Dynamic Scene Understanding: The Role of Orientation Features in Space and Time in Scene Classification

Konstantinos G. Derpanis[1], Matthieu Lecce[1], Kostas Daniilidis[1] and Richard P. Wildes[2]
[1]GRASP Laboratory, University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Computer Science and Engineering, York University, Toronto, ON, Canada
{derpanis, mlecce, kostas}@cis.upenn.edu, wildes@cse.yorku.ca

Abstract

Natural scene classification is a fundamental challenge in computer vision. By far, the majority of studies have limited their scope to scenes from single image stills and thereby ignore potentially informative temporal cues. The current paper is concerned with determining the degree of performance gain in considering short videos for recognizing natural scenes. Towards this end, the impact of multiscale orientation measurements on scene classification is systematically investigated, as related to: (i) spatial appearance, (ii) temporal dynamics and (iii) joint spatial appearance and dynamics. These measurements in visual space, x-y, and spacetime, x-y-t, are recovered by a bank of spatiotemporal oriented energy filters. In addition, a new data set is introduced that contains 420 image sequences spanning fourteen scene categories, with temporal scene information due to objects and surfaces decoupled from camera-induced ones. This data set is used to evaluate classification performance of the various orientation-related representations, as well as state-of-the-art alternatives. It is shown that a notable performance increase is realized by spatiotemporal approaches in comparison to purely spatial or purely temporal methods.

1. Introduction

Natural scene classification is a fundamental challenge in the goal of automated image understanding. Here, "scene" refers to a place where an action or event occurs. The ability to distinguish scenes is very useful, as it can serve to provide priors for the presence of actions [25], surfaces [15] and objects [34] (e.g., for street scenes, it is highly probable to find cars and pedestrians), as well as their locations and scales. Moreover, similar scenes could be retrieved from a database.

A critical challenge to dynamic scene understanding arises from the wide range of naturally occurring phenomena that must be encompassed. Figure 1 shows sample frames from the data set introduced in this paper that highlight such diversity. Although of obvious importance, image motion (i.e., spatial displacement of image elements with time), as arises from the projected movement of scene elements (e.g., cars in "highway" and opening doors in "elevator"), represents a particular instance of the myriad spatiotemporal patterns encountered in the world. Examples of non-motion-related patterns of significance include, non-textured regions (e.g., sky in "sky-cloud"), flicker (i.e., pure temporal intensity change, e.g., fire in "forest fire" and lightning in "lightning storm"), and dynamic texture (e.g., as typically associated with stochastic phenomena, such as the turbulence in "rushing river" and waves in "ocean").

Figure 1. Sample frames of all scene categories from the data set introduced in this paper; see Sec. 3 for details.

The present paper is concerned with investigating the early representation (i.e., the building blocks) of image sequences for the purpose of recognizing natural scene categories. Numerous studies have reported success in classifying scenes through bypassing object recognition and segmentation processes and instead rely on the global layout (the schema or *gist*) of aggregated statistics of early visual cues, such as the power spectrum, orientation and color, e.g., [26, 11, 36].

An emerging theme of previous work in (spatial)

scene recognition, is that gradient-based features that capture orientation (e.g., GIST, SIFT, HOG, etc.) yield a rich set of cues. Consonant with these findings, the current paper extends such ideas to the temporal domain by adopting a representation that naturally integrates both spatial appearance and dynamic information according to local measurements of 3D, x-y-t, orientation structure that are accumulated over fixed subregions of visual spacetime and then related via their global layout. In particular, each spacetime subregion is associated with a distribution of measurements indicating the relative presence of a particular set of spatiotemporal orientations.

Several early studies considered a small number of broad scene categories from image stills, such as indoor vs. outdoor and city vs. landscape (e.g., [11, 35]). More recently, focus has been placed on distinguishing categories numbering in the tens [10, 20] and hundreds [38]. As noted above, common among most approaches to scene recognition is the use of early visual features (but see [21] for an object-centric approach), e.g., spectral features [33, 26], local orientation [11, 14, 29, 20, 31] and color [13, 36], that are in turn aggregated over the entire image (e.g., [29]) or within fixed image subregions (e.g., [11, 20]) to drive recognition.

Others have considered modeling scenes hierarchically through intermediate substrates built upon early visual features. These representations have come in the form of semantic descriptions (e.g., water vs. sky [4, 37] and ruggedness vs. openness [26]) and latent theme models [10, 24, 3, 28].

