

Spacetime Texture Representation and Recognition Based on a Spatiotemporal Orientation Analysis

Konstantinos G. Derpanis and Richard P. Wildes

Abstract—This paper is concerned with the representation and recognition of the observed dynamics (i.e., excluding purely spatial appearance cues) of spacetime texture based on a spatiotemporal orientation analysis. The term “spacetime texture” is taken to refer to patterns in visual spacetime, (x, y, t) , that primarily are characterized by the aggregate dynamic properties of elements or local measurements accumulated over a region of spatiotemporal support, rather than in terms of the dynamics of individual constituents. Examples include image sequences of natural processes that exhibit stochastic dynamics (e.g., fire, water and windblown vegetation) as well as images of simpler dynamics when analyzed in terms of aggregate region properties (e.g., uniform motion of elements in imagery, such as pedestrians and vehicular traffic). Spacetime texture representation and recognition is important as it provides an early means of capturing the structure of an ensuing image stream in a meaningful fashion. Toward such ends, a novel approach to spacetime texture representation and an associated recognition method are described based on distributions (histograms) of spacetime orientation structure. Empirical evaluation on both standard and original image data sets show the promise of the approach, including significant improvement over alternative state-of-the-art approaches in recognizing the same pattern from different viewpoints.

Index Terms—Spacetime texture, image motion, dynamic texture, temporal texture, time-varying texture, textured motion, turbulent flow, stochastic dynamics, distributed representation, spatiotemporal orientation

1 INTRODUCTION

1.1 Motivation

Many commonly encountered visual patterns are best characterized in terms of the aggregate dynamics of a set of constituent elements, rather than in terms of the dynamics of the individuals. Several examples of such patterns are shown in Fig. 1. In the computer vision literature, these patterns have appeared collectively under various names, including, turbulent flow/motion [1], temporal textures [2], time-varying textures [3], dynamic textures [4], and textured motion [5]. Typically, these terms have been used with reference to image sequences of natural processes that exhibit stochastic dynamics (e.g., fire, turbulent water and windblown vegetation). In the present work, the broader class that includes stochastic as well as simpler phenomena (e.g., orderly pedestrian crowds, vehicular traffic, and even scenes containing purely translating surfaces) when viewed on a regional basis is considered in a unified fashion. The term “spacetime texture” is used herein in reference to this broad set to avoid confusion with previous terms that focused on various subsets and thus grouped dynamic



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

patterns (e.g., smooth and stochastic patterns) in an artificially disjoint manner.

The ability to discern dynamic patterns based on visual processing is of significance to a number of applications. In the context of surveillance, the ability to recognize dynamic patterns can serve to isolate activities of interest (e.g., biological movement and fire) from distracting background clutter (e.g., windblown vegetation and changes in scene illumination). Further, pattern dynamics can serve as complementary cues to spatial appearance-based ones to support the indexing and retrieval of video. Also, in the context of guiding the decisions of intelligent agents, the ability to discern certain critical dynamic patterns may

• K. Derpanis and R. Wildes are with the Department of Computer Science and Engineering and Centre for Vision Research (CVR), York University, CSEB 1003, 4700 Keele Street, Toronto, Canada.
E-mail: {kosta, wildes}@cse.yorku.ca.

serve to trigger corresponding reactive behaviours (e.g., flight and pursuit).

The goal of the present paper is the introduction of a unified approach to representing and recognizing a diverse set of dynamic patterns with robustness to viewpoint and with 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 primarily on observed dynamics (i.e., excluding purely spatial appearance cues). For such 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., flicker, single motion, multiple motions and scintillation). Correspondingly, each dynamic pattern is 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 captured by a bank of spatiotemporal filters and recognition is performed by matching such distributions. In computing these distributions, the local measurements are aggregated over the region of concern, consistent with considerations of homogeneity often invoked in visual texture analysis [6]. Owing to this aggregation, the spatiotemporal layout of pattern structure is ignored.

1.2 Related work

Various representations have been proposed for characterizing spacetime textures for the purpose of recognition [7]. One strand of research explores physics-based approaches, e.g., [8]. These methods derive models for specific dynamic patterns (e.g., water) 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 type of approach is that the derived models are highly focused on specific patterns, and thus lack generalization to other classes.

Motivated by successes in spatial texture research, approaches have been proposed that uniformly treat a diverse set of dynamic patterns based on aggregate statistics of local descriptors. A seminal example of this approach was based on extracting first- and second-order statistics of motion flow field-based features, assumed to be captured by estimated normal flow [2]. This work was followed-up by numerous proposed variations of normal flow, e.g., [9], and optical flow-based features, e.g., [10]. There are two main drawbacks related to this strand of research. First, normal flow is correlated with spacetime texture spatial appearance [11]. Thus, in contrast to the

goal of the present work, 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 for stochastic dynamics. Rather than capturing dynamic information alone, others have proposed aggregate measurements of local thresholded values to capture the joint photometric-dynamic pattern structure [12].

A recent research trend is the use of statistical generative models to jointly capture the spatial appearance and dynamics of a pattern. Recognition is realized by comparing the similarity between the estimated model parameters. Several variants of this approach have appeared, including: autoregressive (AR) models [13], [14], [15], [5] and multi-resolution schemes [1], [3]. By far the most popular of these approaches for recognition is the joint photometric-dynamic, AR-based Linear Dynamic System (LDS) model, proposed in [14], which has formed the basis for several recognition schemes [4], [16], [17], [18]. Although impressive recognition rates have been reported ($\sim 90\%$), most previous efforts have limited experimentation to cases where the pattern samples are taken from the exact same viewpoint. As a result, much of the performance is highly tied to the spatial appearance captured by these models rather than the underlying dynamics [16], [17]. To address issues with variability in viewpoint, the joint photometric-dynamic LDS model has also been formulated within the bag-of-features framework [18]. Most closely related to the present paper are the LDS variants that have eschewed the appearance component of the model altogether and have instead restricted attention to the dynamic component for recognition [16], [17], as captured by the hidden state space. Significantly, a comparative study of many of the proposed LDS approaches, both joint photometric-dynamic and dynamic-only, 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 [17].

In the current work, spatiotemporal oriented energy filters serve in defining the representation of observed dynamics. Previous research has used similar operators for image sequence analysis toward various ends, including optical flow estimation [19], [20], [21], activity recognition [22], [23], [24], [25], low-level pattern categorization [26], tracking [27], spacetime stereo [28] and grouping [29], [30]. Further, distributions of purely spatially oriented measurements have played a prominent role in the analysis of static visual texture [6], [31] and form the basis of state-of-the-art recognition methods [32], [33], [34]. Moreover, psychophysical evidence suggests that spacetime orientation plays a role in human discrimination of temporally stochas-

tic visual patterns [35], [36]. Psychophysical investigations also have shed light on spatiotemporal perceptual metamers, where physically distinct dynamic patterns are not distinguished by humans [37]. Nevertheless, it appears that the present work is the first to use spatiotemporal orientation as the computational basis for the representation and recognition of spacetime texture. A preliminary version of this research appeared previously [38].

