Video Segmentation

CVPR 2014 Tutorial

http://www.supervoxel.com/CVPR14_videoseg/

Jason J. Corso
Chenliang Xu
University of Michigan
jjcorso@eecs.umich.edu

Matthias Grundmann
Google Research
grundman@cc.gatech.edu

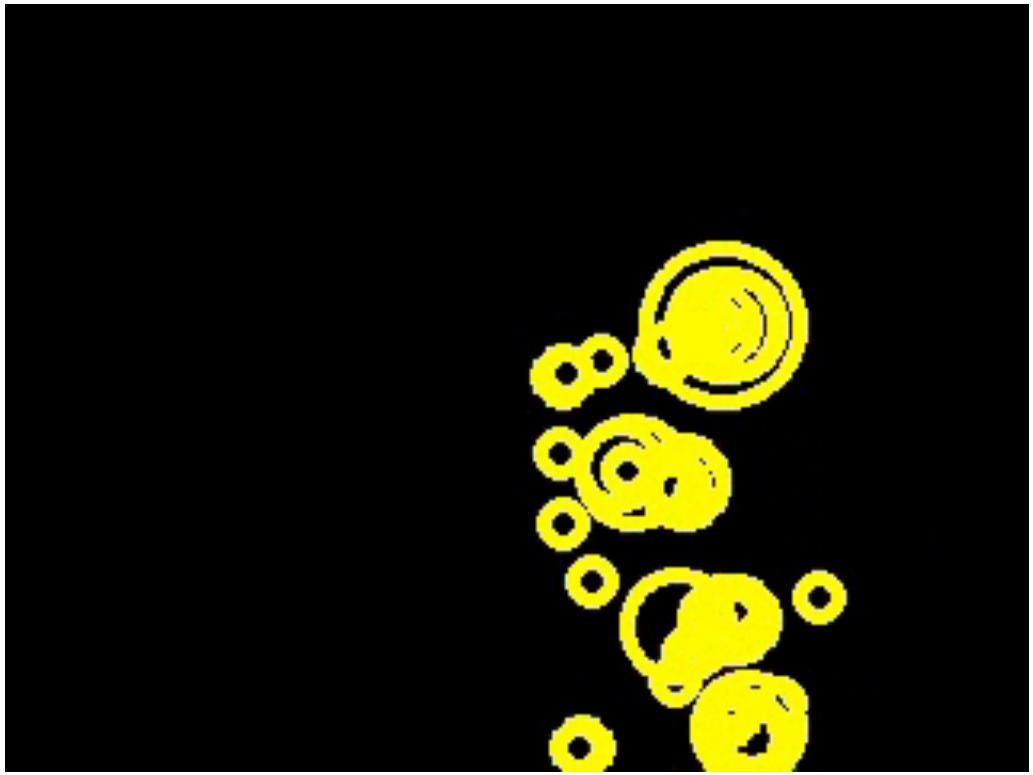
Irfan Essa
Georgia Tech
irfan@cc.gatech.edu

ELECTRICAL ENGINEERING & COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN

Research at Google

Georgia Tech College of Computing

Representing Video Content



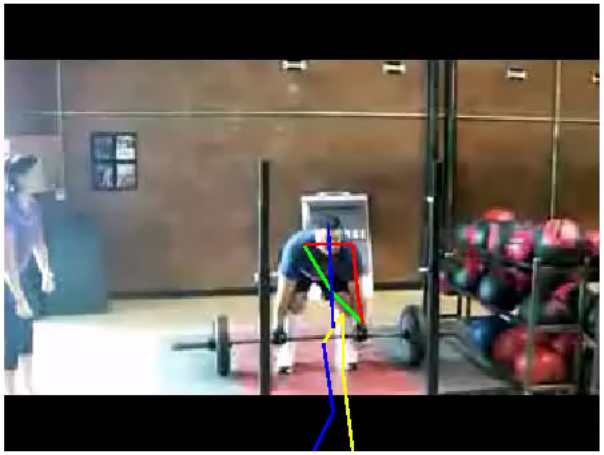
Method: Laptev. “On Space-Time Interest Points.” IJCV 64(2/3):107-123. 2005.

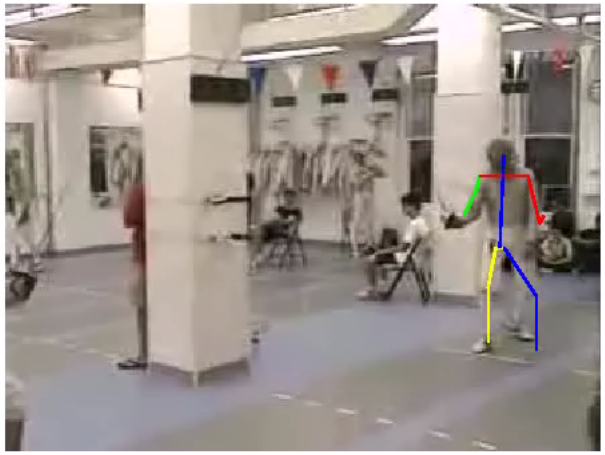
Representing Video Content

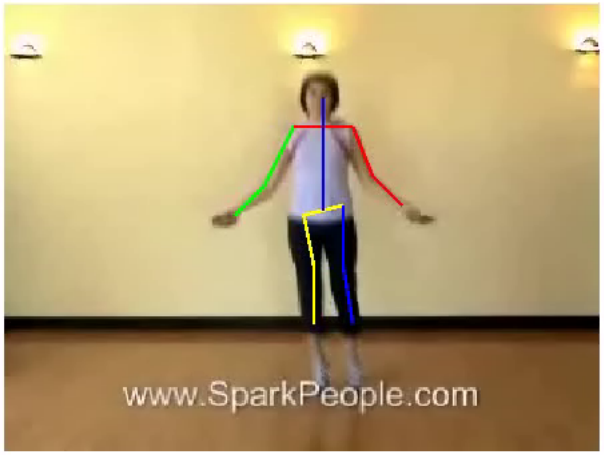
A *good* representation is paramount to *good* high-level video understanding.

Sources: Maas 1971 with Johansson; downloaded from Youtube.

Not There Yet...

