Dynamic Scene Classification Using Redundant Spatial Scenelets

Liang Du and Haibin Ling, *Member, IEEE*

***Abstract*—Dynamic scene classification started drawing an increasing amount of research efforts recently. While existing arts mainly rely on low level features, little work addresses the need of exploring the rich spatial layout information in dynamic scene. Motivated by the fact that dynamic scenes are characterized by both dynamic and static parts with spatial layout priors, we propose to use redundant spatial grouping of a large number of spatio-temporal patches, named *scenelet*, to represent a dynamic scene. Specifically, each scenelet is associated with a category-dependent scenelet model to encode the likelihood of a specific scene category. All scenelet models for a scene category are jointly learned to encode the spatial interactions and redundancies among them. Subsequently, a dynamic scene sequence is represented as a collection of category likelihoods estimated by these scenelet models. Such presentation effectively encodes the spatial layout prior together with associated semantic information, and can be used for classifying dynamic scenes in combination with a standard learning algorithm such as k-nearest neighbor or linear SVM. The effectiveness of our approach is clearly demonstrated using two dynamic scene benchmarks and a related application for violence video classification. In the nearest neighbor classification framework, for dynamic scene classification, our method outperforms previous state-of-the-arts on both the Marryland “in the wild” dataset and the “stabilized” dynamic scene dataset. For violence video classification on a benchmark dataset, our method achieves a promising classification rate of 87.08%, which significantly improves previous best result of 81.30%.**

***Index Terms*—Dynamic scene, Redundant spatial grouping**

I. Introduction

As a fundamental challenge in automated visual understanding, natural scene understanding provides basis for many high level vision tasks, such as object analysis [6], [36], action recognition [35], activity understanding [5], [28] and robotic control [11]. Recently, a significant amount of efforts have been devoted to dynamic scene classification [8], [14], [35], [44], [49]. Understanding dynamic scenes is very important for many practical vision applications like robot navigation and safety systems, *e.g.*, camera monitoring spatio-temporal events like forest fires or avalanches.

While existing methods typically rely on low level visual features, we are interested in the rich spatial layout information revealed by the middle level features, *i.e.*, spatio-temporal sub-volumes within a video volume. Although higher level recognition has gained increasing popularity for many vision tasks [3], [4], [17], [21], [22], [30], [40], [42], [45], [46], [51], its application to dynamic scene is under-explored.

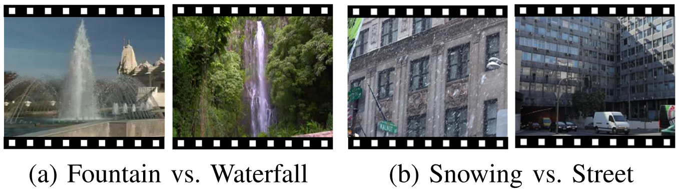


Fig. 1. Features to distinguish dynamic scenes may have different spatial layout and motion properties. (a) “Fountain” and “waterfall” rely on the static background to distinguish them. (b) “Snowing” and “street” differ mainly by the motion patterns.

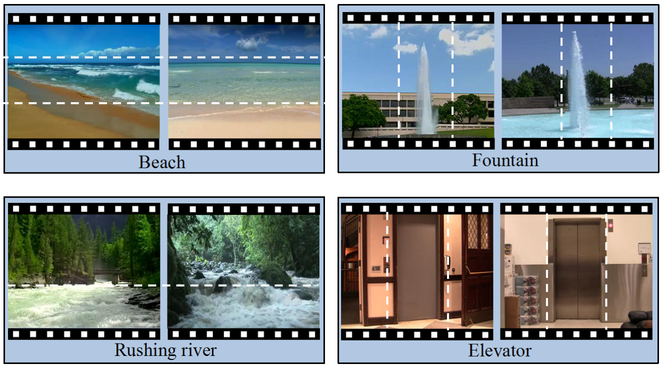


Fig. 2. Videos from the same dynamic scene class frequently share similar spatial layout. For example, a “beach” scene can often be partitioned into upper, middle and bottom parts with coherent appearance and semantic meaning, *i.e.*, sky, ocean and sand. Similar observations can be found in other categories such as “fountain”, “rushing river”, and “elevator”.

A property of dynamic scenes is that they can be characterized by either the dynamic part or the static part, or both. This differs from video-based activity recognition for which the dynamic parts play the key roles [16], [21], [43], [48], [50]. For instance, the “fountain” and “waterfall” scenes (Fig. 1(a)) might be similar in terms of the dynamic parts but can be distinguished easily by their static backgrounds. On the other hand, to distinguish the “snowing” scene from the “street” scene (Fig. 1(b)), the motion related features are more discriminative. Another property of dynamic scenes is that scenes from the same category often share similar spatial layout (*e.g.*, Fig. 2), and some portions of the videos from the same category and same spatial position are often similar, in terms either semantics or appearance, or both. In fact, such property has been exploited for static scene analysis (*e.g.*, [31]). These properties suggest that middle level representation for dynamic scene should cover both static and motion portions and should exploit the spatial layout priors.

With the above motivation, we develop a middle level

Liang Du and Haibin Ling are with the Department of Computer and Information Science, Temple University, Philadelphia, PA, 19111.
E-mail: {liang.du, hbling}@temple.edu

dynamic scene representation including two key components. The first component is the *redundant spatial grouping* (RSG) for decomposing a dynamic scene video into a large number of sub-videos, namely *scenelets*. The redundancy in RSG allows the representation to preserve both of motion and static parts. The second component is the *category-scenelet model* (CSM) such that each model estimates the likelihood of a scenelet belonging to a certain category. To encode the spatial interactions and constraints among scenelets, CSMs for the same category are jointly learned with a group-sparsity constraint that implicitly selects discriminative scenelets for a category. In this way, categorical specific spatial layout priors are effectively encoded in the learned models.

Imposing sparsity constraints for scenelet models across different spatial positions actually performs a discriminative patch pruning, which is a crucial step for many middle level representations (*e.g.*, [3], [21], [40], [45]), though previously not for dynamic scene classification. The information carried in each scenelet model can be viewed as a model for a scene part at a specific location, *e.g.*, "upper-left part of a beach". Fig. 3 shows the flowchart of the proposed method. Experimental results show that this strategy outperforms the baseline methods of independent training or using a generic model for all scenelets with the same scene label in Section IV. To evaluate the effectiveness of the proposed method, it is applied to dynamic scene classification on two benchmark dynamic scene datasets [8], [44] and to a recently proposed video analysis task named violence video classification [20]. In all the experiments, the proposed representation produces significant performance gains over previously proposed solutions.

To summarize, we make the following contributions:

- We propose to use redundant spatial scenelets to exploit middle level information for dynamic scene recognition.
- We propose to jointly learn category-scenelet models by exploiting both categorical supervision and spatial priors and interactions among the scenelets.
- Our approach has registered new better results on two popular dynamic scene classification benchmarks and a violence video classification benchmark.

