Real-Time Coarse-to-fine Topologically Preserving Segmentation

Jian Yao[1], Marko Boben[2], Sanja Fidler[1], Raquel Urtasun[1],
[1]University of Toronto. [2]University of Ljubljana.

Superpixels have become an established image preprocessing step to significantly reduce the complexity of higher-level computer vision techniques. Typically their role is to partition the image into a tractable set of "units" in which the pixels should have similar appearance and consistent depth value if applicable. A useful superpixel algorithm should ideally run fast, possibly real-time, and guarantee a reliable, regular, and topologically coherent image partitioning. While a majority of existing work does not satisfy most of these requirements, particularly speed, some of the recent work reported real-time running times [1, 2].

In this paper, we build on [3] and propose a much more efficient optimization algorithm that results in an order of magnitude less updates (speed-up). Inspired by the SEEDS algorithm [2] our method uses a coarse-to-fine energy update strategy, which allows the optimization to reach better energy minima than [3] when employing even a single iteration as shown in Fig. 2.

More formally, let $s_p \in \{1,\cdots,M\}$ be the assignment of pixel p to a superpixel, and let $\mathbf{s} = (s_1,\cdots,s_N)$ be the set of all random variables representing the segmentation, with N the size of the image. Following [3], we formulate the segmentation problem with an objective function similar to k-means clustering, where we want superpixels that are coherent in appearance (E_{col}) but that have also regular shape (E_{pos}) and reasonable superpixel boundary length (E_b). We additionally add constraints (E_{size}) on the size of the superpixel to prevent tiny superpixels and topology contraints (E_{topo}) to enforce that each superpixel be a connected component (i.e. topology preservation). Let μ_i be the mean position of the i-th superpixel and let c_i be its mean color. Our Markov random field (MRF) energy is then defined as

$$E_{mono}(\mathbf{s},\mu,\mathbf{c}) = E_{col}(\mathbf{s},\mathbf{c}) + \lambda_{pos}E_{pos}(\mathbf{s},\mu) + \lambda_b E_b(\mathbf{s}) + E_{topo}(\mathbf{s}) + E_{size}(\mathbf{s}) \quad (1)$$

with $\mathbf{c} = (c_1,\cdots,c_M)$, $\mu = (\mu_1,\cdots,\mu_M)$ the set of centers and mean positions for all superpixels.

We use block coordinate gradient decent to minimize the energy Eq. 1. This is done by maintaining a queue, which is initialized with the blocks at the boundary on the coarsest level. The blocks in the queue are then iteratively popped out and discarded if minimizing the energy does not change their current assignment. If the assignment changes, the new boundary blocks are pushed to the bottom of the priority queue. When the queue is empty, the optimization will continue on the finer level. The process will not stop until the optimization on the pixel level is completed. During the computation at each level, each block is initialized to be a regular grid and the mean color and position are computed for each block. Then we first check whether changing the label of the block would violate the connectivity. If it does not violate, we solve optimally for the block assignment by minimizing the energy in Eq. 1. This is done by simply trying all assignments from the 4-neighboring blocks. If the block has changed assignment, we update the mean position and color using the incremental mean equation for the two superpixel involved (the one that this block belonged to before, as well as the one in the new assignment). We illustrate this process in Fig. 1.

We then augment this algorithm to perform joint segmentation and stereo estimation. Following slanted-plane methods [3], we represent the disparity of a superpixel with a slanted plane $\theta_i = (A_i, B_i, C_i)$ and reason about segmentation in the left image. The disparity of a pixel belonging to the i-th superpixel can then be computed by $d(\mathbf{p},\theta_i) = A_i p_x + B_i p_y + C_i$,. We further define $o_{i,j} \in \{co, hi, lo, ro\}$ to be a discrete variable that reasons about the type of occlusion boundary between adjacent superpixels i and j, with the states representing whether the boundary is co-planar, hinge, the i-th plane is in front, or behind. Let $\theta = (\theta_1,\cdots,\theta_M)$ be the set of plane parameters for all superpixels and let $\mathbf{o}$ be the set of all occlusion variables. Additionally, let f_i be an outlier flag for the i-th pixel, and let $\mathbf{f}$ the set of flags for all pixels. We define the energy of the joint segmentation and stereo as the sum of energies encoding the monocular energy (E_{mono}) as well as consistency with the stereo image evidence (E_{disp}), a prior on the complexity of the boundaries (E_{prior}) and smoothness (E_{smo}) between slanted planes of neighboring super pixels. Thus

$$E_{total}(\mathbf{s},\mu,\mathbf{c},\theta,\mathbf{o},\mathbf{f}) = E_{mono}(\mathbf{s},\mu,\mathbf{c}) + E_{stereo}(\mathbf{s},\theta,\mathbf{o},\mathbf{f}) \quad (2)$$

with the energy related to stereo defined as

$$E_{stereo}(\mathbf{s},\theta,\mathbf{o},\mathbf{f}) = \lambda_{disp}E_{disp}(\mathbf{s},\theta,\mathbf{f}) + \lambda_{smo}E_{smo}(\mathbf{s},\theta,\mathbf{o}) + \lambda_{prior}E_{prior}(\mathbf{o}) \quad (3)$$

The minimization of Eq. 2 is similar to that in monocular case. The difference is that after updating $\mathbf{s}$, we have to estimate the boundary type $\mathbf{o}$, the outlier flags $\mathbf{f}$ and plane parameters θ. The boundary variables are optimized one at a time by selecting the label with minimum cost. The outlier flags are optimized given the current plane parameters. The planes are then fitted using least squares taking into account the non-outlier pixels only.

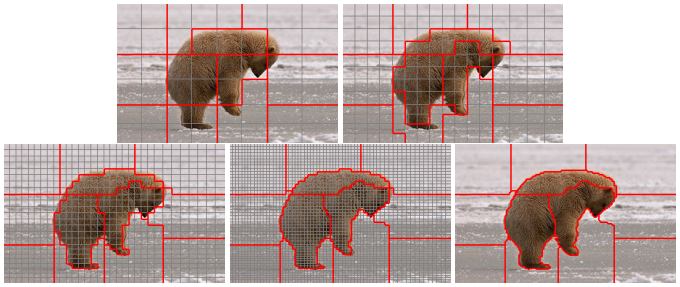

Figure 1: Coarse-to-fine boundary-level updates start at the coarse level (top-left) and proceeds to the finest level iteratively. The final result, defined on the finest (pixel) level, is shown on the bottom-right.

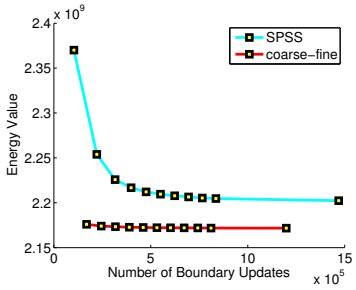

Figure 2: Energy as a function of the number of boundary updates in KITTI dataset.

We demonstrate the effectiveness of our approach in two important settings: unsupervised segmentation of RGB images, as well as joint segmentation and stereo estimation via slanted planes. We evaluate and compare our approach to state-of-the-art superpixel algorithms on the BSD and KITTI benchmarks. Our approach significantly outperforms the baselines in the segmentation metrics and achieves the lowest error on the stereo task.

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *PAMI*, 34(11):2274–2282, 2012.

[2] M. Van den Bergh, G. Roig, X. Boix, S. Manen, and L. Van Gool. Online video superpixels for temporal window objectness. In *ICCV*, 2013.

[3] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *ECCV*, 2014.

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