Segmentation of Low-Level Motion Features to Infer Occlusion Boundaries and Local Depth Ordering

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

1. Motivation

Motion in videos is a powerful cue to aid in scene understanding, by identifying the boundaries and the depth ordering of occluding objects. It can help to separate objects using their intrinsic motion, or parallax-induced motion at different depths. Most existing work rely on the computation of the optical flow, grouped into similar regions according to a parametric (*e.g.* affine) motion model. Two limitations ensue from this approach. First, the computation of the optical flow, despite recent advances, remains computationally expensive and relies on assumptions (*e.g.* brightness constancy or rigidly moving objects) that may not hold true. Secondly, parametric motions may similarly be limited to simple scenes with translating objects. More complex cases include deformable objects, repetitive motions, etc. In this work, we consider the use of **motion energies**, directly obtained from convolutions of the video with spatiotemporal filters, as an alternative image feature to optical flow for motion segmentation. We represent the motion of a region of the video with distributions of such motion energies, thereby alleviating the limitations of parametric motion models.

The combination of motion and appearance cues to improve boundary detection has mostly been addressed through supervised training [4]. This has some disadvantages related to the availability of suitable training data and to possible annotation bias in what actually constitutes relevant boundaries. Instead, we are interested in establishing a learning-free baseline, and we rather hypothesize that a **segmentation** framework is a suitable paradigm for grouping motions, similarly as it is for grouping static color and textures cues. Most work on video segmentation extends image segmentation techniques, and the joint grouping of appearance and motion features in complex scenes is still an open problem. [3], for example, briefly mentions the use of histograms of optical flow, though the improvement in performance was not rigorously evaluated.

Our contributions consist in *(i)* the integration of low-level, filter-based motion features into an existing segmentation framework [3], and *(ii)* an empirical evaluation demonstrating that this approach constitutes a competitive alternative to existing flow-based motion segmentation techniques, at a fraction of their computational demands.

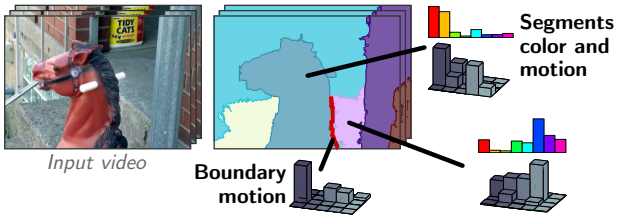


Figure 1. We use distributions of low-level, filter-based motion features as an alternative to optical flow. Beyond segmentation, motion is used to assign a boundary to the most similar of its two adjacent segments, then identified as foreground.

2. Proposed approach

Our approach to identify motion is based on existing work on steerable spatiotemporal filters [2, 1]. 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

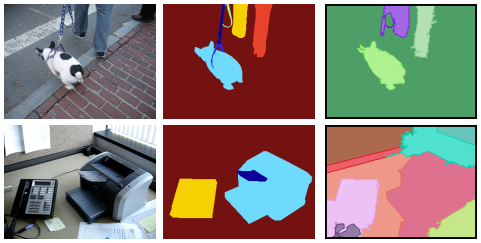
$$E_{\hat{\theta}}(x,y,t) \;=\; (G2_{\hat{\theta}} * \mathcal{V})^2 + (H2_{\hat{\theta}} * \mathcal{V})^2 \;. \qquad (1)$$

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

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

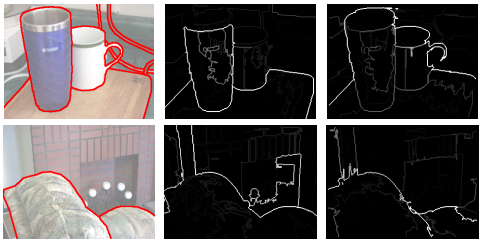
where $N=2$ is the order of the derivative of the filter, and $\hat{\theta}_i$ are filter orientations whose response lie in the plane

This work is supported by EPSRC Grant EP/K02339X/1.