We proceed as follows: In Section II, we give a brief review of existing studies related to our method. In Section III, we present the proposed redundant spatial grouping and the algorithm for training scenelet models. The application of the proposed method for dynamic scene classification and violence video classification are presented in Section IV. Finally, we conclude the paper in Section V.

II. Related Work

A. Static Scene Recognition

Recognizing scenes from static images has been studied intensively and various methods have been shown to be successful, such as [1], [2], [13], [22], [27], [29], [37], [38], [41], [53], [54], to name a few. In [37], a discriminative holistic scene representation, *spatial envelope*, was proposed to represent the "gist" of a natural scene. Besides the holistic scene representation, encoding local features with spatial layout information has achieved impressive performance [1], [27]. In [27], dense SIFT features are organized through spatial pyramid matching for recognizing natural scene categories. In [1], textons are used as "visual words" to represent local features. More recently, middle level features have gained popularity for static scene recognitions [22], [23], [29]. In [29], a bank of object detectors is trained on extra datasets and the responses of the detectors are used as features for scene representation. In [22], distinctive parts are learned to represent image scenes. Recently, deep learned features have also shown to be effective in scene recognition [15], [60].

Our work is partly inspired by middle level feature representations of static scene recognitions [22], [29]. However, compared with the research on static scene recognition, study on dynamic scene recognition is new. One issue is the scarcity of large dynamic scene datasets. While there are a decent number of large static scene datasets (*e.g.*, ImageNet [7], and MIT 67 Indoor [38]) for training base object detectors [29] or mining distinctive parts [22] for static scenes, only a few benchmark datasets are available for dynamic scene recognition. This constraint makes middle level representation a challenging problem for dynamic scene recognition.

B. Dynamic Scene Recognition

Recently, dynamic scene recognition has attracted researchers' attentions [8], [14], [35], [44], [49]. A closely related topic is dynamic texture classification [9], which focuses more on temporally repeating patterns. In [35], the histogram of optical flow (HOF) is used to model scene dynamics. The results of dynamic scene recognition are used as priors for action recognition. In [44], the authors propose to use chaotic system parameters as features for dynamic scene classification. In order to study the role of orientation features in dynamic scenes, Derpanis *et al.* [8] use a set of Gaussian derivative filters to obtain orientation features. In [8], a video sequence is partitioned following the canonical spatial pyramid paradigm. In [19], optical flow features are extracted from each pair of consecutive frames, and quantized into discrete visual flow words. In contrast, in our method, spatio-temporal blocks (scenelets) are obtained by combining any number of neighboring elementary patches whenever they form a cuboid. This leads to a more redundant spatial layout division. In addition, each spatio-temporal patch is treated equally in the final classification for [8]. In [24], wavelet domain multi-spectrum fractal analysis is proposed for dynmic texture classification. In contrast, in our methods, the importance of each redundant scenelet is learned via a group sparse algorithm. In [47], slow feature analysis (SFA) is applied to dynamic scene recognition. In [14], a forest-based classifier is used with spatio-temporal descriptor for dynamic scene recognition. In [49], five dimensional motion features are used for dynamic scene recognition. In [18], a method based on bag of visual word framework was proposed by using spacetime energies and color feature to study dynamic scene recognition problem. Almost all these works on dynamic scene classification are devoted to exploit low level cues or models. The role of middle level features is under-explored for dynamic scene recognition.

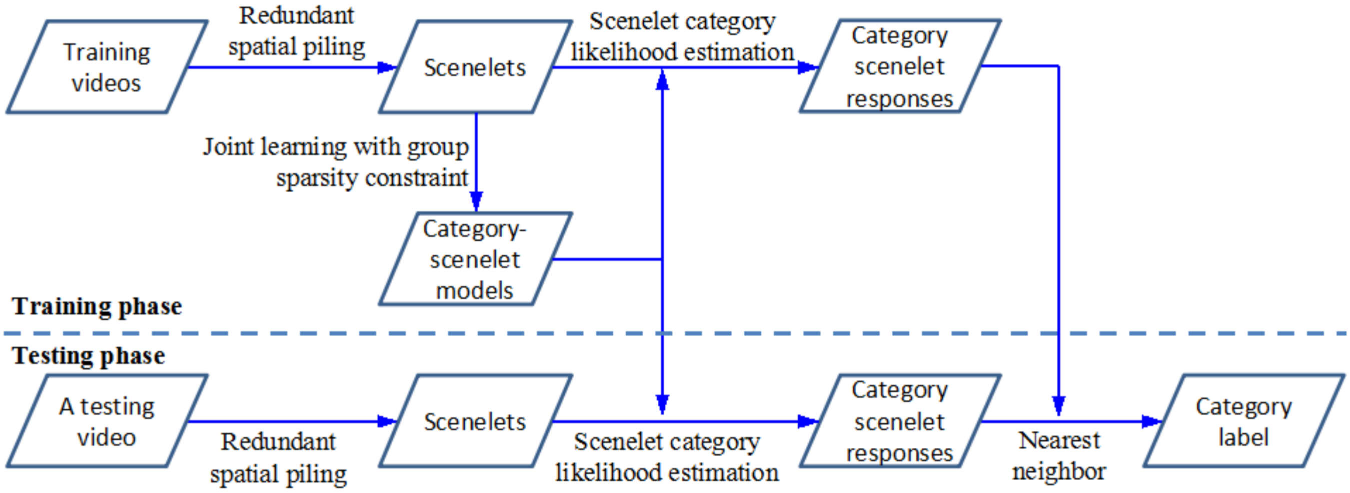

Training videos
Redundant spatial piling
Scenelets
Scenelet category likelihood estimation
Category scenelet responses
Joint learning with group sparsity constraint
Category-scenelet models
Training phase
Testing phase
A testing video
Redundant spatial piling
Scenelets
Scenelet category likelihood estimation
Category scenelet responses
Nearest neighbor
Category label

Fig. 3. Framework of the proposed dynamic scene classification approach.

C. Middle Level Feature Representation

Recently, middle level representation has been proven effective in many vision tasks, *e.g.*, human detection [3], object detection [52], [57], [58], static scene classification [29], [45] and activity recognition [21], [50]. Generally, middle level features are extracted by models (*e.g.*, e-SVM detectors [34]) trained from discriminative patches. The core component of these methods is to harvest and prune representative and discriminative patches [10], [45].

Our work is most related to recent studies in middle level video representation such as [21], [39], [50], which also use spatio-temporal patches to represent videos. Besides the obvious differences in applications, our method differs from them in the way we sample spatio-temporal patches and train middle level models. Instead of mining discriminative patches by ranking methods [21] or using motion saliency cues [50], scenelets are generated at by redundantly grouping local parts. Each scenelet in the proposed method is not only associated with a scene class but also a spatial layout information in the video volume. In addition, instead of using e-SVM detectors [34] or patch template [50] to model middle level patches, scenelet models are jointly modeled by a logistic regression formulation, in which group sparsity is employed to enforce sparsity of scenelet models across different spatial positions. In this way, categorical specific spatial layout priors are encoded in the learned models.

