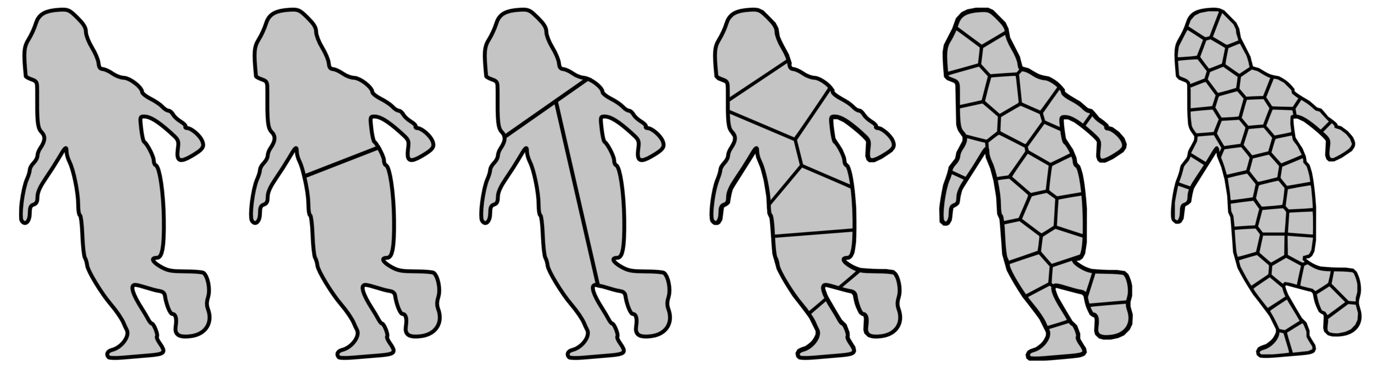
A Video Representation Using Temporal Superpixels

Jason Chang, Donglai Wei, & John W. Fisher III

Temporal Superpixels

A superpixel [4] is a set of pixels that are “local, coherent, and which preserve most of the structure necessary for segmentation.” Generic oversegmentations are shown below, with superpixels shown to the right.



Superpixels are often used as a proxy for pixels, and features within a superpixel (e.g. color) can often be assumed to be a constant. We believe that temporal superpixels (TSPs) can be used similarly. For example, one may want to approximate the motion of a TSP to be a constant translation. We believe a TSP representation can help in many computer vision tasks such as optical flow, object tracking, and video segmentation.

