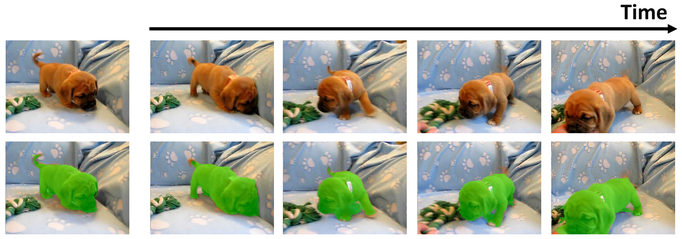
Supervoxel-Consistent Foreground Propagation in Video

Suyog Dutt Jain and Kristen Grauman
University of Texas at Austin
http://vision.cs.utexas.edu/projects/videoseg/

Problem

Automatic propagation of foreground segmentation in videos from a **single/multiple labeled frame (s)**.

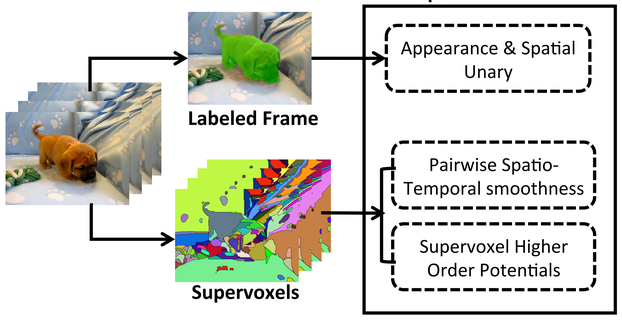


Existing methods [Tsai 2010, Fathi 2011, Vijayanarasmihan 2012] **can only enforce local consistency** in space and time (using pairwise connections).

Robust foreground propagation requires **capturing long range dependencies** as object evolves in shape over time.

Our Idea

Higher order potentials for supervoxels to discover long-range coherent regions.

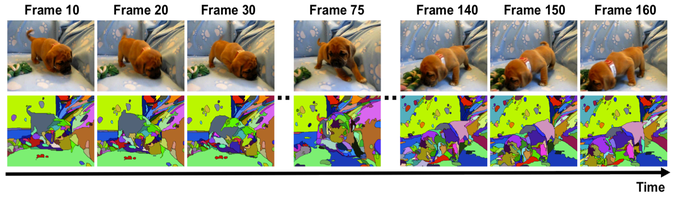


Enforce **long term temporal consistency** using higher order potentials defined over **supervoxel based cliques.**

Supervoxel cliques often **span long and broader areas in space and time**, hence better capture object's long term evolution in shape and appearance.

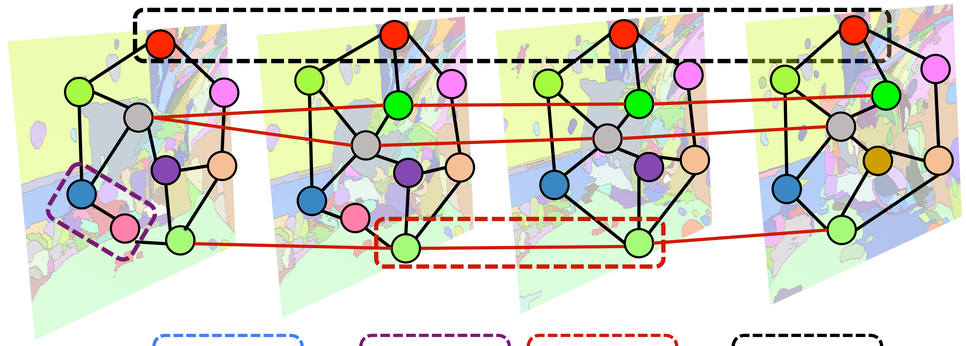
Supervoxels

Supervoxels are space-time regions computed with a bottom-up unsupervised video segmentation algorithm.



