Bags of Spacetime Energies for Dynamic Scene Recognition

Christoph Feichtenhofer[1,2] Axel Pinz[1] Richard P. Wildes[2]
[1]Institute of Electrical Measurement and Measurement Signal Processing, TU Graz, Austria
[2]Department of Electrical Engineering and Computer Science, York University, Toronto, Canada
{feichtenhofer, axel.pinz}@tugraz.at wildes@cse.yorku.ca

Abstract

This paper presents a unified bag of visual word (BoW) framework for dynamic scene recognition. The approach builds on primitive features that uniformly capture spatial and temporal orientation structure of the imagery (e.g., video), as extracted via application of a bank of spatiotemporally oriented filters. Various feature encoding techniques are investigated to abstract the primitives to an intermediate representation that is best suited to dynamic scene representation. Further, a novel approach to adaptive pooling of the encoded features is presented that captures spatial layout of the scene even while being robust to situations where camera motion and scene dynamics are confounded. The resulting overall approach has been evaluated on two standard, publically available dynamic scene datasets. The results show that in comparison to a representative set of alternatives, the proposed approach outperforms the previous state-of-the-art in classification accuracy by 10%.

1. Introduction

In the last decade, research in image classification and object recognition is dominated by three general steps: (i) In the *feature extraction* step, low-level descriptors are extracted from interest points or densely from regular locations. (ii) The *coding* step generates intermediate visual words that transform local features into more effective representations for the underlying task. (iii) The *pooling* step accumulates encoded features over regions to embed weak geometric information even while maintaining important properties of spatial invariance.

This paper proposes a novel approach to dynamic scene recognition, Bags of Spacetime Energies (BoSE), within the framework indicated in Fig. 1. The approach combines primitive features based on local measurements of spatiotemporal orientation, careful selection of encoding technique and a dynamic pooling strategy that in empirical evaluation outperforms the previous state-of-the-art in dynamic scene recognition by a significant margin.

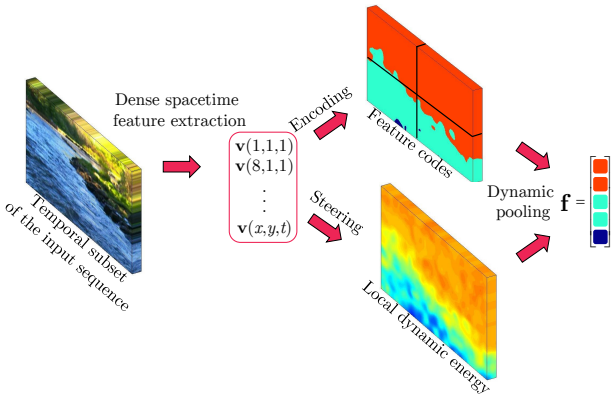

Figure 1. Proposed Representation for Dynamic Scene Recognition. First, spatiotemporal oriented primitive features are extracted from a temporal subset of the input video. Second, features are encoded into a mid-level representation learned for the task and also steered to extract dynamic pooling energies. Third, the encoded features are pooled via a novel dynamic spacetime pyramid that adapts to the temporal image structure, as guided by the pooling energies. The pooled encodings are concatenated into vectors that serve as the final representation for online recognition.

Considerable research has addressed the challenge of recognizing scenes from single images (*e.g.*, [8, 14, 15, 17]). In contrast, relatively little attention has been paid to dynamic scene recognition from image sequences, even though temporal information should support additional representational richness for visual scene classification. A possible reason for this limited research in dynamic scenes was the lack of a substantial database; however, this limitation has now been addressed, as two dynamic scene video databases have appeared [6, 22]. Even as it stands, however, the dynamic scene recognition literature can be reviewed within the tripartite framework of primitive feature extraction, coding and pooling.

Various primitive features have been investigated for dynamic scenes, including, flow vectors [16], linear dynamical systems [22], chaotic invariants [22], spatiotemporal orientations [6, 9] and slowly varying spatial orientations [24]. Here, systematic comparisons suggest that spatiotemporal

orientations provide particularly strong primitives for dynamic scene recognition [6, 9].

The coding step has been less investigated in dynamic scene recognition, with most recognition approaches based more directly on pooling of primitive features. An exception [24] is work that used extant approaches to dictionary building via unsupervised sampling [21] combined with subsequent soft assignment for final encoding [12]; however, this work lacked any systematic evaluation of alternative encoding strategies. Here, it is worth noting that classification performance can be greatly impacted by the choice of coding approach [4]; so, careful selection based on systematic study is appropriate.

The dominant approach to pooling throughout the dynamic scene recognition literature is that of spatial pyramids [6, 9, 22, 24], which employ predefined hierarchically gridded aggregation regions to capture coarse geometric layout in a multiresolution fashion. A limitation of this approach is that its predefined aggregation regions do not adapt to the dynamics of a video sequence, *e.g.*, as regions of interest move to different relative positions with the passage of time and when camera and scene motion are confounded. While recent advances in pooling are more flexible (*e.g.* [3, 10, 13]), they do not adapt dynamically to the time varying information that is present in a given dynamic scene and their utilty will thereby be limited.

