University at Buffalo The State University of New York | REACHING OTHERS

CVPR 2012

Evaluation of Super-Voxel Methods for Early Video Processing

Chenliang Xu and Jason J. Corso
Department of Computer Science and Engineering - SUNY at Buffalo, Buffalo, NY

Links: PDF | LIBSVX Web | LIBSVX Code

Objective: The basic position of this paper is that supervoxels have great potential in advancing video analysis methods, as superpixels have for image analysis. To that end, we perform a thorough comparative evaluation of five supervoxel methods. We have also released the underlying methods' code and the benchmark.

What Makes a Good Supervoxel Method?

Why supervoxels?

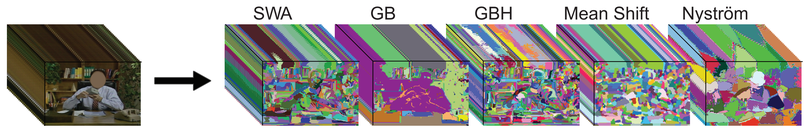
Images have many pixels; videos have more. Supervoxels have strong promise for early video processing: voxels are an artifact of the digital sampling process and not a natural representation, and there are many voxels in a video making sophisticated computational methods intractable.

Traits of a good supervoxel method.

- ***Spatiotemporal uniformity, or conservatism***, prefers compact and uniformly shaped supervoxels in space and time.
- ***Spatiotemporal boundary and preservation***: supervoxels should follow object and scene boundaries when they are present and should be stable when they are not present.
- ***Computation and performance***: computing supervoxels should reduce the overall amount of computation required and not decrease task performance.
- ***Parsimony***: the above properties should be maintained with as few supervoxels as possible.

Supervoxel Methods Evaluated:

We broadly sample the methodology-space, and intentionally select the methods with differing qualities for supervoxel segmentation in our analysis.



- ***Segmentation by Weighted Aggregation (SWA)*** solves the well-known normalized cut criterion approximately by sequentially computing a hierarchy of coarser segmentations. It applies algebraic multigrid techniques and recomputes affinity between regions at multiple scales in the hierarchy. The method was originally proposed by Sharon et al. CVPR 2000.

- ***Graph-based (GB)*** is a spatiotemporal extension of the Felzenszwalb and Huttenlocher (IJCV 2004) segmentation method, which iteratively computes a minimum spanning forest over the pixel lattice by merging similar regions.

- ***Graph-based Hierarchical (GBH)*** extends the GB method to sequentially compute a hierarchy of minimum spanning forests: the input graph at a level is the minimum spanning forest at the next finer level down. The method was proposed by Grundmann et al. CVPR 2010.

- ***Meanshift*** is a nonparametric mode-seeking method; we use Paris and Durand's (CVPR 2007) implementation that takes a Morse theory interpretation of the mean shift as a topological decomposition of the feature space.

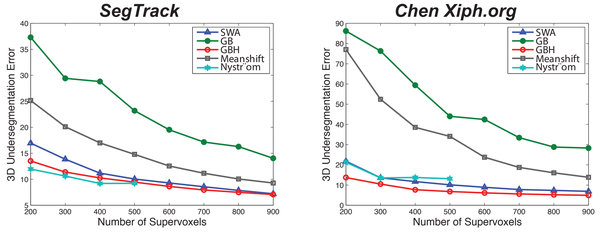
- ***Nyström*** approximately solves the normalized cut eigenproblem; each voxel is embedded into a low-dimensional eigenspace and then k-means clustering computes the final partitioning (Fowlkes et al. PAMI 2004).

The Supervoxel Benchmark and Quantitative Results:

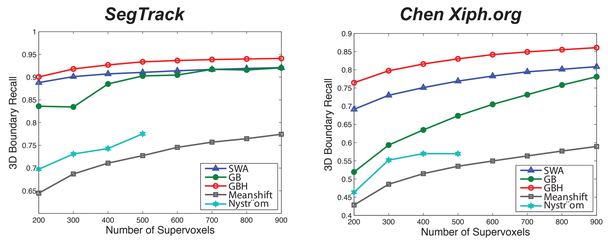
We propose a novel supervoxel benchmark that is not tied to any particular application but rather evaluates the desiderata described earlier. We evaluate the benchmark on three data sets:

- ***GaTech***: unlabeled videos.
- ***SegTrack***: labeled with a single foreground object.
- ***Chen Xiph.org***: fully labeled with region segmentations.

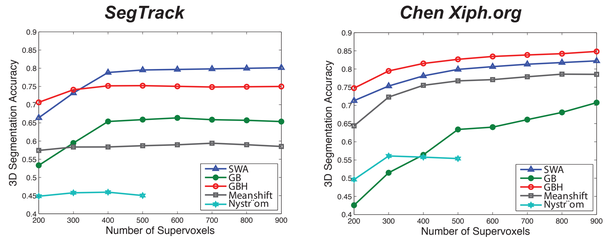
- ***3D Undersegmentation Error*** measures what fraction of voxels exceed the volume boundary of the ground-truth segment when mapping the supervoxels onto it.