While the vast majority of the literature has centered on scene recognition from image stills, two notable exceptions have appeared [25, 30], where histograms of optical flow (HOF) [25] and chaotic system parameters [30] are used to model scene dynamics. A drawback of optical flow is that it is limited in the complexity of patterns it can capture, as local non-translational image motions, such as, multiple motions at an image point, temporal flicker (e.g., lightning) and dynamic textures (e.g., dynamic water), violate the underlying assumptions of the flow computation, e.g., brightness constancy.

Linear dynamical systems (LDS) have been proposed as models to the restricted class of video patterns termed dynamic textures [9]. While LDS models have shown promise on dynamic texture classification, their application to classifying the wider set of patterns found in dynamic scenes has been shown to perform poorly [30]. More closely related to the present research is previous work on dynamic textures that have made use of spatiotemporal oriented energy features to capture pattern structure [8]. Indeed, the present research makes use of the same primitive filtering operations to derive orientation features; however, it significantly differs in three ways. First, the two efforts are concerned with very different problem domains, dynamic texture vs. dynamic scene analysis. In texture analysis one typically is concerned with a single relatively uniformly structured region; whereas, in scene analysis it is typical for several regions of differing type and their inter-relationships to be of concern. Second, the previous work applied its filtering at a single spatiotemporal scale; whereas, the present work employs multiscale filtering to capture a more detailed feature set. Third, the previous work aggregated its filter responses across an entire image; whereas, the present work aggregates over subregions defined over a spatiotemporal grid to maintain scene layout information not available in the previous work. As noted above, maintenance of scene layout is of much greater concern in scene analysis in comparison to texture analysis.

In addition to dynamic texture analysis [8], measures of spatiotemporal oriented energy have been used previously to capture a wide range of dynamic patterns, including image motion [1, 12, 32], semi-transparency [6] and human actions [5, 16, 7]. Moreover, energy measurements of purely spatial orientation are popular in the analysis of static scenes [26] and have a long history in the analysis of (static) visual texture [2]. Nevertheless, it appears that the present work is the first to use spatiotemporal oriented energies as the computational basis for the representation of information in the context of recognizing dynamic natural scenes.

For evaluation, two data sets previously have been introduced based on natural scenes culled from "in-the-wild" sources (i.e., cinematic movies [25] and amateur footage from the Internet [30]). While these sources are appealing because they are readily available and representative of the type of footage that certain applications would be expected to process, there remain significant drawbacks in experimenting with such data (cf., [27]). Most prominent is the inclusion of significant (distracting) camera motion and scene cuts; thus, temporal information is confounded by both scene-related dynamics and camera-related motion. Consequently, it is difficult to tell whether success of a particular approach arises as it captures critical aspects of scenes vs. temporal regularities introduced by extraneous sources (e.g., camera movement). Furthermore, for failures there is no clear indication of the source (e.g., failure of representation to capture critical scene-related dynamics or lack of invariance to camera motion).

Contributions: The present paper makes three main contributions. First, the impact of multiscale orientation measurements on scene classification are system-

atically investigated, as related to: (i) spatial appearance, (ii) temporal dynamics and (iii) joint spatial appearance and dynamics. These orientation measurements are realized as distributions capturing oriented structure in visual space, x-y, and spacetime, x-y-t, as recovered by a bank of spatiotemporal oriented filters. The spatiotemporal orientation measurements capture a wide range of dynamic patterns in natural scenes, both motion (e.g., object movement) and more complicated dynamics (e.g., flickering prominent in fire, lightning and water) as well as purely spatial pattern (e.g., static surface texture). While spatiotemporal filters have been used before for analyzing image sequences in a variety of contexts, they have not been applied to the recognition of natural scenes. Second, given the aim of exploring temporal information present in natural scenes, a new data set is introduced that contains 420 videos spanning fourteen scene categories. Emphasis is placed on isolating temporal scene information due to objects and surfaces from camera-induced ones, as present in previous data sets [25, 30]. Third, a detailed empirical evaluation on both extant public data and the data introduced in this paper is provided that demonstrates overall strong performance of jointly modeling spatial appearance and dynamics via multiscale oriented energies. These comparisons are conducted with representations focused on static appearance alone as well as the previous state-of-the-art in joint modeling of appearance and dynamic information.

2. Methodology

There are two key parts to the analysis of dynamic scene recognition considered in this paper: (i) a representation based on the global layout of local spacetime orientation measurements that are aggregated across image subregions; (ii) a match measure between any two samples under consideration. Section 2.1 provides a summary of the oriented energy measurements used in this work to systematically evaluate the relative impact of spatial appearance [26], temporal/dynamic [8] and joint spatial appearance and dynamic information, on scene classification. Section 2.2 extends the spatiotemporal oriented energy model by introducing two scale parameters that determine the spatiotemporal details captured by the oriented energy representation and the layout of the energies.

