Supplementary Materials: Towards Segmenting Consumer Stereo Videos: Benchmark, Baselines and Ensembles

Wei-Chen Chiu[†], Fabio Galasso[‡], Mario Fritz[†]

[†] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[‡] OSRAM Corporate Technology

1 Content of this supplemental material

We illustrate in this supplemental pdf document:

- Derivatives of the performance proxies $\hat{P}$ (Referred from Section 6.2.1 in the paper).
- Additional examples of the proposed EASVS approach in comparison to baselines.

We enclose in the supplemental video:

- Videos and annotations for the stereo videos sequences in the Consumer Stereo Videos Segmentation Challenge (CSVSC) (Referred from Section 3 in the paper);
- Sample video results comparing the proposed efficient and adaptive stereo video segmentation algorithm (EASVS) with the state-of-the-art (cf. Section 7 in the paper).

2 Derivatives of Performance Proxies $\hat{P}$

2.1 Review of Transfer-Cut [1]

For convenience of the reviewer, we first report additional details of [1]. By the use of *transfer-cut*, [1] connects the pool of segmentations to the pixels of the image (the work is originally defined for image segmentation). In the original procedure, given K pooled segmentation results composed of N^k superpixels, the pairwise distance matrix between superpixels from the k-th pooled result is given by:

$$A^k = \begin{bmatrix} w^k_{1,1} & w^k_{1,2} & \cdots & w^k_{1,N^k} \\ w^k_{2,1} & w^k_{2,2} & \cdots & w^k_{2,N^k} \\ \vdots & \vdots & w^k_{I,J} & \vdots \\ w^k_{N^k,1} & w^k_{N^k,2} & \cdots & w^k_{N^k,N^k} \end{bmatrix} \tag{1}$$

Let us define then a block matrix S_Y to stack all pairwise distance matrices from all K algorithms:

$$S_Y = \begin{bmatrix} A^1 & & & \\ & A^2 & & \\ & & \ddots & \\ & & & A^K \end{bmatrix} \tag{2}$$

which will have size $N_{sp} \times N_{sp}$, where $N_{sp} = \sum^{K} N^{k}$.

Let us then define the distance matrix S_{XY} between the superpixel i and the pixel j for all algorithms as:

$$S_{XY}(i,j) = \alpha^{k} \text{ ,if pixel } j \in \text{ superpixel } i \text{ from algorithm } k \tag{3}$$

where its size is in size of $N_{sp} \times N_{p}$, and N_p is the total number of pixels.

Then the total distance matrix H with size $(N_p + N_{sp}) \times N_{sp}$ can be written in

$$H = \left[S_{XY} | S_{Y} \right]^{\top} \tag{4}$$

Let us define

$$D_X = \mathbf{diag}(H\mathbf{1}) \tag{5}$$

and the graph between all the superpixels from pooled segmentations by the transfer-cut:

$$W_Y = H^{\top} \cdot D_X^{-1} \cdot H \qquad \text{(size: } N_{sp} \times N_{sp}) \tag{6}$$

To look into the details, the parametric forms of elements in W_Y can be written into:

$$\begin{aligned} &\text{diagonal} \Rightarrow \sum^{M} \frac{(\alpha^{k})^{2}}{\sum^{K} \alpha^{k}} + \sum_{J=1}^{N^{k}} \frac{(w_{I,J}^{k})^{2}}{\sum_{Z=1}^{N_k} w_{J,Z}^{k}} \\ &\qquad \text{where } M \text{ is the number of pixels} \in \text{superpixel } I, \\ &\qquad \text{and superpixel } I \in k\text{-th pool.} \end{aligned} \tag{7}$$

$$\text{off-diagonal} \Rightarrow \begin{cases} \sum^{M} \frac{(\alpha^{k})^{2}}{\sum^{K} \alpha^{k}} + \sum_{Z=1}^{N^{k}} \frac{w_{I,Z}^{k} \cdot w_{Z,J}^{k}}{\sum_{L=1}^{N^{k}} w_{Z,L}^{k}} \\ \quad \text{if both superpixels } I \text{ and } J \in k\text{-th pool.} \\ \sum^{M} \frac{\alpha^{k} \cdot \alpha^{k'}}{\sum^{K} \alpha^{k}} \\ \quad \text{if superpixel } I \in k\text{-th pool} \\ \qquad \text{and } J \in k' \neq k\text{-th pool.} \end{cases} \tag{8}$$

where M is the number of pixels $\in I \cap J$

2.2 Derivatives of $\mathcal{G}^{Q}$

The *minimally overlapping superpixels* allow an efficient segmentation model (cf. Section 5.2 and Figure 4 in the paper). In order to use the minimally overlapping superpixels for learning (cf. Section 6.2), we need to:

- express the pairwise affinities between minimally overlapping superpixels in terms of the parameters α and β;
- write the full derivatives of the affinity matrix among minimally overlapping superpixels.

