Dynamic Texture Recognition Based on Distributions of Spacetime Oriented Structure

Konstantinos G. Derpanis and Richard P. Wildes
Department of Computer Science and Engineering
York University
Toronto, Ontario, Canada
{kosta,wildes}@cse.yorku.ca

Abstract

This paper addresses the challenge of recognizing dynamic textures based on their observed visual dynamics. Typically, the term dynamic texture is used with reference to image sequences of various natural processes that exhibit stochastic dynamics (e.g., smoke, water and windblown vegetation); although, it applies equally well to images of simpler dynamics when analyzed in terms of aggregate region properties (e.g., uniform motion of elements in traffic video). In this paper, a novel approach to dynamic texture representation and an associated recognition method are proposed. The approach pursued here recognizes dynamic textures based on matching distributions (histograms) of spacetime orientation structure. Empirical evaluation on a standard database with controls to remove the effects of identical viewpoint demonstrates that the proposed approach achieves superior performance over alternative state-of-the-art methods.

1. Introduction

1.1. Motivation

A readily observable set of visual phenomena encountered in the natural world are dynamic patterns that are due to the temporal variation (e.g., movement) of a large number of individual elements. Several examples are depicted in Fig. 1. Such patterns primarily are characterized by the aggregate dynamic properties of elements or local measurements taken over a region of spatiotemporal support, rather than in terms of the dynamics of individual constituents. In the computer vision literature, these patterns have appeared collectively under various names, including, turbulent flow/motion [24], temporal textures [28], time-varying textures [2], dynamic textures [31], and textured motion [38]; the term dynamic texture will be used herein. Most typically, the term "dynamic texture" has been used with reference to images of natural processes that exhibit stochastic dynamics (e.g., fire, turbulent water and windblown vegetation); however, it can apply equally well to simpler phenomena when analyzed in terms of aggregate regional properties (e.g., orderly pedestrian crowds and vehicular traffic).

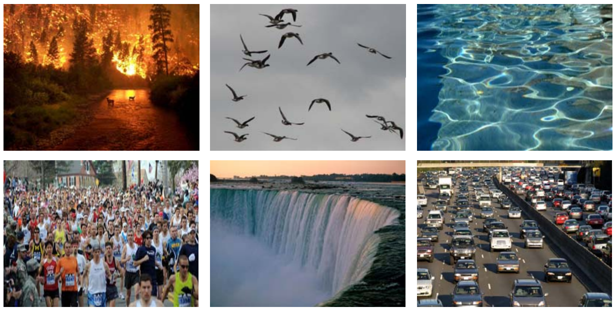

Figure 1. Examples of dynamic textures in the real world. (left-to-right, top-to-bottom) Forest fire, flock of birds in flight, water, crowd of people running, waterfall and vehicular traffic.

The ability to recognize dynamic textures based on visual processing is of significance to a number of applications, including, video indexing/retrieval, surveillance and environmental monitoring where they can serve as keys, isolate background clutter (e.g., fluttering vegetation) from activities of interest and detect various critical conditions (e.g., fires), respectively.

The goal of the present work is the development of a unified approach to representing and recognizing a diverse set of dynamic textures with robustness to viewpoint and ability to encompass recognition in terms of semantic categories (e.g., recognition of fluttering vegetation without being tied to a *specific* view of a *specific* bush). Toward that end, an approach is developed that is based solely on observed dynamics (i.e., excluding purely spatial appearance).

For present purposes, local spatiotemporal orientation is of fundamental descriptive power, as it captures the first-order correlation structure of the data irrespective of its origin (i.e., irrespective of the underlying visual phenomena), even while distinguishing a wide range of dynamic patterns of interest (e.g., single motion, multiple motions and scintillation). Correspondingly, each dynamic texture will be associated with a distribution (histogram) of measurements that indicates the relative presence of a particular set of 3D orientations in visual spacetime, (x, y, t), as measured by a bank of spatiotemporal filters, and recognition will be performed by matching such distributions. Interestingly, the distribution of oriented spacetime structure has been shown to be an important discriminating factor in human perception studies of dynamic textures [41, 42].

1.2. Related work

Over the past three decades various representations have been directed at characterizing dynamic textures for the purpose of recognition [11]. In this section, several representative strands of research are reviewed.

One strand of research explores physics-based approaches (e.g., [25]). These methods derive models for specific dynamic textures (e.g., water [25]) based on a first-principles analysis of the generating process. With the model recovered from input imagery, the underlying model parameters can be used to drive inference. Beyond computational issues, the main disadvantage of this class of approaches is that the derived models are highly focused on specific dynamic textures, and thus lack generalization to larger classes of dynamic textures.