1.3 Contributions

In the light of previous research, the contributions of the present work are as follows. First, a broad set of dynamic patterns are considered that subsume those generally considered disjointly, including both motion and more irregular, stochastic spatiotemporal patterns. This broader set is termed “spacetime texture”. The key unifying attribute of these patterns is that they can be distinguished by their underlying spacetime orientation structure. Second, a particular spatiotemporal filtering formulation is developed for measuring spatiotemporal oriented energy and is used for representing and recognizing spacetime textures based primarily on their underlying dynamics. While spacetime filters have been used before for analyzing image sequences, they have not been applied to the recognition of spacetime textures in the manner proposed. Third, empirical evaluation on a standard data set shows that the proposed approach achieves superior performance over state-of-the-art methods. Fourth, to evaluate the proposed approach on the wider set of patterns encompassed by spacetime textures, a new data set is introduced containing 610 challenging natural videos. The experiments on this data set further demonstrates the efficacy of the proposed approach to modeling and recognition.

2 TECHNICAL APPROACH

2.1 Orientation in visual spacetime

The local orientation (or lack thereof) of a pattern is a salient characteristic. Figure 2 (top and middle rows) illustrates the significance of this structure in terms of describing a range of dynamic patterns in image sequence data (cf. [26]). With reference to Fig. 2, image velocity corresponds to a three-dimensional orientation in (x, y, t) [39], [40], [41], [19], [20]; indeed, motion represents a prominent instance of dominant oriented patterns with static patterns corresponding 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. A lesser known instance of a (approximately) single oriented pattern are image sequences of rain and hard snow streaks [42], [43]. Here, spacetime orientation lies perpendicular to the temporal axis with the spatial orientation component

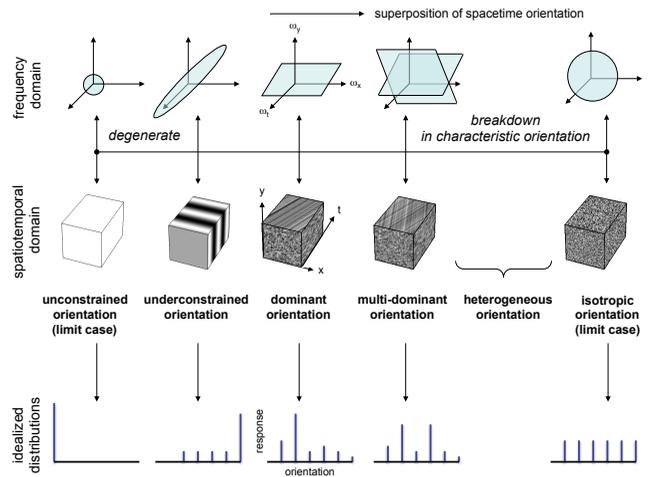


Fig. 2. Spacetime texture spectrum. The top and middle rows depict prototypical patterns of spacetime textures in the frequency and spacetime domains, resp. The horizontal axis indicates the amount of spacetime oriented structure superimposed in a pattern, with increasing amounts given along the rightward direction. The bottom row depicts the distribution (i.e., seven bin histogram) of the relative spacetime oriented structure (or lack thereof) present in each pattern. The first histogram bin captures lack of structure. The remaining histogram bins from left-to-right correspond to spacetime orientations selective for static, rightward motion, upward motion, leftward motion, downward motion and flicker structure. In practice, a larger set of orientations are employed; in this figure, a reduced set is presented for the sake of compact illustration.

corresponding to the streaks’ spatial orientation [42]. In the frequency domain, the energy of these patterns lies approximately on a plane through the temporal frequency axis, where the orientation of the spectral plane is related to the spatial orientation of the streaks [43]; for an illustration of the visual pattern induced by rain in the frequency domain, see Fig. 3 [43].

Beginning with a single spacetime oriented pattern and superimposing spacetime orientations spanning a narrow range about the initial oriented pattern yields “optical snow” [44] and “nowhere-static” [15]. Optical snow arises in many natural situations where the imaged scene elements are restricted to a single direction of motion but vary in speed (e.g., camera translating across a static scene containing a range of depths and vehicular traffic scenes [45], where the speeds may vary but the direction of motion is generally uniform). In the frequency domain, the energy approximately corresponds to a “bow tie” signature formed by the superposition of planes. In contrast, “nowhere-static” patterns do not impose such local directionality constraints (e.g., camera translating over a scene exhibiting stochastic movement, such as windblown flowers and cheering crowds). In the frequency domain, one can think of the energy as corresponding to the superposition of several “bow ties”, each sharing a common

central spacetime orientation. These two patterns and single oriented patterns are collectively referred to as “dominant oriented” patterns herein.

To the left of the dominant oriented pattern reside two degenerate cases corresponding to patterns where the recovery of spacetime orientation is underconstrained (e.g., the aperture problem and pure temporal luminance flicker) and completely unconstrained (e.g., blank wall). In the frequency domain, the energy of the partially specified case corresponds to a line through the origin; in the case of the aperture problem, the spectral line orientation is a function of both the spatial appearance and dynamics [46], given by the component of the pattern velocity along the spatial gradient, while for flicker, the line lies strictly along the temporal frequency axis [26]. In the limit, a region can totally lack any spatiotemporal contrast (unconstrained case) and the frequency domain correlate is isolated in the low-frequency portion of the spectrum (ideally consisting of a DC component only).

Starting again from a single spacetime oriented pattern and superimposing an additional spacetime orientation yields a multi-dominant oriented pattern (e.g., semi-transparency [46] and translucency [47]). 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 isotropic structure (e.g., “television snow”), where no discernable orientations dominate the local region [26]; 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 multi-dominant and isotropic structure lie various complicated phenomena that arise as multiple spacetime oriented structures (e.g., motions) are composited; as noted in Sec. 1.1, these patterns have been called temporal texture [2], dynamic texture [14] and various other terms. 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). These patterns are collectively referred to as “heterogeneous oriented” herein.

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, in the remainder of this paper an attempt is made to recognize and categorize visual spacetime texture using the outputs of local orientation operators aggregated over a region of interest.

2.2 Distributed spacetime orientation

The desired spacetime orientation decomposition is realized using a bank of broadly tuned 3D Gaussian third derivative filters, $G_{3_{\hat{\theta}}} \equiv \partial^3 k \exp[-(x^2 + y^2 + t^2)]/\partial \hat{\theta}^3$, with the unit vector $\hat{\theta}$ capturing the 3D direction of the filter symmetry axis and k a normalization factor. (Filtering details are provided elsewhere [48].) 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 spacetime texture sample under analysis, to yield the following energy measurement for the region

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

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 $\hat{\theta}$. More specifically, this follows from Parseval’s theorem [49] that specifies the phase-independent signal energy in the frequency passband of the Gaussian derivative:

$$E_{\hat{\theta}} \propto \sum_{(\omega_x, \omega_y, \omega_t)} |\mathcal{F}\{G_{3_{\hat{\theta}}} * I\}(\omega_x, \omega_y, \omega_t)|^2, \quad (2)$$

where $(\omega_x, \omega_y, \omega_t)$ denotes the spatiotemporal frequency coordinate, and \mathcal{F} the Fourier transform¹.

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 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 [40]. 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 [50]. Let each plane be parameterized in terms of its unit

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

normal, $\hat{\mathbf{n}}$; a set of equally spaced $N + 1$ directions within the plane are given as

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

with

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

where $\hat{\mathbf{e}}_x$ denotes the unit vector along the ω_x -axis² and $0 \leq i \leq N$. In the case where the spacetime orientation is defined by image velocity $(u_x, u_y)^\top$, the normal vector is given by $\hat{\mathbf{n}} = (u_x, u_y, 1)^\top / \|(u_x, u_y, 1)^\top\|$.