Here we show more details and complete equations for both parts.

Let us extend the theory in the previous section to stereo videos (modeled on the left view frames) and voxels. As discussed in the paper, this implies changing the elements of matrices S_{XY} and S_Y from pixels and pooled superpixels to minimally overlapping superpixels.

First, the β edges used in our submission to denote the similarities between the pooled superpixels I, J is expanded into the similarities between minimally overlapping superpixels i, j by the graph expansion approach [2]:

$$w^k_{ij} = \begin{cases} w^k_{I,J} & \text{if } I \neq J, i \in I, j \in J \\ \sum_{Z \neq I} w^k_{I,Z} & \text{if } i, j \in I, i \neq j \\ 0 & \text{if } I = J \end{cases} \tag{9}$$

Note that $w^k_{I,J} = e^{-(\beta^1 d^{k_1}_{I,J} + \cdots + \beta^C d^{k_C}_{I,J})}$ in our proposed scheme, where $d^{k_c}_{I,J}$ is the distance between superpixels I and J from the k-th pooled output based on c-th feature.

Then we have the parametric forms of the elements for H' as:

$$\begin{aligned} \text{diagonal} \Rightarrow \frac{(\alpha^k)^2}{\sum^K \alpha^k} + \sum_{j=1}^{N^m} \frac{(w^k_{ij})^2}{\sum_{z=1}^{N_m} w^k_{j,z}} \\ \text{off-diagonal} \Rightarrow \frac{(\alpha^k)^2}{\sum^K \alpha^k} + \sum_{z=1}^{N^m} \frac{w^k_{iz} \cdot w^k_{zj}}{\sum_{l=1}^{N^m} w^k_{zl}} \end{aligned} \tag{10}$$

where N^m is number of minimally overlapping superpixels. The reduced graph $\mathcal{G}^Q$ is then easily to derived from H' by the graph reduction [3] to group the edges of identical minimally overlapping superpixels in H'. And we denote the elements in $\mathcal{G}^Q$ by w^Q_{ij}. (Please note that we can also build up the graph H' based on the voxels as nodes then reduced to $\mathcal{G}^Q$, which will follow the story of Equation 1 in the main paper. Here we directly use minimally overlapping superpixels in H' for clarity.)

The reduced graph $\mathcal{G}^Q$ is the similarity matrix W which we base on to compute the spectral properties NCut and Trace$_R$ for the performance proxies, as in Equation 5 of the main paper. And the derivatives of NCut and Trace$_R$ shown in the Equation 7 of the main paper are:

$$\begin{aligned} \frac{\partial(\text{NCut})}{\partial\theta} &= \sum_{r=1}^{R} \frac{-e_r^\top \frac{\partial W}{\partial\theta} e_r e_r^\top D e_r + e_r^\top W e_r e_r^\top \frac{\partial D}{\partial\theta} e_r}{(e_r^\top D e_r)^2} \\ \frac{\partial(\text{Trace}_R)}{\partial\theta} &= \text{trace}(V^\top \frac{\partial L(\theta)}{\partial\theta} V) \end{aligned} \tag{11}$$

where V denotes the subspace spanned by the first R eigenvectors of L.

According to the formulations mentioned above, the derivatives $\frac{\partial w^Q_{i,j}}{\partial\alpha^k}$ and $\frac{\partial w^Q_{i,j}}{\partial\beta^c}$ for entries of W can be derived by a sequence of chain rules. The degree matrix D is diagonal matrix, where we can represent its elements on the diagonal by:

$$D_{II} = \sum_{J=1}^{N^m} w^Q_{I,J} \tag{12}$$

Then the derivatives of D's elements are:

$$\begin{aligned}\frac{\partial D_{II}}{\partial \alpha^k} &= \sum_{J=1}^{N^m} \frac{\partial w_{I,J}^Q}{\partial \alpha^k} \\ \frac{\partial D_{ii}}{\partial \beta^c} &= \sum_{J=1}^{N^m} \frac{\partial w_{I,J}^Q}{\partial \beta^c}\end{aligned} \tag{13}$$