III. Redundant Spatial Scenelets for Dynamic Scene Classification

A. Overview

A flowchart for the our study, including both model learning and testing phases, is given in Fig. 3. In particular, given a dynamic scene video, it is first decomposed into a set of scenelets through redundant spatial grouping, and then *category-scenelet models* are applied on these scenelets to get a collection of category likelihood estimation per scenelet and category. Such a representation, named *categorical-scenelet response matrix* or CSR, is then used for scene classification. For each category, the category-scenelet models of all scenelets are jointly learned to explore the spatial interactions and constraints among them.

B. Redundant Spatial Grouping

Our goal is to generate a large number of representative scenelets. The scenelets of the proposed method are generated by the *redundant spatial grouping* of elementary spatio-temporal cells. In this way, the proposed representation can redundantly cover all spatial positions. It accords with our observation that dynamic scenes are characterized by both dynamic foreground and (relatively) static backgrounds.

Specifically, given an input video V, its scenelet representation is a set of n_S spatio-temporal patches, named *scenelets*, denoted as

$$\mathbb{V} = \{V_i : i = 1, \ldots, n_S\} \ . \tag{1}$$

The decomposition is achieved as following: we first partition V spatially into a $N \times N$ non-overlapping spatio-temporal cells with equal size. Then, any grouping of neighboring patches, including the patches by themselves, is treated as a scenelet as long as it forms a cuboid. This process is similar to the over-complete repceptive field in [25]. This operation will lead to a large number of spatially redundant scenelets. Fig. 4 shows an example of redundant spatial partition for $N = 3$, producing 36 scenelets in total. Note that different scenelets can be composed of different numbers of elementary spatio-temporal cells.

In theory, temporal divided patches may also help in discriminating scene classes, at a cost of increased computation complexity. In practice, only relatively short videos are available (*e.g.*), which renders temporal division unreliable. On the other hand, since spatio-temporal features (*e.g.*, SOE or Gabor3D, see Section IV) are included to derive the CSR calculation, they can compensate the temporal variations in videos. For these reasons, we only use redundant spatial grouping and leave the temporal one.

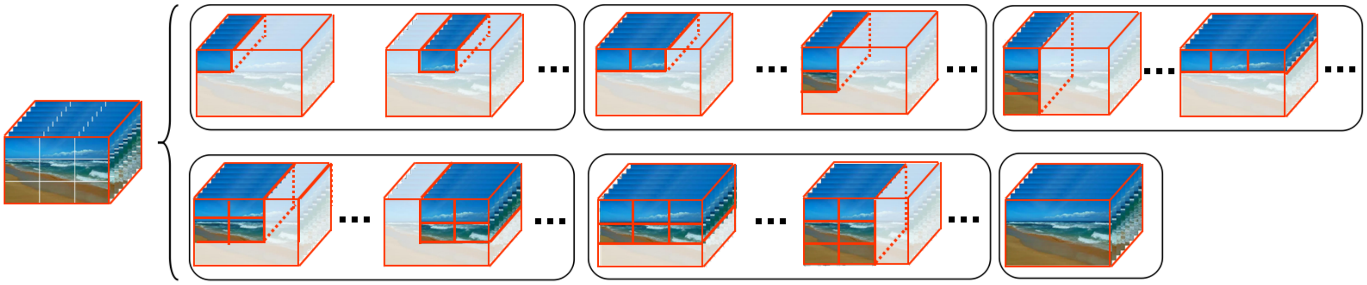


Fig. 4. Example of scenelets from redundant spatial grouping a 3×3 spatial cells. From left to right and from top to bottom, the numbers of cells in each group of scenelets are 1,2,3,4,6 and 9.

C. Joint Learning of Category-Scenelet Models

The scenelets from the c-th dynamic scene category and i-th spatial grouping are treated as the same class of scenelets which is indexed by (c,i). The *category-scenelet model* (CSM) of scenelet class (c,i) can be jointly trained with group sparsity constraints with other scenelet models.

We denote the CSM model of the c-th scene category and i-th scenelet by $f_{c,i}(\mathbf{x}_i)$, which estimates the likelihood of the i-th scenelet belonging to the c-th category. The function $f_{c,i}(\mathbf{x}_i)$ is modeled as a logistic function of input feature vector $\mathbf{x}_i$, *i.e.*,

$$f_{c,i}(\mathbf{x}) = \frac{1}{1+\exp\left(-\ell_{c,i}(\mathbf{x})\right)}, \tag{2}$$

where $\ell_{c,i}(\mathbf{x}) = \mathbf{w}_{c,i}^{\top}\mathbf{x} + b_{c,i}$, and $\mathbf{w}_{c,i} \in \mathbb{R}^d, b_{c,i} \in \mathbb{R}$ are the model coefficients. Our CSM representation is determined by the two groups of parameters denoted by

$$W = [W_1 W_2 \ldots W_{n_C}] \in \mathbb{R}^{d \times n_C n_S}$$

where $B = (b_{c,i}) \in \mathbb{R}^{n_C \times n_S}$, $W_c = [\mathbf{w}_{c,1}\mathbf{w}_{c,2}\ldots\mathbf{w}_{c,n_S}] \in \mathbb{R}^{d \times n_S}$ collects all weight vectors of all scenelets classifiers for c-th category.

Given a training set of n_V sample videos with extracted features and annotated labels $\{(\mathbb{X}^{(s)}, y^{(s)}) : s = 1,2,\ldots,n_V\}$, where $y^{(s)} \in \{1,\ldots,n_C\}$ are the categorical labels. The model for each category of scenelets are learned in a one-vs-all fashion. For the c-th scene category, by taking the negative logarithm on the loss of sum of likelihood and adding a mix-norm regularizer, we reach the following optimization problem:

$$\min_{W_c,\mathbf{b}_c} \sum_{s=1}^{n_V}\sum_{i=1}^{n_S} \ln\Big(1+\exp\big(-\delta(c,y^{(s)})\ell_{c,i}(\mathbf{x}_i^{(s)})\big)\Big) + \gamma\|W_c\|_{2,1}, \tag{3}$$

where $\delta(c,y) = \begin{cases} 1, & y=c \\ -1, & y \neq c \end{cases}$ is the indicator function that converts labels into binary; $\mathbf{b}_c \in \mathbb{R}^{n_S}$ is the vector whose elements are bias terms for the corresponding models; γ is weight parameter balancing between regularization term and the loss term. Learning middle level features with weak supervision (*i.e.*, only categorical level information but not patch level supervision) has been proven to be beneficial [17], [45].

The $\ell_{2,1}$ norm in (3) is defined by $\|W_c\|_{2,1} = \sum_{i=1}^{n_S}\|\mathbf{w}_{c,i}\|_2$. It enforces column-wise sparisty, *i.e.*, sparsity over scenelets across different spatial positions. This regularization implicitly performs a scenelet pruning during the joint model learning.