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

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

with $\hat{\theta}_i$ one of $N + 1 = 4$ directions, (3), and each $E_{\hat{\theta}_i}$ calculated via the oriented energy filtering, (1). Note that the discounting of spatial appearance presented here differs from that used elsewhere in using spatiotemporal oriented filtering for optical flow estimation [19], which employs a nonlinear optimization.

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 the four diagonals, 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,

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

where M denotes the number of spacetime orientations considered and ϵ is a constant introduced as a noise floor. As applied to the 27 oriented, appearance

marginalized energy measurements, (5), Eq. (6) produces a corresponding set of 27 normalized, marginalized oriented energy measurements. To this set an additional measurement is included that explicitly captures lack of structure via normalized ϵ ,

$$\hat{E}_\epsilon(\mathbf{x}) = \epsilon / \left(\sum_{j=1}^M \tilde{E}_{\hat{\mathbf{n}}_j}(\mathbf{x}) + \epsilon \right), \quad (7)$$

to yield a 28 dimensional feature vector. (Note that for regions where oriented structure is less apparent, the summation in (7) will tend to 0; hence, \hat{E}_ϵ approaches 1 and thereby indicates relative lack of structure.) This ensemble of (normalized) energy measurements, $\{\hat{E}_{\hat{\mathbf{n}}_i}\} \cup \hat{E}_\epsilon$, is taken as a distribution with spatiotemporal orientation, $\hat{\mathbf{n}}_i$, as variable. In practice, the set of measurements are maintained as a histogram.

Figure 2 (bottom) illustrates a set of idealized histogram signatures realized from (1) - (7). Unconstrained orientation regions (i.e., regions devoid of structure) give rise to a peak response in \hat{E}_ϵ . Unconstrained orientation regions can yield a variety of signatures. Regions containing pure flicker (shown), give rise to a peak response in the orientation selective for flicker/infinite motion. In the case of the aperture problem (not shown), significant responses in multiple orientations will arise in the distribution, since multiple planes are consistent with the spectral line structure of this pattern. Furthermore, this distribution will depend on both the dynamics and spatial appearance of the pattern (i.e., the velocity component along the spatial gradient). Dominant orientation regions (e.g., motion) gives rise to a significant response in only one component of the decomposition, corresponding to a particular orientation. Multi-dominant oriented regions (e.g., semi-transparency and translucency) give rise to significant responses in multiple components of the decomposition, corresponding to the individual single oriented patterns superimposed. Heterogeneously oriented regions (not shown) give rise to a wide variety of distributions, depending on the particulars of the observed phenomena (e.g., fire vs. water vs. fluttering leaves). Finally, isotropic regions give rise to an approximately equal magnitude across all spacetime orientation components.

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

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.

arises with temporal variation (motion, flicker, etc.). Fourth, owing to the broad tuning of the filters, a parsimonious covering of the spacetime orientation space is made possible. More specifically, while the filters have peak responses for particular spacetime orientations and scales, they remain responsive to intermediate instances of these parameters. As a result, the representation is tolerant to small degrees of viewpoint change that manifest as shifts in spacetime orientation and scale. Finally, the representation is efficiently realized via linear (separable convolution and pointwise addition) and pointwise non-linear (squaring and division) operations; thus, efficient computations are realized [48], including real-time realizations on GPUs [25].

2.3 Spacetime orientation distribution similarity

Given the spacetime oriented energy distributions of a query and database with entries represented in like fashion, the final step of the approach is recognition. To compare two histograms, denoted x and y , there are a variety of similarity measures that can be used [51]. In evaluation, the L_1 , L_2 , Earth Mover’s Distance (EMD) [51] (with L_1 and L_2 ground distances) and Bhattacharyya coefficient [52] were considered.

Finally, for any given distance measure, a method must be defined to determine the classification of a given probe relative to the database entries. To make the results between the proposed approach and the various recognition results reported elsewhere [17], [18] comparable, the same Nearest-Neighbour (NN) classifier [53] 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 [54] for dynamic texture classification [55] and thus provides a useful lower-bound on performance. Another motivation for this choice of simple classifier is the desire to evaluate the utility of the proposed representational substrate without confounding performance with classifier sophistication.

3 EMPIRICAL EVALUATION

3.1 Data sets

The performance of the proposed approach to spacetime texture recognition is evaluated on two data sets capturing various subsets of spacetime textures. The first is a standard data set that captures the heterogeneous pattern subset of spacetime textures, as defined in Sec. 2.1. The second is a new data set that captures a diverse set of spacetime textures containing both motion and non-motion-related dynamic patterns, including heterogeneous textures.



Fig. 3. Sample frames from the UCLA data set. (left-to-right, top-to-bottom) Candle, fire, rising smoke, boiling water, fountain, spurting spray style fountain, waterfall and windblown vegetation.

TABLE 1
YUVL spacetime texture data set summary. The number of samples per category are given in parentheses. N/A denotes not applicable.

Basic-Level	Subordinate-Level
<i>unconstrained</i> (16)	N/A
<i>underconstrained</i> (77)	<i>flicker</i> (45)
	<i>aperture problem</i> (32)
	<i>single oriented</i> (229)
<i>dominant</i> (293)	<i>non-single oriented</i> (64)
<i>multi-dominant</i> (85)	N/A
<i>heterogeneous and isotropic</i> (139)	<i>wavy fluid</i> (35)
	<i>stochastic</i> (104)

3.1.1 UCLA data set

For the purpose of evaluating the proposed approach on the heterogeneous pattern subset of spacetime textures, recognition performance was tested on the standard UCLA data set [4], which concentrates on exactly this subset as it has been the focus of much previous work in the area. The data set is comprised of 50 scenes, including, boiling water, fire, fountains, rippling water and windblown vegetation. Each scene is represented by four greyscale image sequences. Critically, all four scene exemplars are captured at a single camera viewpoint, as a result the same area of the scene is captured in each sequence. In total there are 200 sequences, each sequence consisting of 75 frames of size 110×160 . Figure 3 shows sample frames from the data set.

3.1.2 York University Vision Lab (YUVL) spacetime texture data set

To evaluate the proposed approach on spacetime textures, a new data set³ was collected for each of the categories of spacetime texture described in Sec. 2.1 and illustrated in Fig. 2. The data set contains a total of 610 spacetime texture samples. The videos were obtained from various sources, including a Canon HF10 camcorder and the “BBC Motion Gallery” (www.bbcmotiongallery.com) and “Getty Images” (www.gettyimages.com) online video repositories; the videos vary widely in their resolution, temporal extents and capture frame rates. Owing to the diversity within

3. This data set is available at: <http://www.cse.yorku.ca/vision/research/spacetime-texture>.

and across the video sources, the videos generally contain significant differences in scene appearance, scale, illumination conditions and camera viewpoint. Also, for the stochastic patterns, variations in the underlying physical processes ensure large intra-class variations.

The data set is partitioned in two ways: (i) basic-level and (ii) subordinate-level, by analogy to terminology for capturing the hierarchical nature of human categorical perception [56]. The basic-level partition, summarized in Table 1 (left column), is based on the *number* of spacetime orientations present in a given pattern; for a detailed description of these categories, see Section 2.1. For the multi-dominant category, samples were limited to two superimposed structures (e.g., rain over a stationary background). Note, the basic-level partition is not arbitrary, it follows from the systematic enumeration of dynamic patterns based on their spacetime oriented structure presented in Sec. 2.1. Furthermore, each of these categories have been central to research regarding the representation of image-related dynamics, typically considered on a case-by-case basis, as referenced in Sec. 2.1.