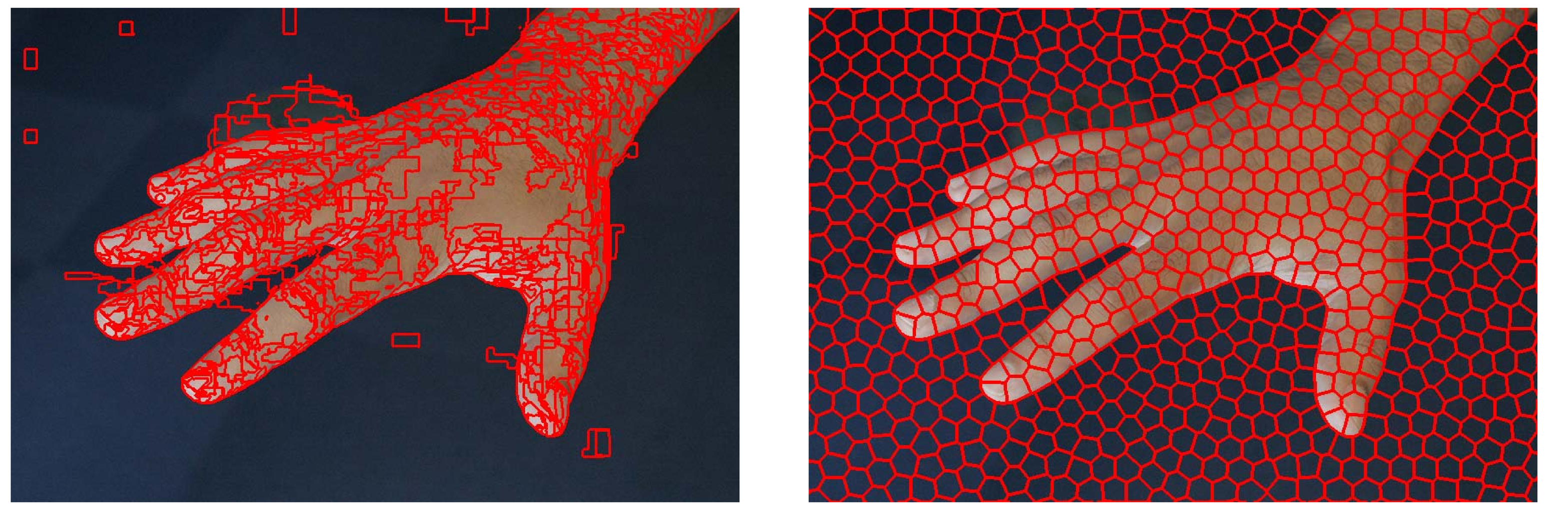
TSP Motivation

Why is a TSP representation needed with the supervoxel papers of [6-7]? TSPs are inherently different than the supervoxel methods of [6-7]. The supervoxel methods obtain accurate oversegmentations of videos, but do not necessary create a good representation. For example, the same supervoxel will often not correspond to the same part of an object in different frames. Additionally, a supervoxel may grow in space before growing in time. In contrast, TSPs are tracked for as long as they can be tracked for through the entire video.

	Supervoxels	TSPs
Video Oversegmentation	Yes	Yes
Model / Estimate Motion	No	Yes
Similarly-sized Superpixels	No	Yes
Specify # of Supervoxels	Yes	No
Specify Average # of Superpixels	No	Yes

We believe it is often better to specify the average number of superpixels per frame, since the number of supervoxels is highly dependent on the video (the number of frames, the amount of motion in the frame, etc.).

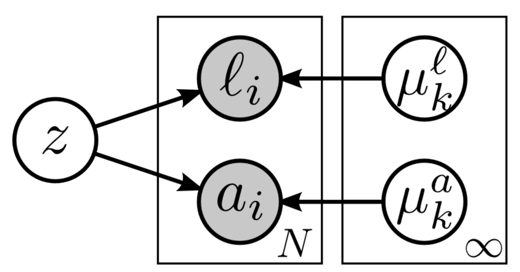
Supervoxels and the TSPs for a single frame are shown below. The unequal sizes of the SVs (one large background, and many small hand superpixels) can be undesirable. For example, approximating the SVs with constant translational flow may be poor.



Supervoxels [?] [?] | Temporal Superpixels

Single Frame

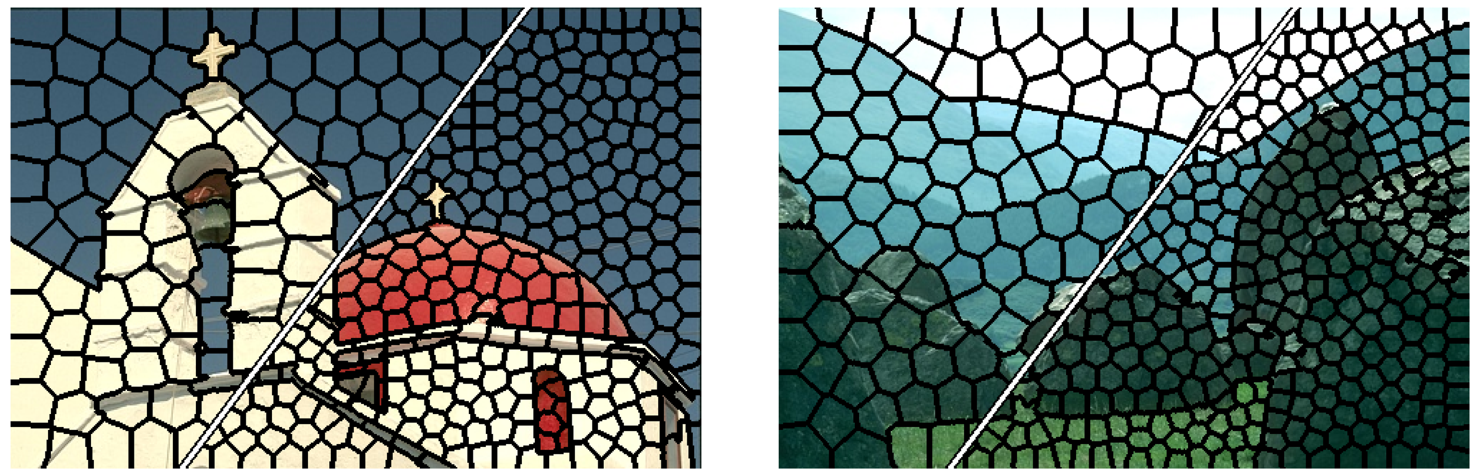
The Simple Linear Iterative Clustering (SLIC) [1] superpixel algorithm is extended to form a generative model for superpixels. Each pixel is modeled using a 5-D feature vector for each pixel is the 3-D color (a) and the 2-D location (ℓ). Each feature is modeled as being Gaussian with known variance, and superpixels are inferred by clustering of the mixture model. The following depicts a graphical model of the mixture model.



