Saliency-Aware Geodesic Video Object Segmentation

Wenguan Wang[1], Jianbing Shen[1,*], Fatih Porikli[2]
[1]Beijing Lab of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, China.
[2]Research School of Engineering, Australian National University, and NICTA Australia.

Unsupervised video object segmentation methods aim at automatically extracting the object from the whole video. Such segmentation has shown to benefit many specific visual tasks, such as video summarization and compression. Several methods [1, 3, 5] explored the notion of what a foreground object should look like in video data. These approaches generate considerable object proposals in every frame and transform the task of video object segmentation into an object region selection problem. More specifically, a clustering process was introduced for finding objects by [1], a constrained maximum weight cliques technique to model the selection process was proposed by [3], and a layered directed acyclic graph based framework was presented by [5]. However, these proposal based techniques have high computational complexity, and their dependency on the large number of proposals leads to much difficulty and complexity of the selection process.

Our goal is to segment the foreground objects from the background in all frames of a given video sequence automatically. Our method is based on the proposed visual saliency detection technique that incorporates several visual cues such as motion boundary, edge and color. We consider two discriminative visual features: spatial edges and temporal motion boundaries as indicators of foreground object locations. By imposing motion continuity, we establish a dynamic location model for each frame. Finally, the spatiotemporal saliency maps, appearance models and dynamic location models are combined into an energy minimization framework to attain both spatially and temporally coherent object segmentation (Fig. 1).

Given an input video sequence $\mathbf{F} = \{F^1, F^2, \cdots\}$, we compute an edge probability map $E_c^k(x_i^k)$ corresponding to k-th frame F^k at pixel x_i^k using [2]. Let V^k be the optical flow field of frame F^k, we then compute the gradient magnitude E_o^k of the optical flow field V^k as $E_o^k = \|\nabla V^k\|$. Let $\mathbf{Y}^k = \{Y_1^k, Y_2^k, \cdots\}$ be the superpixel set of frame F^k. Given the pixel edge map E_c^k, the edge probability of each superpixel Y_n^k is computed as the average value of the pixels with ten largest edge probabilities within Y_n^k. This generates a superpixel edge map$\widehat{E}_c^k$. Similarly, we compute a superpixel optical flow magnitude map $\widehat{E}_o^k$ using E_o^k. Then a spatiotemporal edge probability map E^k is generated as:

$$\mathrm{E}^k = \widehat{E}_c^k \cdot \widehat{E}_o^k. \tag{1}$$

Intra-frame graph construction For frame F^k, we construct an undirected weighted graph $\mathcal{G}^k = \{\mathcal{V}^k, \mathcal{E}^k\}$ with superpixels $\mathbf{Y}^k$ as nodes $\mathcal{V}^k$ and the links between pairs of nodes as edges $\mathcal{E}^k$. The weight w_{mn}^k of the edge $e_{mn}^k \in \mathcal{E}^k$ between adjacent superpixels Y_m^k and Y_n^k is defined as:

$$e_{mn}^k = \|\mathrm{E}^k(Y_m^k) - \mathrm{E}^k(Y_n^k)\|, \tag{2}$$

where $\mathrm{E}^k(Y_m^k)$ and $\mathrm{E}^k(Y_n^k)$ correspond to the spatiotemporal boundary probability of superpixels Y_m^k and Y_n^k, separately. Based on the graph structure, we derive an $|\mathcal{V}^k| \times |\mathcal{V}^k|$ weight matrix W^k, where $|\mathcal{V}^k|$ is the number of nodes in $\mathcal{V}^k$. The (m, n)th element of W^k is: $W^k(m,n) = e_{mn}^k$. For each superpixel Y_n^k, the probability P_n^k for foreground is computed by the shortest geodesic distance to the image boundaries using

$$P_n^k = \min_{T \in \mathbf{T}^k} d_{geo}(Y_n^k, T, \mathcal{G}^k), \tag{3}$$

where $\mathbf{T}^k$ indicate the superpixels along the four boundaries of frame F^k. The geodesic distance $d_{geo}(v_1, v_2, \mathcal{G}^k)$ between any two superpixels $v_1, v_2 \in \mathcal{V}^k$ in graph $\mathcal{G}^k$ is defined as the accumulated edge weights along their shortest path on graph $\mathcal{G}^k$:

$$d_{geo}(v_1, v_2, \mathcal{G}^k) = \min_{C_{v_1,v_2}} \sum_{p=0,1} |W^k \cdot \dot{C}_{v_1,v_2}(p)|, \tag{4}$$

where $C_{v_1,v_2}(p)$ is a path connecting the nodes v_1, v_2 (for $p = 0$ and $p = 1$).