To demonstrate the proposed approach’s ability to make finer categorical distinctions, several of the basic-level categories were further partitioned into subordinate-levels. This partition, summarized in Table 1 (right column), is based on the *particular* spacetime orientations present in a given pattern. Beyond the basic-level categorization of an unconstrained oriented pattern (i.e., unstructured), no further subdivision based on pattern dynamics is possible. Underconstrained cases arise naturally as the aperture problem and pure temporal variation (i.e., flicker). Dominant oriented patterns can be distinguished by whether there is a single orientation that describes a pattern (e.g., motion) or a narrow range of spacetime orientations distributed about a given orientation (e.g., “nowhere-static” and “optical snow”). Note that further distinctions might be made based, for example, on the velocity of motion (e.g., stationary vs. rightward motion vs. leftward motion). In the case of multi-dominant oriented patterns, the initial choice of restricting patterns to two components to populate the database precludes further meaningful parsing. The heterogeneous and isotropic basic category was partitioned into wavy fluid and those more generally stochastic in nature. Further parsing within the heterogeneous and isotropic basic-level category is possible akin to the semantic categorization experiment based on the UCLA data set discussed later (see Section 3.2.3). Since this more granular partition is considered in the UCLA data set, in the YUVL data set only a two-way subdivision of the heterogeneous and isotropic basic-level category is considered.

Figure 4 illustrates the overall organization of the YUVL spacetime texture data set in terms of the basic- and subordinate-level categories.

3.2 Heterogeneous spacetime classification

3.2.1 Viewpoint specific classification

The first experiment largely followed the standard protocol set forth in conjunction with the original investigation of the UCLA data set [4]. The only difference is that unlike [4], 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. (Note that the actual windows used in the original experiments [4] were not reported other than to say that they were selected to, “include key statistical and dynamic features”.) As in [4], a leave-one-out classification procedure was used, where a correct classification for a given texture sequence was defined as having one of the three remaining 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. 5. The highest recognition rate achieved using the proposed spatiotemporal oriented energy approach was 81% with the L_2 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 [4] (and higher rate reported using a SVM classifier, 97.5% [55], again with cropped input), the current results are 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 [16], [17]; whereas, the proposed approach focuses on pattern dynamics due to the spatial appearance marginalization step in the construction of the representation, (5). More specifically, 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 [16]. 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 [17], which is well above the performance of random guessing (about 1%). In the experiments to follow, it will be shown that there are distinct advantages to eschewing the purely spatial appearance attributes as one moves beyond viewpoint specific recognition.

3.2.2 Shift-invariant classification

To remove the effect of identical viewpoint, and thus the appearance bias in the data set, it was proposed that each sequence in the data set be cropped into non-overlapping pairs, with subsequent comparisons only performed between different crop locations

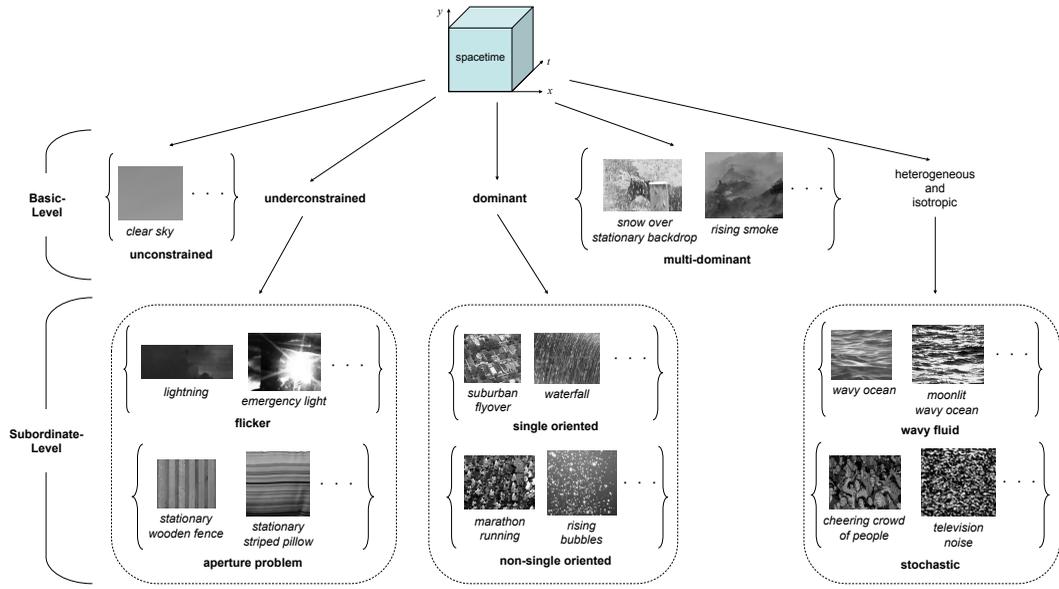


Fig. 4. Organization of the YUVL spacetime texture data set with sample frames for each category.

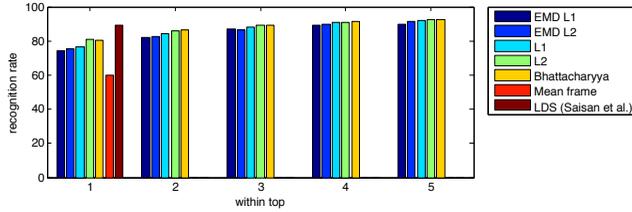


Fig. 5. Viewpoint specific recognition results based on the UCLA data set. EMD $L_{1,2}$, $L_{1,2}$ and Bhattacharyya correspond to the results of the proposed approach with respective distance measures. “Mean frame” refers to using the mean frame of each image sequence with an NN classifier [17]. The previous state-of-the-art result is denoted by LDS as reported in [4]; this result is based on a NN classifier, SVM classifier-based results, as reported in [55], are slightly higher. Previous evaluations do not report matching beyond top 1 [4].

[17]. Recognition rates under this evaluation protocol showed dramatic reduction in the state-of-the-art LDS-based approaches from approximately 90% to 15% [17]; chance performance was $\approx 1\%$. Further, introduction of several novel distance measures yielded slightly improved recognition rates of $\approx 20\%$ [17]. Restricting comparisons 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 data set, as described above [17]. Each sequence was spatially partitioned into left and right halves (window pairs), with a few exceptions. (In

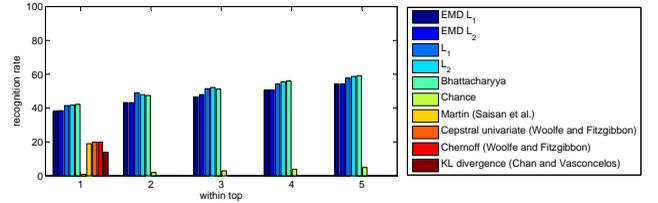


Fig. 6. Shift-invariant recognition results based on the UCLA data set. EMD $L_{1,2}$, $L_{1,2}$ and Bhattacharyya correspond to the results of the proposed approach with respective distance measures. Chance refers to performance based on random guessing. Martin, cepstral univariate, Chernoff and KL divergence results are taken from [17]. Previous evaluations do not report matching beyond top 1 [17].

contrast, [17] 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 scenes are not spatially stationary; therefore, the cropping regimen described above would result in left and right views of different textures for these cases. For instance, in several of the fire and candle samples, one view captures a static background, while the other captures the flame. Previous shift-invariant experiments elected to neglect these cases, resulting in a total of 39 scenes [17]. In the present evaluation, all cases in the data set were retained with special manual cropping introduced to the non-stationary cases to include their key dynamic features; for crop documentation see: 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.)

