Learning to Parse Video into Stable Spatiotemporal Volumes[1]

Thomas Dean
Google Inc.
tld@google.com

We are interested in learning how to exploit continuity, motion and context to account for stable, recoverable, spatiotemporal phenomena embedded in video. While most humans can make sense of still images, for the most part, we need continuity and motion to make sense of the world around us. Humans are also aided by strong priors that allow us to make confident predictions despite ambiguity, noise and occlusion.

The idea of combining top-down prior knowledge and bottom-up cues derived from motion and other low-level features has been around almost as long as research in computer vision, e.g., [10], and has recently seen renewed interest, e.g., [3, 2, 6, 11]. Rather than the traditional tasks of object recognition or image categorization, here we focus on the task of explaining each new frame in a video in terms of a continuously evolving representation of spatiotemporal volumes that account for the complete visual field. For the purpose of this abstract, the primary task of visual inference is to map an existing interpretation in the form of a segmentation of the current frame into a new segmentation that accounts for the fate of all segments in the previous frame and introduces new segments only when necessary.

Assuming there isn't an abrupt change of scene and the pose of the camera hasn't changed significantly, then the extant interpretation should account for most of the current frame if only we can map the surviving segments from the previous frame onto the current frame. Following the work of Borenstein and Ullman, each segment in an interpretation is accounted for by a distribution over smaller fragments that *covers* the segment while respecting segment boundaries [3]. Depending on how such a distribution evolves over time, it can be used to model affine-invariant views of a rigid object or more amorphous, homogeneous regions of texture that correspond to bodies of water or patches of grass.

If there is any disparity in the depth of the objects in the scene and any movement of the objects or the camera, then some segments will change their shape and this change constitutes an anomaly requiring an explanation. If an object appears in the previous frame and is accounted for by matching a particular view, then the same or a competing view of the same object must account for the object in the current frame. We are experimenting with the bag-of-visual-words-distribution method of categorizing fragments described in the work of Russell *et al* [7].

Inference then corresponds to continuously mapping the segments in one frame to those in the next. A *stable spatiotemporal volume* is simply a sequence of segments and associated transformations that explain how each pair of adjacent segments are related to one another in a manner akin to the video sprites of Jojic and Frey [5]. Each segment in a frame either extends some existing spatiotemporal volume or signals the creation of a new one.

We use motion to estimate occlusion boundaries and thereby attempt to explain (and exploit) one class of anomalies. Specifically, we use estimated occlusion boundaries to infer when two segments that consistently move together should be combined and, therefore, the distribution or distributions used to account for such segments have to be revised. Similarly, motion is used to identify when one segment undergoes a change that cannot be explained and thus requires its associated explanatory distributions modified to account for the unexpected variation or perhaps refined to account for two or more independent phenomena.

Occlusion boundaries are notoriously difficult to estimate using traditional optical-flow algorithms that treat such boundaries as a violation of a smoothness assumption. Stein and Hebert [9] present one approach that uses anistropic, edge-respecting, fast-marching diffusion processes to construct SIFT-like features that

[1]Topic: learning, vision Preference: oral

behave reasonably well along occlusion boundaries. Black and Fleet use separate probabilistic models for smooth motion and occlusion boundaries to estimate the relative-depth ordering of neighboring surfaces [1]. However, neither approach is perfect in detecting occlusion boundaries in our target dataset of videos on Google Video and YouTube, and so we combine motion with other cues to resolve ambiguities. In particular, we are experimenting with variants of the methods of Hoiem *et al* [4] and Saxena *et al* [8] which use color, texture and edge-based cues in single frames to infer occlusion boundaries between adjacent *superpixels*.

We then learn to merge superpixels into stable spatiotemporal volumes using an approach inspired by the model described in Borenstein and Ullman [3]. Learning proceeds by iteratively refining the set of explanatory fragments to account for unexplained anomalies until all segments are routinely accounted for. This work is still in a preliminary stage and we are most interested in presenting the basic problem formulation and discussing its merits in a public forum.

References

[1] Michael Black and David Fleet. Probabilistic detection and tracking of motion boundaries. *International Journal of Computer Vision*, 38(3):231–245, 2000.

[2] Eran Borenstein and Jitendra Malik. Shape guided object segmentation. In *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition*, pages 969–976. IEEE Computer Society, 2006.

[3] Eran Borenstein and Shimon Ullman. Class-specific, top-down segmentation. In *Proceedings of the 7th European Conference on Computer Vision*, pages 109–124, London, UK, 2002. Springer-Verlag.

[4] Derek Hoiem, Alexei Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007.

[5] N. Jojic and B. Frey. Learning flexible sprites in video layers. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 199–206, 2001.

[6] Xiaofeng Ren and Jitendra Malik. Learning a classification model for segmentation. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*, volume 1, pages 10–17, 2003.

[7] Bryan Russell, Alexei Efros, Josef Sivic, William Freeman, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1605–1614. IEEE Computer Society, 2006.

[8] Ashutosh Saxena, Sung Chung, and Andrew Ng. 3-D depth reconstruction from a single still image. *International Journal of Computer Vision*, 76(1):53–69, 2007.

[9] Andrew Stein and Martial Hebert. Local detection of occlusion boundaries in video. In *British Machine Vision Conference*, volume 1, pages 407–416, September 2006.

[10] Jay Tenenbaum and Harry Barrow. Experiments in interpretation-guided segmentation. *Artificial Intelligence*, 8:241–277, 1977.

[11] Liming Wang, Jianbo Shi, Gang Song, and I-Fan Shen. Object detection combining recognition and segmentation. In Yasushi Yagi, Sing Bing Kang, In-So Kweon, and Hongbin Zha, editors, *Proceedings of the 8th Asian Conference on Computer Vision*, volume 4843 of *Lecture Notes in Computer Science*, pages 189–199. Springer, 2007.