Video Segmentation with Just a Few Strokes

Naveen Shankar Nagaraja, Frank R. Schmidt and Thomas Brox

[o] Vision
COMPUTER VISION University of Freiburg

SANTIAGO, CHILE DECEMBER 11-18, 2015
ICCV15
International Conference on Computer Vision

Motivation

We address two important challenges in video segmentation:

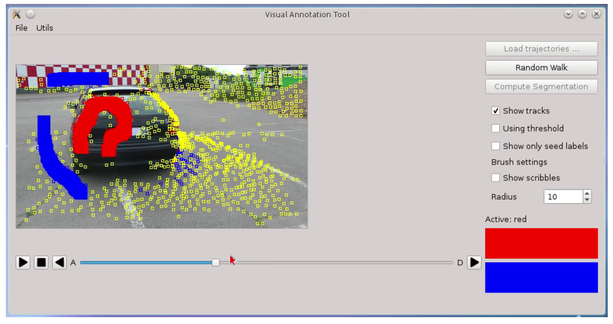
☞ **Disocclusion:** When "new things" appear in a video, do they belong to background or an object that we are interested in?

☞ **Base point disambiguation:** When there is only camera motion and our object of interest does not move, then cues like motion and appearance fail to correctly distinguish between object and background at the *base point* i.e. where the object touches the ground.

Contributions

- ☞ An easy to use tool for interactive segmentation of point trajectories.
- ☞ Elegant combination of user input with motion, appearance, and length cues for dense segmentation.
- ☞ A temporally consistent appearance preserving cue.
- ☞ A case study on quantifying the amount of user interaction.

Sparse Segmentation with Tool



- ☞ **Intuition:** segmenting *point trajectories* is faster for user interaction.
- ☞ User marks scribbles in various frames to alleviate both disocclusion and base point problem.
- ☞ **Segmentation:** Random Walk assigns a soft label to the unlabeled trajectories. Output is a *coarse* object segmentation.
- ☞ With our tool the user can also see a per frame *instant preview* of the expected dense segmentation.

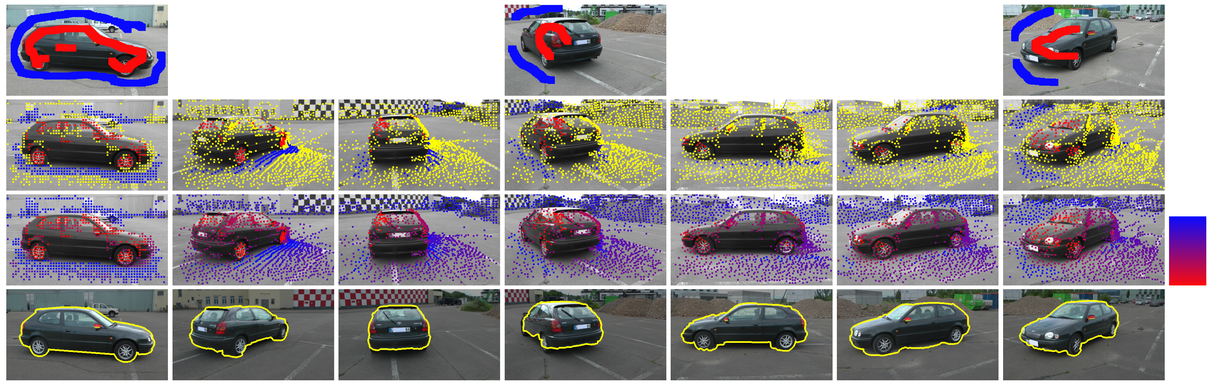
Dataset and Binaries

http://lmb.informatik.uni-freiburg.de/resources/binaries/iVideoSeg/

Contact: nagaraja@cs.uni-freiburg.de

We gratefully acknowledge partial funding from the ERC starting grant Videolearn. This study was also supported by BIOSS Centre for Biological Signalling Studies (EXC 294).