The present paper makes the following three main contributions. 1) A novel dynamic pooling is presented that allows for encoded features to be aggregated adaptively as a function of scene dynamics. It is shown that this approach supports superior performance to alternatives when camera and scene motion are confounded without compromising performance when camera motion is absent. 2) The only systematic evaluation of feature coding techniques applied to dynamic scenes is documented. Here, spatiotemporal oriented features serve as primitives, owing to their strong performance in previous evaluations of dynamic scene recognition features. Recently, the Fisher vector representation [19] has shown state-of-the-art results for a variety of visual tasks (*e.g.*, [5, 14, 18, 23]); in contrast, here it is found that locality constrained linear coding [27] performs particularly well for dynamic scenes. 3) The selected feature, encoding and pooling approaches have been assembled into a complete system for dynamic scene recognition that has been evaluated on two standard datasets. The results show that the system outperforms the previous state-of-the-art by a 10% improvement in classification accuracy.

2. Technical approach

2.1. Primitive feature extraction

The underlying descriptor is based on spatiotemporal measurements that jointly capture spacetime image appearance, dynamics and colour information at multiple scales.

To extract the representation of spacetime orientation, the input volume is filtered with oriented 3D Gaussian third-derivative filters $G_{3D}^{(3)}(\theta_i, \sigma_j) = \kappa \frac{\partial^3}{\partial \theta_i^3} \exp\left(-\frac{x^2+y^2+t^2}{2\sigma_j^2}\right)$, with κ a normalization factor. The responses are pointwise squared and smoothed to yield oriented spacetime energy measurements

$$E(\mathbf{x}; \theta_i, \sigma_j) = G_{3D}(\sigma_j) * |G_{3D}^{(3)}(\theta_i, \sigma_j) * V(\mathbf{x})|^2, \quad (1)$$

where G_{3D} is a three-dimensional Gaussian, $\mathbf{x} = (x, y, t)^\top$ are spacetime coordinates, V is the grayscale spacetime volume formed by stacking all frames in a sequence along the temporal axis and $*$ denotes convolution. Convolution with G_{3D} serves to blur the filter responses, thereby ameliorating phase sensitivity and suppressing noise. Local smoothing is also appropriate, because the responses are directly used for subsequent encoding. This is in contrast to previous work using similar oriented filter responses for dynamic scene recognition which immediately aggregated filter responses over some support region (*e.g.*, [6, 9]).

Thus, for every spacetime location, $\mathbf{x}$, the local oriented energy $E(\mathbf{x}; \theta_i, \sigma_j)$ measures the power of local oriented structure along each considered orientation θ_i and scale σ_j. To uniformly sample the 3D spacetime domain along the minimal set of directions that span orientation for $G_{3D}^{(3)}$ [11], the filter directions are chosen along the vertices of a dodecahedron with antipodal directions identified to yield a set of ten θ_i. Figure 2 shows the spatiotemporal energies for the employed filter orientations on a sequence of a windmill farm. The energies collect dynamic information, see *e.g.*, the dominant energies in Fig. 2(h) capturing the movement of the rotating rotor blades, as well as spatial information, see *e.g.*, the energies in Fig. 2(d), reaching high values for spatial orientation structure on the ground of the scene.

The filter responses (1) are sensitive to image contrast. To achieve invariance to multiplicative contrast variation, the responses are normalized with respect to the sum of all filter responses at a point

$$\hat{E}(\mathbf{x}; \theta_i, \sigma_j) = \frac{E(\mathbf{x}; \theta_i, \sigma_j)}{\sum_{k=1}^{|\theta|} E(\mathbf{x}; \theta_k, \sigma_j) + \epsilon}, \quad (2)$$

where $|\theta| = 10$ denotes the number of orientations and the noise bias, ϵ, avoids numerical instabilities at low overall energies. To explicitly capture lack of oriented spacetime structure, another feature channel

$$\hat{E}_\epsilon(\mathbf{x}; \sigma_j) = \frac{\epsilon}{\sum_{k=1}^{|\theta|} E(\mathbf{x}; \theta_k, \sigma_j) + \epsilon}, \quad (3)$$

is added to the contrast-normalized filter responses of (2). Figure 2(l) shows $\hat{E}_\epsilon(\mathbf{x}; \sigma_j)$ for a windmill sequence, where large responses are seen in the unstructured sky region.

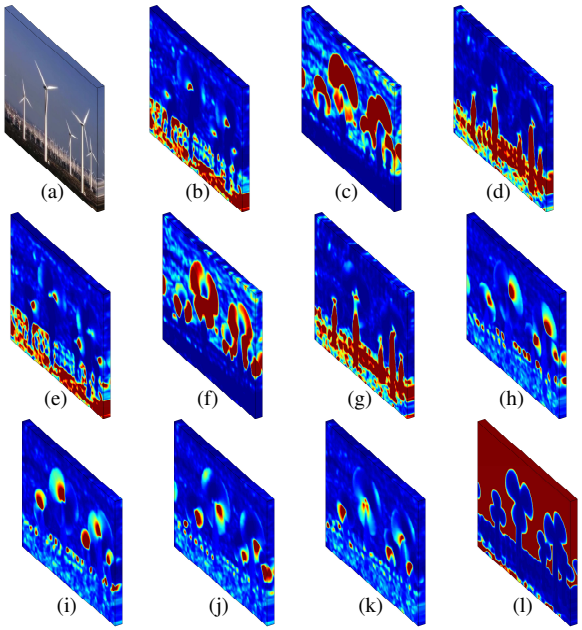

Figure 2. Distribution of spatiotemporal oriented energies for 16 frames of a sequence from the YUPENN dataset (a) [6]. In (b)-(k) the oriented energies, calculated by Gaussian derivative filtering, (1), is shown. (l) illustrates the no structure channel, (3). Hotter colours (*e.g.*, red) indicate larger filter responses.