Changes to SLIC

- Valid *prior* over mean parameters
- *Topology constrained* prior on labels
- *Model order* term on number of unique labels
- *Joint optimization* of labels and parameters (local moves, split moves, merge moves)

Example superpixel segmentations are shown below. We require the user to specify an approximate number of desired superpixels, K.

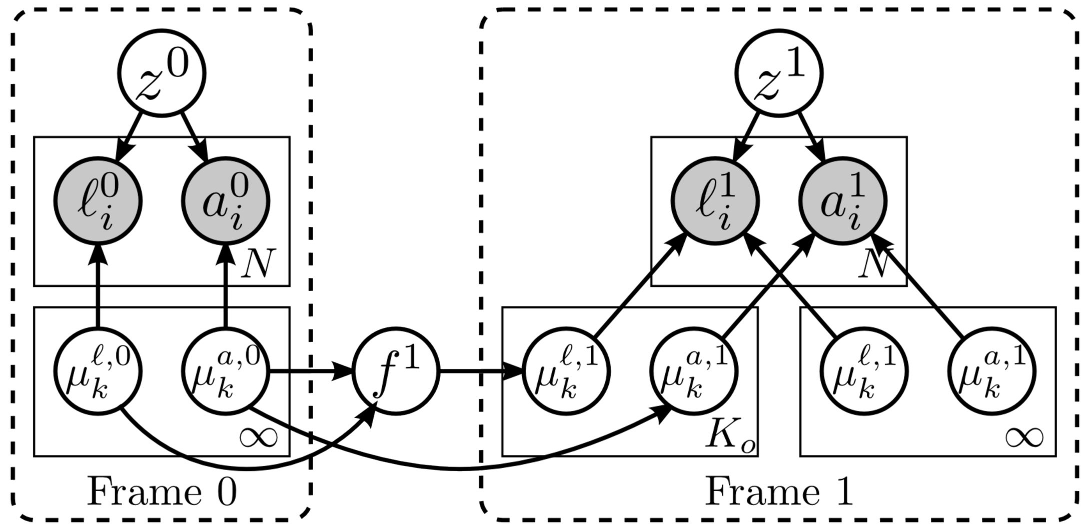


Multiple Frames

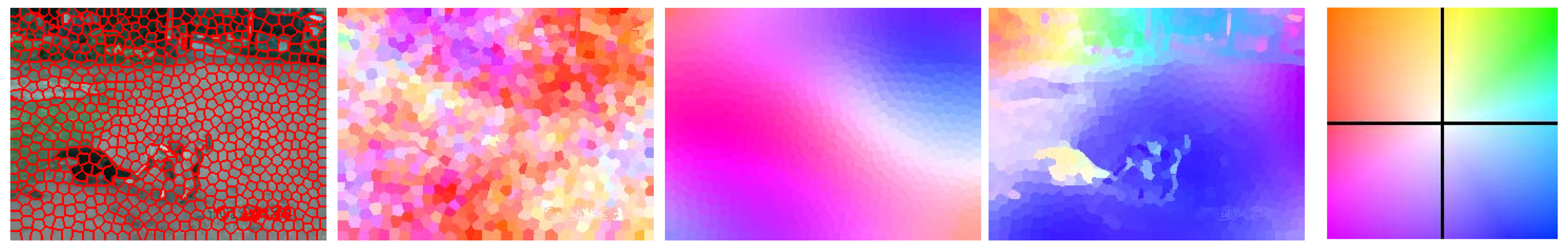
We now extend the single-frame model for superpixels to videos and TSPs. We assume that superpixels close in location and with similar colors should have similar motion. A *Gaussian Process* (GP) with the following *bilateral kernel* is used to model the motion between frames.

$$K(x_i, x_j) = \prod_{d=1}^{5} \exp\left[\frac{(x_i - x_j)^2}{2\sigma_d^2}\right] \qquad x_i = \{a_i, \ell_i\}$$

The appearance means are assumed to evolve independently. The corresponding graphical model is shown below, where f is the GP.