Motivated by spatial texture-related research, an early approach to uniform analysis of a diverse set of dynamic textures was based on extracting first- and second-order statistics of motion flow field-based features, assumed to be captured by estimated normal flow [28]. This work was followed-up by numerous proposed variations of normal flow (e.g., [6]) and optical flow-based features (e.g., [27]). There are two main drawbacks related to this strand of research. First, normal flow is highly correlated with dynamic texture spatial appearance [29]. Thus, in contrast to the goal of the present paper, recognition is highly tuned to a particular spatial appearance. Second, optical flow and its normal flow component are predicated on assumptions like brightness constancy and local smoothness, which are generally difficult to justify in the context of dynamic textures.

A recent trend in dynamic texture research is the use of statistical generative models to jointly model the spatial appearance and dynamics of a pattern. Recognition is realized by comparing the similarity between the underlying model parameters. Several variants of this approach have appeared, including: autoregressive (AR) models [34, 16, 19, 38] and multi-resolution schemes [24, 2]. By far the most popular of these approaches for recognition is the joint photometric-dynamic, AR-based Linear Dynamic System (LDS) model, proposed in [16], which has formed the basis for several recognition schemes [31, 9, 43, 37]. Although impressive recognition rates have been reported ($\sim 90\%$), most previous efforts have limited experimentation to cases where the dynamic texture samples are taken from the exact same viewpoint. As a result, much of the performance is highly tied to the spatial appearance rather than the underlying dynamics [9, 43]. In a few variants, the cited approaches have considered only the dynamics portion of the LDS model for recognition [9, 43]. Significantly, a comparative study of many of the proposed approaches showed that when applied to image sequences with non-overlapping views of the same scene ("shift-invariant" recognition), all yield significantly lower recognition rates ($\sim 20\%$), whether using joint spatial/dynamic or only the dynamic portion of the LDS model [43].

Spatiotemporal oriented energy filters serve in defining the representation employed in the current work. Previous efforts have used similar operators in the analysis of image sequences for various purposes, e.g., optical flow estimation [23, 32, 22], activity recognition [12, 15], low-level pattern categorization [40], tracking [8], spacetime stereo [33] and spacetime grouping [14]. Significantly, it appears that no previous work has used the filter outputs to support dynamic texture recognition, as shown here.

1.3. Contributions

In the light of previous research, the major contributions of the present work are as follows. (i) A unified approach to representing and recognizing dynamic textures is proposed based on their underlying dynamics and thereby enables recognition that is viewpoint robust. Other than [43], there is no previous work addressing view- or shift-invariant dynamic texture recognition based solely on the observed dynamics of the scene. (ii) A particular spacetime filtering formulation is developed for measuring spatiotemporal oriented energy and is used for classifying dynamic textures. While spacetime filters have been used before for analyzing image sequences, they have not been applied to the recognition of dynamic textures in the manner proposed. (iii) Empirical evaluation on a standard dynamic texture database, using non-overlapping viewpoints, demonstrates that the proposed approach achieves superior performance over state-of-the-art methods.

2. Technical approach

There are two key parts to the proposed approach to dynamic texture recognition: First, a representation based on a distribution of spatiotemporal oriented energies; second, a match measure between any two samples under consideration. This section begins by motivating the significance of visual spacetime orientation in the context of dynamic texture analysis. Subsequently, the particulars of the proposed representation and match measure are detailed.

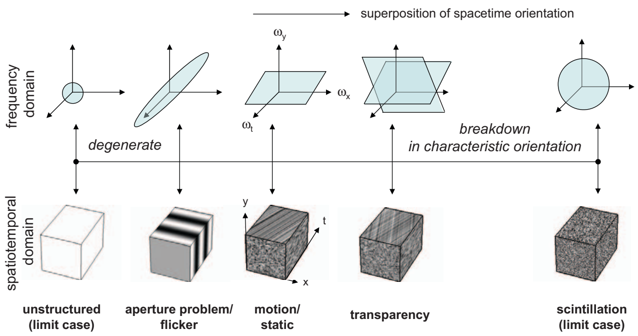
superposition of spacetime orientation
ω_y
ω_x
ω_t
frequency domain
degenerate
breakdown in characteristic orientation
spatiotemporal domain
y
t
x
unstructured (limit case)
aperture problem/ flicker
motion/ static
transparency
scintillation (limit case)

Figure 2. Range of spacetime oriented structure. The bottom and top rows of images depict prototypical patterns of dynamic structure in the spacetime and frequency domains, respectively. The horizontal axis indicates the amount of spacetime oriented structure superimposed in a pattern, with increasing amounts given along the rightward direction.

2.1. Orientation in visual spacetime

The local orientation (or lack thereof) of a pattern is a salient characteristic. Figure 2 illustrates the significance of this structure in terms of describing the range of dynamic patterns in image sequence data (cf. [40]). The horizontal axis indicates the amount of spacetime oriented structure superimposed in a pattern, with increasing amounts given along the rightward direction. The bottom and top rows of images depict prototypical patterns of dynamic structure in the spacetime and frequency domains, respectively. The presence of a single spacetime orientation corresponds to image velocity [18, 1, 39, 23, 32]; static patterns correspond to the special case of zero velocity. In the frequency domain, the energy of these patterns correspond to a plane through the origin, with the planar surface slant indicative of velocity. To the left of the motion pattern reside two degenerate cases corresponding to spacetime orientation that is partially specified (the aperture problem and pure temporal luminance flicker) and completely unspecified (unstructured). In the frequency domain, the energy of the the partially specified case corresponds to a line through the origin; in the special case of flicker, the line lies strictly along the temporal frequency axis. In the limit, a region can totally lack any spatiotemporal contrast (unstructured case) and the frequency domain correlate is isolated in the low-frequency portion of the spectrum.

