Spacetime Forests with Complementary Features for Dynamic Scene Recognition

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

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

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

This paper presents spacetime forests defined over complementary spatial and temporal features for recognition of naturally occurring dynamic scenes. The present work makes three main contributions. (i) A novel descriptor is proposed that integrates complementary information from separate spatial and temporal orientation measurements for aggregation in spacetime pyramids. (ii) A random forest classifier is applied for the first time to dynamic scenes. This spacetime forest allows for automatic determination of the most discriminative features to separate the classes based on appearance and dynamics with computational efficiency. (iii) Video is processed in incremental temporal slices with scale matched preferentially to scene dynamics (in comparison to camera motion). This strategy allows for temporal alignment to be treated as latent in the classifier, efficient processing and robustness to large temporal variation across time (*e.g.* from camera motion), even while capturing intrinsic scene characteristics. Previous dynamic scene research has suffered in the presence of camera motion [1] and has provided little consideration of on-line processing concerns. Figure 1 overviews the framework.

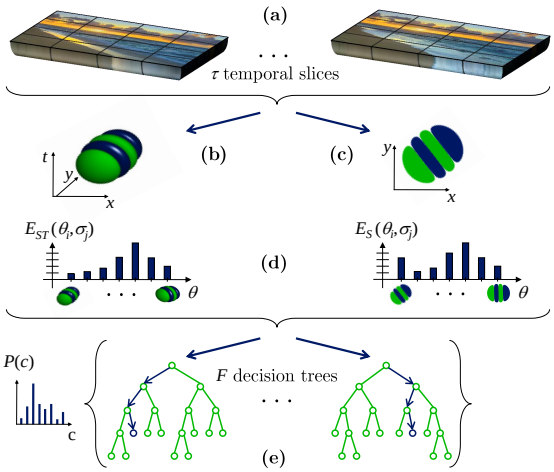

Figure 1: The proposed dynamic scene classification framework. (a) τ temporal slices are created to process the frames of a video in a sliding window approach, divided in spatiotemporal pyramid cuboids. (b,c) The cuboids are filtered by banks of multiscale, σ, oriented filters along equally separated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

Complementary spacetime features. Spatial appearance and temporal dynamics information are extracted via applications of multiscale filter banks that are further tuned for spatial or spatiotemporal orientation. In the spatial domain, 2D Gaussian third derivative filters are applied (Fig. 1(c)) to yield a set of multiscale, multiorientation measurements. Similarly, dynamic information is extracted via application of 3D Gaussian third derivative filters (Fig. 1(b)). To yield measures of dynamic information independent of spatial appearance, the initial spatiotemporal orientation measurements are summed across all orientations consistent with a single frequency domain plane, where motion occurs as a plane through the origin. Previous spacetime filtering approaches to dynamic scene recognition tend to exhibit decreased performance when dealing with scenes captured with camera motion, in comparison to scenes captured with stationary cameras. A likely explanation for this result is that the approaches have difficulty in disentangling image dynamics that are due to camera motion vs. those that are intrinsic to the scenes. Here, it is interesting to note that camera motion often unfolds at coarser time scales (*e.g.*, extended pans and zooms) in comparison to intrinsic scene dynamics (*e.g.*, spacetime textures of water, vegetation, *etc.*); however, previous approaches have made their measurements using relatively coarse temporal scales and thereby failed to exploit this difference. In the present approach this difference in time scale is captured by making use of only fine scales during spatiotemporal filtering, so that they are preferentially matched to scene, as opposed to camera, dynamics. Note, however, that spatial filtering is performed at four coarse scales varying by octave. Furthermore, CIE-LUV colour channels are included to capture complementary chromatic information.

For the purpose of classification of the entire video into a scene class, the local descriptors are summed across time, t, within τ discrete units of equal duration to yield a set of temporally aggregated images, which are referred to as temporal slices. Temporal slicing is motivated by the desire for incremental processing that can allow for efficient, on-line operation. Use of short-term parcelling of the measurements also is well matched with the restriction to use of fine temporal scales during spatiotemporal filtering to favour scene over camera dynamics. Each temporal slice is hierarchically aggregated into histograms to form a spatiotemporal pyramid, analogous to that used previously in dynamic [1] scene analysis. At each level of the pyramid, each temporal slice is broken into 3D cuboids (see Fig. 1(a)), with filter measurements collapsed into histograms within each cuboid, as illustrated in Fig. 1(d). The support of the cuboid at any given level of the pyramid corresponds to the aggregation region in the filtering process. Moreover, the adjacency structure of the cuboids capture the overall scene layout. For each cuboid, the histograms are $L1$-normalized, to represent a distribution of chromatic, multiscale oriented spacetime energy as well as lack of oriented structure. The histograms for all cuboids are concatenated into a final feature vector that comprises the Complementary Spacetime Orientation descriptor (CSO) to characterize a temporal slice of the video.

Spacetime forests. A Random Forest (RF) classifier consisting of F decision trees is used for classification, based on the leaf node at which the corresponding CSO descriptor arrives (see Fig. 1(e)). During training, the temporal alignment of the video slices is treated as latent; correspondingly, each temporal slice of each training video generates its own feature vector. This approach allows leveraging of the high temporal diversity in the spatiotemporal patterns. As some classes may be better represented by specific feature types, the node optimization process in each tree is restricted to a single feature type. The input for the RF is first structured into the three complementary feature channels of the CSO descriptor. Then, the channels are used to train $\frac{F}{3}$ trees each, to best distinguish the classes, based on a particular channel only. Lastly, these complementary trees are merged, to obtain the spacetime forest. During classification, each temporal slice initially is classified individually, with the individual classifications subsequently combined to yield a final classification across all temporal slices available up to a given time.

The approach has been evaluated on two publically available datasets [1, 2]; results show that it achieves a new state-of-the-art in dynamic scene recognition.

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

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