Spacetime Forests with Complementary Features for Dynamic Scene Recognition

Christoph Feichtenhofer[1]
cfeichtenhofer@gmail.com
Axel Pinz[1]
axel.pinz@tugraz.at
Richard P. Wildes[2]
wildes@cse.yorku.ca

[1] Institute of Electrical Measurement and Measurement Signal Processing
Graz University of Technology, Austria

[2] Department of Electrical Engineering and Computer Science
York University
Toronto, Ontario, Canada

Abstract

This paper presents spacetime forests defined over complementary spatial and temporal features for recognition of naturally occurring dynamic scenes. The approach improves on the previous state-of-the-art in both classification and execution rates. A particular improvement is with increased robustness to camera motion, where previous approaches have experienced difficulty. There are three key novelties in the approach. First, a novel spacetime descriptor is employed that exploits the complementary nature of spatial and temporal information, as inspired by previous research on the role of orientation features in scene classification. Second, a forest-based classifier is used to learn a multi-class representation of the feature distributions. Third, the video is processed in temporal slices with scale matched preferentially to scene dynamics over camera motion. Slicing allows for temporal alignment to be handled as latent information in the classifier and for efficient, incremental processing. The integrated approach is evaluated empirically on two publically available datasets to document its outstanding performance.

1 Introduction

Recognizing scene categories in unconstrained video is important for many practical vision applications such as surveillance and safety systems, *e.g.* cameras monitoring spacetime events such as forest fires or avalanches. Although this importance has triggered recent research activity, state-of-the-art classification frameworks are still far from human recognition performance. The amount of information and variability present in images of diverse natural scenes calls for an approach that is able to handle multiple classes, scales and temporal variations, yet still be efficient in training and recognition.

While static scene recognition from single images has been researched extensively (*e.g.* [11, 18, 20, 21, 24, 27, 30, 31, 33]), relatively little research has considered video-based dynamic scene recognition [8, 22, 28], even though the availability of temporal information should provide an additional means for classifing scenes visually. A useful feature type for both static [18, 24, 30] and dynamic [8] scene classification is based on local measurements of orientation. In using local orientation measurements that have been aggregated into texture

patches, these approaches build on research in both static [3] and dynamic [7] texture analysis that use similar primitives. Application of such measurements to dynamic image analysis additionally has been shown useful in a variety of areas, perhaps most related to current concerns are image motion estimation [1, 29] and human action characterization [9]. While their application to dynamic scene recognition has led to state-of-the-art performance, it also has shown notable limitations when confronted with significant camera motion [8].

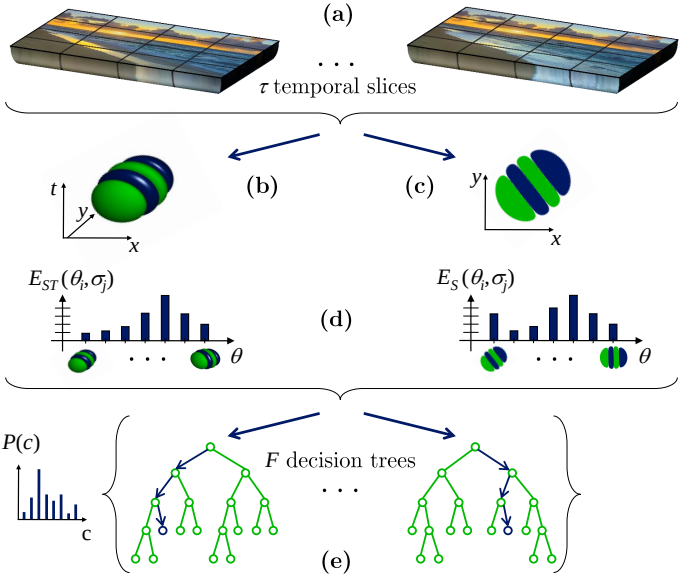


Figure 1: Overview of the proposed dynamic scene classification framework. (a) The input sequence is divided into cuboids using a spatiotemporal pyramid representation. τ temporal slices are created to process the frames in a sliding window approach. (b,c) The cuboids are filtered by banks of multscale, σ, oriented filters along equally separated directions, θ, in image spacetime (b) and space (c) to capture both dynamic and static appearance information. (d) Filter responses cast weighted votes in spacetime orientation cells. (e) The class of each temporal slice is determined via likelihood voting, using a multi-class random forest classifier. Subsequently, all slice-based classifications are combined across the entire input.