Starting again from a motion pattern and superimposing an additional spacetime orientation yields a semi-transparency pattern [20]. Here, two spacetime orientations dominate the pattern description. In the frequency domain, the energy corresponds to two planes, each representative of its respective spacetime orientation. Continuing the superposition process to the limit, yields the special case of scintillation (e.g., "television snow"), where no discernable orientation dominates the local region; nevertheless, significant spatiotemporal contrast is present. In the frequency domain, the energy of this pattern corresponds to an isotropic response throughout. In between the cases of two-motion transparency and scintillation lie various complicated phenomena that arise as multiple spacetime oriented structures (e.g., motions) are composited. Occurrences in the world that give rise to such visual phenomena include those governed by turbulence and other stochastic processes (e.g., dynamic water, windblown vegetation, smoke and fire).

As illustrated above, the local spacetime orientation of a visual pattern captures significant, meaningful aspects of its dynamic structure; therefore, a spatiotemporal oriented decomposition of an input pattern is an appropriate basis for local representation. By extension, aggregated measures of orientation over a region of visual spacetime may be of use in characterizing the region's dynamic texture for recognition. Interestingly, distributions of spatially oriented measurements have played a prominent role in the analysis of static visual texture (see, e.g., [4] for review); however, it appears that the present paper documents the first application of such an approach to *dynamic* visual texture.

2.2. Representation: Distributed spacetime orientation

The desired spacetime orientation decomposition is realized using broadly tuned 3D Gaussian third derivative filters, $G_{3_{\hat{\theta}}}(x, y, t)$, with the unit vector $\hat{\theta}$ capturing the 3D direction of the filter symmetry axis. The responses of the image data to this filter are pointwise rectified (squared) and integrated (summed) over a spacetime region, Ω, that covers the entire dynamic texture sample under analysis, to yield the following energy measurement for the region

$$E_{\hat{\theta}} = \sum_{(x,y,t)\in\Omega} (G_{3_{\hat{\theta}}} * I)^2, \qquad (1)$$

where $I \equiv I(x, y, t)$ denotes the input imagery and $*$ convolution. Notice that while the employed Gaussian derivative filter is phase-sensitive, summation over the support region ameliorates this sensitivity to yield a measurement of signal energy at orientation θ. More specifically, this follows from Parseval's theorem [7] that specifies the phase-independent signal energy in the frequency passband of the Gaussian derivative:

$$E_{\hat{\theta}} \propto \sum_{(\mathbf{k},\omega_t)} |\mathcal{F}\{G_{3_{\hat{\theta}}} * I\}(\mathbf{k}, \omega_t)|^2, \qquad (2)$$

where $\mathbf{k} = (\omega_x, \omega_y)^\top$ denotes the spatial frequency vector, ω_t the temporal frequency and $\mathcal{F}$ the Fourier transform[1].

Each oriented energy measurement, (1), is confounded with spatial orientation. Consequently, in cases where the spatial structure varies widely about an otherwise coherent dynamic region (e.g., single motion across a region with varying spatial texture), the responses of the ensemble of oriented energies will reflect this behaviour and thereby

[1] Strictly, Parseval's theorem is stated with infinite frequency domain support on summation.

are spatial appearance dependent; whereas, a description of pure pattern dynamics is sought. To remove this difficulty, the spatial orientation component is discounted by "marginalization" of this attribute, as follows.

In general, a pattern exhibiting a single spacetime orientation (e.g., image velocity) manifests itself as a plane through the origin in the frequency domain [39]. Correspondingly, summation across a set of x-y-t-oriented energy measurements consistent with a single frequency domain plane through the origin is indicative of energy along the associated spacetime orientation, independent of purely spatial orientation. Since Gaussian derivative filters of order $N = 3$ are used in the oriented filtering, (1), it is appropriate to consider $N+1 = 4$ equally spaced directions along each frequency domain plane of interest, as $N+1$ directions are needed to span orientation in a plane with Gaussian derivative filters of order N [21]. Let each plane be parameterized in terms of its unit normal, $\hat{\mathbf{n}}$; a set of equally spaced $N+1$ directions within the plane are given as

$$\hat{\theta}_i = \cos\left(\frac{2\pi i}{N+1}\right)\hat{\theta}_a(\hat{\mathbf{n}})+\sin\left(\frac{2\pi i}{N+1}\right)\hat{\theta}_b(\hat{\mathbf{n}}),\ 0 \leq i \leq N, \tag{3}$$

with

$$\hat{\theta}_a(\hat{\mathbf{n}}) = \hat{\mathbf{n}} \times \hat{\mathbf{e}}_x / \|\hat{\mathbf{n}} \times \hat{\mathbf{e}}_x\| \quad \hat{\theta}_b(\hat{\mathbf{n}}) = \hat{\mathbf{n}} \times \hat{\theta}_a(\hat{\mathbf{n}}) \tag{4}$$

where $\hat{\mathbf{e}}_x$ denotes the unit vector along the ω_x-axis[2].