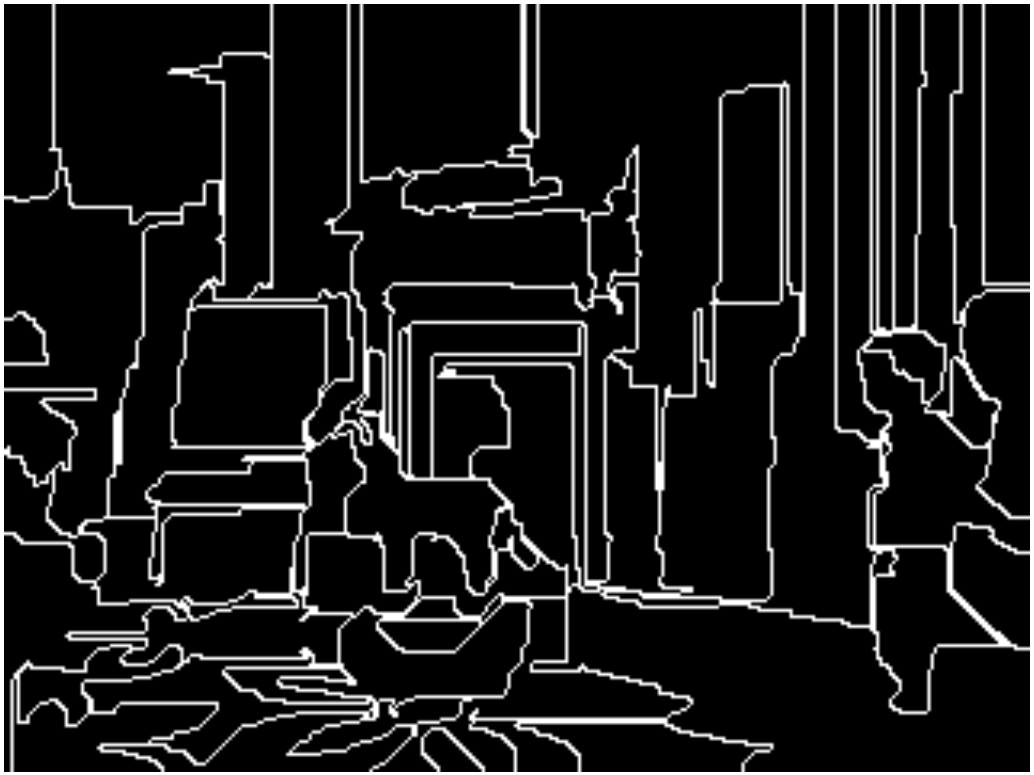






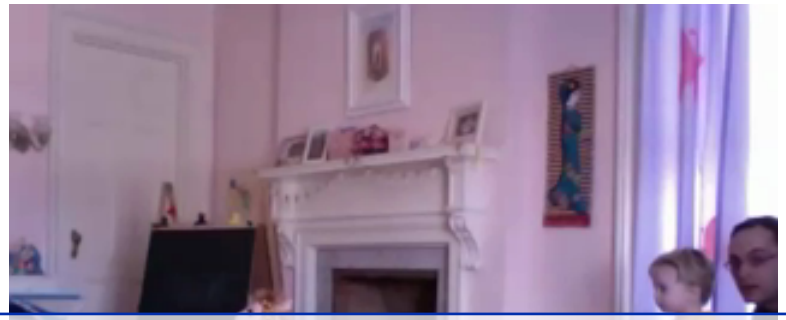
Method: Yang and Ramanan. “Articulated Pose Estimation with Flexible Mixtures-of-Parts.” CVPR 2011.

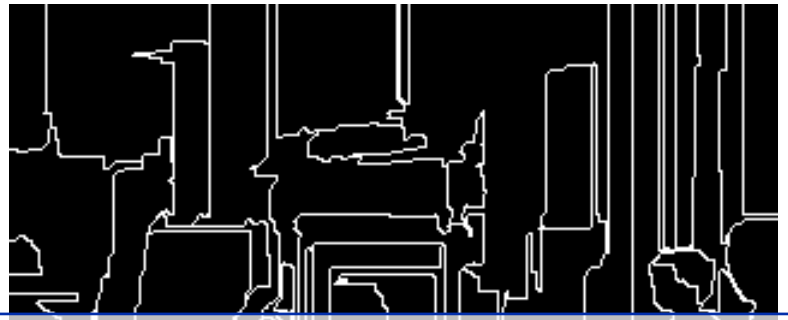
Alas, what makes such a good representation?

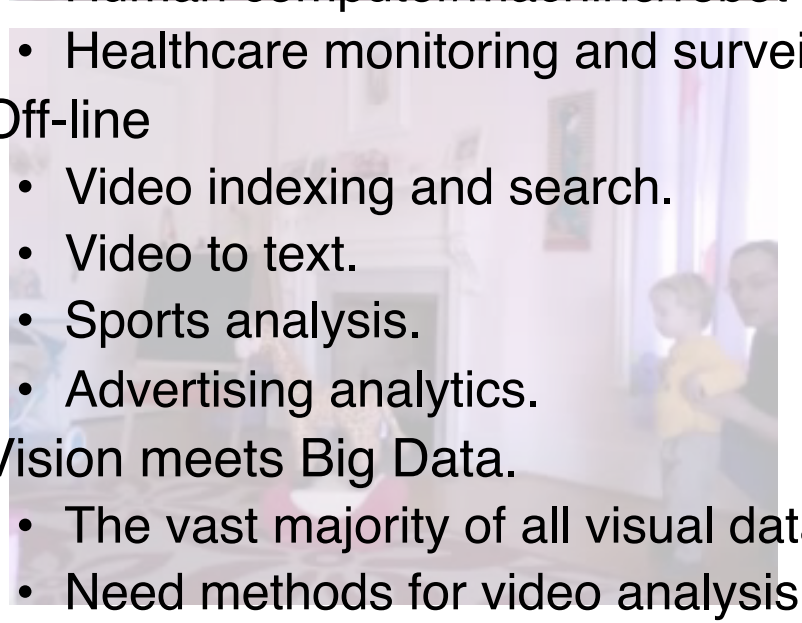


Method: Supervoxel segment boundaries. Xu and Corso CVPR 2012.

Segmentation: Toward a Rich Representation?







Applications of Video Understanding

- Real-time / Interactive
  - Mobile robotic guidance, navigation and manipulation.
  - Human computer/machine/robot interaction and entertainment.
  - Healthcare monitoring and surveillance.
- Off-line
  - Video indexing and search.
  - Video to text.
  - Sports analysis.
  - Advertising analytics.
- Vision meets Big Data.
  - The vast majority of all visual data is video data (YouTube: 72h/min).
  - Need methods for video analysis before we can handle the deluge.

Goals of the Tutorial

1. Spread the word on the advances of recent supervoxel methods—this tutorial is about an alternative representation of video content suitable for various subsequent inquiries.
2. Expose the vision audience to the how these methods can be used as an early step in various video analysis problems.
3. Introduce the software tools we have produced and released that are available to the community.

Distinct Types of Video Segmentation

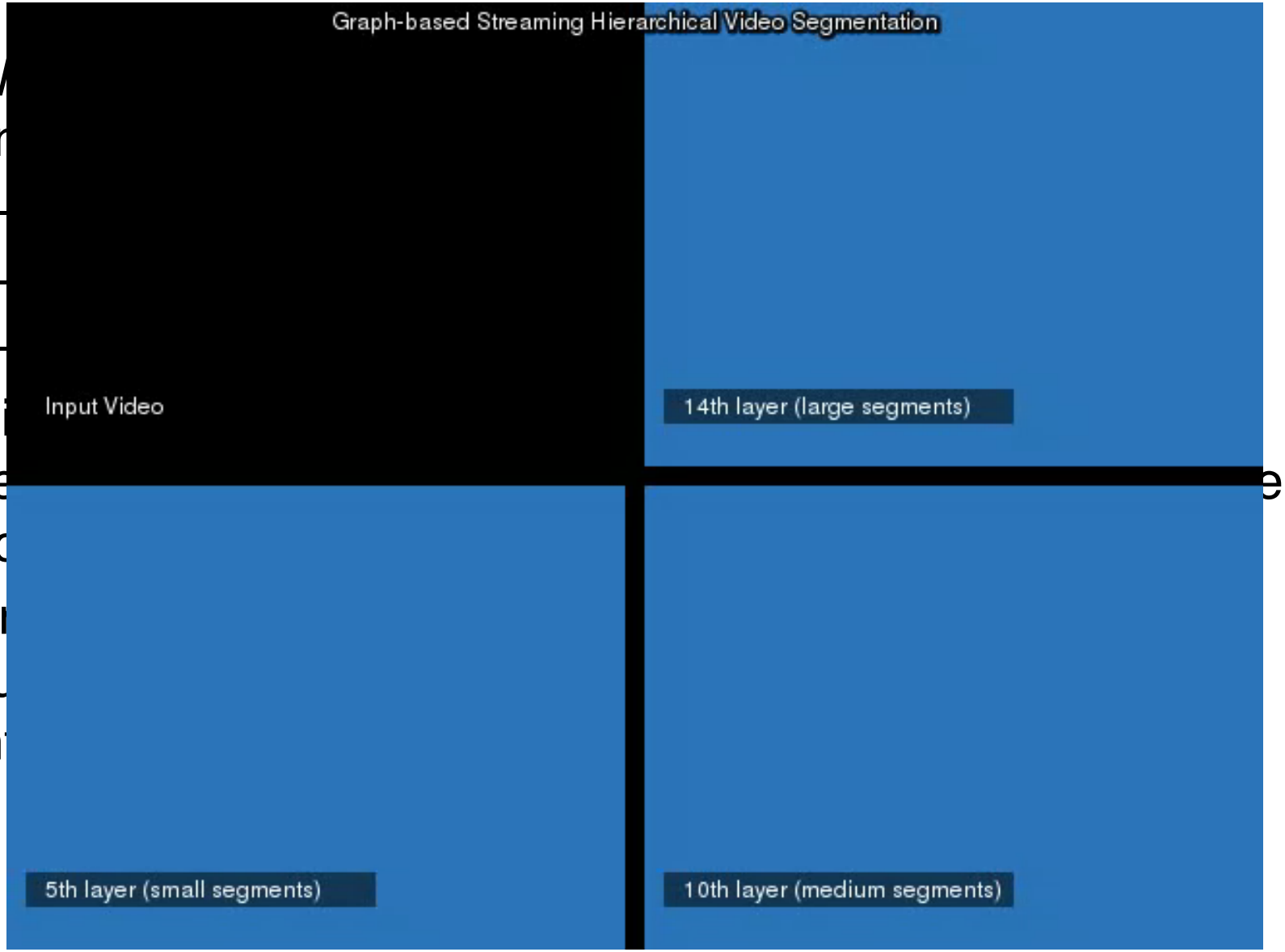
- Shot Segmentation
- Motion Segmentation
- Supervoxel *Over-Segmentation*
- *Video* Segmentation
- Semantic Segmentation