In the light of previous research, the present work makes three main contributions. (i) A novel descriptor is presented that integrates complementary information from separate spatial and temporal orientation measurements for aggregation in spacetime pyramids. Distinct from previous application of spatiotemporal orientation to dynamic scenes [8], separation of spatial and temporal orientation allows those components to be differently weighted in classification. (ii) A random forest classifier is applied for the first time to dynamic scenes. This spacetime forest allows for automatic determination of the most discriminative features to separate the classes based on appearance and dynamics with computational efficiency. The approach allows the classifier to learn different weights for different class discriminations; *e.g.* a beach sequence may be better represented by its motion information, while a forest fire sequence might be better distinguished by its spatial appearance. (iii) Video is processed in incremental temporal slices with scale matched preferentially to scene dynamics (in compar-

ison to camera motion). This strategy allows for temporal alignment to be treated as latent in the classifier, efficient processing and robustness to large temporal variation across time (*e.g.* from camera motion), even while capturing intrinsic scene characteristics. Previous dynamic scene research has suffered in the presence of camera motion [8] and has provided little consideration of on-line processing concerns. Further, the approach has been evaluated on two publically available datasets [8, 28]; results show that it achieves a new state-of-the-art in dynamic scene recognition. Figure 1 overviews the framework.

2 Technical approach

2.1 Complementary spacetime orientation descriptor

This section puts forth a novel descriptor for dynamic scene representation that is based on the complementary combination of several different primitive measurements. Spatially oriented measurements are used to capture static image appearance and are combined with spatiotemporally oriented measurements to capture image dynamics. Filtering operates at multiple scales to capture the multiscale characteristics of natural scenes. Furthermore, colour channels are included to capture complementary chromatic information. Interestingly, evidence from biological systems suggests that they exploit similar complementary feature combination in their visual processing [10, 13, 15, 26, 34].

Spatial and temporal information. Spatial appearance and temporal dynamics information are extracted via applications of multiscale filter banks that are further tuned for spatial or spatiotemporal orientation. In the spatial domain, 2D Gaussian third derivative filters, $G^{(3)}_{2D}(\theta_i,\sigma_j)$ with θ_i denoting orientation and σ_j scale, are applied to yield a set of multiscale, multiorientation measurements according to

$$E_S(\mathbf{x};\theta_i,\sigma_j) = \sum_{\Omega} |G^{(3)}_{2D}(\theta_i,\sigma_j) * \mathcal{I}(\mathbf{x})|^2, \tag{1}$$

where $\mathcal{I}$ is an image, $\mathbf{x} = (x,y)^\top$ spatial coordinates, $*$ convolution, Ω a local aggregation region and subscript S appears on E_S to denote *spatial* orientation.

Similarly, dynamic information is extracted via application of 3D Gaussian third derivative filters, $G^{(3)}_{3D}(\theta_i,\sigma_j)$ with θ_i and σ_j now denoting the 3D filter orientations and scales (resp.), applied to the spacetime volume, $\mathcal{V}$, indexed by $\mathbf{x} = (x,y,t)^\top$, as generated by stacking all video frames of a sequence along the temporal axis, t, to yield

$$E_{ST}(\mathbf{x};\theta_i,\sigma_j) = \sum_{\Omega} |G^{(3)}_{3D}(\theta_i,\sigma_j) * \mathcal{V}(\mathbf{x})|^2, \tag{2}$$

with subscript ST on E_{ST} to denote *spatiotemporal* orientation. Following previous work in spacetime texture analysis [7], the spatiotemporal responses, (2), are further combined to yield measures of dynamic information independent of spatial appearance, as follows. In the frequency domain, motion occurs as a plane through the origin [35]. Therefore, to remove spatial information from the initial spatiotemporal orientation measurements, (2), they can be summed across all orientations consistent with a single frequency domain plane. Let the plane be defined by its normal, $\hat{\mathbf{n}}$, then measurements of orientation consistent with this plane

are given as

$$E_{MST}(\mathbf{x};\hat{\mathbf{n}},\sigma_j) = \sum_{i=0}^{N} E_{ST}(\mathbf{x},\theta_i,\sigma_j), \tag{3}$$

with θ_i one of $N+1$ equally spaced orientations consistent with the frequency domain plane and $N=3$ is the order of the employed Gaussian derivative filters; for details see [7]. Here, the subscript *MST* on E_{MST} serves to denote that the spatiotemporal measurements have been "*marginalized*" with respect to purely spatial orientation.