Now, energy along a frequency domain plane with normal $\hat{\mathbf{n}}$ and spatial orientation discounted through marginalization, is given by summation across the set of measurements, $E_{\hat{\theta}_i}$, as

$$\tilde{E}_{\hat{\mathbf{n}}} = \sum_{i=0}^{N} E_{\hat{\theta}_i}, \tag{5}$$

with $\hat{\theta}_i$ one of $N+1 = 4$ specified directions, (3), and each $E_{\hat{\theta}_i}$ calculated via the oriented energy filtering, (1). In the present implementation, 27 different spacetime orientations, as specified by $\hat{\mathbf{n}}$, are made explicit, corresponding to static (no motion/orientation orthogonal to the image plane), slow (half pixel/frame movement), medium (one pixel/frame movement) and fast (two pixel/frame movement) motion in the directions leftward, rightward, upward, downward and diagonal, and flicker/infinite vertical and horizontal motion (orientation orthogonal to the temporal axis); although, due to the relatively broad tuning of the filters employed, responses arise to a range of orientations about the peak tunings.

Finally, the marginalized energy measurements, (5), are confounded by the local contrast of the signal and as a result increase monotonically with contrast. This makes it impossible to determine whether a high response for a particular spacetime orientation is indicative of its presence or is indeed a low match that yields a high response due to significant contrast in the signal. To arrive at a purer measure of spacetime orientation, the energy measures are normalized by the sum of consort planar energy responses,

$$\hat{E}_{\hat{\mathbf{n}}_i} = \tilde{E}_{\hat{\mathbf{n}}_i} / \left(\sum_{j=1}^{M} \tilde{E}_{\hat{\mathbf{n}}_j} + \epsilon\right), \tag{6}$$

where M denotes the number of spacetime orientations considered and ϵ is a constant introduced as a noise floor. Conceptually, (1) - (6) can be thought of as taking an image sequence, $I(x, y, t)$, and carving its power spectrum into a set of planes, with each plane corresponding to a particular spacetime orientation, to provide a relative indication of the presence of structure along each plane.

The constructed representation enjoys a number of attributes that are worth emphasizing. First, owing to the bandpass nature of the Gaussian derivative filters (1), the representation is invariant to additive photometric bias in the input signal. Second, owing to the divisive normalization (6), the representation is invariant to multiplicative photometric bias. Third, owing to the marginalization (5), the representation is invariant to changes in appearance manifest as spatial orientation variation. Overall, these three invariances allow abstractions to be robust to pattern changes that do not correspond to dynamic pattern variation, even while making explicit local orientation structure that arises with temporal variation (motion, flicker, etc.). Fourth, the representation is efficiently realized via linear (separable convolution, pointwise addition) and pointwise non-linear (squaring, division) operations; thus, efficient computations are realized [13].

Overall, each of the normalized oriented energies can be viewed as expressing the evidence for the presence of a particular, spacetime oriented structure. Taken as an ensemble (distribution), they provide the relative contribution of each spacetime orientation in the decomposition of the dynamic texture signal under consideration. Previously, a similar representation was presented with application to video segmentation [14]. In that earlier effort energy was defined locally; whereas, here it is taken as a regional measurement.

2.3. Recognition: Spacetime orientation distribution similarity

An ensemble of (normalized) energy measurements, $\hat{E}_{\hat{\mathbf{n}}_i}$, is taken as a distribution with spatiotemporal orientation, $\hat{\mathbf{n}}_i$, as variable. (In practice, these measurements are maintained as histograms.) Given the spacetime oriented energy distributions of an input query and database with entries represented in like fashion, the final step of the approach is recognition. In general, to compare two distributions, denoted $\mathbf{x}$ and $\mathbf{y}$, there are several standard similarity measures in the literature that can be used. In evaluation, the following measures were considered. (In the following,

[2] Depending on the spacetime orientation sought, $\hat{\mathbf{e}}_x$ can be replaced with another axis to avoid the case of an undefined vector.

Figure 3. Sample frames from the UCLA dynamic texture database used for evaluation.

individual entries in the employed histogram representation of the distributions are specified via subscripting, e.g., x_i, and summations are taken across all bins.)