The problem in (3) is an ℓ_1/ℓ_q regularized multi-task learning problem, which could be formally defined as follows:

$$\min_{W \in \mathscr{R}^p} f(W) = l(W) + \lambda\omega(W). \tag{4}$$

where $l(\cdot)$ is convex loss dependent on training samples and

$$\omega(W) = \sum_{i=1}^{s} ||w_i||_q. \tag{5}$$

is the ℓ_1/ℓ_q norm.

Though not convex, efficient solutions have been proposed recently. We follow the algorithm in [32] in our solution.

An accelerated gradient using ℓ_1/ℓ_q Euclidean projection is used to solve the ℓ_1/ℓ_q regularized problem efficiently.

More details can be referred to [12], [32], [59].

Now that we have a dynamic scene video V decomposed into scenelets $\mathbb{V}$, and a set of learned CSMs $\{f_{c,i}(.) : 1 \le c \le n_C, 1 \le i \le n_S\}$, we can represent V by applying all CSMs to the scenelets in $\mathbb{V}$. We call such representation *categorical-scenelet response matrix* (CSR). In particular, CSR for V is a matrix denoted by $A = (a_{c,i}) \in \mathbb{R}^{n_C \times n_S}$, such that $a_{c,i} = f_{c,i}(\mathbf{x}_i)$ is an estimated category likelihood.

Each scenelet model may correspond to a partial dynamic scene (*e.g.*,"concrete pavement" for scene "street"), a dynamic scene per se (*e.g.*, scene "sky clouds"), or perhaps a random but informative spatio-temporal patch in a video. The impurity of each type of scenelets might cause some ambiguities in the detection. This implies that the corresponding scenelet model is not an exact category detector. It is a noisy model for spatial specific component of a dynamic scene category. The model does not perform classification tasks by itself. The discriminative ability of individual scenelet is boosted by aggregating all scenelet responses into a CSR descriptor, which is used as the final feature. Actually, allowing certain level of ambiguity and avoiding hard decisions before the very final step are favored in many semantic modeling approaches. It is a manifestation of the data processing theorem which advocates to postpone hard decisions until the very last stage of processing [33].

In the rest of the paper, we call the proposed dynamic scene classification algorithm again as **CSR**. By contrast, in our experiments we also include two baseline variations: one trains all categorical-spatial likelihood functions independently by replacing the $\ell_{2,1}$ norm with the ℓ_2 norm and the other trains a generic scene classifier using the whole video for estimating category likelihood of all scenelets. These two variations are referred as $\mathbf{CSR}_{\ell_2}$ and $\mathbf{CSR}_b$ ('b' for baseline) respectively.

D. Implementation Details

To describe a scenelet, we use two spatio-temporal features, Gabor3D and SOE [8], as the low level features in all experiments. Note that other features are also possible for our framework, and some of them are actually tested with our framework. We choose Gabor3D and SOE for their popularity in dynamic scene classification.

Gabor3D (G3D). The extraction of video structures in different orientations is done by filtering the video using a bank of 3D Gabor filters with different orientations: $G_3(\sigma,\theta)\otimes V_i$, where V_i is a subsequence and $G_3(\sigma,\theta)$ is the 3D Gabor filter with scale σ and orientation θ. In practical implementation, the convolution is firstly performed on the whole sequence and then, features are extracted from each filtered subsequences. The concatenation of the histogram of the magnitudes of the filtered videos in different orientations and scales are used as feature. In experiments, the scale is set to be fixed and 8 orientations ($\theta = \{\pi n/8\}_{n=0}^{7}$) is used. The size of Gabor filter is $27\times 27\times 27$. The magnitudes of the filtered video in each channel are binned into a 300 dimension histogram. Therefore, this amount to a final feature vector of $2,400$ dimensional.

Spatiotemporal oriented energy features (SOE) [8]. This feature is defined as the normalized response of Gaussian filtered video volumes. *i.e.*

$$E_{\hat{\theta}_i,\sigma_j} = \frac{E_{\hat{\theta}_i,\sigma_j}}{\varepsilon + \sum_{\hat{\theta}\times\sigma} E_{\hat{\theta},\sigma}}, \tag{6}$$

where $E_{\hat{\theta},\sigma}$ is the local energy measurement defined as

$$E_{\hat{\theta},\sigma} = \sum_{\mathbf{x}} \Omega(\mathbf{x})[G_{K_{\hat{\theta},\sigma}}(\mathbf{x}) * I(\mathbf{x})]^2 \ , \tag{7}$$

$\Omega(\mathbf{x})$ is a mask for the aggregation region and $G_{K_{\hat{\theta},\sigma}}$ the K-th derivative of the Gaussian with scale σ, and $\hat{\theta}$ is the direction of the filter's axis of symmetry.

For redundant spatial grouping, we fix N to 5 which results in 5×5 basic spatio-temporal cells and 225 scenelets in total (*i.e.*, $n_S = 225$). In practice, we find the 5×5 basic spatio-temporal cells provide sufficiently fine granularity for dynamic scene classification while effectively balancing the computational cost. The influence of parameter N in our method is shown in Table III.

IV. Experiments

We evaluate the proposed representation on three benchmark datasets. For two dynamic scene datasets, both nearest neighbor classifier and linear SVM classifier are tested for comparison following the protocols of the state-of-the-art dynamic scene methods [8], [18], [44], [47]. In particular, for nearest neighbor classification, let $A^{(1)}, A^{(2)}$ be CSRs from two scene videos, we use the Frobenius form of their difference, *i.e.*, $|A^{(1)} - A^{(2)}|_F$, as the metric. For experiment on violence video classification, SVM classifier with RBF kernel is used as in the state of the art [20].

A. Evaluation on Public Dynamic Scene Datasets

Datasets. We evaluate our proposed CSR on two benchmarking datasets, *i.e.*, the Maryland "in-the-wild" scenes dataset [44] and the "Stabilized" dynamic scenes dataset [8].

Maryland "in-the-wild" scenes dataset [44]. This dataset contains thirteen dynamic scene classes with ten color videos each class. The videos were collected from Internet video sharing sites, *e.g.*, Youtube (www.youtube.com). This dataset is very challenging that the videos therein are less constraint with large variation in illumination, frame rate, viewpoint, image scale and various degrees of camera-introduced motions. Fig. 5 shows some example scene frames of the dataset.

"Stabilized" dynamic scenes dataset [8]. This is a new dataset introduced by Derpanis *et al.* for the purpose of evaluating their video orientation descriptor by excluding influences from camera motions. There are in total fourteen dynamic scene classes with thirty color videos each. Fig.6 shows some example scene frames of the dataset.

Classification results. We follow the same leave-one-video-out setting in [8]. Results of both nearest neighbor classifier and linear SVM are reported for comparison. The proposed CSR representation can achieve better performance than the state of the arts, demonstrating the discriminative ability of the proposed representation.