Finally the derivative of generalized Laplacian matrix $L = D^{-1} \cdot W$ in equation 11 is given by the chain rule:

$$\frac{\partial L}{\partial \theta} = D^{-1} \cdot \frac{\partial W}{\partial \theta} + \frac{\partial D^{-1}}{\partial \theta} \cdot W \tag{14}$$

References

1. Li, Z., Wu, X.M., Chang, S.F.: Segmentation using superpixels: a bipartite graph partitioning approach. In: CVPR. (2012)
2. Agarwal, S., Branson, K., Belongie, S.: Higher order learning with graphs. In: ICML. (2006)
3. Galasso, F., Keuper, M., Brox, T., Schiele, B.: Spectral graph reduction for efficient image and streaming video segmentation. In: CVPR. (2014)
4. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph-based video segmentation. In: CVPR. (2010)
5. Ochs, P., Brox, T.: Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In: ICCV. (2011)
6. Hickson, S., Birchfield, S., Essa, I., Christensen, H.: Efficient hierarchical graph-based segmentation of rgbd videos. In: CVPR. (2014)

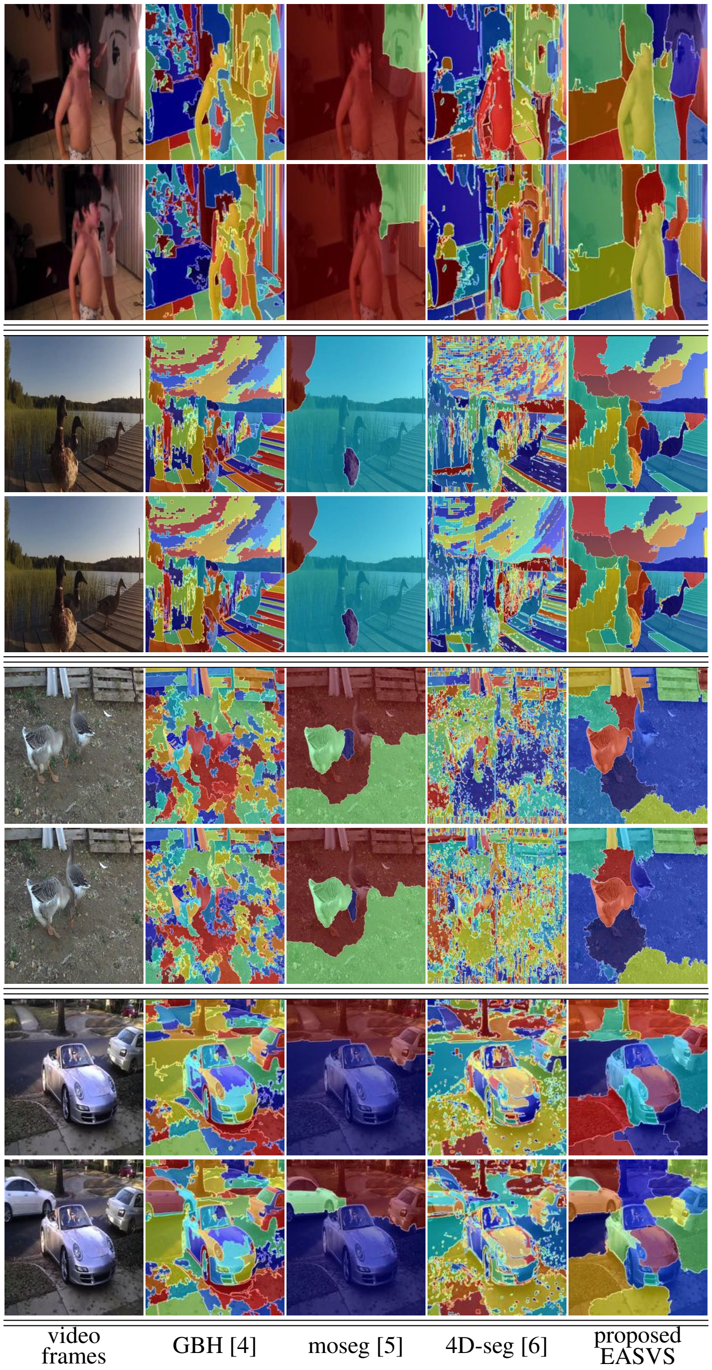

Table 1: Additional examples of the proposed EASVS compared to baselines.