Tutorial Plan

1:00 – 1:30	Introduction	Jason
1:30 – 2:00	Graph-Based Hierarchical Segmentation	Matthias
2:00 – 2:30	Segmentation by Weighted Aggregation	Jason
2:30 – 3:00	Other Methods/Topics	Jason
3:00 – 3:30	Coffee Break	
3:30 – 4:15	Applications of Video Segmentation	Irfan & Matthias
4:15 – 4:45	LIBSVX and Evaluation	Chenliang
4:45 – 5:00	Wrap-Up	All

Supervoxels: A Complementary "Feature"?

- W
  ur
  -
  -
  -
- Di
  se
  do
- Pr
  su
  in



What makes a good spatial segmentation method?

- Rationale for oversegmentation
  - Pixels are not natural elements in images.
  - The number of pixels is very high.
- **Spatial uniformity** – prefers compact and uniformly shaped superpixels.
  - Embeds basic Gestalt principles of continuity, closure, etc.
- **Spatial boundary preservation –** as superpixel boundaries should align with perceptual boundaries when present and should be stable when they are not.
- **Computation –** the overall computational cost for a particular application should be reduced via superpixels.
- **Performance –** the overall performance of a method should be increased.
- **Parsimony –** The above properties should be maintained with as few superpixels as possible.

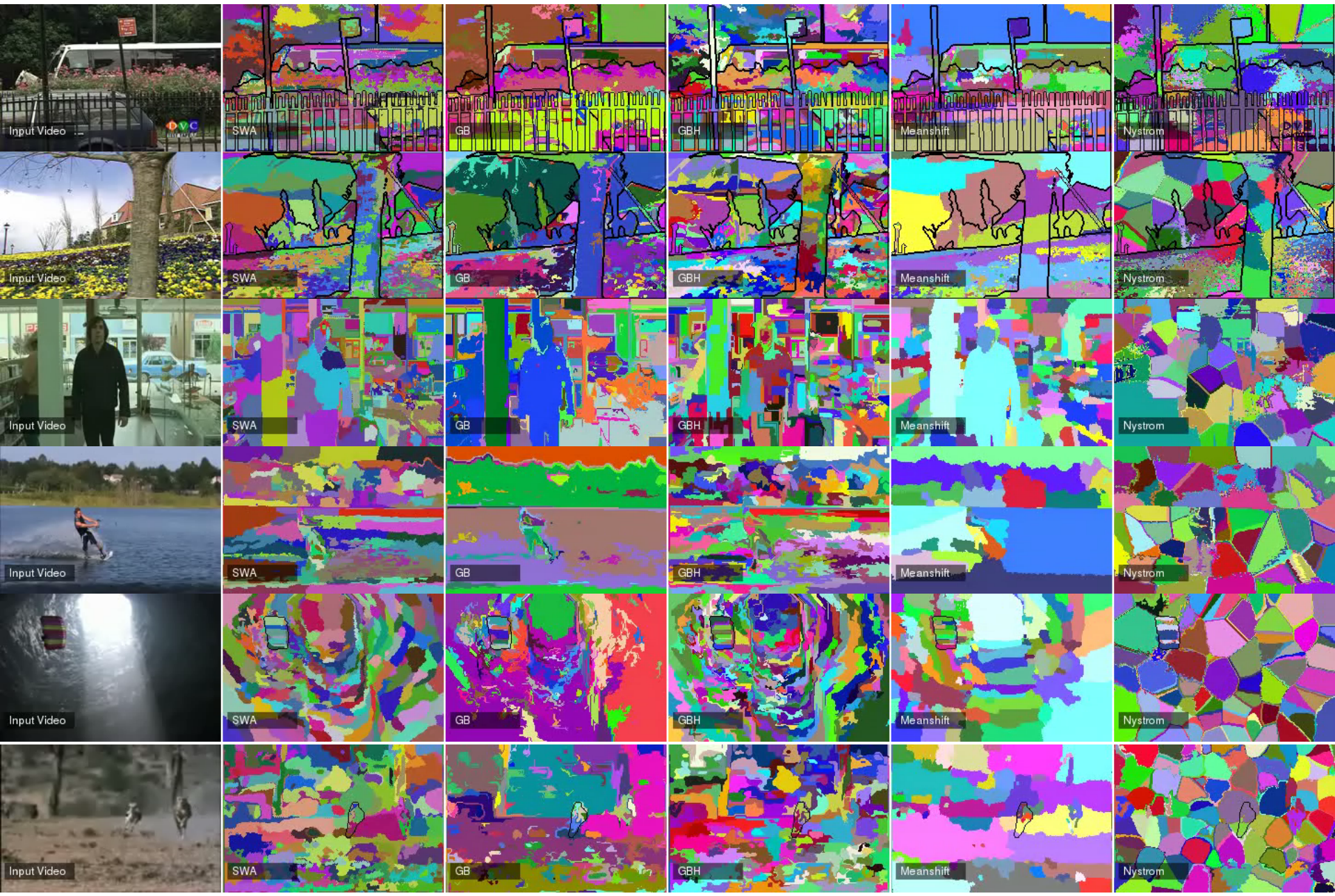
What makes a good space-time segmentation method?

- Rationale for oversegmentation
  - Voxels are not natural elements in video.
  - The number of voxels is very high.
- **Spatiotemporal uniformity** – prefers compact and uniformly shaped supervoxels.
  - Embeds basic Gestalt principles of continuity, closure, etc.
- **Spatiotemporal boundary preservation –** as supervoxel boundaries should align with perceptual boundaries when present and should be stable when they are not.
- **Computation –** the overall computational cost for a particular application should be reduced via supervoxels.
- **Performance –** the overall performance of a method should be increased.
- **Parsimony –** The above properties should be maintained with as few supervoxels as possible.

Evaluating Standard Methods.

- Meanshift
  - Fukunaga and Hostetler, Comaniciu and Meer, Wang et al.
- Graph-based / Minimum Spanning Forest
  - Felzenswalb and Huttenlocher.
  - Arguably the most popular superpixel method.
- Hierarchical graph-based
  - Grundmann et al.
- Nyström normalized cuts.
  - Shi and Malik, Fowlkes et al.
- Segmentation by weighted aggregation
  - Sharon et al., Corso et al.