2.1. Spatiotemporal oriented energy features

In the current investigation, orientation features are used to describe subregions of an imaged scene that are derived via application of an orientation tuned filter bank. In particular, the employed filtering operations follow previous work [8], where it was used instead for dynamic texture analysis and without concern for multiscale analysis that is employed in the current work.

The spacetime orientation measurements are constructed by filtering using a set of Gaussian derivative filters, pointwise squaring and summation over a given spacetime region,

$$E_{\hat{\theta},\sigma} = \sum_{\mathbf{x}} \Omega(\mathbf{x})[G_{N_{\hat{\theta},\sigma}}(\mathbf{x}) * I(\mathbf{x})]^2, \qquad (1)$$

where $\mathbf{x} = (x, y, t)^\top$ denotes the spatiotemporal image coordinates, $I(\mathbf{x})$ the input image sequence, $*$ convolution, $\Omega(\mathbf{x})$ a mask defining the aggregation region and $G_{N_{\hat{\theta},\sigma}}(\mathbf{x})$ the Nth derivative of the Gaussian with scale σ and $\hat{\theta}$ the direction of the filter's axis of symmetry, and care taken to normalize the filters to ensure that their energy across scale is constant [22].

The initial definition of local energy measurements, (1), is confounded by local image contrast that appear as a multiplicative constant in the set of energies. To remove contrast-related information, the energy measures, (1), are normalized by the ensemble of oriented responses,

$$\hat{E}_{\hat{\theta}_i,\sigma_j} = E_{\hat{\theta}_i,\sigma_j} / \left(\epsilon + \sum_{\hat{\theta}\times\sigma\in\mathbb{S}} E_{\hat{\theta},\sigma} \right), \qquad (2)$$

where $\mathbb{S}$ denotes the set of considered multiscale oriented energies, (1), and ϵ is a constant that serves as a noise floor. In addition, a normalized ϵ is computed, as in (2), to explicitly capture lack of texture within the region. (Note that regions where texture is less apparent, e.g., region of sky, the summation in the denominator approaches zero; hence, the normalized ϵ approaches one and thereby indicates lack of structure.)

The normalized oriented energy responses, (2), form the local basis for analyzing dynamic scenes in this paper, as they jointly capture static spatial and dynamic structure in imagery. The GIST representation [26] for scene classification can be seen as a spatial analogue of the presented spatiotemporal oriented energy method, in its use of purely spatial oriented energy measurements in (1) to characterize local image structure.

As an approach to understanding the contribution of just the dynamic component, it is possible to discount the impact of purely spatial appearance on the presented oriented energy filtering via a marginalization process [8]. The local spacetime orientation structure of a visual pattern (remaining after marginalization) has been previously shown to capture significant, meaningful aspects of its dynamics [8]. As examples: Purely spatial pattern structure (e.g., surface texture) is captured by orientations that are parallel to the image plane; whereas, dynamic attributes of the scene (e.g., velocity and flicker) are captured by orientations extending into time. In particular, to remove the de-

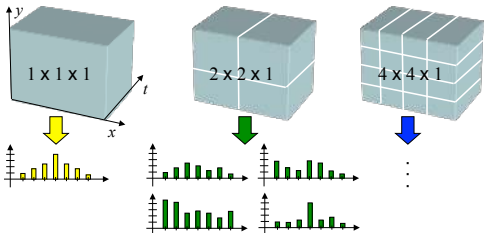

Figure 2. Outer scale and scene layout examples. The input image sequence is spatially subdivided. Outer scale is determined by the spatiotemporal support of the individual subdivided regions. Relative position of subdivided regions captures scene layout. The illustrated histograms correspond to the energy distributions, (2), within each image sequence subdivision.

pendence on the spatial orientation component, linear combinations of the initial energy measures, (1), supporting a single spacetime orientation are taken, given by the unit normal $\hat{\mathbf{n}}$, corresponding to its frequency domain plane. (Recall, a pattern exhibiting a single spacetime orientation, e.g., velocity, manifests as a plane through the origin in the frequency domain [1].) The resulting energy measures are expressed as:

$$\tilde{E}_{\hat{\mathbf{n}},\sigma} = \sum_{i=0}^{N} E_{\hat{\theta}_i,\sigma}, \tag{3}$$

where $\hat{\theta}_i$ represents one of $N+1$ equal spaced orientation tunings consistent with direction $\hat{\mathbf{n}}$ and N the order of the Gaussian derivative filter; for details see [8]. To complete this filtering process, the appearance marginalized responses, (3), are normalized for contrast, (2).