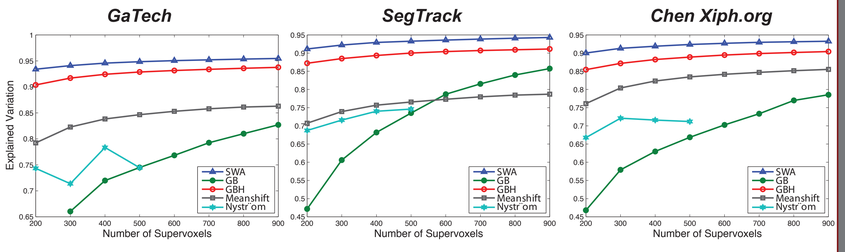
- ***3D Boundary Recall*** measures the spatiotemporal boundary detection: for each segment in the ground-truth and supervoxel segmentations, we extract the within-frame and between-frame boundaries and measure recall.



- ***3D Segmentation Accuracy*** measures what fraction of a ground-truth segment is correctly classified by the supervoxels: each supervoxel should overlap with only one object/segment.

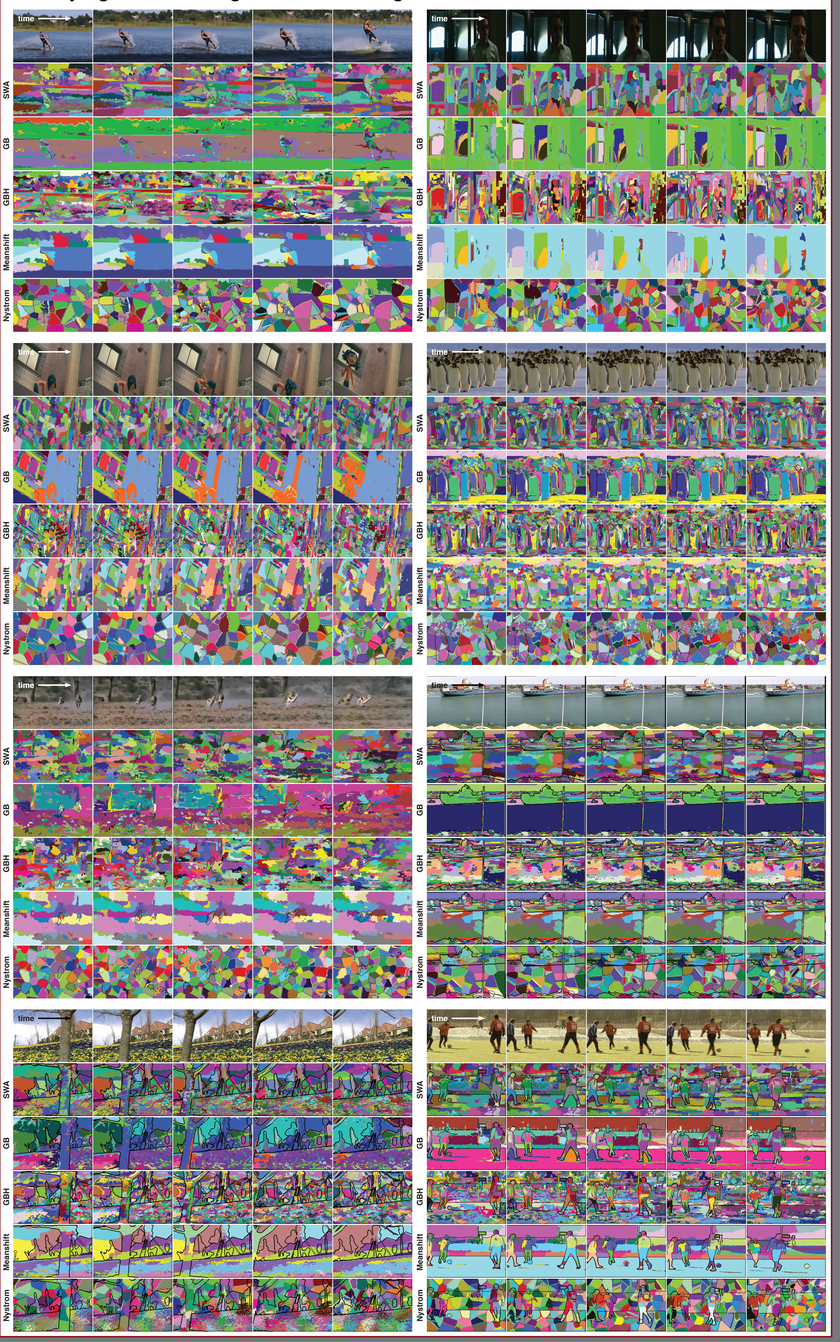


- ***Explained Variation,*** a human-independent metric, measures the difference between the image intensities and the mean-statistics of each supervoxel region; i.e., how well the original video is "compressed" by the supervoxel regions.



Take Away Message and Visual Examples:

Overall, the two best-performing methods are GBH and SWA. The common distinction setting these two methods apart is that they reevaluate region similarity at varying levels during hierarchical segmentation.



Acknowledgements. This work was partially supported by the National Science Foundation CAREER grant (IIS-0845282), the Army Research Office (W911NF-11-1-0090), the DARPA Mind's Eye program (W911NF-10-2-0062), and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069 The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, DARPA, ARO, NSF or the U.S. Government.