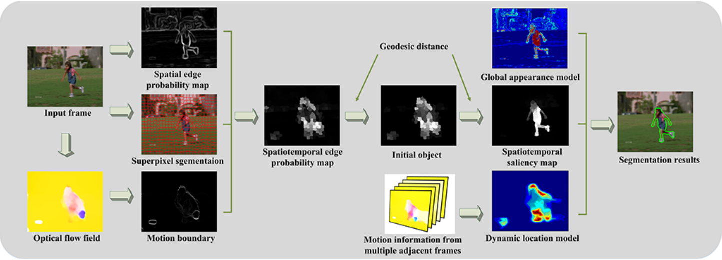

Figure 1: Overview. Input frame is over-segmented into superpixels and a spatiotemporal edge map is produced by the combination of static edge probability map and optical flow gradient magnitude.

Inter-frame graph construction For each pair of subsequent frame F^k and F^{k+1}, an undirected weighted graph $\mathcal{G}'^k = \{\mathcal{V}'^k, \mathcal{E}'^k\}$ is constructed. The nodes $\mathcal{V}'^k$ consist of all the superpixels $\mathbf{Y}^k$ of frame F^k and all the superpixels $\mathbf{Y}^{k+1}$ of frame F^{k+1}. For each frame, a self-adaptive threshold is used to decompose frame F^k into background regions $\mathbf{B}^k$ and object-like regions $\mathbf{U}^k$ through the object probability map P^k. This threshold σ^k for frame F^k is computed by $\sigma^k = \mu(P^k)$, where $\mu(\cdot)$ computes the mean probability of all pixels within frame F^k by probability map P^k. Additionally, the background information of previous frame offers valuable prior. Therefore, we define the background regions $\mathbf{B}^k$ of k-th frame as:

$$\begin{aligned} \mathbf{B}^k = &\ \{Y_n^k | P_n^k \le \sigma^k\} \cup \{Y_n^k | Y_n^k \text{ is temporally connected to } \mathbf{B}^{k-1}\}, \\ \mathbf{U}^k = &\ \mathbf{Y}^k - \mathbf{B}^k, \end{aligned} \tag{5}$$

We formulate video object segmentation as a pixel labeling problem with two labels (foreground and background). Each pixel $x_i^k \in \mathbf{X}^k$ can take a label $l_i^k \in \{0,1\}$, where 0 corresponds to background and 1 corresponds to foreground. A labelling $\mathbf{L} = \{l_i^k\}_{k,i}$ of pixels from all frames represents a segmentation of the video. Similarly to other segmentation works [1, 4], we define an energy function for labeling $\mathbf{L}$ of all the pixels:

$$\begin{aligned} \mathcal{F}(\mathbf{L}) = &\sum_{k,i} \mathcal{U}_i^k(l_i^k) + \lambda_1 \sum_{k,i} \mathcal{A}_i^k(l_i^k) + \lambda_2 \sum_{k,i} \mathcal{L}_i^k(l_i^k) \\ &+ \lambda_3 \sum_{(i,j)\in \mathbf{N}_s} \mathcal{V}_{ij}^k(l_i^k, l_j^k) + \lambda_4 \sum_{(i,j)\in \mathbf{N}_t} \mathcal{W}_{ij}^k(l_i^k, l_j^{k+1}), \end{aligned} \tag{6}$$

where spatial pixel neighborhood $\mathbf{N}_s$ consists of eight spatially neighboring pixels within one frame, temporal neighborhood $\mathbf{N}_t$ consists of the forward-backward nine neighbors in adjacent frames, and i, j index the pixels.

Our source code will be publicly available online .

[1] Yong Jae Lee, Jaechul Kim, and K. Grauman. Key-segments for video object segmentation. In *ICCV*, 2011.

[2] Marius Leordeanu, Rahul Sukthankar, and Cristian Sminchisescu. Efficient closed-form solution to generalized boundary detection. In *ECCV*, 2012.

[3] Tianyang Ma and Longin Jan Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *CVPR*, 2012.

[4] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM TOG*, 23(3), 2004.

[5] Dong Zhang, Omar Javed, and Mubarak Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, 2013.

*Corresponding author: Jianbing Shen (shenjianbing@bit.edu.cn). This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

http://github.com/shenjianbing/videoseg15