Minkowski-Form distance [17]:

$$d_{L_p}(\mathbf{x}, \mathbf{y}) = \left(\sum_i |x_i - y_i|^p \right)^{1/p} \tag{7}$$

Bhattacharyya coefficient (similarity on hyper-sphere) [5]:

$$s_B(\mathbf{x}, \mathbf{y}) = \sum_i \sqrt{x_i y_i} \tag{8}$$

Earth Mover's Distance (EMD) [30]:

$$d_{EMD}(\mathbf{x}, \mathbf{y}) = \sum_i \sum_j c_{i,j} f_{i,j} \tag{9}$$

where $f_{i,j}$, is the set of flows that minimizes the overall distance, (9), subject to the following set of constraints,

$$\sum_i f_{i,j} = y_j, \qquad \sum_j f_{i,j} = x_i, \quad \text{and} \quad f_{i,j} \geq 0. \tag{10}$$

To complete the definition of the EMD, the ground distance, $c_{i,j}$, between histogram bins must be defined. In the empirical evaluation, L_1 (Manhattan) and L_2 (Euclidean) distances, (7), were applied to the spacetime orientation space. The flow values are determined by solving a linear programming problem.

Finally, for any given distance measure, a method must be defined to determine the classification of a given probe relative to the database entries. In order to make the results between the proposed approach and the various recognition results reported elsewhere [43] comparable, the same Nearest-Neighbour (NN) classifier [17] was used in the experiments to be presented. Although not state-of-the-art, the NN classifier has been shown to yield competitive results relative to the state-of-the-art Support Vector Machine (SVM) classifier [35] for dynamic texture classification [10] and thus provides a useful lower-bound on performance.

3. Empirical evaluation

3.1. Database

For the purpose of evaluating the proposed approach, recognition performance was tested on the standard UCLA dynamic texture database [31]. The database is comprised of 50 dynamic texture scenes, including, boiling water, fire, fountains, rippling water and windblown vegetation. Each scene is given in terms of four greyscale image sequences; for each scene, all four example sequences are captured with the same viewing parameters (e.g., identical viewpoint). In total there are 200 sequences. Each sequence consists of 75 frames of size 110×160. Figure 3 shows sample frames from the data set. The remainder of this section documents three recognition experiments that were conducted to evaluate the performance of the proposed approach on the UCLA database.

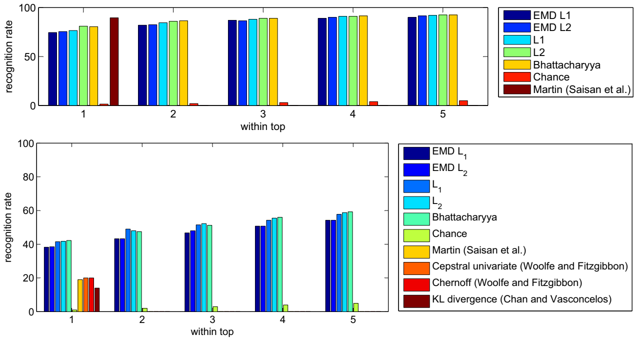

Figure 4. UCLA database recognition results. (top) Viewpoint specific results. Previous state-of-the-art result is denoted by Martin as reported in [31]; this result is based on a NN classifier, SVM classifier-based results, as reported in [10], are slightly higher. All other results correspond to the proposed representational approach under various distance measures. (bottom) Shift-invariant results. The Martin, cepstral univariate, Chernoff and KL divergence results are taken from [43]. All other results correspond to the proposed representational approach under various distance measures. Previous reports [31, 43] do not provide results for matching beyond top 1.

3.2. Viewpoint specific recognition

The first experiment largely followed the standard protocol set forth in conjunction with the original investigation of the UCLA database [31]. The only difference is that unlike [31], where careful manual (spatial) cropping was necessary to reduce computational load in processing, such issues are not a concern in the proposed approach and thus cropping was avoided altogether. (Note that the actual windows used in the original experiments [31] were not reported other than to say that they were selected to, "include key statistical and dynamic features".) As in [31], a correct detection for a given dynamic texture sequence was defined as having one of the three other dynamic texture sequences of its scene as its nearest-neighbour. Thus, the recognition that is tested is *viewpoint specific* in that the correct answer arises as a match between two acquired sequences of the same scene from the same view.

Results are presented in Fig. 4 (top). The highest recognition rate achieved using the proposed spacetime oriented energy approach to representing dynamic texture was 81% with the L_1 and Bhattacharyya measures. Considering the closest five matches, classification improved to 92.5%. Although, below the state-of-the-art NN benchmark of 89.5% using cropped input imagery [31] (and higher rate reported using a SVM classifier, 97.5% [10], again with cropped input), the current results are quite competitive given that the benchmark setting AR-LDS approaches are based on a joint photometric-dynamic model, with the photometric portion playing a pivotal role [9, 43][3]; whereas, the proposed approach focuses strictly on pattern dynamics due to the spatial appearance marginalization step in the construction of the representation, (5). In subsequent experiments, it will be shown that there are distinct advantages to eschewing the purely spatial appearance attributes as one moves beyond viewpoint specific recognition.