Uses image and motion features to produce a coarse to fine over-segmentation of the video volume (e.g., Grundmann et al.)

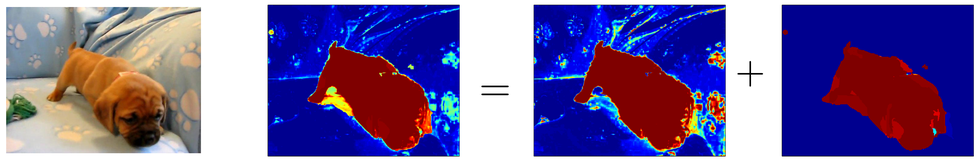
Approach



Appearance & Spatial Prior | Spatial Cliques | Temporal Cliques | Supervoxel Cliques

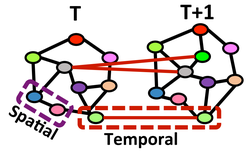
$$E(\mathcal{Y}) = \underbrace{\sum_{(t,i)\in\mathcal{X}} \Phi_t^i(y_t^i)}_{Unary\ potential} + \underbrace{\sum_{\substack{[(t,i),(t',j)]\in\mathcal{E} \\ t'\in\{t,t+1\}}} \Phi_{t,t'}^{i,j}(y_t^i, y_{t'}^j)}_{Pairwise\ potential} + \underbrace{\sum_{v\in\mathcal{S}} \Phi_v(y_v)}_{Higher\ order\ potential}$$

Unary Potential $\Phi_t^i(y_t^i) = \underbrace{\lambda_{app}\, A_t^i(y_t^i)}_{Appearance\ prior} + \underbrace{\lambda_{loc}\, L_t^i(y_t^i)}_{Spatial\ prior}$



Pairwise Potential $\Phi_{t,t'}^{i,j}(y_t^i, y_{t'}^j) = \delta(y_t^i \neq y_{t'}^j)\exp(-\beta_p \mathcal{D}(x_t^i, x_{t'}^j))$

- Contrast-sensitive smoothness potentials (based on feature distances).
- Encourages spatial and temporal smoothness in label propagation.



Higher Order Potential

- We adopt the Robust P^n model (Kohli et al. 2008) for supervoxel based higher order cliques.
- Using this, we assign **soft preferences for label consistency** among all superpixel nodes which belong to the same supervoxel.
- We modulate the higher order potential using supervoxel color consistency.

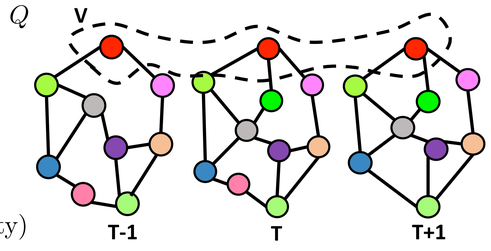
$$\Phi_v(y_v) = \begin{cases} N(y_v)\frac{1}{Q}\gamma_{\max}(v) & \text{if } N(y_v) \leq Q \\ \gamma_{\max}(v) & \text{otherwise} \end{cases}$$

$$N(y_v) = \min(|y_v = -1|, |y_v = +1|)$$

$$\gamma_{\max}(v) = |y_v| \exp(-\beta_h \sigma_v)$$

σ_v : total RGB variance in supervoxel v.