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

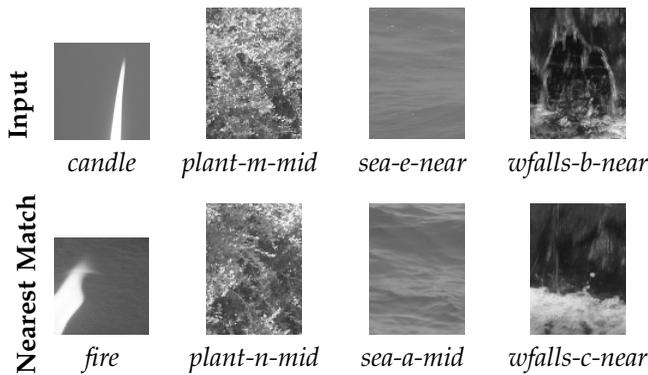


Fig. 8. Examples of several misclassifications from the shift-invariant recognition experiment. From a semantic perspective, the inputs and their respective nearest match are equivalent. The text below each figure, indicating the scene, refers to the filename prefix used in the UCLA data set.

between non-overlapping views. A correct detection for a given texture sequence was defined as having one of the four sequences from the other views of its scene as its nearest-neighbour.

The results for the second experiment are presented in Fig. 6. In this scenario the proposed approach achieved its top classification rate of 42.3% with the Bhattacharyya measure, significantly outperforming the best result of 20% reported elsewhere [17]. Considering the closest five matches, classification improved to $\approx 60\%$. 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. Figure 7 provides several successful classification examples.

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 8 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 data set and consequently has a great impact on the overall classification rate.

Finally, recall that the results reported elsewhere [17] 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 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 previous evaluations [17]. This manipulation was found to have negligible impact on the recognition results.

3.2.3 Semantic category classification

Examining the UCLA data set, one finds that many of the scenes (50 in total) are capturing semantically

TABLE 2

Summary of semantic reorganization of UCLA data set. “filename prefix” refers to the filename prefix of the original scenes in the UCLA data set. The number of samples per category are given in parentheses.

Category	Filename Prefix	Description
<i>flames</i> (16)	candle, fire	flames
<i>fountain</i> (8)	fountain-c	spurting fountain
<i>smoke</i> (8)	smoke	smoke
<i>water</i>	boiling,	turbulent
<i>turbulence</i> (40)	water	water dynamics
<i>water waves</i> (24)	sea	wave dynamics
<i>waterfalls</i> (64)	fountain-{a,b}, wfalls	water flowing down surfaces
<i>windblown</i>	flower,	fluttering
<i>vegetation</i> (240)	plant	vegetation

TABLE 3

Confusion matrix for seven semantic categories of heterogeneous spacetime texture. Results are based on the Bhattacharyya coefficient.

		Classified						
		<i>flames</i>	<i>fountain</i>	<i>smoke</i>	<i>w. turbulence</i>	<i>water waves</i>	<i>waterfall</i>	<i>w. vegetation</i>
Actual	<i>flames</i> (total 16)	12			1		2	1
	<i>fountain</i> (8)		8					
	<i>smoke</i> (8)			6				
	<i>w. turbulence</i> (40)				34		6	
	<i>water waves</i> (24)					24		
	<i>waterfall</i> (64)						51	11
	<i>w. vegetation</i> (240)	3	1		2			234

equivalent categories. As examples, different scenes of fluttering vegetation share fundamental dynamic similarities, as do different scenes of water waves vs. fire, etc.; indeed, these similarities are readily apparent during visual inspection of the data set as well as the shift-invariant confusions shown in Fig. 8. In contrast, the usual experimental use of the UCLA data set 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 data set overlooks the fact that there are fundamental similarities between different scenes and views of the same semantic category rooted in their exhibited image dynamics.

In response to the observations above, the next experiment reorganizes the UCLA data set into the seven semantic categories summarized in Table 2 (reorganization done by authors). Evaluation on this data set was conducted using the same procedure outlined for the shift-invariant experiment to yield *semantic category* recognition.

The semantic category recognition results based on the closest match are shown as a confusion table in Table 3. The overall classification rate in this scenario is 92.3% with the Bhattacharyya measure. As with the previous experiment, inspection of the confusions

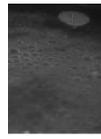
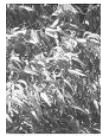
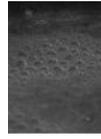
Input						
Nearest Match						
Scene	<i>fire</i>	<i>smoke</i>	<i>boiling-b-near</i>	<i>plant-d-near</i>	<i>sea-a-mid</i>	<i>wfalls-d-near</i>

Fig. 7. Example correct classifications for shift-invariant recognition experiment. In each subfigure, the first row shows an example frame from an input sequence and the second row shows an example frame from the corresponding nearest match in the database. The text below each row indicates the scene and corresponds to the filename prefix used in the UCLA data set.

reveals that they typically are consistent with their apparent dynamic similarities (e.g., waterfall and turbulence confusions, smoke and flames confusions).

While the presented partitioning is reasonably consistent with the semantics of the depicted patterns, alternative categorical organizations might be considered. Elsewhere, an eight class semantic partition of the UCLA data set was presented with focus on evaluating viewpoint invariance [18]. This partitioning only considered 88 hand selected video sequences from the data set taken from different viewpoints (50% for training and 50% for testing). Under this alternative partitioning, the proposed approach achieved an overall correct classification rate of 73%. The number of exemplars per category correctly classified are as follows (number of testing exemplars given on right): boiling water 0/4, fire 4/4, flowers 6/6, fountain 5/8, sea 2/6, smoke 2/2 water 5/6 and waterfall 8/8. Inspection of the confusions reveals that they typically are consistent with their apparent dynamic similarities (e.g., all cases of boiling water misclassified as water and all fountain and sea misclassifications confused amongst each other). In comparison, approaches that capture joint photometric-dynamic aspects of heterogeneous spacetime texture, both holistically [4] and within a bag-of-features framework [18], have been reported elsewhere [18] to achieve between 52% and 70% on this partitioning using the same nearest-neighbour classifier considered here.

The results on the two semantic partitions collectively provide strong evidence that the proposed approach is extracting information relevant for delineating heterogeneous spacetime textures along semantically meaningful lines while the spatial appearance and viewpoint of the pattern may vary; moreover, that such distinctions can be made based on dynamic information without inclusion of spatial appearance.

3.3 Spacetime texture classification

3.3.1 Basic-level classification

As with the previous set of evaluations on heterogeneous spacetime texture classification performance,

TABLE 4
Confusion matrix for the five basic-level categories of spacetime texture. Results are based on the Bhattacharyya coefficient.

		Classified				
		<i>unconstrained</i>	<i>underconstrained</i>	<i>dominant</i>	<i>multi-dominant</i>	<i>hetero. and isotropic</i>
Actual	<i>unconstrained</i> (total 16)	16				
	<i>underconstrained</i> (77)	76			1	
	<i>dominant</i> (293)	1	278		5	9
	<i>multi-dominant</i> (85)	2	6	71		6
	<i>heterogeneous and isotropic</i> (139)	1	3	3		132

the same leave-one-out classification procedure was used to evaluate the performance of the proposed spacetime texture recognition approach using the YUVL spacetime texture data set. Overall results are presented in Fig. 10 (a). The highest recognition rate achieved using the proposed spacetime oriented energy approach was 94% with the L_1 , L_2 and Bhattacharyya similarity measures. In the remainder of this section, discussion will be limited to results based on the Bhattacharyya coefficient measure; results based on the alternative distance measures are generally slightly lower. Considering the closest three matches, classification improved to 98.9%. Class-by-class results are presented in Fig. 10 (b) and Table 4. In the case of the class-by-class results, nearest-neighbour recognition rates ranged between 83.5% to 100%. Considering the closest three matches, recognition rates improved, ranging between 96.5% to 100%. Figure 9 provides an example correct classification for each basic-level category.