3.3. Shift-invariant recognition

To remove the effect of identical viewpoint, and thus the appearance bias in the database, it was proposed that each sequence in the database be cropped into non-overlapping pairs, with subsequent comparisons only performed between different crop locations [43]. Recognition rates under this evaluation protocol showed dramatic reduction in the state-of-the-art LDS-based approaches from approximately 90% to 15% [43]; chance performance was $\sim 1\%$. Further, introduction of several novel distance measures yielded slightly improved recognition rates of $\sim 20\%$ [43]. Restricting comparisons to between non-overlapping portions of the original image sequence data tests *shift-invariant* recognition in that the "view" between instances is spatially shifted. As a practical point, shift-invariant recognition arguably is of more importance than viewpoint specific, as real-world imaging scenarios are unlikely to capture a scene from exactly the same view across two different acquisitions.

The second experiment reported here closely follows previous shift-invariant experiments using the UCLA database, as described above [43]. Each sequence was spatially partitioned into left and right halves (window pairs), with a few exceptions. (In contrast, [43] manually cropped sequences into 48×48 subsequences; again, the location of the crop windows were not reported.) The exceptions arise as several of the imaged dynamic textures are not spatially stationary; therefore, the cropping regimen described above would result in left and right views of different dynamic textures for these cases. For instance, in several of the fire and candle samples, one view would capture a static background, while the other would capture the flame. Previous shift-invariant experiments elected to neglect these cases, resulting in a total of 39 scenes [43]. In the present evaluation, all cases in the database were retained with special manual cropping introduced to the non-stationary cases to include their key dynamic features; documentation of crop windows is available at: www.cse.yorku.ca/vision/research/spacetime-texture. (In experimentation, it was found that dropping these special cases entirely had negligible impact on the overall result.)

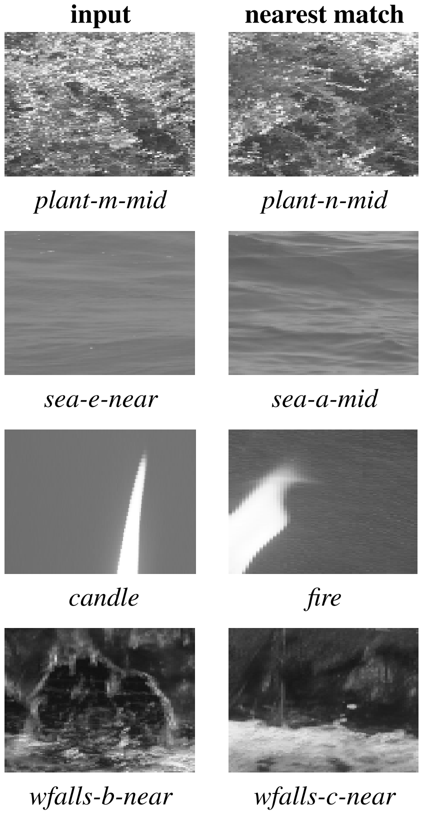

Figure 5. Examples of several misclassifications from the shift-invariant recognition experiment. From a semantic perspective, the inputs and their respective nearest match are equivalent. The text below each figure, indicating the dynamic texture scene, refers to the filename prefix used in the UCLA database.

Overall, the current experimental design yielded a total of 400 sequences, as each of the original 200 sequences were divided into two non-overlapping portions (views). Comparisons were performed only between different views. A correct detection for a given dynamic texture sequence was defined as having one of the four dynamic texture sequences from the other views of its scene as its nearest-neighbour.

The results for the second experiment are presented in Fig. 4 (bottom). In this scenario the proposed approach achieved a 42.3% classification rate, significantly outperforming the the best result of 20% reported in [43]. Con-

[3] Given that the image sequences of each scene in the UCLA database were captured from the exact same viewpoint and that the scenes are visually distinctive based on image stills alone, it has been conjectured that much of the early reported recognition performance was driven mainly by spatial appearance [9]. Subsequently, this conjecture was supported by showing that using the mean frame of each sequence in combination with a NN classifier yielded a 60% classification rate [43].

sidering the closest five matches, classification improved to $\sim 60\%$. This strong performance owes to the proposed spatiotemporal oriented energy representation's ability to capture dynamic properties of visual spacetime without being tied to the specifics of spatial appearance.

Interestingly, close inspection of the results shows that many of the misclassifications for the proposed approach arise between different scenes of semantically the same material, especially from the perspective of visual dynamics. Figure 5 shows several illustrative cases. For example, the most common "confusion" arises from strong matches between two different scenes of fluttering vegetation. Indeed, vegetation dominates the database and consequently has a great impact on the overall classification rate.

Finally, recall that the results in [43] used carefully chosen windows of spatial size 48×48; whereas, the results reported here for the proposed approach are based on simply splitting the full size dynamic textures in half. To control against the impact of additional spatiotemporal support, the proposed approach was also evaluated on cropped windows of similar size to [43]; there was negligible impact on the recognition results.