Previous evaluations [6, 9, 22] showed that integrating colour cues is useful for dynamic scene categorization. Colour information is incorporated in the present spacetime primitives via the addition of three smoothed colour measurements,

$$C_m(\mathbf{x};\sigma_j) = G_{3D}(\sigma_j) * V_m(\mathbf{x}), \tag{4}$$

where m is one of the three CIE-LUV colour channels, *i.e.* $m \in \{L, U, V\}$ and all other notation is by analogy with the filtering formula (1).

Overall, each point, $\mathbf{x}$, in the spatiotemporal image volume, V, yields a locally defined, primitive feature vector, $\mathbf{v}(\mathbf{x})$, that is formed by concatenating the normalized, multiscale orientation measurements, (2), with the measures of unstructuredness, (3), and colour, (4). Notably, owing to the separability and steerabilty of the underlying filtering operations, these features can be extracted with modest computational expense.

2.2. Coding

A variety of different coding procedures exist to convert primitive local features, $\mathbf{v}(\mathbf{x}) \in \mathbb{R}^D$, into more effective intermediate-level representations, $\mathbf{c}(\mathbf{x}) \in \mathbb{R}^K$, for classification purposes and the choice can significantly impact performance [4]. In the present case, the primitive features are given in terms of the feature vector constructed in the previous subsection. To best mediate between these primitives and dynamic scene classification, a systematic empirical evaluation of a representative set of four contemporary coding techniques has been performed. In all cases, the encoding is with respect to an unsupervised trained codebook $\mathbf{B} \in \mathbb{R}^{D\times K}$.

As a baseline, vector quantization (VQ) is considered. In this case, each local feature $\mathbf{v}(\mathbf{x})$, is assigned the nearest codeword in $\mathbf{B}$, based on the minimum Euclidean distance in the D-dimensional feature space. The use of only a single codeword does not incorporate distances in the feature space, discards much descriptive information and is sensitive to noise [2, 27]. To improve on these limitations recent research has concentrated on retaining more information from the original features and representatives from these coding methods therefore are considered for incorporation into the proposed dynamic scene recognition system.

Largely two categories of improved approach have emerged. One category expresses features as combinations of (sparse) codewords (*e.g.*, [2, 3, 10, 27]). The other category considers differences between the original features and the codewords (*e.g.*, [19, 20, 30]). Correspondingly, two particularly strong performers from each of these two categories are considered [4]. Local linear coding (LLC) [27] is selected as the representative of the first category. In LLC, each local feature $\mathbf{v}(\mathbf{x})$ is encoded by $M \ll K$ nearest codewords in $\mathbf{B}$. Fisher vectors (FV) [19] are considered as a representative of the second category. An FV models mean and covariance gradients between a set of features $\{\mathbf{v}(\mathbf{x})\}$ and a generative model. The model is learned on training descriptors by using a Gaussian mixture model (GMM). Additionally, a recent improvement to Fisher vectors (IFV) is considered to see what gains result [20].

For VQ and LLC, the codebook entries are learned by quantizing the extracted descriptors from training sequences with K-means. In the case of Fisher vector coding, a GMM is fit to the training descriptors.

2.3. Dynamic pooling

When pooling the encoded features, $\mathbf{c}(\mathbf{x})$, from dynamic scenes, those that significantly change their spatial location across time should be pooled adaptively in a correspondingly dynamic fashion. For example, global image motion induced by a camera pan could cause the image features to move with time and pooling that is tied to finely specified image location will fail to capture this state of affairs. Similarly, when regions change their spatial relations with time, pooling should adapt. In such situations, a lack of appropriately dynamic pooling will degrade recognition performance, as features pooled at one location will have moved to a different location at a subsequent time and thereby be at risk of improper matching. Significantly, this challenge persists if the pooling positions are hierarchically arranged [15] or even more adaptively defined [3, 10, 13, 25, 26], but without explicit attention to temporal changes in pooling regions.

In contrast, features that retain their image positions over time (*i.e.*, static patterns) can be pooled within finer, predefined grids, *e.g.*, as with standard spatial pyramid matching (SPM) [15]. Indeed, even highly dynamic features that retain their overall spatial position across time (*i.e.*, temporally stochastic patterns, such as fluttering leaves on a bush and other dynamic textures) can be pooled with fine grids. Thus, it is not simply the presence of image dynamics that should relax finely gridded pooling, but rather the presence of larger scale coherent motion (*e.g.*, as encountered with global camera motion).

2.3.1 Dynamic pooling energies

In response to the above observations, a set of dynamic energies have been derived that favour orderless pooling (*e.g.*, global aggregation) when coarse scale image motion dominates and spatial pooling (as in an SPM scheme) when a visual word is static or its motion is stochastic but otherwise not changing in overall spatial position. These energies are used as pooling weights applied to the locally encoded features so that they can be pooled in an appropriate fashion.