[Xu and Corso CVPR 2012]



[Xu and Corso CVPR 2012]

LIBSVX: Library and Benchmark

- We implemented a set of quantitative evaluation benchmarks to assess these five methods against the properties discussed earlier.
  - 3D undersegmentation error.
  - 3D boundary recall.
  - 3D segmentation accuracy.
  - Explained variation (human independent).
- Three data sets

	Human Annotation	No. Videos	Mean FPV
SegTrack	Single Object	6	41
GaTech	None	15	86
Chen Xiph.org	Full Scene Segments	8	85

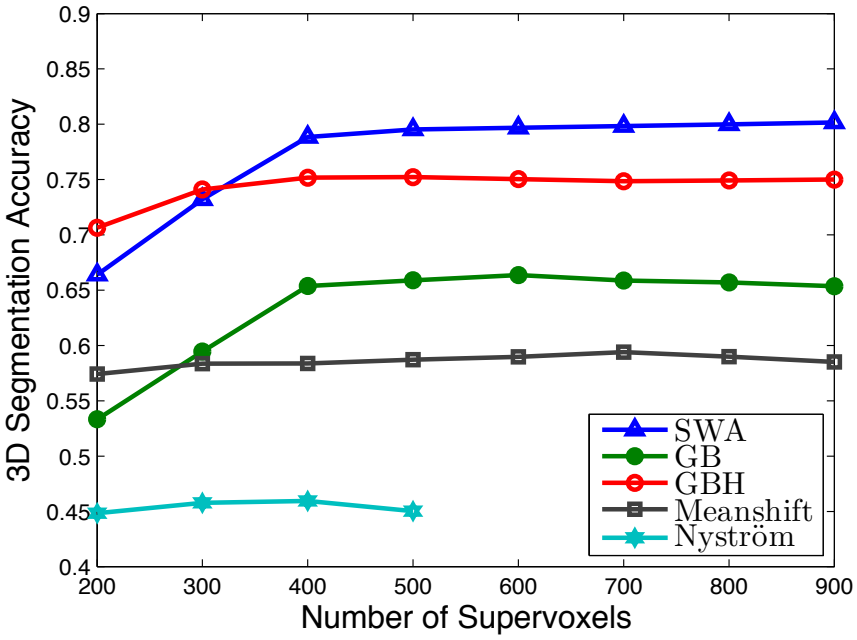
[Xu and Corso CVPR 2012]

Chen Xiph.org Video Example

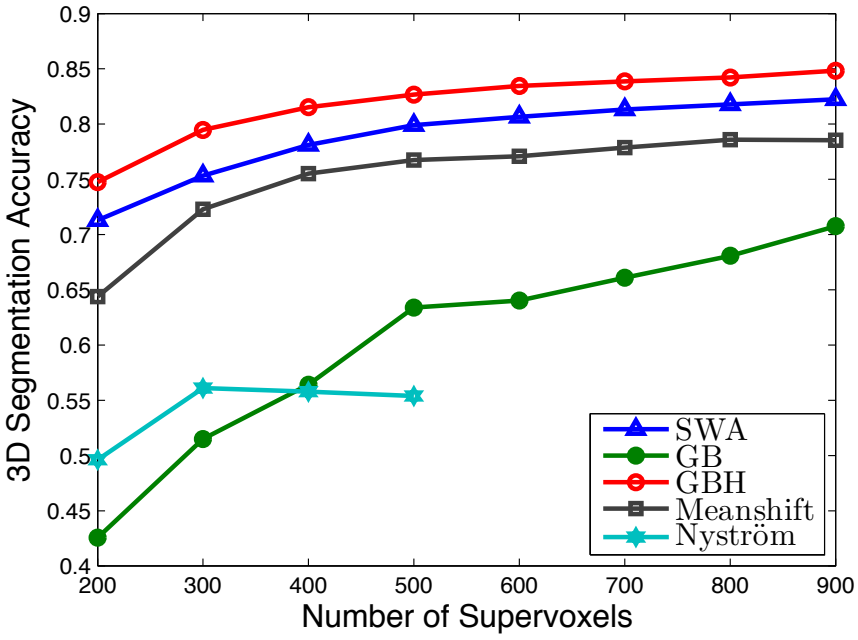
[Xu and Corso CVPR 2012]