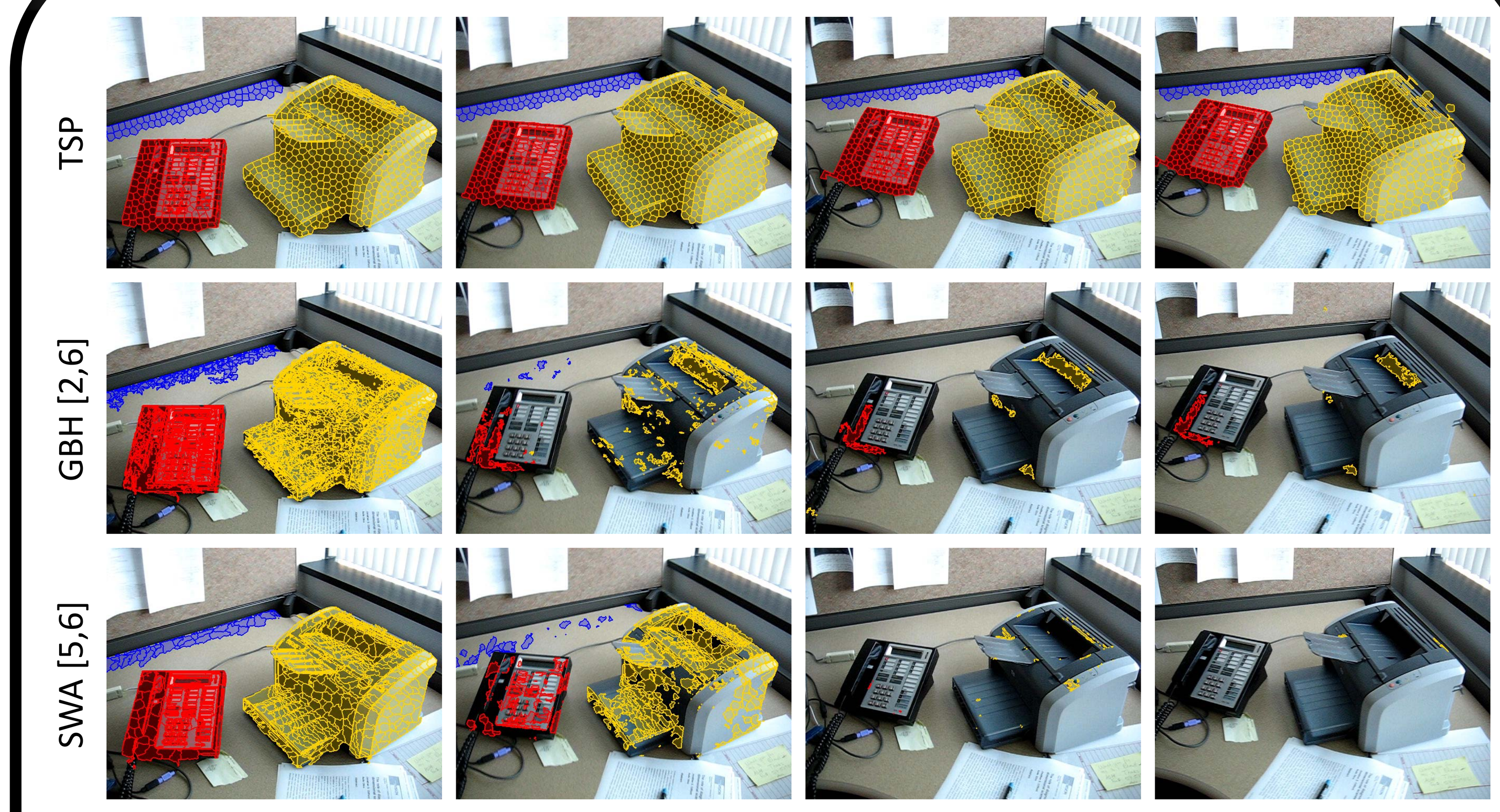
The following images show a sample from different GP priors. We consider GPs using a kernel similar to the Horn-Schunk [3] optical flow regularization, one that depends only on location, and the bilateral kernel.