The energies are based on the primitive feature measurements (1) and therefore provide an intuitive and efficient strategy to enhance pooling. To capture the desired coarse scale spatiotemporal information the 3D Gaussian third derivative responses are aggregated as

$$E^{\mathcal{R}}(\mathbf{x};\theta_i,\sigma_j) = \sum_{\mathbf{x}\in\mathcal{R}} |G^{(3)}_{3D}(\theta_i,\sigma_j) * V(\mathbf{x})|^2, \qquad (5)$$

where $\mathcal{R}$ is a rectangular spacetime region defined by $\{\mathcal{R}_x, \mathcal{R}_y, \mathcal{R}_t\}$ and centred at $\mathbf{x}$. A rectangular aggregation region is used to be consistent with the ultimate pooling grids. Next, since interest is in capturing image dynamics, irrespective of spatial orientation, the spacetime energies in equation (5) are steered and then combined across all orientations consistent with a single spacetime orientation (*e.g.*, motion direction), as specified by the unit normal, $\hat{\mathbf{n}}$, corresponding to its frequency domain plane [28]. To span orientation space in a plane, 4 directions are needed for a Gaussian 3$^{\text{rd}}$ derivative [11]. So, dynamic energies for direction $\hat{\mathbf{n}}$ are given by

$$E^{\mathcal{D}}(\mathbf{x};\hat{\mathbf{n}},\sigma_j) = \sum_{i=1}^{4} E^{\mathcal{R}}(\mathbf{x};\hat{\theta}_i,\sigma_j), \qquad (6)$$

with $\hat{\theta}_i$ denoting equally spaced orientations consistent with $\hat{\mathbf{n}}$; for details see [7]. In the present context, directions are considered corresponding to motion along the leftward, rightward, upward, downward and four diagonal directions as well as static (zero velocity), which are denoted in the following as $l, r, u, d, ru, rd, lu, ld$ and s, respectively.

The directional spacetime energies, (6), are not sufficient for distinguishing between so called coherent motion (*e.g.*, as exemplified by large scale motion resulting from camera movement) and incoherent motion (*e.g.*, as exemplified by stochastic dynamic textures) [1, 29]. The desired pooling energies are meant to capture coherent motion and that can be accomplished by combing the directional energies in opponent-motion channels as follows

$$\begin{aligned}
E^{\mathcal{D}}_{|r-l|}(\mathbf{x};\sigma_j) &= |E^{\mathcal{D}}(\mathbf{x};\hat{\mathbf{n}}_r,\sigma_j) - E^{\mathcal{D}}(\mathbf{x};\hat{\mathbf{n}}_l,\sigma_j)| \\
E^{\mathcal{D}}_{|u-d|}(\mathbf{x};\sigma_j) &= |E^{\mathcal{D}}(\mathbf{x};\hat{\mathbf{n}}_u,\sigma_j) - E^{\mathcal{D}}(\mathbf{x};\hat{\mathbf{n}}_d,\sigma_j)| \\
E^{\mathcal{D}}_{|ru-ld|}(\mathbf{x};\sigma_j) &= |E^{\mathcal{D}}(\mathbf{x};\hat{\mathbf{n}}_{ru},\sigma_j) - E^{\mathcal{D}}(\mathbf{x};\hat{\mathbf{n}}_{ld},\sigma_j)| \\
E^{\mathcal{D}}_{|lu-rd|}(\mathbf{x};\sigma_j) &= |E^{\mathcal{D}}(\mathbf{x};\hat{\mathbf{n}}_{lu},\sigma_j) - E^{\mathcal{D}}(\mathbf{x};\hat{\mathbf{n}}_{rd},\sigma_j)|
\end{aligned} \qquad (7)$$

to yield a set of dynamic energies representing coherent image motion in 4 equally spaced directions (horizontal ($r-l$), vertical ($u-d$) and two diagonals ($ru-ld$ and $lu-rd$)). In contrast to the individual motion direction consistent energy samples from (6), the opponent motion channels explicitly capture coherent motion across various directions. For example, a spatial region with a stochastically moving spacetime pattern, *e.g.* the leaves of a tree in the wind can exhibit large motions in several specific directions $\hat{\mathbf{n}}$; however, after taking the absolute arithmetic difference from opponent directions, the coherent motions, (7), of such stochastic spacetime texture patterns are approximately zero. On the other hand, regions that are dominated by a single direction of motion (*i.e.* coherent motion regions) will yield a large response in the most closely matched channel.

The coherent motion energies are ℓ_1 normalized together with the static energy channel that indicates lack of coarse motion

$$\hat{E}^{\mathcal{D}}_{\Lambda_k}(\mathbf{x};\sigma_j) = \frac{E^{\mathcal{D}}_{\Lambda_k}(\mathbf{x};\sigma_j)}{\sum_{i\in\Lambda} E^{\mathcal{D}}_{\Lambda_i}(\mathbf{x};\sigma_j) + \epsilon}, \forall k \in \Lambda, \qquad (8)$$

to yield a point-wise distribution of static, coherent, as well as unstructured energy via the normalized ϵ indicating homogeneous regions

$$\begin{aligned}
\hat{E}^{\mathcal{D}}_{\epsilon}(\mathbf{x};\sigma_j) &= \frac{\epsilon}{\sum_{i\in\Lambda} E^{\mathcal{D}}_{\Lambda_i}(\mathbf{x};\sigma_j) + \epsilon}, \\
\text{with } \Lambda &= \{s, |r-l|, |u-d|, |ru-ld|, |lu-rd|\}.
\end{aligned} \qquad (9)$$