As noted in Sec. 1, previous spacetime filtering approaches to dynamic scene recognition tend to exhibit decreased performance when dealing with scenes captured with camera motion, in comparison to scenes captured with stationary cameras. A likely explanation for this result is that the approaches have difficulty in disentangling image dynamics that are due to camera motion vs. those that are intrinsic to the scenes. Here, it is interesting to note that camera motion often unfolds at coarser time scales (*e.g.*, extended pans and zooms) in comparison to intrinsic scene dynamics (*e.g.*, spacetime textures of water, vegetation, *etc.*); however, previous approaches have made their measurements using relatively coarse temporal scales and thereby failed to exploit this difference. In the present approach this difference in time scale is captured by making use of only fine scales, σ, during spatiotemporal filtering, (2), so that they are preferentially matched to scene, as opposed to camera, dynamics.

The orientation measurements, (1) and (3), can be taken as providing measures of the signal energy along the specified directions, θ_i. This interpretation is justified by Parseval's theorem [25], which states that the sum of the squared values over the spacetime domain is proportional to the sum of the squared magnitude of the Fourier components over the frequency domain; in the present case, the squared values of the orientation selective filtering operations are aggregated over the support regions, Ω.

Owing to the bandpass nature of the Gaussian derivative filters, the oriented energy features are invariant to additive photometric variations (*e.g.*, as might arise from overall image brightness change in imaged scenes). To further provide for invariance to multiplicative photometric variations, each orientation selective measurement in (1) and (3) is normalized with respect to the sum of all filter responses at that point according to

$$\hat{E}_S(\mathbf{x},\theta_i,\sigma_j) = \frac{E_S(\mathbf{x},\theta_i,\sigma_j)}{\sum_i^N E_S(\mathbf{x},\theta_i,\sigma_j)+\varepsilon} \tag{4}$$

for the purely spatially oriented measurements, (1), and similarly for the dynamic measurements, (3), to yield a correspondingly normalized set of measurements, $\hat{E}_{MST}$. Note that ε is a small constant added to the sum of the energies over all orientations. This bias operates as a noise floor and avoids numerical instabilities at low overall energies. Further, normalized εs are added to the set of filtering results, calculated as $\hat{\varepsilon}_S = \varepsilon / \left(\sum_i^N E_S(\mathbf{x},\theta_i,\sigma_j)+\varepsilon\right)$, to capture lack of spatial orientation structure in a region and an analogously defined $\hat{\varepsilon}_{MST}$ to capture lack of spatiotemporal structure. For example, notice that for regions that are devoid of oriented structure, the sum in the numerator will be dominated by ε so that the ratio will tend to 1 and thereby be indicative of lack of (orientation) structure.

Chromatic information. Chromatic information can greatly influence (static) object and scene recognition [32] and also has proven useful in previous work on dynamic scene recognition [8, 28]. Correspondingly, chromatic information is incorporated in the present dynamic scene descriptor by adding three more measurements at each point in the image sequence taken as CIE-LUV colour space observations [32].

Temporal slice-based aggregation. The complementary spacetime orientation measurements presented so far are defined pointwise across a video sequence. For the purpose of classification of the entire video into a scene class, the local descriptors are summed across time, t, within τ discrete units of equal duration to yield a set of temporally aggregated images, which are referred to as temporal slices. Temporal slicing is motivated by the desire for incremental processing that can allow for efficient, on-line operation. Use of short-term parceling of the measurements also is well matched with the restriction to use of fine temporal scales during spatiotemporal filtering to favour scene over camera dynamics. During classification (Sec. 2.2) each temporal slice initially is classified individually, with the individual classifications subsequently combined to yield overall classification for the video.

Having established temporal slices for an input video, the complementary measurements are processed in successive temporal slices across the video. Each slice is hierarchically aggregated into histograms to form a spatiotemporal pyramid, analogous to that used previously for static [18] and dynamic [8] scene analysis. At each level of the pyramid, each temporal slice is broken into $X \times Y \times T$ 3D cuboids (see Fig. 1(a)), with filter measurements collapsed into histograms within each cuboid, as illustrated in Fig. 1(d). The support of the cuboid at any given level of the pyramid corresponds to its outer scale [17]; indeed, it corresponds to the aggregation region Ω in the filtering equations, (1) and (2). Moreover, the adjacency structure of the cuboids capture the overall scene layout. For each cuboid, the histograms are $L1$-normalized, to represent a distribution of chromatic, multiscale oriented spacetime energy and lack of oriented structure (via $\hat{\varepsilon}$). Let $M_\theta, M_{\hat{\mathbf{n}}}$, M_{σ_θ} and $M_{\sigma_{\hat{\mathbf{n}}}}$ be the number of spatial orientations, spatiotemporal orientations and their (inner) scales considered in the multiscale oriented filtering operations, resp. Then, the dimension of each histogram is the quantity $(M_\theta + 1) \times M_{\sigma_\theta} + (M_{\hat{\mathbf{n}}} + 1) \times M_{\sigma_{\hat{\mathbf{n}}}} + 3$, with 1 added to the number of orientations due to $\hat{\varepsilon}$, and 3 the number of colour channels. The histograms for all cuboids are concatenated into a final feature vector, $\mathbf{v}$, that comprises the Complementary Spacetime Orientation descriptor (CSO) to characterize a temporal slice of the video.

