Bags of Spacetime Energies for Dynamic Scene Recognition

Christoph Feichtenhofer1,2  Axel Pinz1  Richard P. Wildes2
1Institute of Electrical Measurement and Measurement Signal Processing, TU Graz, Austria
2Department of Electrical Engineering and Computer Science, York University, Toronto, Canada
{feichtenhofer, axel.pinz}@tugraz.at  wildes@cse.yorku.ca

Abstract

This work 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. 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, publicly available dynamic scene datasets. The results show that the proposed approach outperforms the previous state-of-the-art in classification accuracy by 10%.

1. Introduction

Use of temporal image sequences (e.g., video) to support scene recognition has potential to improve performance over single images owing to the additional information available, e.g., regarding scene dynamics. A key challenge in exploiting such information is disentangling intrinsic scene dynamics from camera motion incurred during capture. 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.

2. Bags of Spacetime Energies

Primitive feature extraction. The underlying descriptor is based on spatiotemporal orientation measurements that jointly capture spacetime image appearance, dynamics and colour information at multiple scales [1]. To extract the representation, input imagery is filtered with oriented 3D Gaussian third-derivatives \( G_{3D}(\theta_i, \sigma_j) \). The responses are pointwise squared and blurred to yield oriented spacetime energy measurements

\[
E(\mathbf{x}; \theta_i, \sigma_j) = G_{3D}(\sigma_j) \ast |G_{3D}^{(3)}(\theta_i, \sigma_j) \ast V(\mathbf{x})|^2,
\]

where \( G_{3D} \) is a 3D Gaussian, \( \mathbf{x} = (x, y, t)^T \) are spacetime coordinates, \( V \) is the grayscale spacetime volume formed by stacking all frames in a sequence along the temporal axis and \( \ast \) denotes convolution. To achieve invariance to multiplicative contrast variation, the responses, (1), are normalized with respect to the sum of all filter responses at a point. Finally, colour information is incorporated in the present spacetime primitives via the addition of three smoothed LUV colour measurements.

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. 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 primi-
tives and dynamic scene classification, a systematic empirical evaluation of a representative set of four contemporary coding techniques has been performed: Vector quantization (VQ), locality-constrained linear coding (LLC), Fisher vectors (FV) and its improved version with power- and $\ell_2$-normalization (IFV).

**Dynamic pooling.** When pooling the encoded features, $c(x)$, from dynamic scenes, those that significantly change their spatial location across time should be pooled adaptively in a correspondingly dynamic fashion. 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). 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).

In response to the above observations, a set of dynamic pooling energies, $E^D$, have been derived that favour orderless pooling (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.

In the present context, a set of dynamic energies representing coherent image motion in 4 equally spaced directions (horizontal $(r - l)$, vertical $(u - d)$ and two diagonals $(r u - l d$ and $l u - r d)$, as well as a static channel $(s + e)$ that indicates lack of coarse motion are employed. The dynamic pooling energies for a temporal subset of a street sequence are shown in Figure 2.

3. Experimental evaluation

The proposed Bags of Spacetime Energies (BoSE) system is evaluated on the Maryland [4] and YUPENN [1] dynamic scene recognition datasets. In Table 1 the full proposed BoSE system is compared to several others that previously have shown best performance: GIST + histograms of flow (HOF), GIST + chaotic dynamic features (Chaos) [4], spatiotemporal oriented energies (SOE) [1], slow feature analysis (SFA) [5] and complementary spacetime orientation (CSO) features [3]. For both datasets, BoSE performs considerably better than the previous state-of-the-art, CSO [3], with an improvement of 10% or better on both datasets. BoSE is able to best represent the videos, by mod-

![Figure 2. 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.](image)

<table>
<thead>
<tr>
<th>Approach</th>
<th>HOF+GIST</th>
<th>Chaos+GIST</th>
<th>SOE</th>
<th>SFA</th>
<th>CSO</th>
<th>BoSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maryland</td>
<td>33.08</td>
<td>58.46</td>
<td>43.08</td>
<td>60.00</td>
<td>67.69</td>
<td>77.69</td>
</tr>
<tr>
<td>YUPENN</td>
<td>68.33</td>
<td>22.86</td>
<td>80.71</td>
<td>85.48</td>
<td>85.95</td>
<td>96.19</td>
</tr>
</tbody>
</table>

Table 1. Classification accuracy for different representations.

4. Conclusion

The insights of this work have an 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 video retrieval or indexing. Details of the approach are documented in the full version of the paper [2] at [http://www.cse.yorku.ca/vision/publications/FeichtenhoferPinzWildesCVPR2014.pdf](http://www.cse.yorku.ca/vision/publications/FeichtenhoferPinzWildesCVPR2014.pdf).

References


---

1See corrected results at [http://webia.lip6.fr/theriault/sfa.html](http://webia.lip6.fr/theriault/sfa.html)