Segmentation of Dynamic Scenes with Distributions of Spatiotemporally Oriented Energies

Damien Teney
d.teney@bath.ac.uk
Matthew Brown
m.brown@bath.ac.uk

Media Technology Research Centre
Department of Computer Science
University of Bath
Bath, UK

Overview

In video segmentation, disambiguating appearance cues by grouping similar motions or dynamics is potentially powerful, though non-trivial. Dynamic changes of appearance can occur from rigid or non-rigid motion, as well as complex dynamic textures. While the former are easily captured by optical flow, phenomena such as a dissipating cloud of smoke, or flickering reflections on water, do not satisfy the assumption of brightness constancy, or cannot be modelled with rigid displacements in the image. To tackle this problem, we propose a robust representation of image dynamics as histograms of motion energy (*HoME*) obtained from convolutions of the video with spatiotemporal filters. They capture a wide range of dynamics and handle problems previously studied separately (motion and dynamic texture segmentation). They thus offer a potential solution for a new class of problems that contain these effects in the same scene. Our representation of image dynamics is integrated in a graph-based segmentation framework [3] and combined with colour histograms to represent the appearance of regions. In the case of translating and occluding segments, the proposed features additionally serve to characterize the motion of the boundary between pairs of segments, to identify the occluder and inferring a local depth ordering. The resulting segmentation method is completely model-free and unsupervised, and achieves state-of-the-art results on the SynthDB dataset for dynamic texture segmentation, on the MIT dataset for motion segmentation, and reasonable performance on the CMU dataset for occlusion boundaries.

Proposed approach

Our approach to identify motion is based on existing work on steerable spatiotemporal filters [1, 2]. Similarly to 2D filters used to identify 2D structure in images (*e.g.* edges), these 3D filters can reveal structure in the spatiotemporal video volume. We employ Gaussian second derivative filters $G2_{\hat{\theta}}$ and their Hilbert transforms $H2_{\hat{\theta}}$. They are both steered to a spatiotemporal orientation parameterized by the unit vector $\hat{\theta}$ (the symmetry axis of the $G2$ filter). They are convolved with the video volume $\mathcal{V}$ of stacked frames, and give an energy response

$$E_{\hat{\theta}}(x,y,t) \;=\; (G2_{\hat{\theta}} * \mathcal{V})^2 + (H2_{\hat{\theta}} * \mathcal{V})^2 \quad . \tag{1}$$

In the frequency domain, a pattern moving in the video with a certain direction and velocity correspond to a plane passing through the origin. We obtain a representation of image dynamics by measuring the energy along a number of those planes, obtained by summing responses of filters consistent with the orientation of each plane. The resulting **motion energy** ME along the plane of unit normal $\hat{n}$ is given by

$$ME_{\hat{n}}(x,y,t) \;=\; \sum_{i=0}^{N} E_{\hat{\theta}_i}(x,y,t) \quad , \tag{2}$$

where N=2 is the order of the derivative of the filter, and $\hat{\theta}_i$ are filter orientations whose response lie in the plane specified by $\hat{n}$ (see [1] for details). This provides a representation of *dynamics* only, marginalizing the filter responses over appearance. The measurements $ME_{\hat{n}_i}$ can be compared to the extraction of optical flow, since each $\hat{n}_i$ specifies a particular orientation and velocity (*e.g.* patterns moving rightwards at 2 pixels per frame). The complete set of measurements $ME_{\hat{n}_i}$ is potentially capable of representing multiple, superimposed motions at a single location, offering definitive advantages over optical flow. Using the observation that motion- and color-based segmentation are two intrinsically similar problems, we adapt the segmentation algorithm of [3] to use our representation of motion. In addition to the original color histograms that represent the appearance of regions, we similarly accumulate our features into motion histograms (as in [3]). These motion histograms have 2 dimensions, corresponding to the (spatial) orientations and (spatiotemporal) velocities of the different $\hat{n}_i$ considered. The agglomerative segmentation iteratively produces results at decreasing levels of granularity.

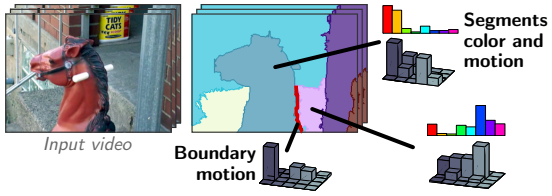

Figure 1: We represent dynamics in regions of the video with histograms of motion energies (HoME) measured at various space-time orientations. They are combined with colour histograms in a graph-based segmentation framework [3]. Post segmentation, HoMEs are additionally used to compare the motion of boundaries with their adjacent segments'. We thereby identify the occluders and infer a local depth ordering.

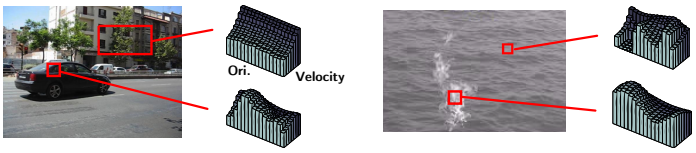

Figure 2: Actual HoMEs of real sequences, visualized as 2D histograms, of image (spatial) orientations and (spatiotemporal) velocities (lighter colours represent higher velocities; a limited set of velocities is represented for compactness). **(Left)** The background is mostly static with a uniform range orientations, whereas the moving car produces a single mode in the histogram. **(Right)** The sea waves exhibit multiple motion modes; the upwards motion of the flame is more simply defined.

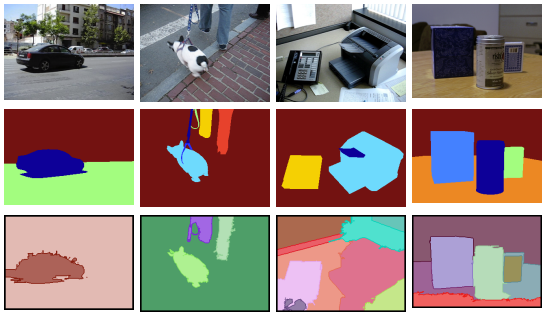

Figure 3: Motion segmentation (MIT dataset); input frame, ground truth, and segmentation. Different objects are correctly segmented, whether from their intrinsic motion (first two examples) or different relative motion induced by parallax and a translating camera (last two examples).

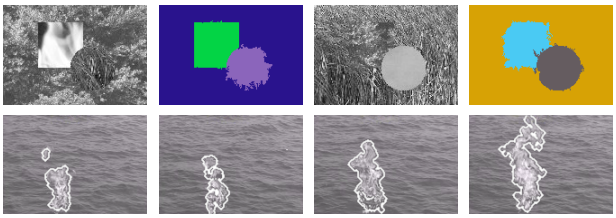

Figure 4: Segmentation of dynamic textures (SynthDB dataset). Static appearance of different textures may be very similar, and image dynamics are then crucial to distinguish them.

[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] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(9):891–906, 1991.

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