In summary, (1)-(2) yield a distribution/histogram indicating the relative presence of a particular set of spacetime orientations in the input imagery and spatial orientations when filtering is restricted to the spatial domain. Significantly, the derived measurements are invariant to additive and multiplicative bias in the image signal, due to the bandpass nature of (1) and the normalization, (2), resp. Invariance to such biases provides a degree of robustness to various potentially distracting photometric effects (e.g., overall scene illumination, sensor sensitivity). Application of these filters thereby provides an integrated approach to capturing both the local spatial and temporal structure of imagery. Further, to study the descriptive power of dynamic information alone, the initial filter responses, (1), can be marginalized for purely spatial appearance, (3), prior to normalization, (2).

2.2. Spatiotemporal scale: Inner vs. Outer

Within the spatiotemporal oriented energy representation, (1)-(2), one can identify two types of scale parameters, namely, the inner and outer scale [17] that serve distinct roles. Notably, these parameters were both limited to a single scale in previous work [8].

The inner scale, corresponding to the Gaussian filter standard deviation, σ, in the energy computation, (1), determines the range of spatiotemporal details captured by the representation. The outer scale, given by $\Omega(x, y, t)$ in (1), specifies the spatiotemporal scale of the support region for aggregating measurements. Limiting the outer scale to the entire image sequence itself (cf., [8]) disregards the spatiotemporal layout of dynamic structure and thus ignores a potentially diagnostic cue for scene classification. Similar to previous work (e.g., [11, 20, 19]), coarse spatiotemporal layout information is introduced by subdividing the spatial dimensions of the image sequence at increasingly finer outer scales; the oriented energy representation is computed within each region separately and the grid arrangement of the subdivisions captures scene layout (see Fig. 2).

2.3. Classification

To emphasize the relative strengths of the orientation representations to tease out critical scene regularities, while not confounding success with classifier sophistication, a Nearest Neighbour (NN) classifier was used in all evaluations. The set of normalized oriented energy measurements, (2), within each outer scale form a histogram. Preliminary investigation considered a variety of (dis)similarity measures, e.g., Bhattachayya, L_1, L_2 and χ^2, that yielded little difference in classification performance. The Bhattacharyya coefficient provided slightly better overall performance, similar to previous work [8]. Consequently, only results for this measure are presented here. The final global similarity between two scenes is realized as the sum across the histogram similarities computed for each outer scale,

$$s(\mathbf{u}, \mathbf{v}) = \sum_i \sum_j \sqrt{u_{i,j} v_{i,j}}, \tag{4}$$

where $\mathbf{u}$ and $\mathbf{v}$ denote the scene descriptors, (2), i and j index over the outer scale partition of the scene and the individual entries in the histograms, resp., and the innermost summation is the Bhattacharyya coefficient.

2.4. Implementation details

In the presented experiments, 3D Gaussian third derivative filters, G_3, were used to realize the spatiotemporal oriented energy representation; alternative oriented filters are also applicable, e.g., Gabor [12]. Ten spacetime orientations were selected as they correspond to the minimal spanning set for G_3 [12]. To uniformly sample 3D, the particular orientations were taken as the corners of a dodecahedron with antipodal directions identified. Each filter was computed over three inner scales, σ. Outer scale was realized by aggregation over regions, Ω, that resulted from di-

viding the image sequences into $4 \times 4 \times 1$ grids for capturing spatial layout, unless otherwise noted. The grid choice was made to allow for direct comparison to previously reported results [30]. For the appearance marginalized energy measures, (3), 27 spacetime orientations were used (realized through linear combinations of the G_3 basis set). The orientations selected correspond 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), as they were found useful for dynamic texture analysis [8].

3. Empirical evaluation

In addition to the purely spatial (i.e., GIST [26]) and spatiotemporal orientation representations described in Sec. 2.1, several alternative approaches are compared. First, to capture spatial appearance, a simple three bin color histogram model consisting of averaged RGB values [13] is compared; for an extensive evaluation of spatial representations, see [38]. For the purpose of determining the relative merits of color as a cue, GIST was computed on the intensity image alone. Second, given the intense research activity in capturing temporal information via optical flow, comparison is made to the histogram of optical flow (HOF) [25] recovered using a recent global flow implementation [23]; here a normalized 25 bin histogram consisting of eight quantized flow directions vs. three magnitudes and an additional bin capturing approximately zero velocity is computed at three inner scale levels and fused together to yield the final descriptor. Each of the representations described so far are computed over a $4 \times 4 \times 1$ outer scale parceling of the video. Third, a feature representation adapted from the chaotic dynamic systems literature is compared [30]. This 9600-dimensional representation was recently shown empirically to outperform many existing substrates in the literature (e.g., LDS) in application to dynamic scene recognition; results are based on the same code and parameters as [30]. Comparative results for additional dynamic representations (e.g., LDS) are available elsewhere [30]. All approaches were evaluated on a recently introduced data set containing "in-the-wild" type scene footage and the new data set introduced in this paper containing scenes captured from stationary cameras.