2.2 Spacetime Forests

Random Forests (RFs) are an ensemble of F decision trees $\{\mathcal{T}_k\}_{k=1}^{F}$ learned with random features. Each decision tree is used independently to classify the input feature vector, $\mathbf{v}$, based on the leaf node at which the corresponding feature vector arrives. Hence, the leaf nodes of all trees hold the posterior distribution $P(c|\mathbf{v})$ over the classes $c \in \{1, \cdots, C\}$. Previously, RF classifiers have been applied to a variety of vision tasks, *e.g.*, [4, 19, 23, 36]. Detailed descriptions of RFs are available elsewhere, *e.g.*, [2, 5, 6].

In the present work, RFs are employed for their ability to combine several cues for multiclass prediction, as well as their increased speed in the training and testing processes over traditional classifiers. Here, the classes correspond to different dynamic scenes (*e.g.*, beach vs. city, *etc.*) and the feature vectors correspond to the CSO descriptors defined in the previous subsection. In this subsection, a particular instantiation of RFs, termed Space-Time Random Forests (STRF) are defined.

Randomized Training. During training, the temporal alignment of the video slices is treated as latent; correspondingly, each temporal slice of each training video generates its own feature vector according to procedures of Sec. 2.1. This approach allows leveraging of the high temporal diversity in the spatiotemporal patterns.

Each tree is constructed by drawing a random bootstrap sample from the training set. Bootstrapping in the training stage allows maximum training efficiency [4] and avoids overfitting [2] to generalize well in practice. Further randomness is introduced in the node optimization step by selecting a random subset m of the feature vector's dimension to be considered in a random threshold test for determination of the best split for each node. Here, the split is selected as that which maximizes the information I in the final class distribution, after splitting into a left (L) and right (R) node:

$$I = H(Q) - \sum_{i \in \{L,R\}} \frac{|Q^i|}{|Q|} H(Q^i), \tag{5}$$

where $H(Q) = -\sum_{c \in C} p(c) log(p(c))$ is the Shannon entropy, $p(c)$ the proportion of classes in Q belonging to class c and $|\cdot|$ denotes the size of the set at a given node. Tests are selected for all nodes recursively until the growing process is stopped when no further information gain is achieved. The final class distributions at the leaves are calculated as the ratio between the number of feature vectors of each class and the total number of features which reach the leaf after training.

As some classes may be better represented by specific feature types, the node optimization process in each tree is restricted to a single feature type. To best separate the classes with the CSO descriptor, the input for the RF is first structured into the three complementary feature channels: spatial orientation, (marginalized) spatiotemporal orientation and colour. Then, the channels are used to train $\frac{F}{3}$ trees each, to best distinguish the classes, based on a particular channel only. Lastly, these complementary trees are merged, to obtain the spacetime forest $\{\mathcal{T}_k\}_{k=1}^{F}$.

Recognizing Dynamic Scenes. For classification, the feature vectors, $\mathbf{v}^\tau$, of scenes to be recognized are again decomposed into the three distinct channels and sent simultaneously through the respective complementary trees until the corresponding leaves are reached. Here, τ is the temporal slice of the input volume where the feature is extracted. Each tree gives a classification by voting for the class labels according to the class distribution $p_k(c|\mathbf{v}^\tau)$ of the leaf which is reached by $\mathbf{v}^\tau$ in tree k.

