Learning Similarity Metrics for Dynamic Scene Segmentation

Damien Teney[1], Matthew Brown[2], Dimitry Kit[2], Peter Hall[2]
[1]The Robotics Institute, Carnegie Mellon University. [2]Department of Computer Science, University of Bath.

We study the segmentation of videos of arbitrary dynamic scenes, focussing on dynamic textures such as water, fire, or swaying trees. These phenomena are commonplace in videos of natural scenes, but are poorly represented in general-purpose segmentation benchmarks, which mainly involve rigid or smooth non-rigid motion. Dynamic textures exhibit complex appearance and motion patterns, that usually have semantics beyond a simple consistency metric. A sequence depicting trees, for example, may contain smaller and larger branches, some static and others swaying in the wind, that would generally be assigned separate segments by unsupervised segmentation methods, while they ideally should all be part of one "tree" dynamic texture. Such mid-level interpretations of a scene are beyond the uniform priors of most "supervoxels" and unsupervised video segmentation methods. To overcome these limitations, we present more appropriate features, together with a supervised algorithm to learn such information from annotated, ground truth segmentations.

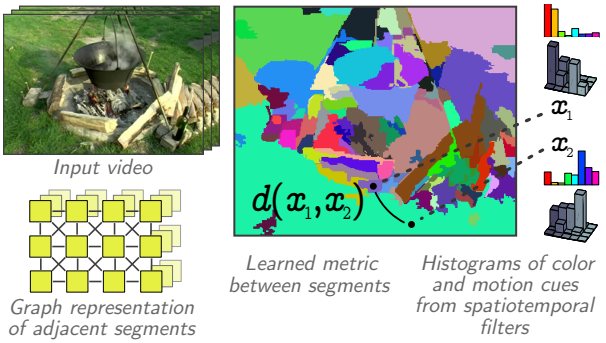

Figure 1: We extend the hierarchical graph-based segmentation technique and apply it to dynamic scene segmentation by learning distance metrics over motion and appearance features. Segments are iteratively merged, spatially and temporally, to form larger and larger segments. Merges occur between the most "similar" segments, as determined by our learned metric.

Our contribution consists of two parts. **First**, we use the responses to a bank of spatiotemporal filters to capture the appearance and motion characteristics of dynamic textures. Similarly to 2D filters that can capture structure in static images (*e.g.* edges), these 3D filters capture structure in the video volume, such as moving patterns [1]. The filters in the bank are tuned to various scales, orientations, and speeds, and provide a rich set of features that characterize image appearance and dynamics. Different dynamic textures often present different motion statistics. Water ripples in a pond, for example, will exhibit more low-frequency motion than ocean waves. These examples can be treated as a single "water" class, or as seperate phenomena depending on the labels in the training set, which justifies a learning approach. Therefore and **secondly**, we show how to learn a metric between descriptors made of histograms of color and filter-based motion cues, that allows supervised video segmentaton in a hierarchical, graph-based framework [2] (Fig. 1). We use ground truth segmentations to generate pairwise constraints between descriptors, and learn a metric that predicts whether segments should be merged or not during the segmentation. If training segments are provided with semantic labels, we use additional pairwise constraints between segments of ("same-" or "different-class") to learn a metric that explicitly separates multiple semantic classes. This allows a variety of training scenarios. The general task of video segmentation encompasses a number of specific applications that can benefit from learned mid-level models and we additionally evaluate the applicability of our method to the more classical tasks of motion segmentation and object segmentation from motion boundaries.

Practically, our work extends the hierarchical graph-based video segmentation technique of Grundmann *et al.* [2], which constructs a hierarchy of supervoxels of decreasing granularity. This algorithm is intrinsically suitable for a variety of tasks where the size of the desired segments is not fixed or known a priori. Although it already proved very successful in a number of benchmarks, the algorithm is still limited to the grouping of pixels with a uniform prior on appearance and/or motion. Another practical, but significant drawback is its large memory requirements (handled *e.g.* with block-processing), especially with the rich motion features that we propose. We address both of these issues by learning a metric between segment descriptors in a supervised setting, and *jointly* optimizing dimensionality reduction of the descriptors to dramatically reduce the memory requirements of the original algorithm. Note that this is not equivalent to first projecting the data with *e.g.* PCA and then learning a metric on low-dimensional descriptors, which may lose discriminative information. We rather optimize both tasks with a single objective function, thus fully exploiting the available supervision [3].

The technical contributions of the paper are threefold. (1) A new method to improve hierarchical video segmentation with supervised learning. We optimize a metric between segment descriptors over labelled training data, using a large-margin formulation suitable for hierarchical segmentation, unlike existing algorithms designed for nearest-neighbour classification. (2) We characterize the appearance and the motion within segments with a novel set of features based on spatiotemporal filters, that allows segmenting videos with arbitrary motions and complex dynamic textures. (3) We provide a method to optimize dimensionality reduction together with the learned metric, which drastically reduces the memory requirements of the original segmentation algorithm. We evaluated our contributions on a number of tasks, including dynamic texture segmentation benchmarks, and classical rigid motion segmentation datasets. Please consult the paper for extensive evaluations and discussions of these results.

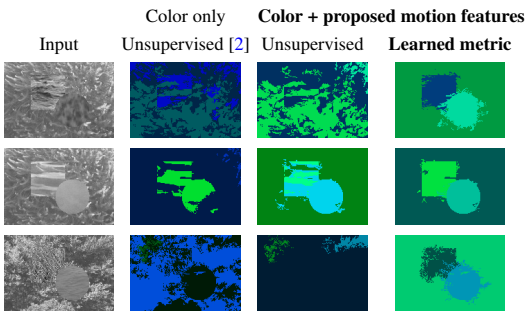

Figure 2: We obtained state-of-the-art results on SynthDB dynamic texture segmentation benchmark.

Figure 3: Segmentation of dynamic textures in complex scenes (Dyntex dataset).

[1] K. G. Derpanis and R. P. Wildes. Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(6): 1193–1205, 2012.

[2] M. Grundmann, V. Kwatra, M. Han, and I. A. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, pages 2141–2148, 2010.

[3] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher Vector Faces in the Wild. In *BMVC*, 2013.