Maryland "in-the-wild" scenes data set: This data set contains thirteen dynamic scene classes with ten color videos per class; see Fig. 3 for representative imagery. The average dimensions of the videos are 308×417 (pixels) $\times$ 617 (frames). The videos were collected from Internet-based video hosting sites, e.g., YouTube (`www.youtube.com`). The set captures large variations in illumination, frame rate, viewpoint, image scale and various degrees of camera-induced motion (e.g., panning and jitter) and scene cuts.

Figure 3. Maryland "in-the-wild" scenes data set [30]. (left-to-right, top-to-bottom) avalanche, boiling water, chaotic traffic, forest fire, fountain, iceberg collapse, landslide, smooth traffic, tornado, volcanic eruption, waterfall, waves and whirlpool.

"Stabilized" dynamic scenes data set: This new data set is introduced to emphasize scene specific temporal information over short time durations due to objects and surfaces rather than camera-induced ones, as predominant in the Maryland data set. This improves the understanding of the task of concern. The data set is comprised of fourteen dynamic scene categories each containing 30 color videos; see Fig. 1 for representative imagery. The average dimensions of the videos are 250×370 (pixels) $\times$ 145 (frames). The videos were obtained from various sources, including footage captured by the authors using a Canon HFS20 camcorder and online video repositories, such as YouTube, BBC Motion Gallery (`www.bbcmotiongallery.com`) and Getty Images (`www.gettyimages.com`). Owing to the diversity within and across the video sources, the videos contain significant differences in image resolution, frame rate, scene appearance, scale, illumination conditions (e.g., diurnal) and camera viewpoint. Importantly, video samples were restricted to those from a stationary camera and without scene cuts. In practice, small degrees of camera motion can be handled via image stabilization prior to feature extraction. The "Stabilized" dynamic scenes data set is available at: `www.cse.yorku.ca/vision/research/dynamic-scenes.shtml`.

"In-the-wild" scene recognition: The first experiment followed the same leave-one-video-out protocol set forth with the original investigation of the "in-the-wild" scenes data set [30]. Results are summarized in Table 1 (a); those based on feature combinations (e.g., HOF+GIST, etc.) were realized as a weighted sum of the similarities between the individual features listed in the table. In all combination cases the weight given

(a)	Spatial		Temporal			Spatiotemporal		
Scene classes	Color [13]	GIST [26]	HOF [25]	Chaos [30]	MSOE	Chaos+ GIST+ Color	HOF+ GIST	SOE
avalanche	50	10 (50)	0	30	10	40	30 (20)	10 (10)
b. water	30	60 (60)	40	30	50	40	50 (50)	60 (50)
c. traffic	20	70 (40)	20	50	90	70	40 (30)	80 (80)
f. fire	70	10 (60)	0	30	10	40	30 (50)	40 (40)
fountain	50	30 (20)	10	20	10	70	20 (20)	10 (10)
i. collapse	0	10 (20)	10	10	10	50	10 (20)	20 (10)
landslide	10	20 (30)	20	10	30	50	20 (20)	50 (50)
s. traffic	50	40 (30)	30	20	70	50	30 (30)	60 (70)
tornado	60	40 (60)	0	60	80	90	40 (40)	60 (60)
v. eruption	30	30 (40)	0	70	10	50	30 (20)	10 (30)
waterfall	20	50 (30)	20	30	30	10	20 (20)	10 (20)
waves	40	80 (80)	40	80	80	90	80 (80)	80 (80)
whirlpool	10	40 (30)	30	30	30	40	20 (30)	40 (40)
Avg. (%)	34	38 (43)	17	36	39	52	32 (33)	41 (42)