3.4. Semantic category recognition

Examining the UCLA database, one finds that many of the scenes (50 in total) are capturing semantically equivalent categories. As examples, different scenes of fluttering vegetation share fundamental dynamic texture similarities, as do different scenes of water waves vs. fire, etc; indeed, these similarities are readily apparent during visual inspection of the database as well as the shift-invariant confusions shown in Fig. 5. In contrast, the usual experimental use of the UCLA database relies on distinctions made on the basis of particular scenes, emphasizing their spatial appearance attributes (e.g., flower-c vs. plant-c vs. plant-s) and the video capture viewpoint (i.e., near, medium and far). This parceling of the database overlooks the fact that there are fundamental similarities between different scenes and views of the same semantic category.

In response to the observations above, the final reported experiment reorganizes the database into the following semantic categories (reorganization done by authors, original scenes designated by filename prefix and the total number of sequences given in parentheses): *flames* (16) candle and fire, all instances depict flames; *fountain* (8) fountain-c, depicts spurting spray style fountain; *smoke* (8) smoke; *(water) turbulence* (40) boiling and water, all depict turbulent dynamics; *(water) waves* (24) sea, all depict wave dynamics; *waterfalls* (64) fountain-a, fountain-b and wfalls, fountains that show water flowing down walls thereby similar to waterfalls and hence grouped together; *(windblown) vegetation* (240) flower and plant, all depict fluttering vegetation. Although alternative categorical organizations might be considered, the present one is reasonably consistent with the semantics of the depicted patterns. Evaluation on this data set was conducted using the same procedure outlined for the shift-invariant experiment (Sec. 3.3) to yield *semantic category* recognition.

		Classified						
		flames	*fountain*	*smoke*	*turbulence*	*waves*	*waterfall*	*vegetation*
Actual	*flames* (total 16)	**12**			1		2	1
	fountain (8)		**8**					
	smoke (8)	2		**6**				
	turbulence (40)				**34**		6	
	waves (24)					**24**		
	waterfall (64)				2		**51**	11
	vegetation (240)	3	1		2			**234**

Table 1. Confusion matrix for seven semantic categories.

The semantic category recognition results are shown as a confusion table in Table 1. The overall classification rate in this scenario is 92.3%. As with the previous experiment, inspection of the confusions reveals that they typically are consistent with their apparent dynamic similarities (e.g., waterfall and turbulence confusions, smoke and flames confusions). These results provide strong evidence that the proposed approach is extracting information relevant for delineating dynamic textures along semantically meaningful lines; moreover, that such distinctions can be made based on dynamic information without inclusion of spatial appearance.

4. Discussion and summary

The main contribution of the presented research is the representation of visual spacetime via spatiotemporal orientation distributions for the purpose of recognizing dynamic textures. It has been shown that this tack yields a strong approach to shift-invariant and semantic category-based recognition of dynamic textures. Although the application of spacetime orientation analysis is well documented in the literature for patterns readily characterized as single motion [18, 1, 39, 23, 32] and semi-transparent motion [20, 32, 44, 3], its application to analyzing more complicated phenomena as manifest in dynamic texture patterns, where dominant oriented structure can break down, has received no previous attention.

In this contribution, the dynamic portion of a given texture pattern has been factored out from the purely spatial appearance portion for subsequent recognition. In contrast, LDS-based recognition approaches generally have considered the spatial appearance and dynamic components jointly, which appears to limit performance in significant ways (e.g., weak performance on shift-invariant recognition relative to the proposed approach). The spatial appearance component of these methods is based primarily on a Principal Components Analysis (PCA) that does not represent the current state-of-the-art (e.g., [36]). These observations motivate the future investigation of combining a state-of-the-art appearance-based scheme with the proposed approach to recognizing pattern dynamics.

Although the proposed representation has been presented in terms of oriented filters at a single spatiotemporal scale (i.e., radial frequency), it is an obvious candidate for multi-scale treatment [26]. This extension may serve to support finer categorical distinctions due to characteristic signatures manifesting across scale.

In summary, this paper has presented a unified approach to representing and recognizing dynamic textures based on the underlying pattern dynamics. The approach is based on a distributed characterization of visual spacetime in terms of 3D, (x, y, t), spatiotemporal orientation. Empirical evaluation on a standard database with controls to remove the effects of identical viewpoint shows that the proposed approach achieves superior performance over state-of-the-art methods.

Acknowledgements

G. Doretto provided access to the UCLA dynamic texture database.

References

[1] E. Adelson and J. Bergen. Spatiotemporal energy models for the perception of motion. *JOSA-A*, 2(2):284–299, February 1985.

[2] Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman. Texture mixing and texture movie synthesis using statistical learning. *T-VCG*, 7(2):120–135, April 2001.

[3] S. Beauchemin and J. Barron. The frequency structure of 1D occluding image signals. *PAMI*, 22(2):200–206, February 2000.

[4] J. Bergen. Theories of visual texture perception. In D. Regan, editor, *Vision and Visual Dysfunction*, volume 10B, pages 114–134. Macmillan, NY, 1991.

[5] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distribution. *Bull. Calcutta Math. Soc.*, 35:99–110, 1943.