Table I shows the details comparison of the proposed method with the state of the art. For fair comparison, the left thirteen columns show results using nearest neighbor classifier and the right four columns show results using stronger classifiers (*i.e.*, random forest and linear SVM). The results of the first five columns are quoted from [44]. We can see that G3D and SOE features perform poorly when using the low level feature themselves for classification. By using the proposed representation, performances for both features improve by a large margin. The accuracy for CSR with G3D increases from 21% to 65% and CSR with SOE from 41% to 71%. 71% and 86% are the best published performances on this dataset using nearest neighbor classifier and linear SVM respectively, even though it only uses a single feature channel.

In [47], the SFA-based method achieves an accuracy of 60%, by using stronger classifier (*i.e.*, a linear SVM)[1]. In [14], they proposed to use random forest in classification of dynamic scenes and achieved an accuracy of 68% on this dataset. In [18], by using a combination of spate-time energy feature and color feature, along with a linear SVM classifier, they obtained an accuracy of 69.23%. However, by using linear SVM classifier, we achieve 86%. Fig. 7 presents the confusion matrix for the proposed CSR on this dataset.

Table II shows the comparison results on the "stabilized" dynamic scene dataset. Applying the proposed CSR representation to the SOE feature, our approach outperforms

[1] An erratum on the results of the original paper: http://webia.lip6.fr/~theriaultc/sfa.html

Fig. 5. Sample frames of the Maryland “in-the-wild” scenes dataset [44]. The classes are: avalanche, boiling water, chaotic traffic, forest fire, fountain, iceberg collapse, landslide, smooth traffic, tornado, volcano eruption, waterfall, waves, and whirlpool.

Fig. 6. Examples from the “Stabilized” scenes dataset [8]. The classes are: beach, elevator, forest fire, fountain, highway, lightning storm, ocean, railway, rushing river, sky-clouds, snowing, street, waterfall, and windmill farm.

TABLE I

CLASSIFICATION RATES (%) ON THE MARYLAND “IN THE WILD” DYNAMIC SCENE DATASET. RESULTS OF THE FIRST FIVE COLUMNS ARE FROM [44]. THE LEFT COLUMNS ARE USING NEAREST NEIGHBOR CLASSIFIERS AND THE RIGHT FOUR COLUMNS ARE USING LINEAR SVM, EXCEPT [14] WHICH IS USING RANDOM FOREST. METHODS WITH CSRARE OUR PROPOSED METHODS. THE **BEST** AND SECOND BEST RESULTS ARE HIGHLIGHTED.

Class	LDS GIST	Bag of Words	Mean GIST	Dyn. Chaos	Static +Dynamics	G3D	SOE [8]	CSR$_{\ell_2}$ G3D	CSR$_{\ell_2}$ SOE	CSR$_b$ G3D	CSR$_b$ SOE	CSR G3D	CSR SOE	SFA [47]	STRF [14]	BoSE [18]	CSR SOE
Avalanche	70	30	50	30	40	10	40	40	90	10	40	60	80	60	60	60	80
Boiling water	70	0	30	30	40	30	60	80	80	20	20	100	90	70	80	70	100
Chaotic traffic	10	20	30	50	70	10	80	60	90	20	90	100	90	80	90	90	90
Forest fire	0	30	40	30	40	0	40	50	70	20	60	40	40	10	80	90	90
Fountain	0	10	50	20	70	10	10	90	40	40	10	10	50	50	80	70	80
Iceberg collapse	20	30	40	10	50	0	20	60	40	40	40	40	90	60	60	60	90
Landslide	20	40	20	20	50	10	50	50	50	40	20	30	30	60	30	60	80
Smooth traffic	10	0	40	20	50	20	60	50	40	40	80	80	60	50	50	70	90
Tornado	70	10	70	60	90	20	60	60	40	60	90	90	90	70	80	90	90
Volcano eruption	0	30	30	70	50	20	10	70	90	50	50	80	100	80	70	80	100
Waterfall	0	30	10	30	10	30	10	70	70	40	50	60	60	50	50	100	80
Waves	40	50	70	80	90	10	80	70	70	70	60	80	90	60	80	90	90
Whirlpool	20	30	40	30	40	70	40	50	50	30	60	70	50	80	70	80	60
Average	25	24	40	36	52	21	41	61	63	37	52	*65*	***71***	60	68	*78*	***86***

the state of the art (84% vs. 82%) using nearest neighbor classifier. Moreover, by using the same low level features, the performance gain by using CSR is obvious: for G3D, the performance increases from 40% to 78%, for SOE, the performance increases from 74% to 84%. In [49] and [14], accuracies of 85.61% and 86% are reported. By using linear SVM classifier, our performance is 94%. In [18], by using a combination of spate-time energy feature and color feature, along with a linear SVM classifier, they obtained an accuracy of 96%. Fig. 8 presents the confusion matrix for the proposed CSR on this dataset.

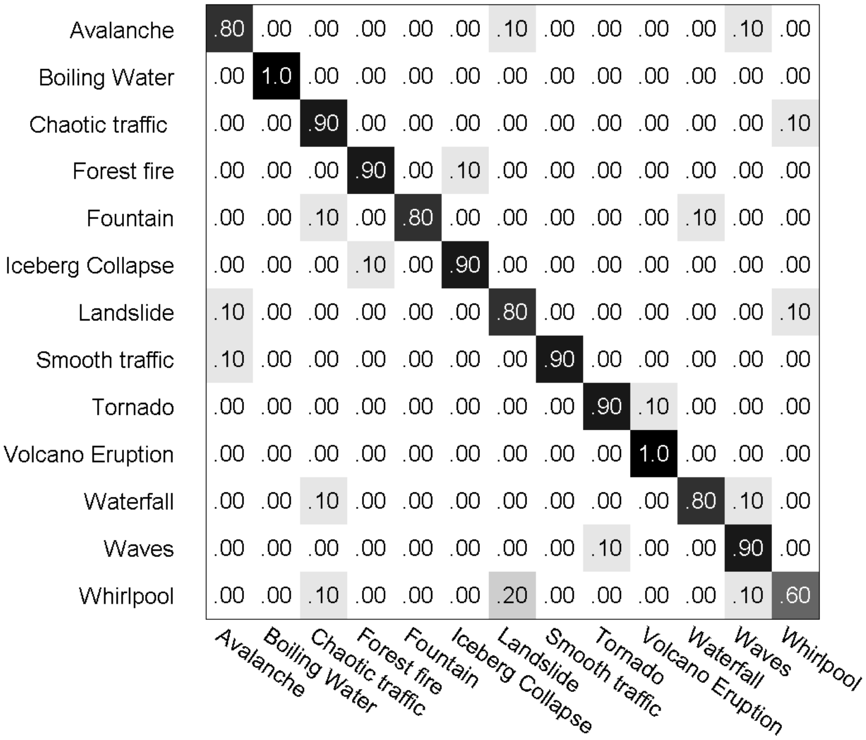
TABLE III

INFLUENCE OF PARTITION PARAMETER N ON PERFORMANCES (%).