(b)	Spatial		Temporal			Spatiotemporal		
Scene classes	Color [13]	GIST [26]	HOF [25]	Chaos [30]	MSOE	Chaos+ GIST	HOF+ GIST	SOE
beach	50	90 (90)	37	27	83	30 (30)	76 (87)	87 (90)
c. street	47	50 (63)	83	17	63	17 (17)	80 (77)	83 (87)
elevator	83	53 (80)	93	40	60	40 (47)	90 (87)	67 (90)
f. fire	47	50 (57)	67	50	60	17 (17)	63 (63)	83 (87)
fountain	13	40 (50)	30	7	40	3 (3)	37 (43)	47 (50)
highway	30	47 (53)	33	17	60	23 (23)	53 (47)	77 (73)
l. storm	83	57 (70)	47	37	87	40 (37)	70 (63)	90 (90)
ocean	73	93 (97)	60	43	97	43 (43)	93 (97)	100 (97)
railway	43	50 (53)	83	3	60	7 (7)	87 (83)	87 (90)
r. river	57	63 (80)	37	3	90	10 (10)	73 (77)	93 (90)
sky	30	90 (93)	83	33	80	43 (47)	87 (87)	90 (93)
snowing	53	20 (20)	57	10	17	10 (10)	40 (47)	33 (50)
waterfall	30	33 (40)	60	10	37	10 (10)	50 (47)	43 (47)
w. farm	57	47 (60)	53	17	47	17 (17)	60 (53)	57 (73)
Avg. (%)	50	56 (65)	59	20	63	22 (23)	69 (68)	74 (79)

Table 1. Comparison of classification rates among the various spatial, temporal and spatiotemporal image representations on the Maryland "in-the-wild" and "stabilized" data sets in (a) and (b), resp. The results for Chaos-related substrates in (a) are reproduced from [30]. Parentheses denote classification rates where Color is additionally considered.

(a)	*Classified*												
Actual	avalanche	b. water	c. traffic	f. fire	fountain	i. collapse	landslide	s. traffic	tornado	v. eruption	waterfall	waves	whirlpool
avalanche	1			1		2	1			1	**3**		1
b. water		**6**	1		1		2						
c. traffic			**8**					2					
f. fire	1			**4**			1		2			1	1
fountain		2	1		1		**3**	1			2		
i. collapse			**4**			2	1		1	1			1
landslide				1	1	2	**5**		1				
s. traffic			1				1	**6**			1	1	
tornado							1	1	**6**	2			
v. eruption	1		**3**	1	2					1		1	1
waterfall	**2**			**2**	1	1	**2**		1		1		
waves				1			1					**8**	
whirlpool			2	1		1	1			1			**4**

(b)	*Classified*												
Actual	Sky	Beach	Ocean	Street	Railway	R. River	Highway	Snowing	Waterfall	Fountain	L. Storm	F. Fire	W. Farm	Elevator
Sky	**27**		1		1	1								
Beach		**26**	3			1								
Ocean			**30**											
Street				**25**		3			1			1		
Railway		1			**26**	2						1		
R. River		1				**28**	1							
Highway		1		1		3	**23**		2					
Snowing				1		6		**10**	4	2		3	1	3
Waterfall						2		2	**13**	7		3	2	1
Fountain								3	9	**14**		4		
L. Storm											**27**	3		
F. Fire								2	1		2	**25**		
W. Farm								3	3	1	1	5	**17**	
Elevator								2	3	4		1		**20**

Table 2. Confusion matrix for SOE ($4 \times 4 \times 1$) on the "in-the-wild" and "stabilized" data sets in (a) and (b), resp. Bold shows top classification for each actual set.

to a feature was set to its average classification accuracy (cf., [38]) and the Bhattacharyya coefficient was used as the similarity measure. The exception is Chaos, where the result is reproduced from the original investigation, which did not publish the weighting factors nor the distance measure used. (Table 1 (a) does not show results for Chaos+GIST without Color, as the original authors do not provide such.)

Best overall results for spatial only information were achieved by GIST (38%), which was further improved when combined with Color (43%). The highest recognition rate among the approaches considering temporal information alone was 39%, achieved by Marginalized Spatiotemporal Oriented Energy (MSOE), (2) combined with (3). Considering the closest three matches, average classification across the entire data set improved to 54% for MSOE.

Under the Spatiotemporal heading in Table 1 (a) several cue combinations are considered. Notice that the Spatiotemporal Oriented Energy (SOE), (2) without (3), is the natural extension of MSOE to include both temporal and spatial information; therefore, no explicit combination of MSOE with a spatial only representation (e.g., GIST) is considered. Here, the fused Chaos feature with GIST and Color provided the best result. The recognition rate of SOE was 41% and 42%, considered alone and when fused with Color, resp., with classification rising to 57% when considering the three closest matches. Interestingly, SOE improves on purely spatial or temporal information taken alone, but not to the same degree as Chaos+GIST. It also is of interest to note that Histogram of Flow (HOF) performs relatively poorly as both purely temporal and spatiotemporal features. Finally, various alternative samplings of both SOE (e.g., oversampling the space of orientations beyond the reported basis set) and HOF (finer and coarser binning of initial flow estimates) did not yield appreciably different results.