Quantitative Evaluation Results

- 3D Segmentation Accuracy



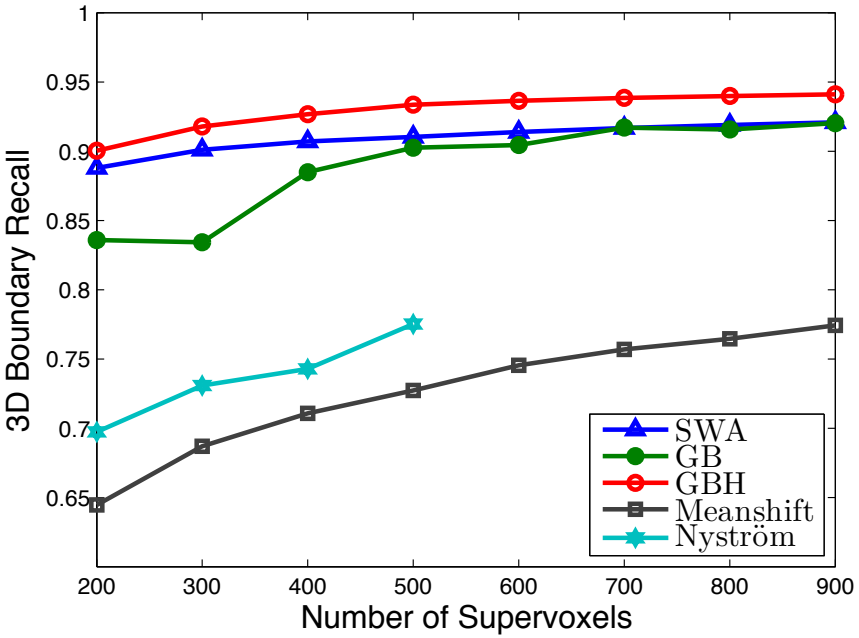
SegTrack

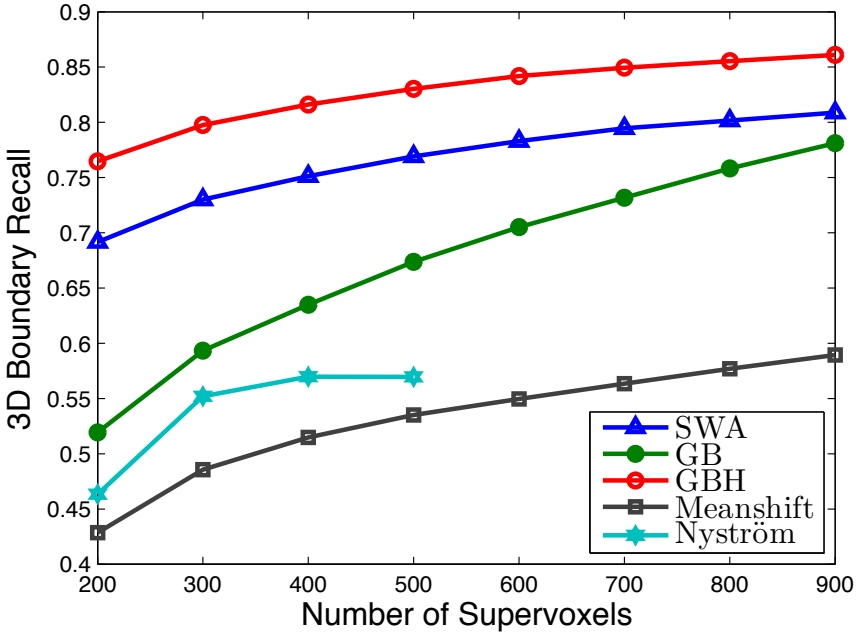


Chen Xiph.org

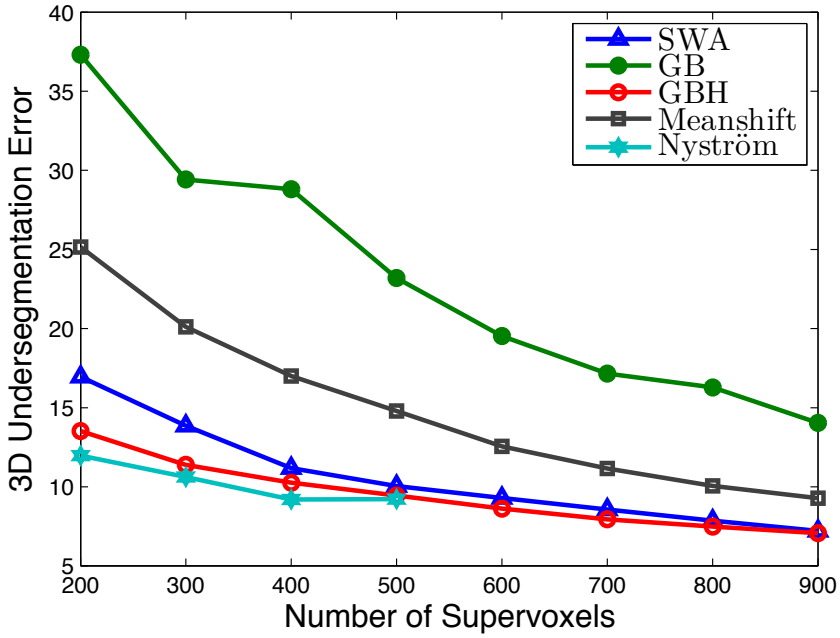
[Xu and Corso CVPR 2012]

- 3D Boundary Recall (Top) and 3D Undersegmentation Error (Bottom)

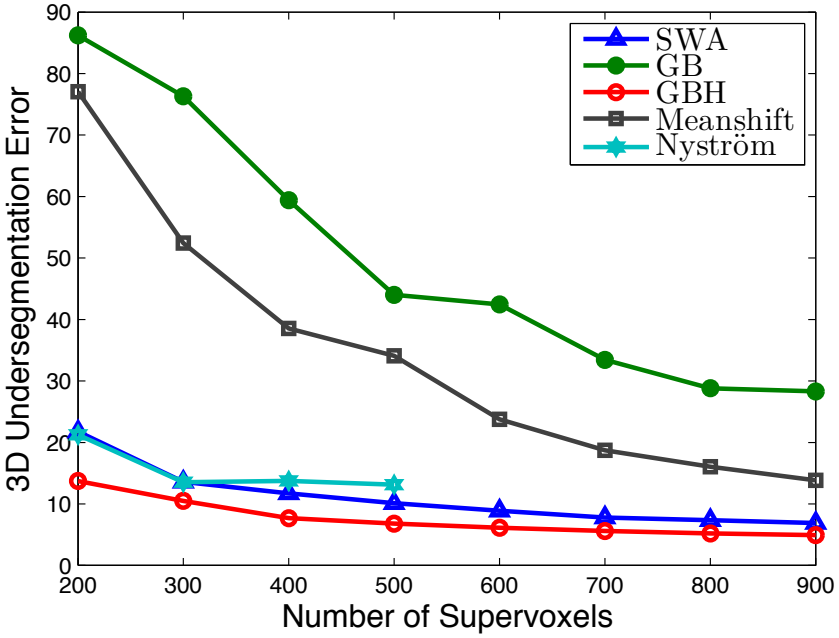




SegTrack

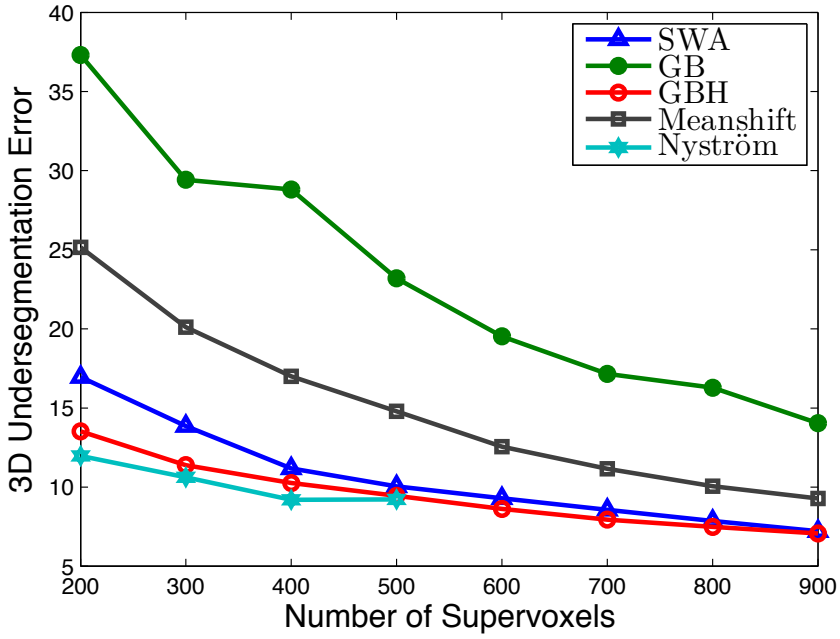


Chen Xiph.org

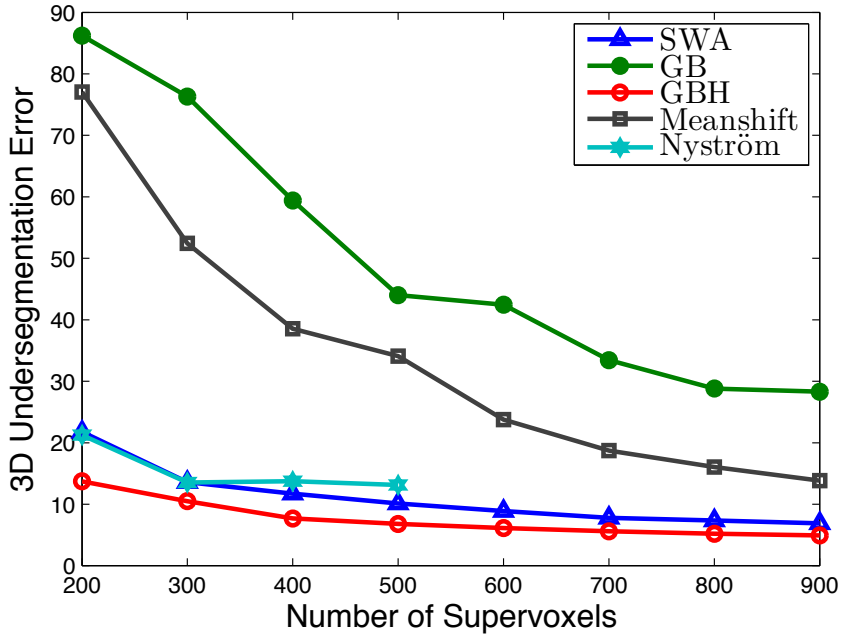


Quantitative Evaluation Results

- 3D Undersegmentation Error



SegTrack

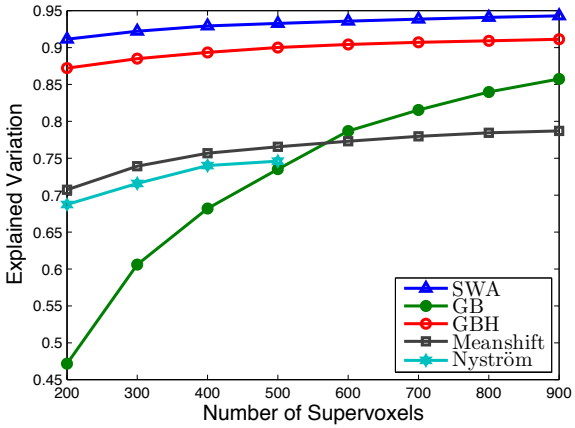


Chen Xiph.org

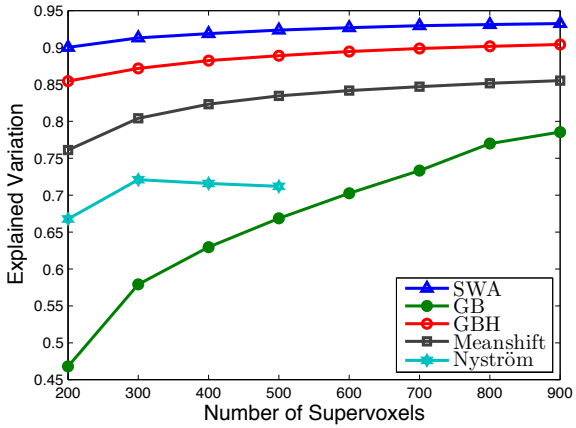
[Xu and Corso CVPR 2012]