Next, since regions without coherent motion or with only fine scale motion (indicated by $\hat{E}^{\mathcal{D}}_s$), as well as homogeneous regions (indicated by $\hat{E}^{\mathcal{D}}_\epsilon$), can be similarly pooled with spatial gridding to capture geometric layout, static energy is summed with unstructured energy as

$$\hat{E}^{\mathcal{D}}_{s+\epsilon}(\mathbf{x};\sigma_j) = \hat{E}^{\mathcal{D}}_s(\mathbf{x};\sigma_j) + \hat{E}^{\mathcal{D}}_\epsilon(\mathbf{x};\sigma_j), \qquad (10)$$

to yield the final set of (coherent) motion directions $\Lambda = \{s+\epsilon, |r-l|, |u-d|, |ru-ld|, |lu-rd|\}$.

Finally, to capture dynamic information across a range of scales, the dynamic pooling energies are extracted with multiple scales, σ_j, and ultimately collapsed across scale as

$$\tilde{E}^{\mathcal{D}}_{\Lambda_k}(\mathbf{x}) = \frac{1}{|\sigma|}\sum_{j=1}^{|\sigma|}\hat{E}^{\mathcal{D}}_{\Lambda_k}(\mathbf{x};\sigma_j),\ \forall k \in \Lambda, \qquad (11)$$

where $|\sigma|$ denotes the number of scales.

The dynamic pooling energies for a temporal subset of a street sequence are shown in Fig. 3. For proper illustration, the temporal support of the largest $G^{(3)}_{3D}$ filter is depicted in Figures 3(a)-3(c). Figure 3(d) depicts the central frame of the filtered sequence and 3(e)-3(i) show the decomposition of the filtered sequence into a distribution of static and coherent motion dynamic energies. Observe that the static+unstructured channel consists of large responses for stationary image structures, *e.g.*, the buildings in the scene, as well as for homogeneous regions such as the sky in the centre of the scene. Whereas the foreground red car's dynamic energy can be decomposed into several coherent motion channels with a large part originating from the horizontal motion channel, *i.e.*, $\tilde{E}^{\mathcal{D}}_{|r-l|}(\mathbf{x})$, shown in Figure 3(f). Note that fine-scale motions, such as the moving cars in the background, are not captured by the coherent motion channels (Fig. 3(f)-3(i)) and therefore exhibit strong responses in the static channel 3(e), which is appropriate as they form (part of) the background dynamic texture.

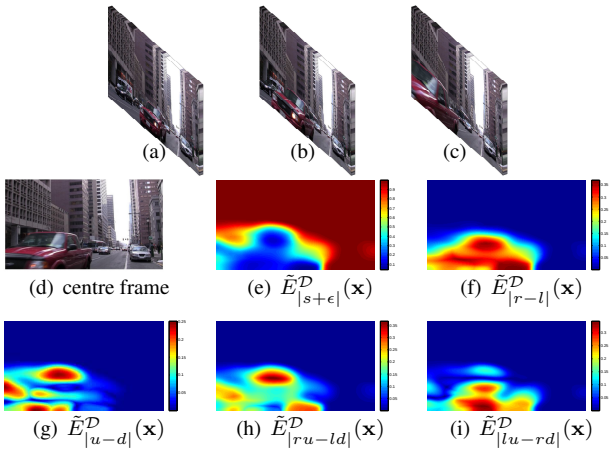

Figure 3. Distribution of Spatiotemporal Oriented Pooling Energies of a Street Sequence from the YUPENN Dataset. (a), (b), (c), and (d) show the first 8, middle 8, last 9, and centre frame of the filter support region. (e)-(i) show the decomposition of the sequence into a distribution of spacetime energies indicating stationarity/homogeneity in (e), and coarse coherent motion for several directions in (f)-(i). Hotter colours (*e.g.*, red) correspond to larger filter responses.

2.3.2 Dynamic spacetime pyramid

The dynamic pooling energies, (11), are used to control construction of a spacetime pyramid that explicitly captures spatial and temporal information as described in the following four steps. First, to keep weak geometry information of the pooled encodings, a traditional three-level spatial pyramid of size $\{(2^l \times 2^l \times 1^l)\}_{l=0}^{2}$ is constructed for each sample point in time within a temporal sliding window, resulting in $M = 21$ regions $\{\mathbf{R}_m\}_{m=1}^{M}$. Second, for pooling at the coarsest level $l = 0$, *i.e.*, in the region without geometric grid, $\mathbf{R}_1$, coded features are globally aggregated. Third, in regions with geometric grids, *i.e.*, $l > 0$ and $\{\mathbf{R}_m | m > 1\}$, the static pooling energies $\tilde{E}^{\mathcal{D}}_{s+\epsilon}$ are used as geometric coefficients, emphasizing the local contribution of each visual word. Fourth, to explicitly pool features favourably from regions with coherent motion, four more channels $\Lambda = \{|r-l|, |u-d|, |ru-ld|, |lu-rd|\}$ are added. Due to the coarse-scale motion of these features, the top pyramid level $l = 0$ is used. Therefore, the final spatiotemporal pyramid encodes a sample point in time in $M + 4 = 25$ channels, with each channel capturing specific spatial and temporal properties of the pooled codewords.