Since the "in-the-wild" data set contains large camera motions and scene cuts, it is difficult to understand whether the performance of approaches depends on their success in capturing underlying scene structure vs. characteristics induced by the camera. This situation is shown in the confusion matrix in Table 2 (a) for SOE, where there is no apparent trend in the failures. To shed light on this question, the next set of experiments tests the same set of approaches on dynamic scenes captured from stationary cameras.

"Stabilized" dynamic scene recognition: The second experiment follows the same procedure as the first. Results are summarized in Table 1 (b). Similar to the first experiment, the best overall results for spatial only information were achieved by GIST (56%). Fur-

			Inner scale			
			0	1	2	all
Outer scale	$1 \times 1 \times 1$	MSOE [8]	52	53	51	55
		SOE	56	54	56	63
	$2 \times 2 \times 1$	MSOE	55	58	58	61
		SOE	66	67	66	69
	$4 \times 4 \times 1$	MSOE	52	57	60	63
		SOE	64	69	69	74
	all	MSOE	53	60	62	63
		SOE	65	70	70	75

Table 3. Impact of inner and outer scales on overall classification on the "stabilized" scenes data set. The "all" row for outer scale is constructed from a weighted sum of the similarities of the individual outer scale levels with the individual weights proportional to the corresponding level's average accuracy. The "all" column for inner scale is constructed as a natural consequence of combining the individual inner scales via the normalization process, (2).

thermore, among the approaches that consider temporal only information, MSOE again attains the highest recognition rate (63%) with classification improving to 81% when considering the closest three matches. Interestingly, HOF performs relatively well in this evaluation, while Chaos is comparatively poor. The highest recognition rate among the approaches considering spatiotemporal information were achieved by SOE (74% and 79% with Color), with classification rising to 90% when considering the three closest matches.

As shown in the confusion matrix for SOE in Table 2 (b), many of the confusions in SOE now have intuitive appeal. For instance, scenes that predominately contain flowing patterns (e.g., "street" vs. "railway" vs. "rushing river") are confused. Furthermore, a cluster of confusions arise among dynamic patterns that contain a significant flicker component (e.g., "snowing" vs. "waterfall" vs. "forest fire" vs. "lighting storm"). This is reasonable because the SOE representation explicitly captures flicker as one of its components. Finally, Table 3 shows the utility of considering multiple inner and outer scales for MSOE and SOE.

Discussion: Overall, different results are observed for the various dynamic scene representations when evaluated on the two data sets. Of the purely dynamic representations, MSOE performs best across both data sets; however, the relative performance of Chaos and HOF switch. The poor performance of HOF on the "in-the-wild" data can be explained by the erratic camera motions and scene cuts that are difficult to capture, even with a state-of-the-art flow estimator. The results for Chaos are more difficult to explain; however, it is interesting that it seems relatively insensitive to the more purely scene dynamics that are present in the stabilized experiment, as it fails to make the necessary inter-class distinctions. MSOE is able to perform well on both data sets, as the structure of the imagery projects in a discriminatory fashion onto its energy sampling.

For the spatiotemporal approaches, it is found that SOE is the best performer on the stabilized data and the second best on in-the-wild data. Also notable is that while SOE has average classification 10% below Chaos+GIST for the in-the-wild case, Chaos+GIST is more than 50% below SOE for the stabilized case. While Chaos is best on in-the-wild data it is worst on the stabilized data; indeed, it is a factor of three below the second best HOF+GIST in terms of average percent correct. This pattern of results suggest that SOE is consistently able to characterize dynamic scenes whether operating in the presence of strictly scene dynamics (stabilized case) or when confronted with overlaid, non-trivial camera motions (in-the-wild case). The alternative approaches considered are less capable of such wide ranging performance.

The results also indicate the inadequacy of conceptualizing dynamic scene recognition simply as dynamic texture recognition, even when common oriented energy features underlie the approaches, e.g., [8]: Table 3 shows that combined spatial and dynamic information (SOE) bests dynamic information alone (MSOE) at all scales. Moreover, maintenance of spatial layout of image subregions (finer outer scale subdivisions) also improves results over texture analysis-based aggregation across the entire image (i.e., $1 \times 1 \times 1$ outer scale). Note, in particular, that application of the most closely related approach from the dynamic texture recognition literature corresponds to the upper-left most cell in Table 3, i.e., with 52% accuracy, compared to the best scale parameters for the proposed approach, which achieve 75%.

Finally, a more general observation is that for both data sets and all approaches, notable performance increase is had by spatiotemporal approaches in comparison to purely spatial or purely temporal methods.

