Causal Video Object Segmentation From Persistence of Occlusions

Brian Taylor, Vasiliy Karasev, Stefano Soatto
UCLA Vision Lab, University of California, Los Angeles, CA 90095

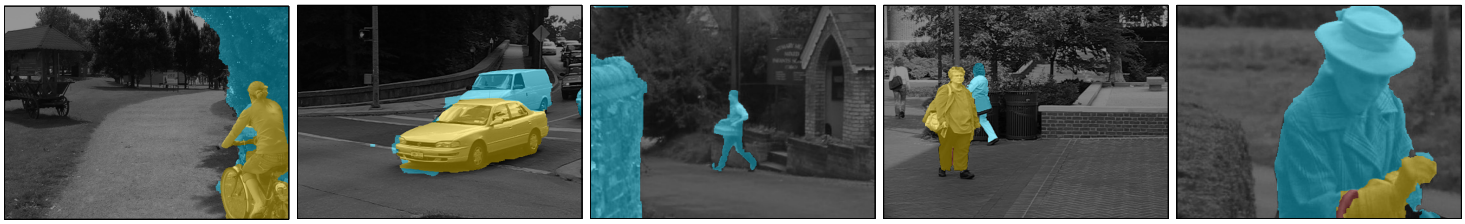

Figure 1: *Sample depth-layers: background $c(x) = 0$ (gray) and foreground layers $c(x) = 1$, $c(x) = 2$, $c(x) = 3$ noted by ■, ■, ■ respectively. On the far right, our algorithm correctly infers that the bag strap is in front of the woman's arm, which is in front of her trunk, which is in front of the background.*

Figure 2: *Sample object labels (connected components of layers). Further details and results are available at* `http://vision.ucla.edu/cvos/`

We seek to partition a video sequence into *"objects"* [1], corresponding to surfaces in space and their depth ordering relative to the viewer. Occlusion phenomena play a key role, as they inform local depth ordering relations: when a surface becomes *occluded*, the image region where it projected becomes occupied by the *occluder*, which is therefore closer to the viewer. If $I_t : D \to \mathbb{R}^3$ is a frame of a video and $D \subset \mathbb{R}^2$ its domain, we seek to partition D into regions, each associated with an integer *depth order*, represented by a function $c_t : D \to \mathbb{Z}_+$ indicating to which layer each pixel belongs: $c_t(x) = 0$ is the background, and larger values indicate foreground regions, or "objects" $c_t(x) = 1, 2, 3, \ldots$. Occlusion relations enable inferring the function c_t by solving a *linear program* [1]. However, errors in determining such occlusion relations have a cascading effect. Furthermore, even when occluded regions can be determined reliably, the *occluder* is non-trivial to determine. Technically, determination of the occluder requires either knowledge of the motion of the occluded region, which is undefined, or knowledge of its partition into regions, which is our goal to begin with.

Thus, our *first contribution* is to design motion and appearance priors to hallucinate motion in the occluded region so as to enable determining *occluders* when sufficient motion is present. Now, sufficient motion is typically *eventually* present, so long as viewer or objects move. However, when inter-frame motion is small, restricting the attention to few adjacent video frames makes determining occluded and occluder regions unnecessarily difficult. Thus, our *second contribution* is a causal framework for temporally integrating video frames to reliably determine occlusion relations. This is done by leveraging on spatio-temporal consistency of the location of objects, by enforcing that objects remain distinct even if they stop moving (and therefore no longer produce occlusion relations). All these cues and constraints can be encoded by a convex penalty function.

Since the function c_t encodes depth ordering relative to the viewer, which can change over time as objects move in front of one another, objects do not naturally enjoy a persistent layer value. However, boundaries between objects persist, unless objects split or merge. Unfortunately, the constraint on the indicator of the segmentation boundary is nonconvex, but once relaxed as a penalty, it can be interpreted as an adjustment of the total variation (TV) regularization weights. This penalty encourages "layer unity" within a convex optimization framework that allows aggregation of the affinity weights over time. We also propagate and accumulate occlusion cues adaptively over time, depending on the amount of inter-frame motion.

The final model that incorporates accumulated occlusion cues, aggregated weights, foreground and layer unity priors is

$$\begin{aligned} c_t = \arg\min_{c_t \geq 0} & \int_D g_t(x) |\nabla c_t(x)| dx + \tau \int_D c_t(x) dx \\ & + \int_D \kappa_t(x) \max\big(0, 1 - c_t(x)\big) dx \\ & + \sum_{\substack{i=1 \\ (y_i^c, y_i) \in \bar{O}_t}}^{N} \lambda_i \max\big(0, 1 - c_t(y_i^c) - c_t(y_i)\big) \end{aligned} \tag{1}$$

The objective consists of weighted TV (the first term, where the weights g_t are described in the paper), regularization on layer values (second term), "foreground prior" (third term, with weights $\kappa_t(x)$), and the penalty encouraging occluder-occluded pairs ($(y^c, y) \in \bar{O}_t$) to lie in different layers—the occluder being closer to the viewer—with weights λ.

Finally, our *third contribution* is to make the solution of the resulting optimization problem efficient. The optimization problem (1) is convex but large enough that off-the-shelf methods cannot solve it without resorting to superpixels or other pre-processing to reduce its dimension. We present an efficient numerical primal-dual scheme based on [2] that allows us to solve (1) on the pixel grid.

Our proposed method is competitive in video object segmentation based on benchmarks such as MoSeg and BVSD, with the advantage of processing the video sequence causally rather than in a batch. Sample outcomes of our scheme are shown in Fig. 1 (depth-layers) and in Fig. 2 (objects).

[1] Alper Ayvaci and Stefano Soatto. Detachable object detection: Segmentation and depth ordering from short-baseline video. *PAMI*, 34(10), 2012.

[2] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1), 2011.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.