Given the resulting set of tree predictions, they are combined in two stages to yield a classification at each temporal instance. First, the prediction results of the forest for the current temporal slice, τ, are calculated as a simple averaging of the leaf distributions p_k in the F trees in the forest

$$P^\tau(c|\mathbf{v}^\tau) = \frac{1}{F} \sum_{k=1}^{F} p_k(c|\mathbf{v}^\tau). \tag{6}$$

Second, to yield a final classification across all temporal slices available up to a given time, the class likelihoods for each slice are treated as temporal predictions and once again combined via averaging

$$P(c|\mathbf{v}) = \frac{1}{\tau} \sum_{l=1}^{\tau} P^l(c|\mathbf{v}^l). \tag{7}$$

The current classification of the video is then given as

$$c = \arg\max_c P(c|\mathbf{v}). \tag{8}$$

2.3 Implementation Details

Video Descriptor. In the current implementation, $M_\theta = 4$ for spatial (1) and 10 for spatiotemporal (2) oriented filtering operations, as those numbers span orientation space for the order and dimensionality of filters employed [12]. Here, it is of note that orientation selective filters other than Gaussian derivatives might have been used (*e.g.*, oriented Gabor filters[16]); however, the chosen filters enjoy the advantage of particularly efficient implementation owing to separability and steerability [12]. In any case, the results of the spatiotemporal filtering, (2), are further combined to capture frequency domain planes, (3), parameterized by $\hat{\mathbf{n}}$ corresponding to motion along the leftward, rightward, upward and downward directions as well as static (zero velocity) and flicker (infinite velocity); *i.e.* $M_{\hat{\mathbf{n}}} = 6$. For each orientation, spatial filtering is performed at $M_{\sigma_\theta} = 4$ different scales, starting at $\sigma = 2$, varying coarser by octave; spatiotemporal filtering is performed at $M_{\sigma_{\hat{\mathbf{n}}}} = 1$ relatively fine scale ($\sigma = 2$) in preference for capturing short term temporal variations. During normalization, (4), $\varepsilon = 500$, which is quite small relative to image energies encountered in practice. The spacetime pyramid is constructed at 4 levels with number of cuboids $(X \times Y \times T) \in \{(8\times8\times1),(4\times4\times1),(2\times2\times1),(1\times1\times1)\}$. Pyramids are constructed for each temporal slice of an input video, with the length of temporal slices set to 16 frames. To represent the colour distribution in each cuboid, a 3 bin histogram of the CIE-LUV colour channels is employed. Other colour spaces also were considered (RGB, HSV, CIE-Lab [37]); however, LUV led to slightly better results in preliminary investigation.

Classifier. Even if the training data exhibits the same number of videos for each class, they may have different durations. Therefore, the classifier would be severely biased towards the classes containing long videos. To compensate for these differences, priors are used in the training stage subsampling process. These are given by the inverse of the number of temporal slices τ of all videos from a specific class c in the training set. For all experiments a multi-class STRF with 500 trees for each of the three feature channels (*i.e.*, spatial orientation, (marginalized) spatiotemporal orientation and colour) of the video descriptor is trained. At each split node, a random sample of m features is drawn for consideration in splitting [5]. In particular, $m = \lfloor \log_2 D \rfloor$, where D is the feature vector dimensionality. The best split of the random tests, determined by Eq. (5), is used to split the node. For training the node splits, each tree uses a random bootstrap sample consisting of two thirds of the training data. The error rate for observations left out of the training data is monitored as out-of-bag error rate and may be used as an unbiased estimate of the classification performance of each tree trained on a bootstrap sample [5].

3 Empirical evaluation

The proposed approach for dynamic scene recognition has been evaluated on the Maryland "In-The-Wild" [28] and YUPENN Dynamic Scenes [8] datasets, consisting of 13 and 14 classes with 10 and 30 videos per class, respectively. The datasets contain videos showing a wide range of natural dynamic scenes (avalanches, traffic, forest fires, waterfalls *etc.*); see Table 1 where complete listings can be read off the left most columns of the tables. A notable difference between the two datasets is that the Maryland dataset includes camera motion, while the YUPENN dataset does not. To be consistent with previous evaluations ([8, 28]), the same evaluation procedure was employed, *i.e.* a leave-one-video-out recognition experiment.

To evaluate systematically the contributions of (i) the video descriptor (CSO), (ii) classifier (STRF), and (iii) the use of temporal slicing and priors for classification, these components are evaluated separately in the remainder of this section.