4. Conclusions and summary

The main contribution of the presented paper is a systematic investigation of the impact of early multiscale orientation measurements on scene classification, as related to: (i) spatial appearance, (ii) temporal/dynamics and (iii) joint spatial appearance and dynamics. Even given the relative simplicity of the spatiotemporal oriented feature set, it is able to achieve consistent, relatively high performance, as compared to other representations considered in this paper. Points of distinction with previous work include: (i) joint consideration of a wider range of patterns, typical of dynamic scenes (e.g., motion, flicker and dynamic texture) and (ii) two multiscale extensions of the basic orientation energy filtering architecture. In addition, a new data set was introduced that highlights various important non-motion-related structures that are commonly encountered in the world. Interesting future ex-

tensions include the investigation of intermediate representations for scene classification, akin to those proposed in the spatial domain literature; for instance, semantic classes or attributes (e.g., [18]) that capture both appearance and dynamics, e.g., describing a region as "fluid-like". It also is of interest to perform a comparative evaluation of dynamic scene recognition algorithms on a data set that captures the same set of scenes with and without camera motion to tease apart further the relative performance of various approaches.

In summary, this paper has presented a systematic analysis of the impact of low-level representations to dynamic scene classification. Key to the investigation are multiscale oriented energy measurements that capture the underlying pattern's spatial appearance, temporal/dynamics and joint spatial appearance and dynamics, and their coarse layout across spacetime. Empirical evaluation on a challenging public dataset and an additional evaluation with control to remove effects of camera motion shows the usefulness of capturing joint spatial appearance and dynamics through oriented energy filtering in comparison to alternative state-of-the-art techniques, especially when scene dynamics are emphasized relative to camera motions.

References

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

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

[3] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *PAMI*, 30:712–727, 2008.

[4] M. Boutell, A. Choudhury, J. Luo, and C. Brown. Using semantic features for scene classification: How good do they need to be? In *ICME*, pages 785–788, 2006.

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

[6] T. Darrell and E. Simoncelli. Separation of transparent motion into layers using velocity-tuned mechanisms. In *MIT TR-244*, 1993.

[7] K. Derpanis, M. Sizintsev, K. Cannons, and R. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *CVPR*, pages 1990–1997, 2010.

[8] K. Derpanis and R. Wildes. Dynamic texture recognition based on distributions of spacetime oriented structure. In *CVPR*, 2010.

[9] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto. Dynamic textures. *IJCV*, 51:91–109, 2003.

[10] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, pages II: 524–531, 2005.

[11] M. Gorkani and R. Picard. Texture orientation for sorting photos at a glance. In *ICPR*, pages I:459–464, 1994.

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

[13] S. Grossberg and T. Huang. ARTSCENE: A neural system for natural scene classification. *Journal of Vision*, 9:1–19, 2009.

[14] A. Guerin-Dugue and A. Oliva. Classification of scene photographs from local orientations features. *PRL*, 21:1135–1140, 2000.

[15] J. Hays and A. Efros. Scene completion using millions of photographs. In *SIGGRAPH*, 2007.

[16] A. Klaeser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, 2008.

[17] J. Koenderink. The structure of images. *B. Cyber.*, 50:363–370, 1984.

[18] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[19] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[20] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[21] L. Li, H. Su, Y. Lim, and L. Fei-Fei. Objects as attributes for scene classification. In *WPA*, 2010.

[22] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer, 1993.

[23] C. Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, MIT, Dept. E.E. and C.S., 2009.

[24] J. Liu and M. Shah. Scene modeling using co-clustering. In *ICCV*, 2007.

[25] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

[26] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42:145–175, 2001.

[27] N. Pinto, D. Cox, and J. DiCarlo. Why is real-world visual object recognition hard? *PLoS Comp. Bio.*, 4:151–156, 2008.

[28] N. Rasiwasia and N. Vasconcelos. Scene classification with low-dimensional semantic spaces and weak supervision. In *CVPR*, 2008.

[29] L. Renninger and J. Malik. When is scene identification just texture recognition? *Vis. Research*, 44:2301–2311, 2004.

[30] N. Shroff, P. Turaga, and R. Chellappa. Moving vistas: Exploiting motion for describing scenes. In *CVPR*, 2010.

[31] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *PAMI*, 29:300–312, 2007.

[32] E. Simoncelli and D. Heeger. A model of neuronal responses in visual area MT. *Vision Research*, 38:743–761, 1996.

[33] M. Szummer and R. Picard. Indoor-outdoor image classification. In *CAIVD*, pages 42–51, 1998.

[34] A. Torralba, K. Murphy, and W. Freeman. Using the forest to see the trees: Object recognition in context. *C. ACM*, 53:107–114, 2010.

[35] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image classification for content-based indexing. *T-IP*, 10:117–130, 2001.

[36] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 32:1582–1596, 2010.

[37] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *IJCV*, 72:133–157, 2007.

[38] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.