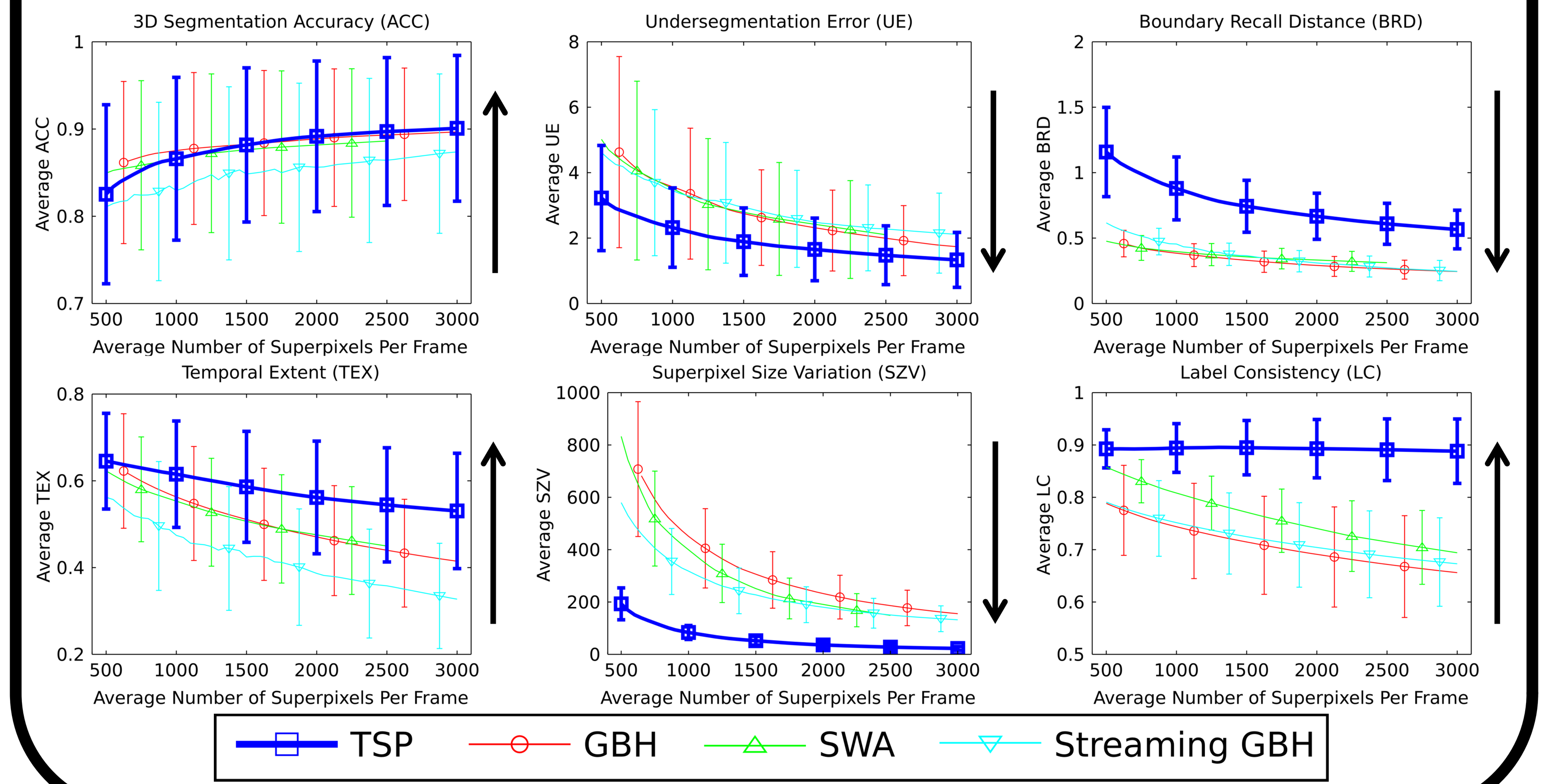


Superpixels | Horn-Schunk | Location-only | Bilateral Kernel

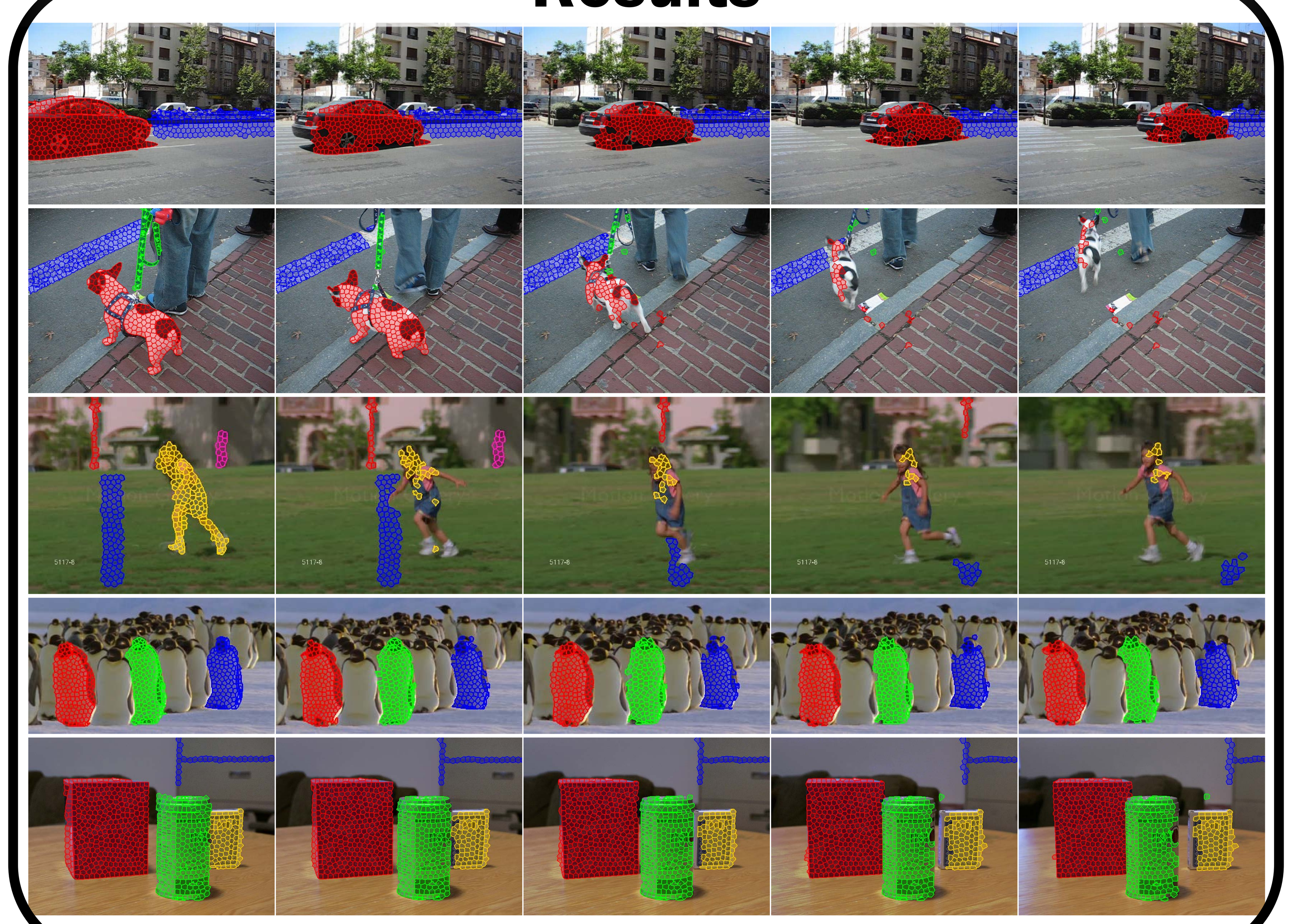
Code available at http://people.csail.mit.edu/jchang7/

TSPs vs. Supervoxels





Results



[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. PAMI 2012.
[2] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph based video segmentation. CVPR 2010.
[3] B. K. P. Horn and B. G. Schunck. Determining optical flow. Artificial Intelligence 1981.
[4] X. Ren and J. Malik Learning a classification model for segmentation. CVPR 2003.
[5] E. Sharon, A. Brandt, and R. Basri. Fast multiscale image segmentation. CVPR 2000.
[6] C. Xu, and J. Corso. Evaluation of super-voxel methods for eary video processing. CVPR 2012.
[7] C. Xu, C. Xiong, and J. Corso. Streaming hierarchical video segmentation. ECCV 2012.