For the sake of comparison, several alternative video descriptors and classifiers are considered that have shown strong performance in previous evaluations [8, 28]. Descriptors considered are GIST [24] + HOF [22], GIST + chaotic dynamic features (Chaos) [28] and Spatiotemporal Oriented Energies (SOE) [8]. All of these approaches include colour histograms [14], as this addition generally increases classification performance [8, 28]. Classifiers considered are Nearest-Neighbor (NN), Support Vector Machine (SVM), Random Forest (RF) and the proposed Spacetime Random Forest (STRF). NN and SVM are included, as they have been employed in previous evaluations [8, 28]; RF is a random forest trained uniformly across all components of the training vectors, which is included for the sake of comparison to the proposed STRF, which trains separate trees for the spatial and temporal orientation as well as colour components. The alternative approaches build their feature vectors by collapsing across *all* temporal frames; for the sake of comparison, the proposed approach is shown processing temporal information in several ways: Use a single temporal slice for classification (RF and STRF, $\tau = 1$); average the CSO feature vectors calculated for individual slices across the entire video (RF, $\tau = all$); combine individual slice classifications across the entire video (STRF, $\tau = all$) according to the proposed averaging (7), *i.e.* the complete proposed approach. Results are shown in Table 1.

In comparison to the most closely related descriptor (SOE[1]) running under the same classifier (RF), it is seen that the proposed CSO features improve overall performance in the presence of camera motion (Maryland dataset) from 43% to 57% recognition accuracy using only a single temporal slice ($\tau = 1$), with an additional boost to 59% when feature vectors are combined across all slices ($\tau = all$). In contrast, when camera motion is not present (YUPENN), the performance of the two feature sets under the same classifier is essentially indistinguishable (81% vs. 82%). These results support the ability of the proposed approach to capture intrinsic scene dynamics with robustness to camera motion. Further allowing the classifier to consider the complementary feature components separately (STRF) shows even better performance whether with $\tau = 1$ slice or with combination across $\tau = all$ slices.

More generally, the proposed approach's 68% accuracy on the Maryland dataset improves on the previous best performer in the presence of camera motion (Chaos+GIST under SVM) by 10%. Further, its accuracy of 86% on the YUPENN dataset sets a new state-of-the-art for that case as well (even with SOE given the advantage of RF-based classification, which it did not enjoy in its original application to dynamic scenes [8], but which is included here for fair comparison). Indeed, while all other compared approaches show high variation in performance between the two datasets, the proposed approach provides the best overall performance in both cases. Moreover, best overall performance is attained even when only a single temporal slice of 16 frames is processed (CSO, STRF, $\tau = 1$).

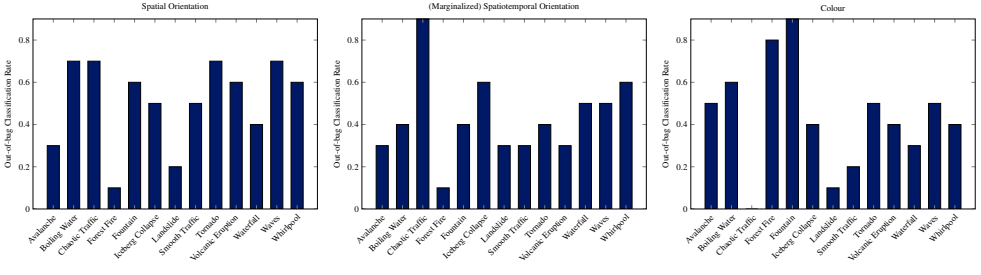
The complementary nature of the CSO descriptor components is illustrated explicitly in Fig. 2(a) and 2(b). The figure shows estimates of cross-validation performance as indicated by the out-of-bag error rate [5] when training the random forest with spatial and temporal orientation as well as colour information separately. It is seen that different classes are better distinguished by different types of information; correspondingly, their combination provided by the proposed approach (Table 1) yields improved classification performance.

[1]Recall that SOE derives from integrated spatiotemporal filtering, (2), without temporal slicing, and without the proposed separation into complementary spatial, (1), and temporal, (3), components.