Figure 11 provides several representative examples of common misclassifications. The corn field was classified as underconstrained, rather than dominant oriented, as labeled in the ground truth. From the figure it is clear that the rows of corn form a nearly vertical spatial pattern (similar to the wooden fence in the

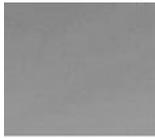
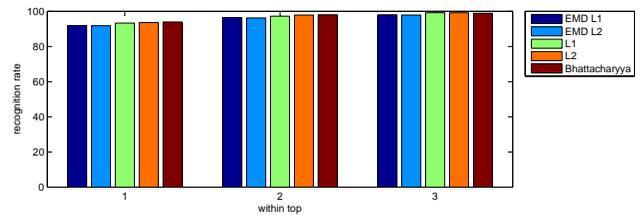
Input	 <i>night sky</i>	 <i>stationary striped pillow</i>	 <i>panning down building</i>	 <i>heavy snow over street</i>	 <i>cheering crowd silhouette</i>
Nearest Match	 <i>clear day sky</i>	 <i>stationary car grille</i>	 <i>panning down cityscape</i>	 <i>heavy snow over Kremlin</i>	 <i>cheering crowd</i>
Basic-Level Category	unconstrained	underconstrained	dominant orientation	multi-dominant orientation	heterogeneous orientation and isotropic structure

Fig. 9. Example correct classifications for basic-level categorization. In each subfigure, the first row shows a frame from an input sequence and the second row shows the corresponding nearest match in the database. The text under each frame provides a description of the video.

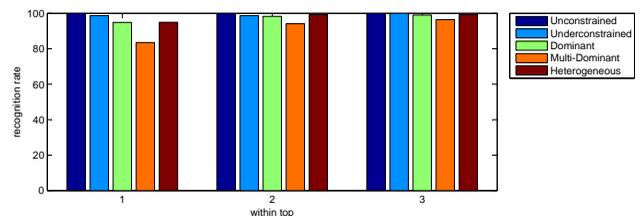
nearest match) and thus could also be considered an instance of the aperture problem (underconstrained). The scene of the people walking down the street was misclassified as heterogeneous oriented, rather than dominant oriented, as labeled. In the input, although there is a small amount of vertical motion downward due to the walking motion, there is also a significant component of bobbing which is visually similar to the nearest match. The corn field and people walking examples highlight the often ambiguous nature of providing ground truth class labels. The scene consisting of falling leaves with a stationary backdrop was misclassified as dominant oriented rather than multi-oriented, as labeled. In this example, the orientation of the nearest sample matched one of the orientation components of the misclassified input. In this case, one of the orientation components of the multi-oriented sequence may lie beyond the resolution (scale) of the filters used; recall that the current analysis is restricted to a single spatiotemporal scale. Such confusions may be addressed via the use of multiple scales of analysis, which is an interesting direction for future research.

3.3.2 Subordinate-level classification

The previous experiment demonstrated strong recognition performance in the context of fairly broad structural categories. This experiment considered finer categorical distinctions using the subordinate-level partition of the YUVL spacetime texture data set. Evaluation on this data (including the unpartitioned categories of “unconstrained” and “multi-dominant”) was conducted using the same leave-one-out procedure outlined for the basic-level experiment above. The subordinate-level recognition results are shown in Fig. 12. Considering only the first nearest-neighbour, class-by-class results ranged between 81.3% to 100%. Considering the closest three matches, recognition



(a) Overall basic-level category results.



(b) Class-by-class basic-level category results.

Fig. 10. Spacetime texture basic-level recognition results. Results for (a) correspond to the proposed approach under various distance measures and (b) to the proposed approach with the Bhattacharyya measure.

rates improved, ranging between 95.3% to 100%. Figure 13 provides an example correct classification for each subordinate-level category.

Figure 14 provides two representative examples of common misclassifications. In the case of the suburb flyover, both the input and nearest match contain roughly the same dominant orientation (same velocity) corresponding to upward motion yet the rising bubbles contain additional deviations. Two possible sources for this misclassification are: (i) the orientation deviations in the rising bubbles sequence are beyond the resolution of the current instantiation of the representation and (ii) the orientation deviations in the bubble sequence are significant, yet the data set does not contain a single-oriented pattern that matches closely with the input (i.e., has the same dominant

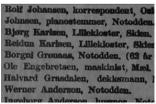
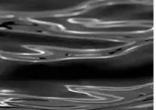
Input	 <i>stationary picket fence</i>	 <i>flashing emergency light</i>	 <i>text scrolling upward</i>	 <i>people riding bicycles</i>	 <i>wavy ocean</i>	 <i>busy trading floor</i>
Nearest Match	 <i>stationary striped wall</i>	 <i>lightning</i>	 <i>panning camera</i>	 <i>crowd walking</i>	 <i>wavy fluid</i>	 <i>commotion of people</i>
Subordinate-Level Category	aperture problem	flicker	single oriented	non-single oriented	wavy fluid	stochastic

Fig. 13. Example correct classifications for subordinate-level categorization. In each subfigure, the first row shows a frame from an input sequence and the second row shows the corresponding nearest match in the database. The text under each frame provides a description of the video.

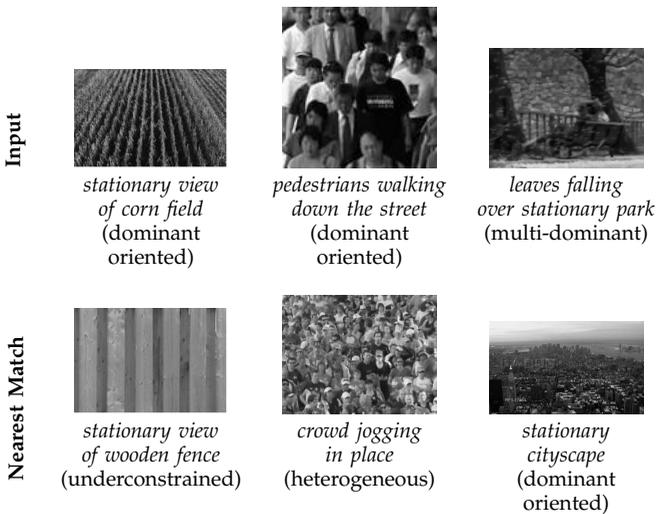


Fig. 11. Example misclassifications for basic-level categorization. In each subfigure, the first row shows a frame from an input sequence and the second row shows the corresponding nearest match in the database. The text under each frame provides a description of the video and basic-level label.