Quantitative Evaluation Results

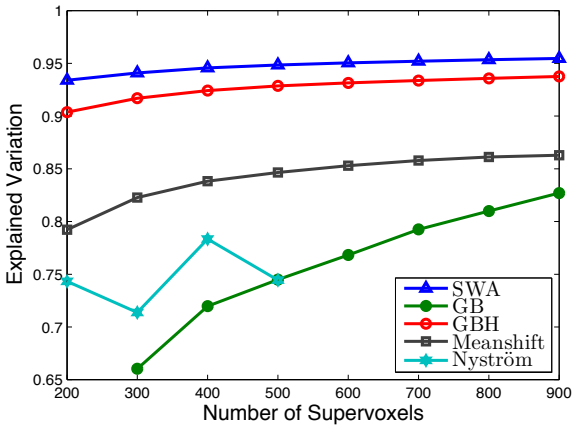
- Explained Variation



SegTrack



Chen Xiph.org



GaTech

[Xu and Corso CVPR 2012]

Key Finding

- SWA and GBH systematically outperform the other three methods, despite sharing some similarities. E.g.,
  - SWA and Nyström both minimize the normalized cut function.
  - GB and GBH are both based on the MST grouping method.
  - SWA, GBH, MeanShift are all hierarchical.
- Why?
- The characteristic that separates them is that they **recompute similarity/affinity at multiple scales** as the hierarchy increases incorporating coarse and fine information.

[Xu and Corso CVPR 2012]

Detailed Methods to the Discussed

- Hence, we will focus on two methods in detail in the tutorial
  - Graph-based hierarchical segmentation
  - Segmentation by Weighted Aggregation
- We will discuss other variants and applications of these.
- Disclaimer: this is not to say there are no other methods we should also be discussing. But, given time and goals, this is what we choose. Examples you may consider reading:
  - Galasso et al. ACCV 12
  - TSP Chang et al. CVPR 13
  - Video Seeds, Van der Bergh et al. ICCV 13
  - Trajectory Binary Partition Tree, Palou and Salembier CVPR13
- However, first, some results on why you should care of supervoxels as a feature.

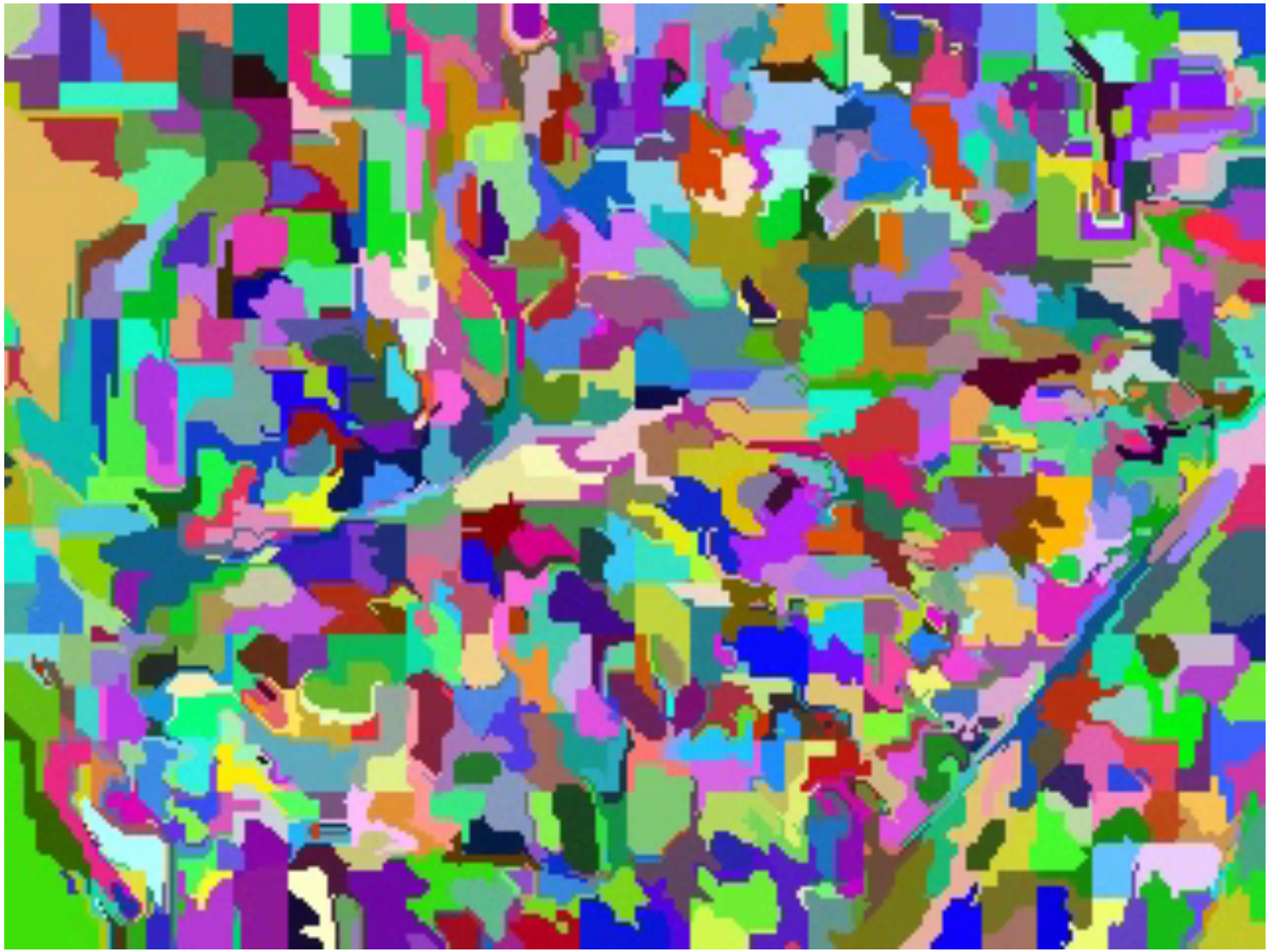
A Study on Human Supervoxel Perception

Are Actor and Action Semantics Retained in Video Supervoxel Segmentation?

Can Humans Perceive Action from Supervoxels?