N	3	4	5	6
Maryland “in-the-wild” scene dataset	33	56	86	79
“Stablized” dynamic scene dataset	44	79	94	93

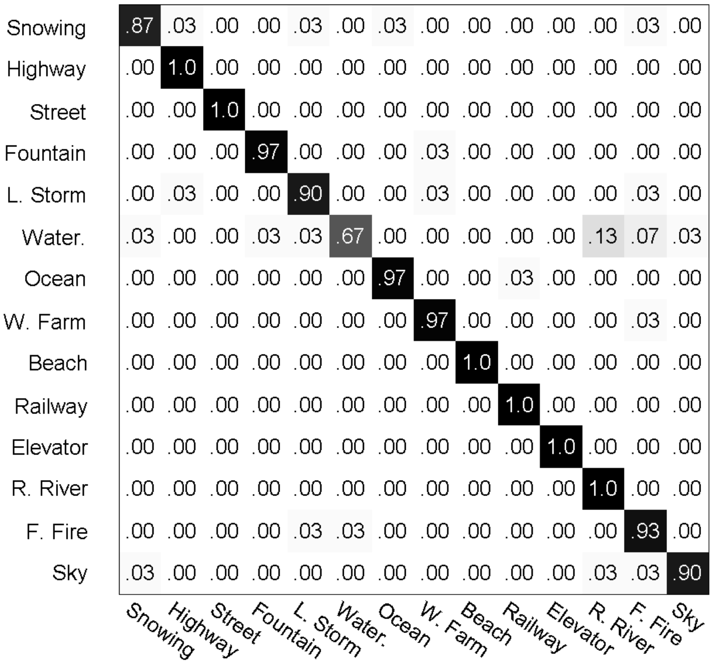
Moreover, we can see that it is advantageous to train our categorical-spatial likelihood functions jointly using $\ell_{2,1}$ regularizer, compared with training the generic scene classifiers (CSR$_b$) or independently training (CSR$_{\ell_2}$).

To investigate the influence of the spatial partition parameter N on the performance of CSR, we test our algorithm with



	Avalanche	Boiling Water	Chaotic traffic	Forest fire	Fountain	Iceberg Collapse	Landslide	Smooth traffic	Tornado	Volcano Eruption	Waterfall	Waves	Whirlpool
Avalanche	.80	.00	.00	.00	.00	.00	.10	.00	.00	.00	.00	.10	.00
Boiling Water	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Chaotic traffic	.00	.00	.90	.00	.00	.00	.00	.00	.00	.00	.00	.00	.10
Forest fire	.00	.00	.00	.90	.00	.10	.00	.00	.00	.00	.00	.00	.00
Fountain	.00	.00	.10	.00	.80	.00	.00	.00	.00	.00	.10	.00	.00
Iceberg Collapse	.00	.00	.00	.10	.00	.90	.00	.00	.00	.00	.00	.00	.00
Landslide	.10	.00	.00	.00	.00	.00	.80	.00	.00	.00	.00	.00	.10
Smooth traffic	.10	.00	.00	.00	.00	.00	.00	.90	.00	.00	.00	.00	.00
Tornado	.00	.00	.00	.00	.00	.00	.00	.00	.90	.10	.00	.00	.00
Volcano Eruption	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.0	.00	.00	.00
Waterfall	.00	.00	.10	.00	.00	.00	.00	.00	.00	.00	.80	.10	.00
Waves	.00	.00	.00	.00	.00	.00	.00	.00	.10	.00	.00	.90	.00
Whirlpool	.00	.00	.10	.00	.00	.00	.20	.00	.00	.00	.00	.10	.60

Fig. 7. Confusion matrix of our method on the Maryland “in the wild” dataset.



	Snowing	Highway	Street	Fountain	L. Storm	Water.	Ocean	W. Farm	Beach	Railway	Elevator	R. River	F. Fire	Sky
Snowing	.87	.03	.00	.00	.03	.00	.03	.00	.00	.00	.00	.00	.03	.00
Highway	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Street	.00	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Fountain	.00	.00	.00	.97	.00	.00	.00	.03	.00	.00	.00	.00	.00	.00
L. Storm	.00	.03	.00	.00	.90	.00	.00	.03	.00	.00	.00	.00	.03	.00
Water.	.03	.00	.00	.03	.03	.67	.00	.00	.00	.00	.00	.13	.07	.03
Ocean	.00	.00	.00	.00	.00	.00	.97	.00	.00	.03	.00	.00	.00	.00
W. Farm	.00	.00	.00	.00	.00	.00	.00	.97	.00	.00	.00	.00	.03	.00
Beach	.00	.00	.00	.00	.00	.00	.00	.00	1.0	.00	.00	.00	.00	.00
Railway	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.0	.00	.00	.00	.00
Elevator	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.0	.00	.00	.00
R. River	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.0	.00	.00
F. Fire	.00	.00	.00	.00	.03	.03	.00	.00	.00	.00	.00	.00	.93	.00
Sky	.03	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.03	.03	.90

Fig. 8. Confusion matrix of our method on the “stabilized” dynamic scene dataset.

different values of N. The results are shown as Table III. From the table, we can see that the performance increases with the number of scenelets until N reaches 5. Moreover, we notice that the performance on the Maryland “in-the-wild” scene dataset decreases more dramatically than that on the “stabilized” scene dataset. This can be explained by the fact that N actually determines the degree of granularity for redundant spatial grouping. A larger N will lead to smaller granularity, which is more susceptible to spatial misalignment within the same dynamic scene. Since the Maryland “in-the-wild” scene dataset has more camera-introduced motions, its performance decreases more (from 86% to 79%) for larger N

TABLE II

Classification rates (%) on the “stabilized” dynamic scene dataset. Results of the first five columns except G3D are from [8]. For SFA [47], the original paper does not provide the per class performances but the overall one. The left columns are using nearest neighbor classifiers and the right four columns are using linear SVM, except [14] which is using random forest. Methods with CSR are our proposed methods. The **best** and second best results are highlighted.

Class	MSOE	Chaos	HOF+	G3D	SOE	SFA	CSR_{ℓ_2}	CSR_{ℓ_2}	CSR_b	CSR_b	CSR	CSR	STRF	BoSE	5DMFV	CSR
		+GIST	GIST		[8]	[47]	G3D	SOE	G3D	SOE	G3D	SOE	[14]	[18]	[49]	SOE
Beach	83	30	76	50	87	-	70	77	77	100	100	100	100	100	98	100
Elevator	60	40	90	16	67	-	83	93	90	100	83	100	100	97	90	100
Forest fire	60	17	63	3	83	-	50	70	60	40	50	73	83	93	80	93
Fountain	40	3	37	73	47	-	73	77	13	43	67	77	47	87	60	97
Highway	60	23	53	23	77	-	50	70	37	67	67	80	73	100	81	100
Lighting storm	87	40	70	93	90	-	90	100	73	87	93	83	93	97	67	90
Ocean	97	43	93	20	77	-	100	100	93	100	100	97 100	90	100	90	97
Railway	60	7	87	27	87	-	87	83	63	97	93	97	93	100	87	100
Rushing river	90	10	73	23	47	-	83	73	70	87	93	93	97	97	95	100
Sky	80	43	87	30	90	-	100	73	67	87	90	93	100	97	92	90
Snowing	17	10	40	33	33	-	80	77	40	73	70	73	57	97	90	87
Street	63	17	80	20	83	-	57	87	77	100	100	100	97	100	97	100
Waterfall	37	10	50	40	43	-	57	63	27	17	13	23	76	83	75	67
Windmill farm	47	17	60	67	77	-	83	87	80	77	83	87	93	100	92	97
Average	63	22	69	40	74	*82*	75	81	62	76	78	***84***	86	***96***	86	*94*

($N = 6$) than that of the “stabilized” scene dataset (from 94% to 93%).