To generate a signature of the spatiotemporal feature codes $\mathbf{c}(\mathbf{x})$, extracted at the same point in time, t, max-pooling is chosen, as it finds the most salient response in a region $\mathbf{R}_m$ and is known to provide robustness to outliers in the collection of encodings to be pooled [2, 21, 27]. The developed dynamic max-pooling operation finds the location $\mathbf{x}_m^{(k)}$ of the codewords that represent the m^{th} channel in the dynamic spacetime pyramid by

$$\mathbf{x}_m^{(k)} = \begin{cases} \arg\max_{\mathbf{x}\in\mathbf{R}_1} \mathbf{c}^{(k)}(\mathbf{x}) & m = 1 \\ \arg\max_{\mathbf{x}\in\mathbf{R}_m} \tilde{E}^{\mathcal{D}}_{|s+\epsilon|}(\mathbf{x})\mathbf{c}^{(k)}(\mathbf{x}) & 2 \le m \le M \\ \arg\max_{\mathbf{x}\in\mathbf{R}_1} \tilde{E}^{\mathcal{D}}_{\Lambda_i}(\mathbf{x})\mathbf{c}^{(k)}(\mathbf{x}) & M+1 \le m \le M+4, \end{cases} \qquad (12)$$

to pool the k^{th} visual codeword by $\mathbf{f}_m^{(k)} = \mathbf{c}^{(k)}(\mathbf{x}_m^{(k)})$, and yield the final description of $\mathbf{R}_m$ by $\mathbf{f}_m \in \mathbb{R}^K$. In (12), the static coefficients, $\tilde{E}^{\mathcal{D}}_{|s+\epsilon|}$, assign higher weights to all features pooled from regions with a spatial grid. Note that these static coefficients $\tilde{E}^{\mathcal{D}}_{|s+\epsilon|}$ are ℓ_1-normalized together with the dynamic energies that indicate coarse scale coherent motion. Therefore, dynamic features with coarse scale motion characteristics are given low weights when pooling in spatial grids of the spacetime pyramid. To explicitly model the visual words with coarse scale dynamics, equation (12) pools features with specific directions. For example, visual words on horizontally moving objects are pooled with high corresponding weights $\tilde{E}^{\mathcal{D}}_{|r-l|}$ to explicitly capture horizontally moving image structures in the dynamic spacetime pyramid.

A global feature vector, $\mathbf{f}$, representing a point in time of the video, is concatenated by stacking the descriptors $\mathbf{f}_m$ of all channels. These feature vectors can then serve as the basis of an on-line recognition scheme when coupled with an appropriate classifier, e.g, a support vector machine (SVM).

2.4. Implementation summary

1. **Sliding window local feature extraction.** The video is processed in a temporal sliding window by dense extraction of normalized oriented spacetime energies, (2), and colour distributions, (4), with the $|\theta| = 10$ filter orientations, one unstructured channel ($\hat{E}_\epsilon$), and three LUV colour channels. All measurements are taken at two scales, $\sigma = \{1, 2\}$ with local filter support of $(x, y, t)^\top \in \{(13, 13, 13)^\top, (25, 25, 25)^\top\}$. The resulting multiscale spacetime orientation features, $\mathbf{v}(\mathbf{x})$, of dimension $D = (|\theta| + 1 + 3) \times |\sigma| = 28$ are extracted densely over a spatiotemporal grid by varying $\mathbf{x}$ in spatial steps of 8 pixels and temporal steps of 16 frames.
2. **Codebook generation.** For VQ and LLC, the codebook entries are learned by quantizing the extracted descriptors from the training sequences with K-means. To maintain low computational complexity, a random subset of features from the training set, consisting of a maximum of 5000 descriptors from each training sequence, are used to learn a visual vocabulary of size $K = 200$ codewords. An approximated nearest neighbour search is used for efficient clustering and encoding. In the case of Fisher vector coding, a GMM with $K_{\text{GMM}} = 50$ mixtures is fitted to the subsampled training descriptors.
3. **Feature coding.** The local spacetime descriptors are encoded either via VQ, LLC, FV, or IFV. The parameters in LLC are set to the default values from the original publication [4, 27]; *i.e.*, the considered neighbouring visual words are set to $M = 5$ and the projection parameter to 10^{-4}.
4. **Feature pooling.** To compare against conventional pooling, an $l = 3$ level SPM is used to maintain weak spatial information of the features extracted in each temporal instance. The resulting 21 pooling regions from spatial grids of size $2^l \times 2^l$ create a $21 \times K = 21 \times 200 = 4200$ dimensional feature vector for VQ and LLC encoding and a $21 \times 2 \times K_{\text{GMM}} \times D = 21 \times 2 \times 50 \times 28 = 58800$ dimensional feature vector for the higher order Fisher encoding. As in the original publications, pooling is performed by taking the average (VQ) or maximum (LLC) of the encoded features. For the proposed dynamic pooling, let V_w and V_h denote the width and height of the spacetime volume in the filtering process (5), then the integration region is set to $\mathcal{R}_x = \frac{V_w}{4}$, $\mathcal{R}_y = \frac{V_h}{4}$, and $\mathcal{R}_t$ is set to the temporal support of the largest filter used in (5). A $25 \times K = 5000$ dimensional feature vector is generated by the dynamic spacetime pyramid that also uses a hierarchical 3-level pyramid with the finest grid size of 4×4 for embedding geometry in 20 of the 25 channels.
5. **Learning and Classification.** Each set of encoded features pooled from the same temporal instance generates a feature vector, $\mathbf{f}$. For training, all feature vectors extracted from the training set are used to train a one-vs-rest SVM classifier. The histogram intersection kernel [15] is used for VQ, while a linear SVM is applied for Fisher and LLC coded features. ℓ_2 normalization is applied to the feature vectors used in the linear SVM. The SVM's regularization loss trade-off parameter C is set after cross validation on the training data. During classification, each feature vector of a test video is classified by the one-vs-rest SVM to yield a temporal prediction. All temporal predictions are subsequently combined to yield an overall classification of the video by the majority of the temporal class predictions.