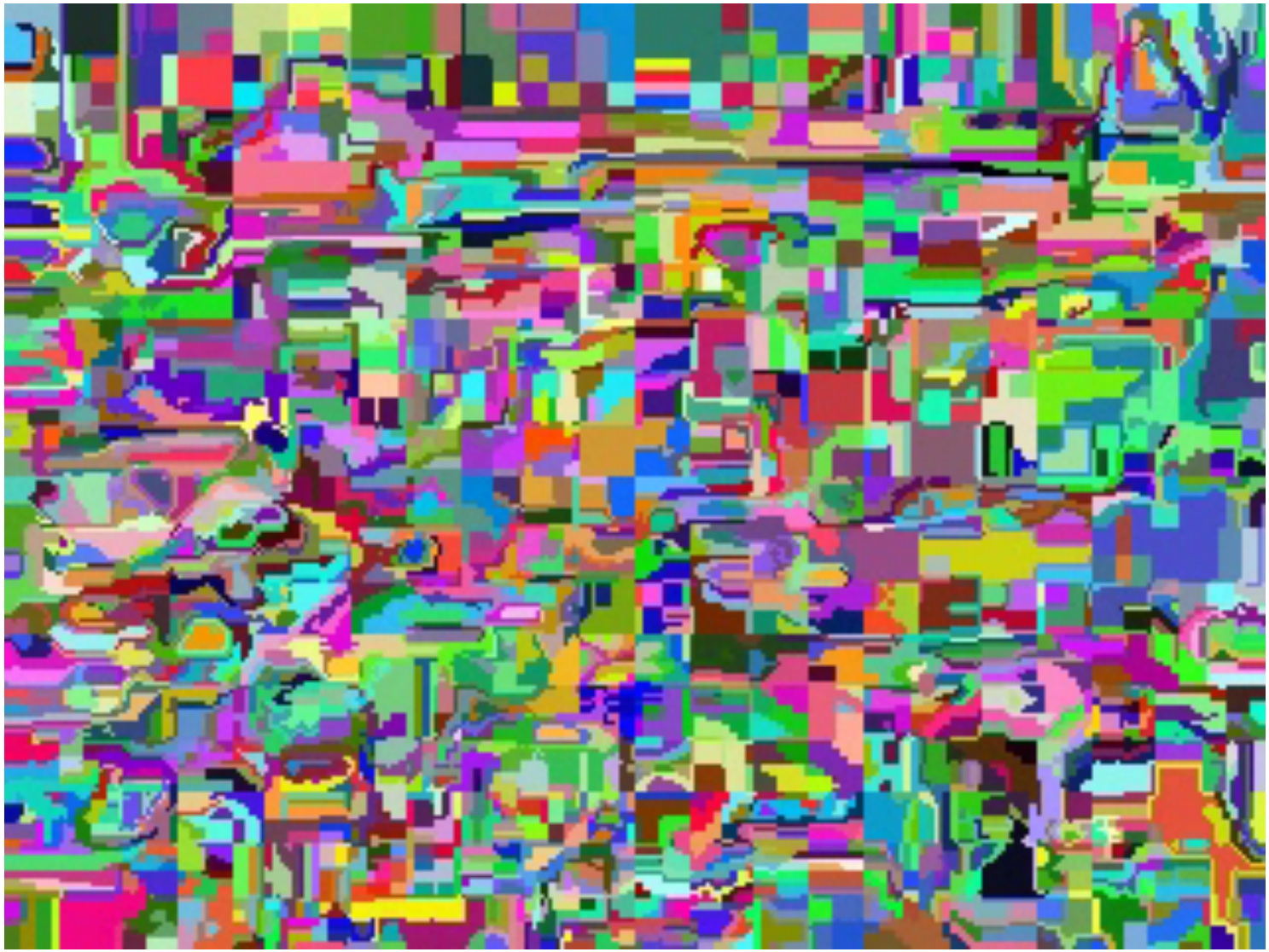
Can Humans Perceive Action from Supervoxels?

Can Humans Perceive Action from Supervoxels?



Can Humans Perceive Action from Supervoxels?

Can Humans Perceive Action from Supervoxels?



Can Humans Perceive Action from Supervoxels?

Study Questions

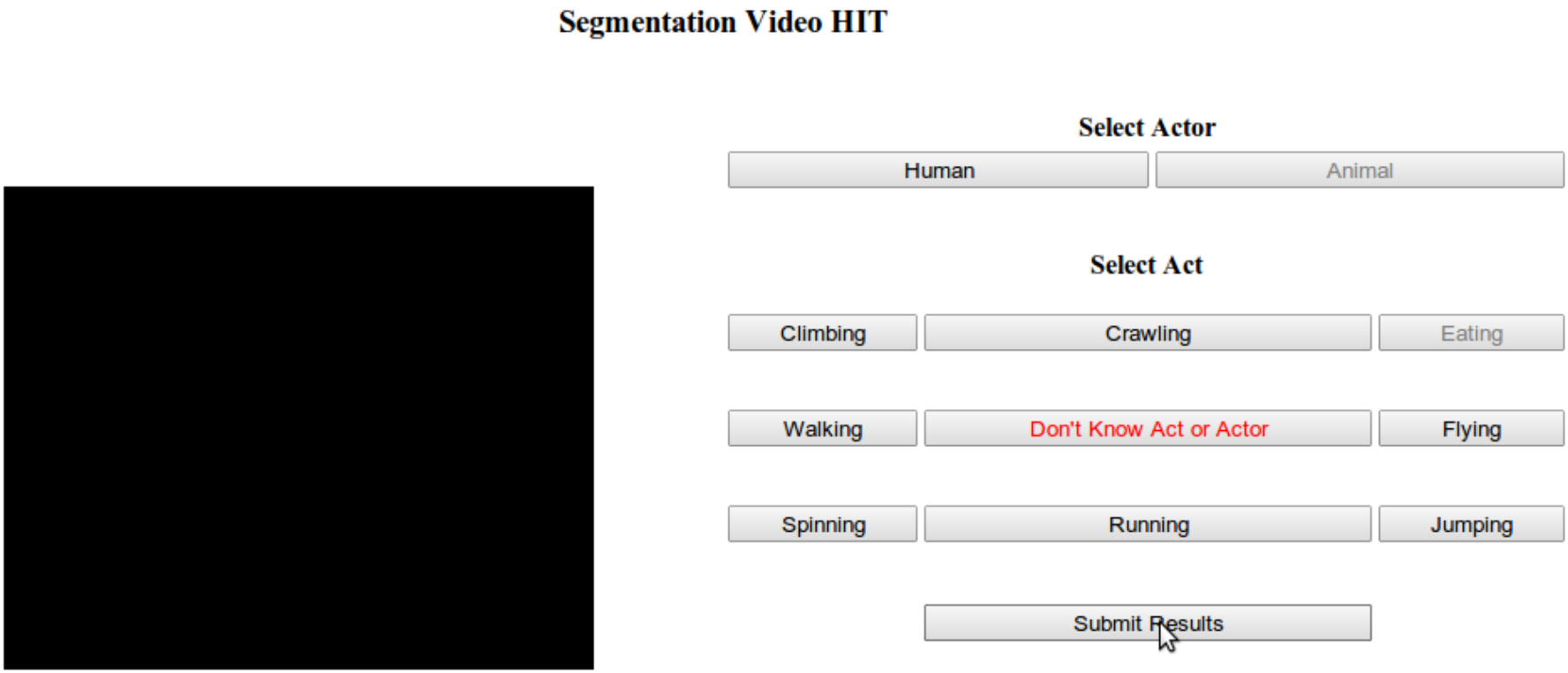
- Primary Question:
  - Do the segmentation hierarchies retain enough information for the human perceiver to recognize
    - Actor? (human or animal)
    - Act? (forced-choice one of eight)
- Secondary Questions:
  - How does the human performance vary with density of the supervoxels?
  - How does the human performance vary with actor?
  - How does human performance vary with static versus moving background?
  - How does speed vary with act? with correctness?

Study Setup: Participant Cohort and Data Set

- Study cohort of 20 college-age participants.
  - No student is studying segmentation.
- Data Set
  - Stratified according to Actors, Acts, and Background
    - **Actors:** human or animal
    - **Background:** static or moving
    - **Acts:** climbing, crawling, eating, flying, jumping, running, spinning, walking.
  - Sample 3 levels of the segmentation hierarchy (coarse, medium, and fine).
  - In total, we have 96 videos
    - 2 actors * 2 backgrounds * 8 acts * 3 levels

Study Setup: The Interface and Instructions

- Web-based interface
- Each participant is shown 32 videos and sees a given (input) video only once (in a single segmentation level).
- Participants never see the input RGB videos.