From the above experiments, we draw safely the conclusion that CSR can improve the discriminability of classifiers for dynamic scene classification. In addition, using the $\ell_{2,1}$ regularizer helps achieve better results than the two baselines.

B. Violence video classification

In this subsection, we apply the proposed middle level representation to an emerging video-based application, *i.e.*, violence video classification. A related topic is abnormal detection [56], which focuses on abnormal behavior of crowds. By using CSR along with a simple nonlinear SVM (*i.e.*, RBF kernel), our method can achieve much better results than the state of the art [20]. Note that, the task is binary (violence vs. nonviolence) and therefore the CSR representation becomes vector.

The goal of violence video classification is to monitor crowded events of outbreak of violence [20]. We use the benchmark dataset assembled in [20] and follow the five-fold cross-validation classification test protocol therein. Fig. 9 shows some examples of violent frames. Violences are often characterized by the interactions between subjects within a video. Both G3D and SOE are tested as our low level features.

We compared our method with the state of the art on violence video classification [20], and many other state-of-the-art techniques on activity recognition[2]: LTP [55], HOG [26], HOF [26] and HNF [26]. The results are reported as mean accuracy and area under ROC curve (AUC) and shown in Table IV. The proposed method outperforms all other methods both in terms of accuracy and AUC. We also notice that in the “Violence” video classification task, G3D outperforms SOE. This could be explained by the fact that the key cue for group “violence” is the quick movements of a group of people, and G3D is more sensitive to this kind of motions.

C. Semantic Interpretability

Considering the middle level nature of CSR, it is of interest to investigate its ability in representing video by semantically meaningful features. For example, an “ocean” scene should pay more attentions to water-related features in CSR. We investigate this property of CSR by plotting the most discriminative type of scenelets except for the scenelet models belonging to itself. The discriminative ability is measured by weights of each type of scenelet for each one-vs-all linear SVM, *i.e.*, the weight of each dimension for CSRs.

Fig. 10 shows the results of two scene categories in the “stabilized” dynamic scene dataset. This figure visualizes the importance of middle level scenelet measured by the weights of one-vs-all linear classifier. Each type of scenelets is visualized as a representative exemplar spatio-temporal patch. A close inspection of the most discriminative scenelets verifies that the proposed representation indeed carries discriminative information. For the “beach” scene, scenelets of “sky clouds” and “ocean” both have high discriminative ability because they are compositional parts for a beach. In addition, the scenelets from the lower part of the “street” scene also show good discriminative abilities. This could be explained by the appearance similarity between the sands on a beach and concrete pavement of a street. For the “waterfall” scene, scenelets from the “fountain” and “rushing river” scenes have the highest discriminative ability. It is intuitively plausible as they resemble “waterfall” in either motion or static context. However, it is surprising that the “forest fire” scene also possesses high discriminative ability. Actually, we have observed that the motion patterns in fire look similar to the motion of waterfall.

V. Conclusions

In this paper, we proposed a middle level feature representation for dynamic scene recognition. In the representation, redundant spatial grouping is used to harvest middle level scenelets which can cover both static and dynamic portions

[2] Re-implemented in [20]

Fig. 9. Examples of violence videos from the violence video dataset [20].

TABLE IV
CLASSIFICATION RESULTS ON THE VIOLENCE DATASET [20]. RESULTS FOR METHODS IN THE FIRST FIVE COLUMNS ARE QUOTED FROM [20].

Method	LTP [55]	HOG [26]	HOF [26]	HNF [26]	ViF [20]	CSR_{ℓ_2} SOE	CSR_b SOE	CSR SOE	CSR_{ℓ_2} G3D	CSR_b G3D	CSR G3D
Accuracy	71.53	57.43	58.53	56.52	81.30	70.95	67.07	78.95	82.20	76.52	**87.08**
AUC	79.86	61.82	57.60	59.94	85.00	78.80	76.21	88.72	90.11	86.17	**93.60**

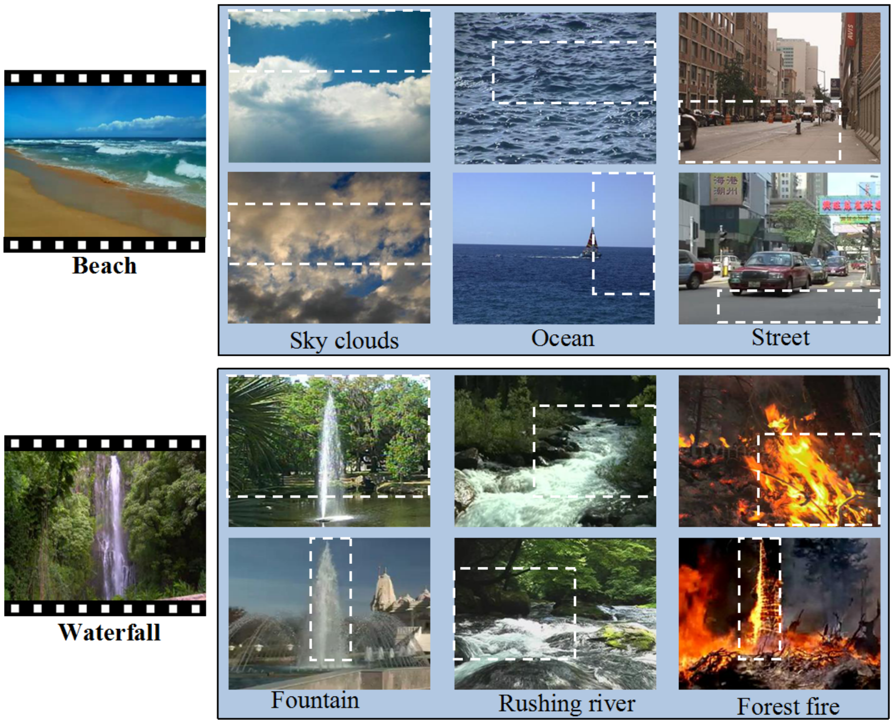


Fig. 10. Examples of dynamic scenes and the most discriminative scenelets. Different scenelet models are visualized by its representative exemplar scenelets (see Sec. IV-C for details).

of dynamic scenes with pre-defined spatial information. The proposed middle level representation can significantly improve the performances of the baseline low level features. In addition, by using the jointly learning of scenelet models via group sparsity regularization, spatial layout information for dynamic scenes is encoded and therefore recognition performances are improved compared with the independent learning. Extensive experiments on two benchmark dynamic scene datasets and a violence video benchmark demonstrate the superiority of the proposed representation in comparison with the state-of-the-art methods.