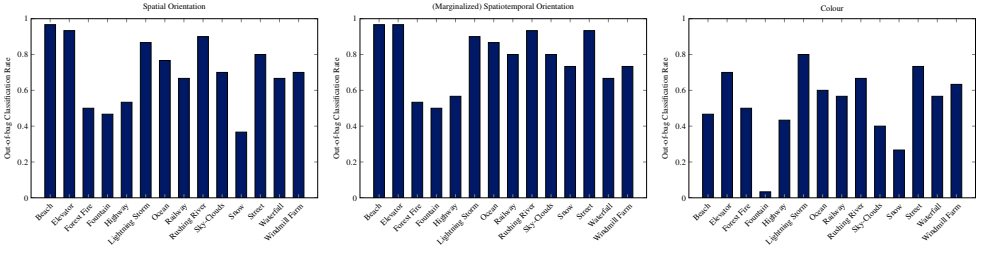
Maryland "In-The-Wild"									
Descriptor	HOF+ GIST	Chaos+ GIST		SOE		CSO (proposed)			
Classifier	NN	NN	SVM	NN	RF	RF	RF	STRF	STRF
Temporal τ	*all*	*all*	*all*	*all*	*all*	1	*all*	1	*all*
Avalanche	0.2	0.4	0.6	0.1	0.4	0.4	0.5	0.6	0.6
Bo. Water	0.5	0.4	0.6	0.5	0.5	0.8	0.8	0.8	0.8
Ch. Traffic	0.3	0.7	0.7	0.8	0.6	0.9	1.0	0.8	0.9
Forest Fire	0.5	0.4	0.6	0.4	0.1	0.3	0.5	0.8	0.8
Fountain	0.2	0.7	0.6	0.1	0.5	0.4	0.5	0.9	0.8
Iceberg Co.	0.2	0.5	0.5	0.1	0.4	0.5	0.4	0.6	0.6
Landslide	0.2	0.5	0.3	0.5	0.2	0.2	0.4	0.2	0.3
Sm. Traffic	0.3	0.5	0.5	0.7	0.3	0.6	0.5	0.6	0.5
Tornado	0.4	0.9	0.8	0.6	0.7	0.8	0.6	0.9	0.8
Volcanic Er.	0.2	0.5	0.7	0.3	0.1	0.5	0.8	0.5	0.7
Waterfall	0.2	0.1	0.4	0.2	0.6	0.5	0.5	0.5	0.5
Waves	0.8	0.9	0.8	0.8	0.5	0.7	0.7	0.6	0.8
Whirlpool	0.3	0.4	0.5	0.4	0.7	0.8	0.5	0.8	0.7
Avg. Perf.	0.33	0.52	0.58	0.42	0.43	0.57	0.59	0.66	0.68

YUPENN Dynamic Scenes data set								
Descriptor	HOF+ GIST	Chaos+ GIST	SOE		CSO (proposed)			
Classifier	NN	NN	NN	RF	RF	RF	STRF	STRF
Temporal τ	*all*	*all*	*all*	*all*	1	*all*	1	*all*
Beach	0.87	0.30	0.90	0.93	0.97	1.00	1.00	1.00
Elevator	0.87	0.47	0.90	1.00	0.97	0.97	0.97	1.00
Forest Fire	0.63	0.17	0.87	0.67	0.80	0.83	0.76	0.83
Fountain	0.43	0.03	0.50	0.43	0.47	0.53	0.40	0.47
Highway	0.47	0.23	0.73	0.70	0.67	0.70	0.67	0.73
Lightning S.	0.63	0.37	0.90	0.77	0.90	0.90	0.93	0.93
Ocean	0.97	0.43	0.97	1.00	0.90	0.90	0.90	0.90
Railway	0.83	0.07	0.90	0.80	0.87	0.90	0.90	0.93
Rushing R.	0.77	0.10	0.90	0.93	0.93	0.93	0.97	0.97
Sky-Clouds	0.87	0.47	0.93	0.83	0.87	0.90	1.00	1.00
Snowing	0.47	0.10	0.50	0.87	0.47	0.43	0.57	0.57
Street	0.77	0.17	0.87	0.90	0.93	0.90	0.97	0.97
Waterfall	0.47	0.10	0.47	0.63	0.67	0.70	0.80	0.76
Windmill F.	0.53	0.17	0.73	0.83	0.93	0.87	0.93	0.93
Avg. Perf.	0.68	0.23	0.79	0.81	0.81	0.82	0.84	0.86

Table 1: Average classification rates for different video descriptor and classifier combinations. The combination of diverse, informative feature channels (CSO) with a suitable classifier (STRF) gives overall best results. See text for details.



(a) Maryland "In-The-Wild"



(b) YUPENN Dynamic Scenes data set

Figure 2: Classification performance measured by the out-of-bag error rate when training the random forest separately with spatial and spatiotemporal orientation as well as colour components of the CSO descriptor. A single temporal slice of each clip is used for training.

Finally, in terms of execution speed, the current unoptimized Matlab implementation of the full proposed approach computes a feature vector for a 16 frame slice in 4 seconds (due to separable and steerable filtering) and takes an additional 5 milliseconds on average to report a class label. This allows state-of-the-art scene classification, being within 2% accuracy of the best performance across both datasets (also attained by the proposed approach, but using a complete set of slices), in nearly real time. Moreover, filtering and random forests are readily parallelizable for GPU implementations.