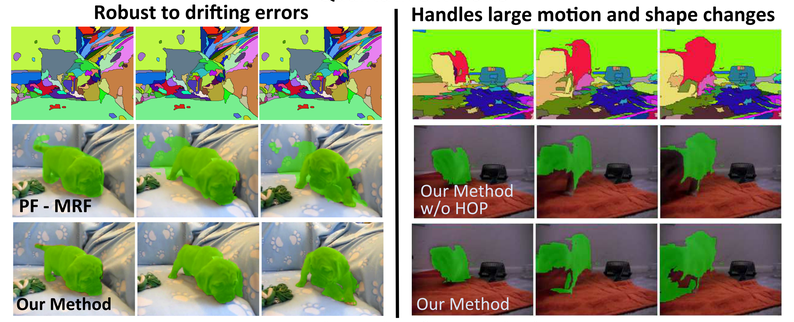
Q : Truncation parameter (controls rigidity)



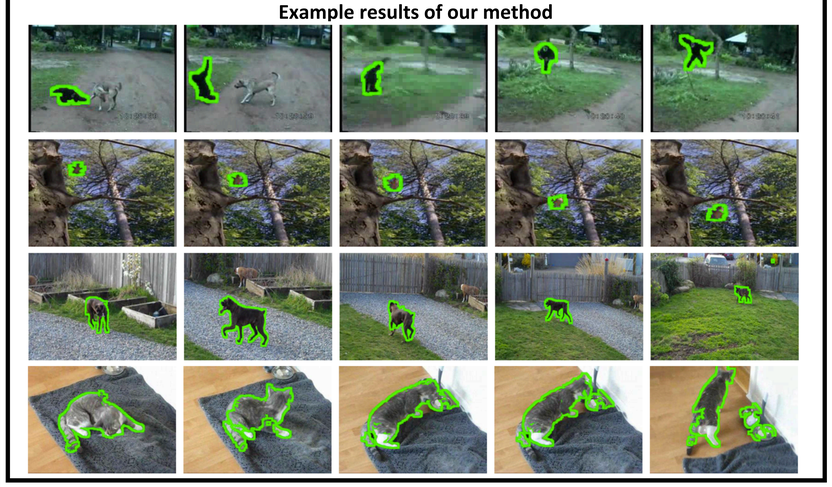
Baselines:

- SVX – Prop (Simple supervoxel propagation)
- SVX – MRF (MRF with supervoxel as nodes)
- PF - MRF (Vijayanarasimhan et al. 2012)
- Ours w/o HOP

Qualitative Results



Example results of our method



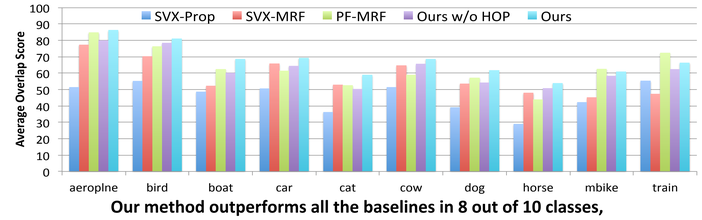
Results

SegTrack Dataset (6 videos, 243 frames):

Video	PF-MRF	Fathi	Tsai	SVX-Prop	SVX-MRF	Ours w/o HOP	Ours
birdfall	405	342	252	453	299	246	**189**
cheetah	1288	711	1142	1832	1202	1287	1170
girl	8575	1206	1304	5402	3950	3286	2883
monkeydog	1225	598	563	1283	737	389	**333**
parachute	1042	251	235	1480	420	258	**228**
penguin	482	1367	1705	541	491	497	**443**

Average pixel error (lower is better).

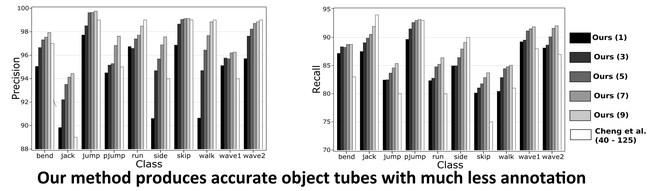
YouTube-Objects Dataset (126 videos, >10k frames):



Our method outperforms all the baselines in 8 out of 10 classes, with gains up to 8 points over the best baseline.

Ground truth pixel level object masks collected using mturk available for download.

Weizmann Dataset (90 videos, > 5k frames)



Our method produces accurate object tubes with much less annotation effort as compared to Cheng et al.