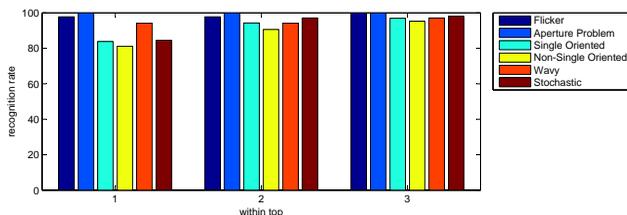


Fig. 12. Spacetime texture subordinate-level recognition results. Class-by-class subordinate-level category results. Results correspond to the proposed approach under the Bhattacharyya measure.

image velocity component). The traffic scene was misclassified as non-single oriented, rather than single oriented, as labeled. One could argue that the traffic scene should also have been labeled as non-single oriented in the ground truth. More generally, most of

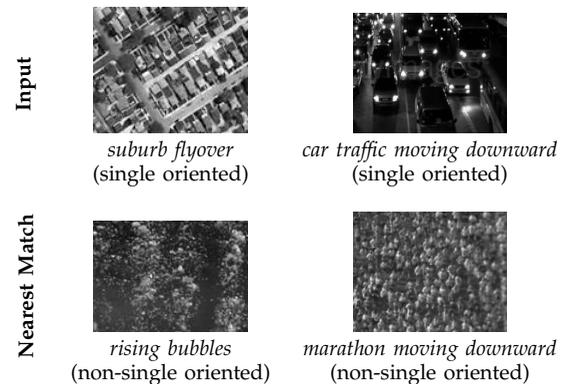


Fig. 14. Example misclassifications for subordinate-level categorization. In each subfigure, the first row shows a frame from an input sequence and the second row shows the corresponding nearest match in the database. The text under each frame provides a description of the video and subordinate-level label.

the misclassifications are of this type (i.e., correspond to matches to “neighbouring” categories). Again, this demonstrates the ambiguous nature of the ground truth labeling task.

Taken together with the results from the basic-level category experiment, the results provide strong evidence that the proposed approach is extracting relevant structural dynamic information to delineate the spectrum of spacetime textures. In particular, the approach is able to represent and recognize patterns encompassing those that traditionally have been treated separately (e.g., motion, dynamic texture, as well as other image dynamics) when considered as aggregate measurements over a region.

4 DISCUSSION AND SUMMARY

There are two main contributions in this paper. First, the definition of texture in the context of dynamics has been broadened to uniformly capture a wide range

of visual spacetime phenomena ranging from deterministic patterns such as motion to more stochastic patterns. The key unifying principle is their underlying first-order spacetime correlation structure. Second, although the application of spacetime oriented filters is well documented in the literature for patterns readily characterized as single motion [40], [41], [19], [20] and semi-transparent motion [46], [20], [57], [58], its application to analyzing more complicated phenomena as manifest in heterogeneous dynamic patterns, where dominant oriented structure can break down, has received no previous attention. Through empirical evaluation it has been shown that this tack yields a strong approach to shift-invariant, viewpoint-invariant and semantic category-based recognition in application to a standard data set. Moreover, similar strong performance was demonstrated in application to a novel spacetime texture data set that encompasses a wider range of dynamic phenomena.

In this contribution, the dynamic portion of a texture pattern has been factored out from the purely spatial appearance portion for subsequent recognition. In contrast, state-of-the-art 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). Furthermore, restricting analysis to the dynamic portion of the LDS model also appears to limit performance significantly. These observations motivate the future investigation of combining a state-of-the-art appearance-based scheme with the proposed approach to recognizing pattern dynamics. The emphasis in comparison has been on the LDS model and its variants (joint photometric-dynamic, dynamics only and bags of LDS) that form the state-of-the-art and that have been evaluated on the common UCLA data set; however, future work can consider additional comparisons, including combination of the proposed approach with appearance and comparison to additional approaches that include both appearance and dynamics). Significantly, the ability to tease apart the spatial appearance and dynamic components of a pattern, rather than jointly (e.g., [4], [12], [18]), is a worthy pursuit in its own right. This approach allows for a compact vocabulary of appearance- and dynamic-only descriptions that can be combined in a variety of ways to realize a rich pattern description.

Although the proposed representation has been presented in terms of oriented filters tuned to a single spatiotemporal scale (i.e., radial frequency), it is an obvious candidate for multi-scale treatment [59]. This extension may serve to support finer categorical distinctions due to characteristic signatures manifesting across scale. Another possible direction of research concerns the use of combinations of second- and higher-order statistics. Higher-order features may

be useful in characterizing pattern dynamics whose oriented structure varies spatiotemporally (i.e., non-stationary spatiotemporal statistics), such as affine motion (varying across space) and acceleration (varying across time); nonetheless, it has been demonstrated that the current approach based on first-order statistics has successfully captured a broad and important set of dynamic patterns.

In summary, this paper has presented a unified approach to representing and recognizing spacetime textures from 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 both standard and original image data sets, including quantitative comparisons with state-of-the-art methods, demonstrates the potential of the proposed approach.

ACKNOWLEDGEMENTS

Portions of this research were funded by an NSERC Discovery Grant to R. Wildes.

REFERENCES

- [1] D. Heeger and A. Pentland, "Seeing structure through chaos," in *Workshop on Motion*, 1986, pp. 131–136.
- [2] R. Nelson and R. Polana, "Qualitative recognition of motion using temporal texture," *CVGIP*, vol. 56, no. 1, pp. 78–89, 1992.
- [3] Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman, "Texture mixing and texture movie synthesis using statistical learning," *T-VCCG*, vol. 7, no. 2, pp. 120–135, 2001.
- [4] P. Saisan, G. Doretto, Y. Wu, and S. Soatto, "Dynamic texture recognition," in *CVPR*, 2001, pp. II:58–63.
- [5] Y. Wang and S. Zhu, "Modeling textured motion: Particle, wave and sketch," in *ICCV*, 2003, pp. 213–220.
- [6] J. Bergen, "Theories of visual texture perception," in *Vision and Visual Dysfunction*, D. Regan, Ed. Macmillan, 1991, vol. 10B, pp. 114–134.
- [7] D. Chetverikov and R. Peteri, "A brief survey of dynamic texture description and recognition," in *CORES*, 2005, pp. 17–26.
- [8] T. Kung and W. Richards, "Inferring "water" from images," in *Natural Computation*, W. Richards, Ed. MIT Press, 1988, pp. 224–233.
- [9] P. Bouthemy and R. Fablet, "Motion characterization from temporal cooccurrences of local motion-based measures for video indexing," in *ICPR*, 1998, pp. I: 905–908.
- [10] Z. Lu, W. Xie, J. Pei, and J. Huang, "Dynamic texture recognition by spatio-temporal multiresolution histograms," in *Workshop on Motion*, 2005, pp. II: 241–246.
- [11] R. Polana and R. Nelson, "Temporal texture and activity recognition," in *Motion-based recognition*, M. Shah and R. Jain, Eds. Kluwer, 1997.
- [12] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *PAMI*, vol. 29, no. 6, pp. 915–928, 2007.
- [13] M. Szummer and R. Picard, "Temporal texture modeling," in *ICIP*, 1996, pp. III: 823–826.
- [14] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto, "Dynamic textures," *IJCV*, vol. 51, no. 2, pp. 91–109, 2003.
- [15] A. Fitzgibbon, "Stochastic rigidity: Image registration for nowhere-static scenes," in *ICCV*, 2001, pp. I: 662–669.
- [16] A. Chan and N. Vasconcelos, "Probabilistic kernels for the classification of auto-regressive visual processes," in *CVPR*, 2005, pp. I: 846–851.
- [17] F. Woolfe and A. Fitzgibbon, "Shift-invariant dynamic texture recognition," in *ECCV*, 2006, pp. II: 549–562.

