1973.

[26] G. O. Lendaris and G. L. Stanley, "Diffraction pattern sampling for automatic pattern recognition," *Proc. IEEE*, vol. 58, pp. 198–216, Feb. 1970.

[27] H. Liu and R. Setiono, "Feature selection via discretization," *IEEE Trans. Knowledge Data Eng.*, vol. 9, pp. 642–645, July/Aug. 1997.

[28] L. MacArther and R. Baron, "Toward an ecological theory of social perception," *Psychol. Preview*, vol. 90, no. 3, pp. 215–238, 1983.

[29] A. Moini, A. Bouzerdoum, K. Eshraghian, A. Yakovleff, X. T. Nguyen, A. Blanksby, R. Beare, D. Abbott, and R. E. Bogner, "An insect vision-based motion detection chip," *IEEE J. Solid-State Circuits*, vol. 32, pp. 279–285, Feb. 1997.

[30] R. C. Nelson and Y. Aloimonos, "Obstacle avoidance using flow field divergence," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 1102–1106, Oct. 1989.

[31] R. C. Nelson and R. Polana, "Qualitative recognition of motion using temporal texture," *CVGIP: Image Understand.*, vol. 56, no. 1, pp. 78–89, 1992.

[32] K. Otsuka, T. Horikoshi, S. Suzuki, and M. Fujii, "Feature extraction of temporal texture based on spatio–temporal motion trajectory," in *Proc. Conf. Pattern Recognition*, Brisbane, Australia, Aug. 1998, pp. 1047–1051.

[33] A. Rosenfeld and A. C. Kak, *Digital Picture Processing*, 2nd ed. New York: Academic, 1982.

[34] S. Runeson and G. Frykholm, "Visual perception of lifted weight," *J. Exp. Psychol.: Hum. Percept. Perform.*, vol. 7, no. 4, pp. 733–740, 1981.

[35] J. Santos-Victor, G. Sandini, F. Curotto, and S. Garibaldi, "Divergent stereo for robot navigation: Learning from bees," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, New York, 1993, pp. 434–439.

[36] M. Szummer and R. W. Picard, "Temporal texture modeling," in *Proc. IEEE Conf. Image Processing*, Lausanne, Switzerland, Sept. 1996, pp. 823–826.

[37] J. I. Thomas, A. Hanson, and J. Oliensis, "Understanding noise: The critical role of motion error in scene reconstruction," in *Proc. DARPA Image Understanding Workshop*, Washington, DC, Apr. 1993, pp. 691–695.

[38] N. Vasconcelos and A. Lippman, "Toward semantically meaningful feature spaces for the characterization of video content," in *Proc. IEEE Conf. Image Processing*, Santa Barbara, CA, Oct. 1997, pp. 78–89.

[39] V. Vinod, "Activity based video shot retrieval and ranking," in *Proc. IEEE Conf. Pattern Recognition*, Brisbane, Australia, 1998, pp. 682–684.

[40] J. Y. A. Wang, E. H. Adelson, and U. Desai, "Applying mid-level vision techniques for video data compression and manipulation," *Proc. SPIE*, vol. 2187, pp. 116–127, Feb. 1994.

[41] J. Weng, T. S. Huang, and N. Ahuja, *Motion and Structure From Image Sequences*. Berlin, Germany: Springer-Verlag, 1991.

[42] J. S. Weszka, C. R. Dyer, and A. Rosenfeld, "A comparative study of texture measures for terrain classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, pp. 269–285, Apr. 1976.

[43] R. P. Wildes and J. R. Bergen, "Qualitative spatiotemporal analysis using an oriented energy representation," in *Proc. 6th Eur. Conf. Computer Vision 2000*, vol. 2, Dublin, Ireland, June 2000, pp. 768–784.

[44] G. S. Young and R. Chellapa, "Statistical analysis of inherent ambiguities in recovering 3-D motion from a noisy flow field," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, pp. 995–1013, Oct. 1992.

[45] S. Zeki, *A Vision of the Brain*. New York: Blackwell, 1993.

[46] S. W. Zucker and D. Terzopoulos, "Finding structure in co-occurrence matrix for texture analysis," *Comput. Graph. Image Process.*, vol. 12, pp. 1286–1308, 1980.

Chin-Hwee Peh was born in Singapore in 1973. He received the B.Eng. and M.Eng. degrees from the National University of Singapore (NUS) in 1998 and 2000, respectively. He is currently pursuing the Ph.D. degree at NUS.

He joined the Department of Electrical and Computer Engineering, NUS, in 2000 as a Research Engineer. His research interests lie in the areas of computer and human vision, with emphasis on three-dimensional motion and shape analysis as well as visual psychophysics.

Loong-Fah Cheong was born in Singapore in 1965. He received the B.Eng. degree from the National University of Singapore and the Ph.D. degree from the Center for Automation Research, University of Maryland at College Park, in 1990 and 1996, respectively.

In 1996, he joined the Department of Electrical and Computer Engineering, National University of Singapore, where he is now an Assistant Professor. His research interests are related to the basic processes in the perception of three-dimensional motion, shape, and their relationship, as well as the application of these theoretical findings to specific problems in navigation and in multimedia systems, for instance, in the problems of video indexing in large databases.