3. Experimental evaluation

The proposed Bags of Spacetime Energies (BoSE) system is evaluated on the Maryland [22] and YUPENN [6] dynamic scene recognition datasets. A leave-one-video-out experiment is used for consistency with previous evaluations in [6, 9, 22, 24]. The structure of the experiments is three-layered. First (Section 3.1), the best encoding method, in the context of dynamic scene understanding, is sought for the proposed spacetime orientation and colour features of Section 2.1. This evaluation includes feature encoding methods that are based on either local codeword statistics (VQ and LLC), or the difference between the codewords and features to encode (FV and IFV). Second, an evaluation of the novel dynamic pooling framework is given in Section 3.2, where, based on the evaluation, LLC encoded features are combined with the proposed dynamic max pooling approach of Section 2.3. Finally, in Section 3.3 the full proposed BoSE system is compared with the state-of-the-art in dynamic scene classification.

3.1. Comparison of feature coding methods

Maryland				YUPENN			
VQ	LLC	FV	IFV	VQ	LLC	FV	IFV
65.38	**69.23**	63.85	66.92	94.52	95.48	91.43	**96.19**

Table 1. Average recognition accuracy with different encoding methods. LLC outperforms IFV in the presence of large intra-class variabilities and temporal variations within the videos (Maryland).

In Table 1, the overall classification performance, averaged over all classes, for the four investigated coding approaches is shown. On both datasets very competitive performance is achieved by the LLC encoding. Especially on the Maryland dataset the higher-order Fisher vector encodings are outperformed by LLC. This is interesting, given that for (static) image classification tasks Fisher vectors generally have been found to provide superior performance to sparse encodings such as LLC [4, 14]. Since most of the

Pooling method	Maryland	YUPENN
max-pooling	69.23	95.48
dynamic max-pooling	**77.69**	**96.19**

Table 2. Dynamic scene recognition accuracy with LLC encoded features for different pooling methods. The proposed dynamic max-pooling allows best performance on data with a high degree of coarse scale motion (Maryland), as well as on dynamic scene sequences captured from static cameras (YUPENN).

videos in the Maryland dataset show significant temporal variation, these results suggest that for highly dynamic data LLC is able to outperform IFV. Based on this outcome, the remainder of the evaluation makes use of LLC encoding.

3.2. Dynamic energy pooling

This section evaluates the performance of LLC encoded spacetime features that are pooled either via conventional max-pooling or via the proposed dynamic max-pooling. In Table 2 the overall classification rate (in %) for variations in the pooling method is reported. The novel dynamic max-pooling leads to best performance on both datasets. Conventional max-pooling is outperformed by a margin of 8.46% and 0.71% for Maryland and YUPENN, resp.

The significant performance gain associated with dynamic max-pooling on the Maryland dataset can be attributed to the severe camera movement that is present in this dataset. Since camera movement generally manifests itself at coarse temporal scales and the proposed dynamic pooling method favours pooling without geometric context within the dynamic pyramid when coarse (coherent) motion is present, it avoids inappropriate spatially gridded pooling when image structure drastically changes its position with time. The approach thereby becomes robust to camera (and other coarse) motions. Interestingly, further investigation showed that performance drops to 66.15% (−3.08%) when the spatial pyramid is not employed at all.

Consideration of the YUPENN results shows that this advantage is had without compromising performance when camera motion is absent: Here, the dynamic pooling allows aggregation at finer levels of the dynamic pyramid to more precisely localize the spatiotemporal image structure. Interestingly, there is even a slight improvement on YUPENN under dynamic max-pooling, which may be due to the fact that coherently moving objects are specifically matched by the dynamic pooling channels in (12). For example, vertically moving visual words from a waterfall sequence will be explicitly matched, since these are favourably pooled within the $\hat{E}^{\mathcal{R}}_{|u-d|}$ channel of the dynamic spacetime pyramid.

3.3. Comparison with the state-of-the-art

The proposed approach is compared to several others that previously have shown best performance: GIST [17] + histograms of flow (HOF) [16], GIST + chaotic dynamic features (Chaos) [22], spatiotemporal oriented energies (SOE) [6], slow feature analysis (SFA) [24], and complementary spacetime orientation (CSO) features [9].