Segmentation (avg. Rand index)	
Color only	73.3 %
Color + Motion$^{(1)}$	77.6 %
Color + Motion$^{(2)}$	**83.2** %
Layers++ [5]	77.5 %
nLayers [6]	82.3 %

Depth order (accuracy, per boundary)	
Motion$^{(1)}$	78.5 %
Motion$^{(2)}$	73.9 %



Input | Ground truth | Color + motion$^{(2)}$ | Input/ground truth | Color only | Color + motion$^{(2)}$

Figure 2. Motion segmentation on the MIT dataset (left and tables) and detection of occlusion boundaries on the CMU dataset (right).

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). We evaluated the use of these measurements in two different ways (noted $^{(1)}$ and $^{(2)}$ in Fig. 2). The first, most akin to optical flow, reduces these measurements to the single orientation of maximum response at each voxel, *i.e.* $\arg\max_{\hat{n}_i} ME_{\hat{n}_i}$. The second consists in the complete set of measurements $ME_{\hat{n}_i}$. This much richer representation of spatiotemporal structure is potentially capable of representing multiple, superimposed motions at a single location, offering definitive advantages over optical flow. Two potential issues must be examined though. Firstly, due to the broad tuning of the $G2$ and $H2$ filters, energy responses arise in a range of orientations around their peak tunings. This propagates to the $ME_{\hat{n}_i}$, whose values are heavily correlated across neighbouring planes. Secondly, the response to a filter is not independent of image contrast. We address these two issues by a non-linear scaling of $ME_{\hat{n}}$, first normalizing w.r.t. the strongest local energy measure, then as to emphasize the actual peak energies at each voxel:

$$ME'_{\hat{n}}(x,y,t) \;=\; ME_{\hat{n}}(x,y,t) \;/\; \max_{\hat{n}} ME_{\hat{n}}(x,y,t) \quad (3)$$

$$ME''_{\hat{n}}(x,y,t) \;=\; e^{\alpha(ME'_{\hat{n}}(x,y,t)-1)} \quad \text{with } \alpha = 1000. \quad (4)$$

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 (Fig. 1). The agglomerative segmentation iteratively produces results at decreasing levels of granularity. The inferred boundaries correspond to the 2 voxel-wide bands wherever two segments are adjacent in the video volume. Each boundary is assigned a **strength** from the highest segmentation level it appears in (Fig. 2, right).

Finally, we determine the local **depth ordering** of adjacent segments, using the observation that an occlusion boundary moves together with the occluding segment. For a given segmentation level, we accumulate motion histograms over each boundary (the set of voxels as defined above), so that it can be directly compared to the motion histogram of the adjacent segments (Fig. 1). The boundary is assigned to the most similar of the two, thus identified as the "most foreground" of the two.

3. Experiments and discussion

We assigned equal weights to color and motion histograms, and chose $\hat{n}_i$ corresponding to 16 spatial orientations and 10 velocities between 0 and 3 px/frame. We obtained excellent results on the **MIT dataset** [5] for motion segmentation (Fig. 2). The average of the maximum Rand indices (over our different levels of segmentation) is superior to a state-of-the-art method. The depth ordering is also well above a random guess. Preliminary experiments on the **CMU dataset** [4] for occlusion boundaries also showed promising results (Fig. 1, and 2 right). This confirms the hypotheses that *(i)* low-level motion cues are a viable alternative to optical flow to segment scene motion, at a fraction of the computational expense, and that *(ii)* a segmentation framework can provide appropriate constraints for model-free, unsupervised groupings of appearance and motion cues in the context of scene understanding.

[1] K. G. Derpanis and R. P. Wildes. Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. *IEEE Trans. PAMI*, 2012. 1, 2

[2] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. PAMI*, 1991. 1

[3] M. Grundmann, V. Kwatra, M. Han, and I. A. Essa. Efficient hierarchical graph-based video segmentation. In *CVPR*, 2010. 1, 2

[4] A. N. Stein and M. Hebert. Occlusion boundaries from motion: Low-level detection and mid-level reasoning. *IJCV*, 2009. 1, 2

[5] D. Sun, E. Sudderth, and M. J. Black. Layered image motion with explicit occlusions, temporal consistency, and depth ordering. *NIPS*, 2010. 2

[6] D. Sun, E. B. Sudderth, and M. J. Black. Layered segmentation and optical flow estimation over time. In *CVPR*, 2012. 2