Overview



Dense Segmentation: Energy and Optimization

Let $\mathcal{S}\colon \Omega \times [1,T] \to \{0,1\}$ denote the whole video segmentation and $S_t := \mathcal{S}(\cdot, t)$ the segmentation of frame I_t.

$$E(\mathcal{S}) := \sum_{t=1}^{T} E^t_{\text{Motion}}(S_t) + E^t_{\text{Appear}}(S_t) + E^t_{\text{Reg}}(S_t)$$

☞ **Motion cue:** This unary term is the soft labeling by the Random Walk. Let $P_0, P_1 \subset \Omega \times [1,T]$ denote the scribble locations in a video with T frames, Ω is the image space. Random Walk's output is a labeling $u_{\text{Motion}}\colon \mathcal{C} \to [0,1]$ to the set of trajectories $\mathcal{C}$. We transform u_{Motion} into the weighted segmentation bias $D_{\text{Motion}}\colon \mathcal{C} \to [-\infty, \infty]$:

$$D_{\text{Motion}}(q) := \frac{1}{u_{\text{Motion}}(q)} - \frac{1}{1 - u_{\text{Motion}}(q)}$$

The per pixel unary cost for frame I_t:

$$f^t_{\text{Motion}}(x) := \begin{cases} +\infty & \text{if } x \in P_0 \\ -\infty & \text{if } x \in P_1 \\ D_{\text{Motion}}(q) & \text{if } \exists q = (t_1, t_2, c) \in \mathcal{C} : \\ & t \in [t_1, t_2] \text{ and } x = c(t) \\ 0 & \text{otherwise.} \end{cases}$$

In addition, for temporal consistency we also use a optical flow based measure for pixels not captured by the trajectories.

$$E^t_{\text{Motion}}(S_t) := \alpha_{\text{Motion}} \cdot \langle f^t_{\text{Motion}}, S_t \rangle + \alpha_{\text{Flow}} \cdot \langle \phi^t_{\mathcal{S}}, c^t_w \rangle$$

☞ **Appearance cue:** we minimize the KL divergence between the colour distributions in successive frames. This enforces temporal consistency globally instead of a local per pixel unary cost (often noisy).

$$E^t_{\text{Reg}}(S_t) := \underbrace{\alpha_{\text{Reg}} \left(\frac{\beta_{\theta_t}}{\text{vol}(S_t)} + \frac{1 - \beta_{\theta_t}}{\text{vol}(1 - S_t)} \right)}_{\text{volume term}} + \underbrace{\langle g^t, |\nabla S_t| \rangle}_{\text{weighted TV}}$$

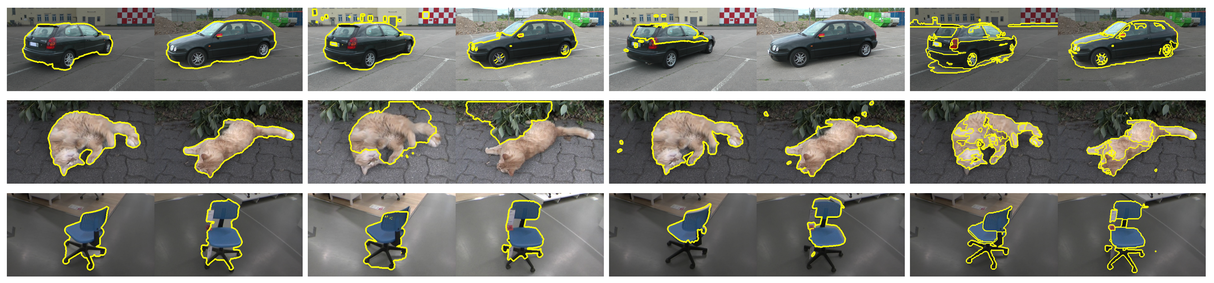
☞ **Volume term:** avoids degenerate segmentation $\emptyset, \Omega$.