Class	HOF+ GIST	Chaos+ GIST	SOE	SFA	CSO	BoSE
Avalanche	20	60	40	60	60	60
Boiling Water	50	60	50	70	80	70
Chaotic Traffic	30	70	60	80	90	90
Forest Fire	50	60	10	10	80	90
Fountain	20	60	50	50	80	70
Iceberg Collapse	20	50	40	60	60	60
Landslide	20	30	20	60	30	60
Smooth Traffic	30	50	30	50	50	70
Tornado	40	80	70	70	80	90
Volcanic Eruption	20	70	10	80	70	80
Waterfall	20	40	60	50	50	100
Waves	80	80	50	60	80	90
Whirlpool	30	50	70	80	70	80
Overall	33.08	58.46	43.08	60.00	67.69	**77.69**

Table 3. Classification accuracy for different representations on the Maryland dataset.

Class	HOF+ GIST	Chaos+ GIST	SOE	SFA	CSO	BoSE
Beach	87	30	93	93	100	100
Elevator	87	47	100	97	100	97
Forest Fire	63	17	67	70	83	93
Fountain	43	3	43	57	47	87
Highway	47	23	70	93	73	100
Lightning Storm	63	37	77	87	93	97
Ocean	97	43	100	100	90	100
Railway	83	7	80	93	93	100
Rushing River	77	10	93	87	97	97
Sky-Clouds	87	47	83	93	100	97
Snowing	47	10	87	70	57	97
Street	77	17	90	97	97	100
Waterfall	47	10	63	73	77	83
Windmill Farm	53	17	83	87	93	100
Overall	68.33	22.86	80.71	85.48	85.95	**96.19**

Table 4. Recognition rates for the best performing approaches on the YUPENN dataset.

Tables 3 (Maryland dataset) and 4 (YUPENN dataset) compare the performance of the final Bags of Spacetime Energies (BoSE) system with the previous state-of-the-art. Here, the BoSE system consists of densely extracted local oriented spacetime energies (5) and colour distributions (4) that are encoded by LLC and pooled via the proposed dynmax-pooling, parameter choices as given in Section 2.4. Note that the reported performance of SFA differs from that given in the original paper [24]. The results presented here are the correct ones; see error report and correct recogition rates (reproduced here) at the SFA website [1].

For both datasets, BoSE performs considerably better than the previous state-of-the-art, CSO [9], with an improvement of 10% or better on both datasets. On the Maryland dataset, the novel BoSE representation achieves an exceptional average accuracy of 78% when coupled with a simple linear SVM classifier. When comparing to other approaches, one striking result is the 100% recognition accuracy for the Waterfall class. The proposed BoSE approach's

[1] http://webia.lip6.fr/~theriaultc/sfa.html

96% accuracy on YUPENN suggests that performance is saturated on this dataset. One remarkable result on this dataset is the 87% recognition rate for the Fountain class, which exhibits huge intra-class variations in the background and only a small amount of common foreground (*i.e.*, the fountain itself). Overall, BoSE is able to best represent the classes, by modelling the visual words with locally encoded spacetime energies that are pooled based on their dynamics.

4. Conclusion

This paper has proposed a generic BoW framework for dynamic scene recognition. Local features are extracted densely in a temporal sliding window via application of multiscale, multiorientation filters followed by smoothing to yield energy responses, as well as multiscale colour cues. Based on an evaluation of several popular feature coding methods, the local spacetime energies are projected into a mid-level representation by using a learned visual vocabulary. Finally, a novel spatiotemporal pooling strategy has been introduced that aggregates the encoded features in a spacetime pyramid representation, based on their dynamics.

The performance of the proposed framework has been verified in rigorous evaluations, where it has been shown that a carefully designed BoW model significantly outperforms the state-of-the-art. A key factor for the success of the system is the novel dynamic spacetime pyramid, which greatly increases performance when camera motion is present, but does not compromise performance when camera motion is absent.

The insights of this paper should have a substantial impact on the design of dynamic scene classification approaches, as they significantly extend the state-of-the-art. More generally, the outstanding performance of the presented spacetime recognition framework suggests application to a variety of other areas, such as event retrieval, video indexing, or object and activity localization.

References

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

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

[3] L. Cao, R. Ji, Y. Gao, Y. Yang, and Q. Tian. Weakly supervised sparse coding with geometric consistency pooling. In *CVPR*, 2012.

[4] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.

[5] R. G. Cinbis, J. Verbeek, and C. Schmid. Segmentation driven object detection with Fisher vectors. In *ICCV*, 2013.

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

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

[8] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

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

[10] J. Feng, B. Ni, Q. Tian, and S. Yan. Geometric ℓ_p-norm feature pooling for image classification. In *CVPR*, 2011.

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

[12] H. Goh, N. Thome, M. Cord, and J. Lim. Unsupervised and supervised visual codes with restricted Boltzmann machines. In *ECCV*, 2012.

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

[14] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013.

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

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

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

[18] D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with Fisher vectors on a compact feature set. In *ICCV*, 2013.

[19] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[20] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, 2010.

[21] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *PAMI*, 29(3):411–426, 2007.

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

[23] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In *BMVC*, 2013.

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

[25] M. M. Ullah, S. N. Parizi, and I. Laptev. Improving bag-of-features action recognition with non-local cues. In *BMVC*, 2010.

[26] J. C. van Gemert. Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In *ICMR*, 2011.

[27] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

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

[29] R. Wildes and J. Bergen. Qualitative spatiotemporal analysis using an oriented energy representation. In *ECCV*, 2000.

[30] X. Zhou, K. Yu, T. Zhang, and T. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010.