[6] P. Bouthemy and R. Fablet. Motion characterization from temporal cooccurrences of local motion-based measures for video indexing. In *ICPR*, pages I: 905–908, 1998.

[7] R. Bracewell. *The Fourier Transform and Its Applications*. McGraw-Hill, 2000.

[8] K. Cannons and R. Wildes. Spatiotemporal oriented energy features for visual tracking. pages I: 532–543, 2007.

[9] A. Chan and N. Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. In *CVPR*, pages I: 846–851, 2005.

[10] A. Chan and N. Vasconcelos. Classifying video with kernel dynamic textures. In *CVPR*, 2007.

[11] D. Chetverikov and R. Peteri. A brief survey of dynamic texture description and recognition. In *CORES*, pages 17–26, 2005.

[12] O. Chomat and J. Crowley. Probabilistic recognition of activity using local appearance. In *CVPR*, pages II: 104–109, 1999.

[13] K. Derpanis and J. Gryn. Three-dimensional nth derivative of Gaussian separable steerable filters. In *ICIP*, pages III: 553–556, 2005.

[14] K. Derpanis and R. Wildes. Early spatiotemporal grouping with a distributed oriented energy representation. In *CVPR*, 2009.

[15] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *PETS*, pages 65–72, 2005.

[16] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto. Dynamic textures. *IJCV*, 51(2):91–109, February 2003.

[17] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2001.

[18] M. Fahle and T. Poggio. Visual hyperacuity: Spatio-temporal interpolation in human vision. *Proceedings of the Royal Society of London - B*, 213:451–477, 1981.

[19] A. Fitzgibbon. Stochastic rigidity: Image registration for nowhere-static scenes. In *ICCV*, pages I: 662–669, 2001.

[20] D. Fleet. *Measurement of Image Velocity*. Kluwer, 1992.

[21] W. Freeman and E. Adelson. The design and use of steerable filters. *PAMI*, 13(9):891–906, September 1991.

[22] G. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer, 1995.

[23] D. Heeger. Model for the extraction of image flow. *JOSA-A*, 2(2):1455–1471, August 1987.

[24] D. Heeger and A. Pentland. Seeing structure through chaos. In *Workshop on Motion*, pages 131–136, 1986.

[25] T. Kung and W. Richards. Inferring "water" from images. In W. Richards, editor, *Natural Computation*, pages 224–233. MIT Press, 1988.

[26] T. Lindeberg. Linear spatio-temporal scale-space. In *Scale-Space*, pages 113–127, 1997.

[27] Z. Lu, W. Xie, J. Pei, and J. Huang. Dynamic texture recognition by spatio-temporal multiresolution histograms. In *Workshop on Motion*, pages II: 241–246, 2005.

[28] R. Nelson and R. Polana. Qualitative recognition of motion using temporal texture. *CVGIP*, 56(1):78–89, July 1992.

[29] R. Polana and R. Nelson. Temporal texture and activity recognition. In M. Shah and R. Jain, editors, *Motion-based recognition*, pages 87–115. Kluwer Academic, 1997.

[30] Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's Distance as a metric for image retrieval. *IJCV*, 40(2):99–121, November 2000.

[31] P. Saisan, G. Doretto, Y. Wu, and S. Soatto. Dynamic texture recognition. In *CVPR*, pages II:58–63, 2001.

[32] E. Simoncelli. *Distributed Analysis and Representation of Visual Motion*. PhD thesis, MIT, 1993.

[33] M. Sizintsev and R. Wildes. Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching. In *CVPR*, 2009.

[34] M. Szummer and R. Picard. Temporal texture modeling. In *ICIP*, pages III: 823–826, 1996.

[35] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[36] M. Varma and A. Zisserman. A statistical approach to material classification using image patch exemplars. *PAMI*, 31(11):2032–2047, November 2009.

[37] S. Vishwanathan, A. Smola, and R. Vidal. Binet-Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *IJCV*, 73(1):95–119, June 2007.

[38] Y. Wang and S. Zhu. Modeling textured motion: Particle, wave and sketch. In *ICCV*, pages 213–220, 2003.

[39] A. Watson and A. Ahumada. A look at motion in the frequency domain. In *Motion Workshop*, pages 1–10, 1983.

[40] R. Wildes and J. Bergen. Qualitative spatiotemporal analysis using an oriented energy representation. In *ECCV*, pages II: 768–784, 2000.

[41] D. Williams and R. Sekuler. Coherent global motion percepts from stochastic local motions. *Vision Research*, 24(1):55–62, 1984.

[42] D. Williams, S. Tweten, and R. Sekuler. Using metamers to explore motion perception. *Vision Research*, 31(2):275–286, 1991.

[43] F. Woolfe and A. Fitzgibbon. Shift-invariant dynamic texture recognition. In *ECCV*, pages II: 549–562, 2006.

[44] W. Yu, K. Daniilidis, S. Beauchemin, and G. Sommer. Detection and characterization of multiple motion points. In *CVPR*, pages I: 171–177, 1999.