4 Conclusions

This paper has presented a novel approach to dynamic scene recognition based on three key ideas. First, different scenes are best characterized in terms of different combinations of spa-

tial, temporal and chromatic information. Correspondingly, the CSO descriptor has been introduced that affords complementary consideration of spatial and spatiotemporal orientation as well as colour information. Second, a particular instantiation of random forests, STRF, has been introduced that allows the complementary components of the CSO descriptor to be exploited during classification. Third, temporal slicing with scale matched to intrinsic scene dynamics has been employed. Matching the scale of spatiotemporal filtering to scene dynamics allows for recognition that is robust to camera motion. Slicing allows for efficient, incremental processing of video as well as treatment of temporal alignment as latent during classification. In empirical evaluation relative to a variety of previous algorithms for dynamic scene recognition, the proposed approach has yielded a new state-of-the-art in dynamic scene recognition accuracy both with and without camera motion. Combined with the inherent computational efficiency of the proposed approach, these recognition results suggest relevance to real-world application scenarios, including video indexing and browsing as well as surveillance and monitoring. Integration with such applications would serve as an interesting direction for future research.

References

[1] Edward H. Adelson and James R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2(2):284–299, 1985.

[2] Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9:1545–1588, 1997.

[3] J. R. Bergen. Theories of visual texture perception. *Vision and Visual Dysfunction*, 10B:114–134, 1991.

[4] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *Proc. ICCV*, 2007.

[5] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Found. Trends. Comput. Graph. Vis.*, 7:81–227, 2012.

[7] K. Derpanis and R. Wildes. Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *PAMI*, 34(6):1193–1205, 2012.

[8] K. Derpanis, M. Lecce, K. Daniilidis, and R. Wildes. Dynamic scene understanding: The role of orientation features in space and time in scene classification. In *Proc. CVPR*, 2012.

[9] K. Derpanis, M. Sizintsev, K. Cannons, and R. Wildes. Action spotting and recognition based on a spatiotemporal orientation analysis. *PAMI*, 35(3):527–540, 2012.

[10] S. Engel, X. Zhang, and B. Wandell. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, 388(6637):68–71, 1997.

[11] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, volume 2, pages 524–531 vol. 2, 2005.

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

[13] Andrei Gorea, Thomas V. Papathomas, and Ilona Kovacs. Motion perception with spatiotemporally matched chromatic and achromatic information reveals a "slow" and a "fast" motion system. *Vision Research*, 33(17):2515 – 2534, 1993.

[14] S. Grossberg and T. R. Huang. Artscene: A neural system for natural scene classification. *Journal of Vision*, 9(4):1–19, 2009.

[15] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254 –1259, 1998.

[16] B. Jahne. *Digital Image Processing, Sixth Edition*. Springer, 2005.

[17] J. Koenderink. The structure of images. *Biological Cybernetics*, 50(8):363–370, 1984.

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

[19] V. Lepetit and P. Fua. Keypoint Recognition using Randomized Trees. *PAMI*, 28(9): 1465–1479, 2006.

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

[21] Noha M. Elfiky, Jordi González, and F. Xavier Roca. Compact and adaptive spatial pyramids for scene recognition. *Image and Vision Computing*, 30(8):492–500, 2012.

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

[23] Frank Moosmann, Bill Triggs, and Frederic Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Proc. NIPS*, 2007.

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

[25] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *Discrete-time signal processing (2nd ed.)*. 1999.

[26] Thomas V. Papathomas, Andrei Gorea, and Bela Julesz. Two carriers for motion perception: Color and luminance. *Vision Research*, 1991.

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

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

[29] E. Simoncelli and D. Heeger. A model of neuronal responses in visual area MT. *Vision Research*, 38(8):743–761, 1996.

[30] M. Szummer and R.W. Picard. Indoor-outdoor image classification. In *IEEE International Workshop on Content-Based Access of Image and Video Database*, 1998.

[31] A. Vailaya, M. A T Figueiredo, A.K. Jain, and Hong-Jiang Zhang. Image classification for content-based indexing. *PAMI*, 10(1):117–130, 2001.

[32] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 32(9):1582–1596, 2010.

[33] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *IJCV*, 72(3):133–157, 2007.

[34] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

[35] B. Watson and A. Ahumada. A look at motion in the frequency domain. In *Proc. of the Motion Workshop*, 1983.

[36] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *Proc. CVPR*, 2006.

[37] G. Wyszecki and W. Stiles. *Color Science*. Wiley, 2000.