- [18] A. Ravichandran, R. Chaudhry, and R. Vidal, "View-invariant dynamic texture recognition using a bag of dynamical systems," in *CVPR*, 2009.
- [19] D. Heeger, "Model for the extraction of image flow," *JOSA-A*, vol. 2, no. 2, pp. 1455–1471, 1987.
- [20] E. Simoncelli, "Distributed analysis and representation of visual motion," Ph.D. dissertation, MIT, 1993.
- [21] G. Granlund and H. Knutsson, *Signal Processing for Computer Vision*. Norwell, Massachusetts: Kluwer, 1995.
- [22] O. Chomat and J. Crowley, "Probabilistic recognition of activity using local appearance," in *CVPR*, 1999, pp. II: 104–109.
- [23] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *PETS*, 2005, pp. 65–72.
- [24] K. Derpanis, M. Sizintsev, K. Cannons, and R. Wildes, "Efficient action spotting based on a spacetime oriented structure representation," in *CVPR*, 2010.
- [25] A. Zaharescu and R. Wildes, "Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing," in *ECCV*, 2010.
- [26] R. Wildes and J. Bergen, "Qualitative spatiotemporal analysis using an oriented energy representation," in *ECCV*, 2000, pp. 768–784.
- [27] K. Cannons, J. Gryn, and R. Wildes, "Visual tracking using a pixelwise spatiotemporal oriented energy representation," in *ECCV*, 2010.
- [28] M. Sizintsev and R. Wildes, "Spatiotemporal stereo via spatiotemporal quadric element (stequel) matching," in *CVPR*, 2009.
- [29] K. Derpanis and R. Wildes, "Early spatiotemporal grouping with a distributed oriented energy representation," in *CVPR*, 2009.
- [30] —, "Detecting spatiotemporal structure boundaries: Beyond motion discontinuities," in *ACCV*, 2009.
- [31] M. Tuceryan and A. Jain, "Texture analysis," in *Handbook of Pattern Recognition and Computer Vision (2nd Edition)*, C. Chen, L. Pau, and P. Wang, Eds. World Scientific Publishing, 1998.
- [32] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *IJCV*, vol. 43, no. 1, pp. 29–44, 2001.
- [33] O. Cula and K. Dana, "3D texture recognition using bidirectional feature histograms," *IJCV*, vol. 59, no. 1, pp. 33–60, 2004.
- [34] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *IJCV*, vol. 62, no. 1-2, pp. 61–81, 2005.
- [35] D. Williams and R. Sekuler, "Coherent global motion percepts from stochastic local motions," *Vision Research*, vol. 24, no. 1, pp. 55–62, 1984.
- [36] D. Williams, S. Tweten, and R. Sekuler, "Using metamers to explore motion perception," *Vision Research*, vol. 31, no. 2, pp. 275–286, 1991.
- [37] S. Treue, K. Hol, and H. Rauber, "Seeing multiple directions of motion - Physiology and psychophysics," *Nature Neuroscience*, vol. 3, pp. 270–276, 2000.
- [38] K. Derpanis and R. Wildes, "Dynamic texture recognition based on distributions of spacetime oriented structure," in *CVPR*, 2010.
- [39] M. Fahle and T. Poggio, "Visual hyperacuity: Spatio-temporal interpolation in human vision," *Proceedings of the Royal Society of London - B*, vol. 213, no. 1193, pp. 451–477, 1981.
- [40] A. Watson and A. Ahumada, "A look at motion in the frequency domain," in *Motion Workshop*, 1983, pp. 1–10.
- [41] E. Adelson and J. Bergen, "Spatiotemporal energy models for the perception of motion," *JOSA-A*, vol. 2, no. 2, pp. 284–299, 1985.
- [42] K. Garg and S. Nayar, "Vision and rain," *IJCV*, vol. 75, no. 1, pp. 3–27, 2007.
- [43] P. Barnum, S. Narasimhan, and T. Kanade, "Analysis of rain and snow in frequency space," *IJCV*, vol. 86, no. 2-3, pp. 256–274, 2010.
- [44] M. Langer and R. Mann, "Optical snow," *IJCV*, vol. 55, no. 1, pp. 55–71, 2003.
- [45] K. Derpanis and R. Wildes, "Classification of traffic video based on a spatiotemporal orientation analysis," in *WACV*, 2011, pp. 606–613.
- [46] D. Fleet, *Measurement of Image Velocity*. Kluwer, 1992.
- [47] K. Derpanis and R. Wildes, "The structure of multiplicative motions in natural imagery," *PAMI*, vol. 32, no. 7, pp. 1310–1316, 2010.
- [48] K. Derpanis and J. Gryn, "Three-dimensional nth derivative of Gaussian separable steerable filters," in *ICIP*, 2005, pp. III: 553–556.
- [49] R. Bracewell, *The Fourier Transform and Its Applications*. McGraw-Hill, 2000.
- [50] W. Freeman and E. Adelson, "The design and use of steerable filters," *PAMI*, vol. 13, no. 9, pp. 891–906, 1991.
- [51] Y. Rubner, C. Tomasi, and L. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *IJCV*, vol. 40, no. 2, pp. 99–121, 2000.
- [52] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distribution," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–110, 1943.
- [53] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley, 2001.
- [54] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.
- [55] A. Chan and N. Vasconcelos, "Classifying video with kernel dynamic textures," in *CVPR*, 2007.
- [56] E. Rosch and C. Mervis, "Family resemblances: Studies in the internal structure of categories," *Cognitive Psychology*, vol. 7, pp. 573–605, 1975.
- [57] W. Yu, K. Daniilidis, S. Beauchemin, and G. Sommer, "Detection and characterization of multiple motion points," in *CVPR*, 1999, pp. I: 171–177.
- [58] S. Beauchemin and J. Barron, "The frequency structure of 1D occluding image signals," *PAMI*, vol. 22, no. 2, pp. 200–206, 2000.
- [59] T. Lindeberg, "Linear spatio-temporal scale-space," in *Scale-Space*, 1997, pp. 113–127.



Konstantinos Derpanis (Member, IEEE) received the BSc (Honours) degree in computer science from the University of Toronto, Canada, in 2000, and the MSc and PhD degrees in computer science from York University, Toronto, Canada, in 2003 and 2010, resp. He received the Canadian Image Processing and Pattern Recognition Society (CIPPRS) Doctoral Dissertation Award 2010 Honorable Mention. He spent the summer of 2005 at Sarnoff Corporation in Princeton,

New Jersey, as an intern developing a robust stereo vision system for an automotive-related application. Currently, he is a postdoctoral researcher in the GRASP Laboratory and the Department of Computer and Information Science at the University of Pennsylvania. His major field of interest is computer vision with an emphasis on low-level modeling of image dynamics and human motion understanding.



Richard Wildes (Member, IEEE) received the PhD degree from the Massachusetts Institute of Technology in 1989. Subsequently, he joined Sarnoff Corporation in Princeton, New Jersey, as a Member of the Technical Staff in the Vision Technologies Group, where he remained until 2001. In 2001, he joined the Department of Computer Science and Engineering at York University, Toronto, where he is an Associate Professor and a member of the Centre for Vision Research.

Honours include receiving a Sarnoff Corporation Technical Achievement Award, the IEEE D.G. Fink Prize Paper Award for his Proceedings of the IEEE publication "Iris recognition: An emerging biometric technology" and twice giving invited presentations to the US National Academy of Sciences. His main areas of research interest are computational vision, as well as allied aspects of image processing, robotics and artificial intelligence.