Superpixel-based Video Object Segmentation using Perceptual Organization and Location Prior

Daniela Giordano[1], Francesca Murabito[1], Simone Palazzo [1], Concetto Spampinato [1]
[1]Department of Electrical, Electronics and Computer Engineering, University of Catania (Italy).

Object segmentation in videos is a fundamental task for many computer vision problems, from object tracking to behaviour understanding to event detection. Recent approaches employ explicit foreground modeling and superpixel-oriented spatio-temporal segmentation [3], but they 1) rely on optical flow, which is computationally expensive, 2) group superpixels together by pure spatio-temporal appearance similarity, without exploiting real-world object features; 3) consider video object segmentation as a single-objective optimization problem, while, in practice, it is a multi-objective optimization problem.

We present an approach for object segmentation in videos which is able to work with fast moving and multimodal backgrounds, highly deformable and/or articulated objects and with different video qualities, without making any assumptions on how objects look like or move but adopting general properties of real-world objects. A key distinctive element of our method is the capability to quickly identify candidate motion regions, as those where significant variations on superpixel segmentation [1] in consecutive frames have been observed, without relying on optical flow but by assessing similarity between spatio-temporal neighbor superpixels, as in Fig. 1.

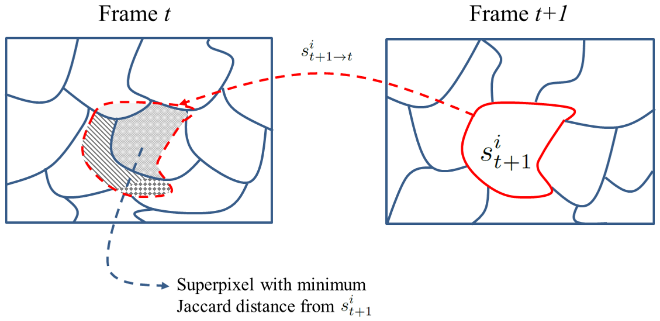

Figure 1: **Motion superpixel identification**. Superpixel s^i_{t+1} at frame $t+1$ is backprojected on frame t; it is marked as "motion superpixel" if the minimum Jaccard distance between overlapping superpixels is above a threshold.

However, motion superpixel segmentation does not take into account visual appearance or how well groups of superpixels geometrically fit together. Therefore, these location priors are combined with the latest foreground maps (to allow segmenting objects which become temporarily stationary), and for each of the resulting *motion regions* a local accurate segmentation subtask based on appearance similarity and perceptual organization is defined, which takes into account motion and non-motion superpixels intersecting the motion region's bounding box. Depending on the size of the object, the number of superpixels involved in each subtask is small enough to solve many such problems efficiently.

We pose each local segmentation task as an energy minimization problem, where higher costs correspond to segmentations which do not satisfy similarity and perceptual contiguity. Formally, given the set of superpixels $S = \{s_1, \dots, s_N\}$ and a set of corresponding labels $\mathcal{L} = \{l_1, \dots, l_N\}$, where $l_i \in \{0 : \text{background}, 1 : \text{foreground}\}$, the overall energy function is:

$$E(\mathcal{L}) = A(\mathcal{L}) + P(\mathcal{L}) \tag{1}$$

$$A(\mathcal{L}) = \sum_{l_i \in \mathcal{L}} a_1(l_i) + \sum_{(l_i,l_j) \in \mathcal{N}(\mathcal{L},S)} a_2(l_i,l_j) \tag{2}$$

$$P(\mathcal{L}) = \sum_{(l_i,l_j) \in \mathcal{N}(\mathcal{L},S)} p(l_i,l_j) \tag{3}$$

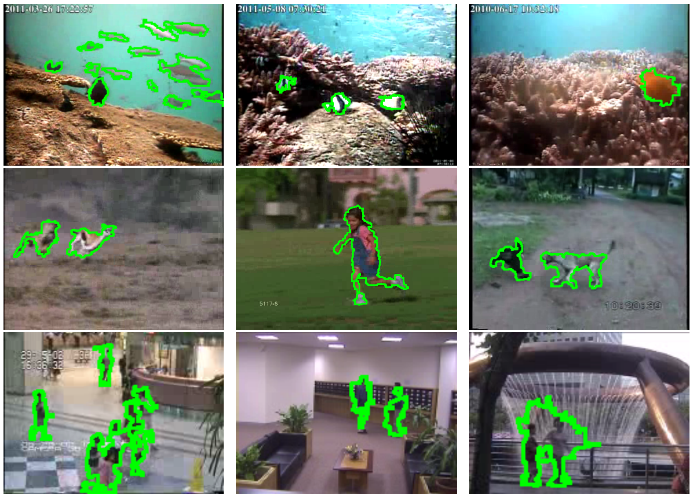

Figure 2: Example results on the Underwater, SegTrack and I2R datasets.

In the formula, $A(\mathcal{L})$ and $P(\mathcal{L})$ respectively represent the overall appearance and perceptual organization energies, $\mathcal{N}(\mathcal{L},S)$ is the set of all pairs of neighbor superpixels, and the potentials $a_1(\cdot)$, $a_2(\cdot,\cdot)$ and $p(\cdot,\cdot)$ enforce our design principles on visual similarity and perceptual organization.

Unary potential $a_1(\cdot)$ indicates whether a superpixel is best associated to the foreground or the background, based on the similarity of each pixels to the current background/foreground models.

Binary potential $a_2(\cdot,\cdot)$ defines the cost of assigning different labels to two neighbor superpixels, based on their color similarity. We estimate the similarity of two superpixels as the probability that their union is generated by the same (background or a foreground) color distribution.

Binary potential $p(\cdot,\cdot)$ defines a similar cost, based on how well they fit together from a perceptual and geometrical point of view. To estimate this quantity, we employ a variant of the approach proposed by [2]: the resulting potential function is a combination of terms encoding *boundary complexity*, *symmetry*, *continuity*, and *attachment strength*.

Each local segmentation task, defined by the above potentials, is then solved iteratively in order to "discover" large objects of which only a part had been detected during motion superpixel estimation.

Our method was evaluated on three public datasets, characterized by slow object motion, camera motion, small objects and cluttered scenes (Fig. 2). We show that the proposed method outperforms or reaches comparable results as state-of-the-art approaches, while keeping processing times low (up to about 0.2 sec/frame at 320×240 resolution). This is remarkable given that [3] takes 3 sec/frame on the same datasets, including optical flow computation but without actually performing better the proposed approach.

[1] Radhakrishna Achanta, Appu Shaji, and Kevin Smith. SLIC superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence*, 6(1):1–8, 2012.

[2] Chang Cheng, Andreas Koschan, Chung-Hao Chen, David L Page, and Mongi a Abidi. Outdoor scene image segmentation based on background recognition and perceptual organization. *IEEE Transactions on Image Processing*, 21(3):1007–19, March 2012.

[3] Anestis Papazoglou and Vittorio Ferrari. Fast Object Segmentation in Unconstrained Video. *2013 IEEE International Conference on Computer Vision*, pages 1777–1784, December 2013.

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