☞ **Weighted TV:** weighted total variation corresponding to the geodesic contour length.

Optimization

☞ Since $E(\mathcal{S})$ is not convex, we use the Fast Trust Region[12] framework for optimization.

☞ Minimization on a GPU takes between 1 to 3 seconds per frame (∼2.0 Megapixels).



Ours | Godec et al.[11] | Papazoglou et al.[20] | Grundmann et al.[14]

Standard Benchmarks

We use the **P**ascal **O**verlap **M**easure (POM) between ground truth GT and segmentation S

$$\text{POM}(\text{GT}, S) := \frac{1}{T} \sum_{i=1}^{T} \frac{|\text{GT}_i \cap S_i|}{|\text{GT}_i \cup S_i|}.$$

Youtube Objects Dataset - Large dataset with 10 object categories. Ground truth from Jain et al.[15] was used.

Category	[11]	[20]	[15]	[19]	OF	Ours
Average	46.2	54.8	66.6	15.5	60.3	**74.1**

SegTrack2 - 14 low resolution videos with motion blur, rapid and articulated motion.

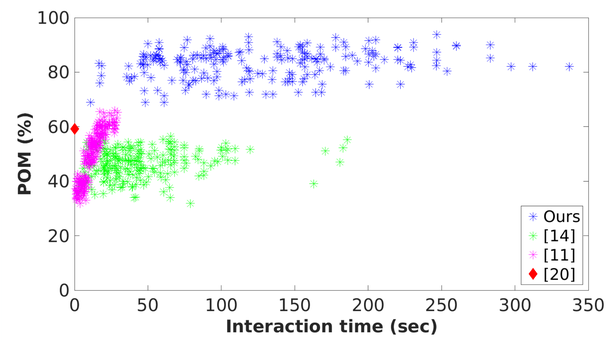
Method	[11]	[20]	[19]	OF	Ours
Average	41.3	53.5	8.0	40.1	**69.6**

Quantifying Interaction

☞ **Dataset:** 24 ego-motion videos across 4 categories - car,cat,chair,dog

☞ **Interaction data:** 5 different users

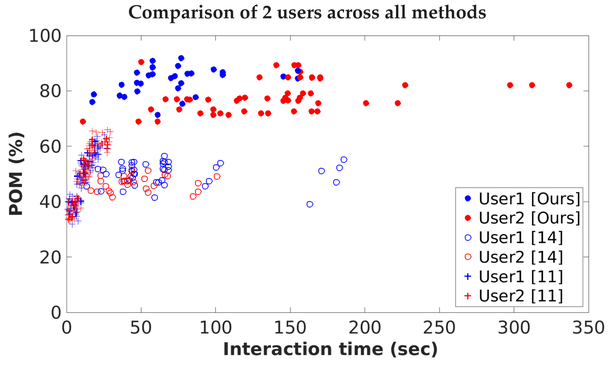
☞ **Comparisons:** ♦ unsupervised, ∗ superpixel based method, and ∗ bounding box tracker



☞ Each dot corresponds to an iteration per user per method. An iteration refers to one round of user interaction.

☞ Our approach gives a high quality segmentation with reasonable user effort. ☺

Comparison of 2 users across all methods



☞ Trained user achieves a high quality segmentation quicker than an untrained user.

☞ All approaches saturate below 100%.

References

[11] Godec et al., "Hough-based Tracking of Non-rigid Objects", *CVIU 2013*
[12] Gorelick et al., "Fast Trust Region for Segmentation", *CVPR 2013*
[14] Grundmann et al., "Efficient Hierarchical Graph-based Video Segmentation", *CVPR 2010*
[15] Jain et al., "Supervoxel-consistent Foreground Propagation in video", *ECCV 2014*
[19] Ochs et al., "Segmentation of Moving Objects by Long Term Video Analysis", *PAMI 2014*
[20] Papazoglou et al., "Fast Object Segmentation in Unconstrained Video", *ICCV 2013*