ACKNOWLEDGEMENT

We sincerely thank the anonymous reviewers for valuable comments and suggestions. This work was supported in part by NSF Grants 1350521, 1218156, and 1449860.

REFERENCES

[1] S. Battiato, G. Farinella, G. Gallo, and D. Ravì. Exploiting textons distributions on spatial hierarchy for scene classification. EURASIP J. Image Video Process. 2010.

[2] A. Bosch, A. Zisserman, and X. Munoz. Scene Classification Using a Hybrid Generative/Discriminative Approach, *PAMI*, 2008.

[3] L. Bourdev, and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *ICCV*, 2009.

[4] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010.

[5] L. Cao, Y. Mu, A. Natsev, S. Chang, G. Hua, J. Smith. Scene aligned pooling for complex video recognition. In *ECCV*,2012.

[6] X. Cao, X. Wei Y. Han, and X. Chen. An Object-Level High-Order Contextual Descriptor Based on Semantic, Spatial, and Scale Cues. IEEE Transactions on Cybernetics,45(7): 1327-1339, 2015.

[7] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[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 *CVPR*, 2012.

[9] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto. Dynamic textures. *IJCV*, 2003.

[10] C. Doersch, A. Gupta, A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013.

[11] B. Doroodgar, Y. Liu, and G. Nejat. A Learning-Based Semi-Autonomous Controller for Robotic Exploration of Unknown Disaster Scenes While Searching for Victims. IEEE Transactions on Cybernetics,44(12):2719-2732, 2014.

[12] A. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, 2007.

[13] G. Farinella, D. Rav, V. Tomaselli, M. Guarnera, S. Battiato. Representing scenes for real-time context classification on mobile devices. *Pattern Recognition*, 2014.

[14] C. Feichtenhofer, A. Pinz, and R. Wildes. Spacetime forests with complementary features for dynamic scene recognition. In *BMVC*, 2013.

[15] C. Farabet, C. Couprie, L. Najman, L., and Y. LeCun. Learning Hierarchical Features for Scene Labeling. PAMI, 2013.

[16] A. Fathi, and G. Mori. Action recognition by learning mid-level motion features. In *CVPR*, 2008.

[17] B. Fernando, E. Fromont, and T. Tuytelaars. Mining mid-level features for image classification. *IJCV*, 2014.

[18] C. Feichtenhofer, A. Pinz, and R. Wildes. Bags of spacetime energies for dynamic scene recognition. In *CVPR*, 2014.

[19] W. Fu, J. Wang, H. Lu, and S. Ma. Dynamic scene understanding by improved sparse topical coding. Pattern Recognition 46(7):1841-1850, 2013.

[20] T. Hassner, Y. Itcher, and O. Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012.

[21] A. Jain, A. Gupta, M. Rodriguez, and L. Davis. Representing videos using mid-level discriminative patches. In *CVPR*, 2013.

[22] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: distinctive parts for scene classification.

[23] R. Kwitt, N. Vasconcelos, and N. Rasiwasia. Scene recognition on the semantic manifold. In *ECCV*, 2012.

[24] H. Ji, X. Yang, H. Ling, and Y. Xu. Wavelet Domain Multi-fractal Analysis for Static and Dynamic Texture Classification. IEEE Transactions on Image Processing 22(1):286–299, 2013.

[25] Y. Jia, C. Huang, and T. Darrell. Beyond spatial pyramids: receptive field learning for pooled image features. In *CVPR*, 2012.

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

[27] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2006.

[28] L. Li, and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007.

[29] L. Li, H. Su, E. Xing, and L. Fei-Fei. Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification. In *NIPS*, 2010.

[30] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *CVPR*, 2013.

[31] D. Lin, and J. Xiao. Characterizing layouts of outdoor scenes using spatial topic processes. In *ICCV*, 2013.

[32] J. Liu, and J. Ye. Efficient L1/Lq Norm Regularization. *Technical Report*, 2010.

[33] D. MacKay. Information theory, inference and learning algorithms. *Cambridge University Press*, 2003.

[34] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *ICCV*, 2011.

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

[36] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *NIPS*, 2003.

[37] A. Oliva, and A. Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *IJCV*, 2001.

[38] A. Quattoni, and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[39] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discriminative action parts from mid-level video representations. In *CVPR*, 2012.

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

[41] N. Rasiwasia, and N. Vasconcelos. Holistic Context Models for Visual Recognition. *PAMI*, 34(5):902–917, 2012.

[42] S. Sadanand, and J. Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[43] L. Shao, X. Zhen, D. Tao and X. Li. Spatio-Temporal Laplacian Pyramid Coding for Action Recognition. IEEE Transactions on Cybernetics,44(6):817-827, 2014.

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

[45] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[46] J. Sun, and J. Ponce. Learning discriminative part detectors for image classification and cosegmentation. In *ICCV*, 2013.

[47] C. Theriault, N. Thome, and M. Cord. Dynamic scene classification: Learning motion descriptors with slow features analysis. In *CVPR*, 2013.

[48] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea. Machine recognition of human activities: a survey. *IEEE T-CSVT*, 18(11):1473-1488, 2008.

[49] A. Vasudevan, S. Muralidharan, S. Chintapalli, and S. Raman. Dynamic scene classification using spatial and temporal cues. In *ICCV Workshops*, 2013.

[50] L. Wang, Y. Qiao, and X. Tang. Motionlets: mid-level 3D Parts for human motion recognition. In *CVPR*, 2013.

[51] L. Wang, Y. Qiao, and X. Tang. Mining motion atoms and phrases for complex action recognition. In *ICCV*, 2013.

[52] X. Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for Generic Object Detection. In *ICCV*, 2013.

[53] J. Xiao, K. Ehinger, A. Oliva and A. Torralba. Recognizing Scene Viewpoint using Panoramic Place Representation. In *CVPR*, 2012.

[54] N. Xie, H. Ling, W. Hu, and X. Zhang. Use Bin-Ratio Information for Category and Scene Classification. In CVPR, 2010.

[55] L. Yeffet, and L. Wolf. Local trinary patterns for human action recognition. In *ICCV*, 2009.

[56] Y. Yuan J. Fang and Q. Wang. Online Anomaly Detection in Crowd Scenes via Structure Analysis. IEEE Transactions on Cybernetics,45(3):562-575, 2015.

[57] X. You, Q. Li, D. Tao, and W. Ou. Mingming Gong:Local Metric Learning for Exemplar-Based Object Detection. IEEE Transactions on Circuits System and Video Technology. 24(8):1265-1276, 2014.

[58] X. You, R. Wang, and D. Tao. Diverse Expected Gradient Active Learning for Relative Attributes. IEEE Transactions on Image Processing 23(7): 3203-3217, 2014.

[59] J. Zhou, J. Chen and J. Ye. MALSAR: Multi-tAsk Learning via StructurAl Regularization. Arizona State University, 2012.

[60] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. In NIPS, 2014.