[Xu, Doell,… Corso ICSC 2013 and IJSC In Press]

Study Results: Actor Discrimination

- High actor discrimination rate: 82.4% overall accuracy.

	un	hu	an
unknown	0	0	0
human	0.11	0.86	0.03
animal	0.17	0.05	0.78

[Xu, Doell,… Corso ICSC 2013 and IJSC In Press]

Study Results: Act Discrimination

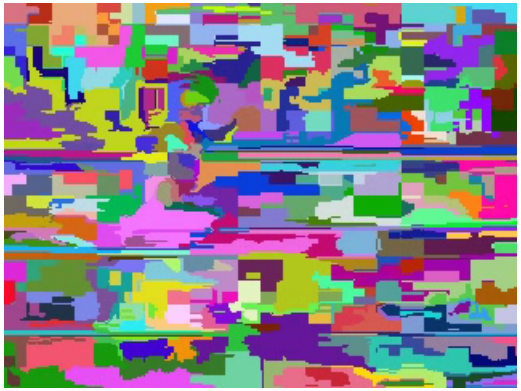
- Overall act discrimination rate: 70.5%.

	un	wl	sp	rn	jm	ea	cl	cr	fl
unknown	0	0	0	0	0	0	0	0	0
walking	0.11	0.57	0.12	0.12	0	0.01	0.01	0.04	0
spinning	0.15	0.06	0.65	0.03	0	0	0.01	0.04	0.06
running	0.01	0.07	0.07	0.79	0.04	0	0	0.01	0
jumping	0.19	0.01	0.04	0.09	0.57	0	0	0.01	0.09
eating	0.19	0	0	0	0	0.76	0.04	0	0.01
climbing	0.06	0.01	0	0	0.03	0	0.90	0	0
crawling	0.20	0.03	0	0.06	0.01	0	0.01	0.69	0
flying	0.19	0.03	0.01	0	0.01	0.01	0.03	0.03	0.70

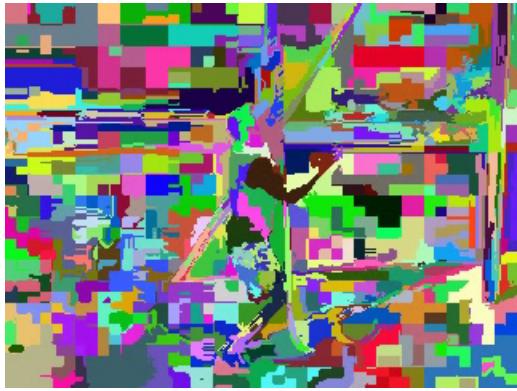
Study Results: Action Discrimination

- Dominant unidirectional motion.

	un	wl	sp	rn	jm	ea	cl	cr	fl
unknown	0	0	0	0	0	0	0	0	0
walking	0.11	0.57	0.12	0.12	0	0.01	0.01	0.04	0
spinning	0.15	0.06	0.65	0.03	0	0	0.01	0.04	0.06
running	0.01	.07	0.0 7	0.79	.04	0	0	0.01	0
jumping	0.19	0.01	0.04	0.09	0.57	0	0	0.01	0.09
eating	0.19	0	0	0	0	0.76	0.04	0	0.01
climbing	0.06	.01	0	0	0.03	0	0.90	0	0
crawling	0.20	0.03	0	0.06	0.01	0	0.01	0.69	0
flying	0.19	0.03	0.01	0	0.01	0.01	0.03	0.03	0.70

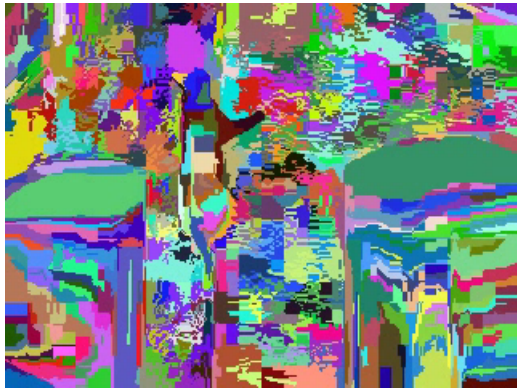


Human_Running



Human_Climbing

Animal_Running



Animal_Climbing

[Xu, Doell,… Corso ICSC 2013 and IJSC In Press]

Study Results: Action Discrimination

- Dominant unidirectional motion.

	un	wl	sp	rn	jm	ea	cl	cr	fl
unknown	0	0	0	0	0	0	0	0	0
walking	0.11	0.57	0.12	0.12	0	0.01	0.01	0.04	0
spinning	0.15	0.06	0.65	0.03	0	0	0.01	0.04	0.06
running	0.01	.07	0.0 7	0.79	.04	0	0	0.01	0
jumping	0.19	0.01	0.04	0.09	0.57	0	0	0.01	0.09
eating	0.19	0	0	0	0	0.76	0.04	0	0.01
climbing	0.06	.01	0	0	0.03	0	0.90	0	0
crawling	0.20	0.03	0	0.06	0.01	0	0.01	0.69	0
flying	0.19	0.03	0.01	0	0.01	0.01	0.03	0.03	0.70

Human_Running

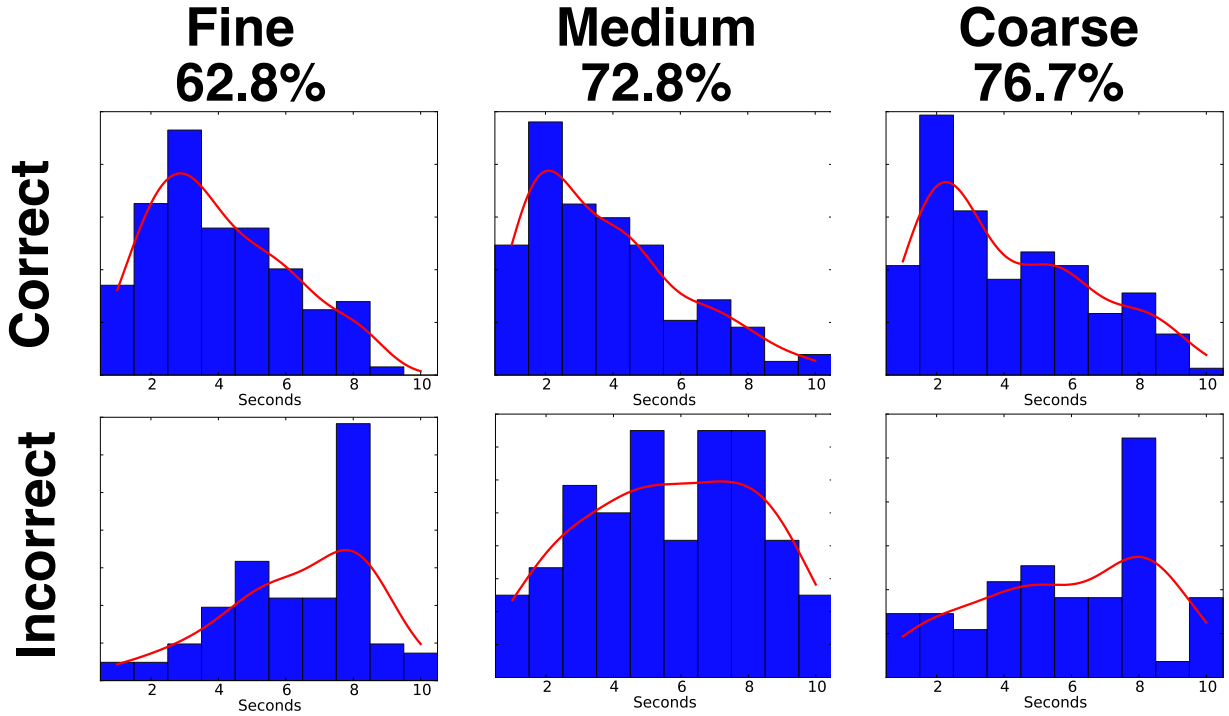
Human_Climbing

Animal_Running

Animal_Climbing

[Xu, Doell,… Corso ICSC 2013 and IJSC In Press]

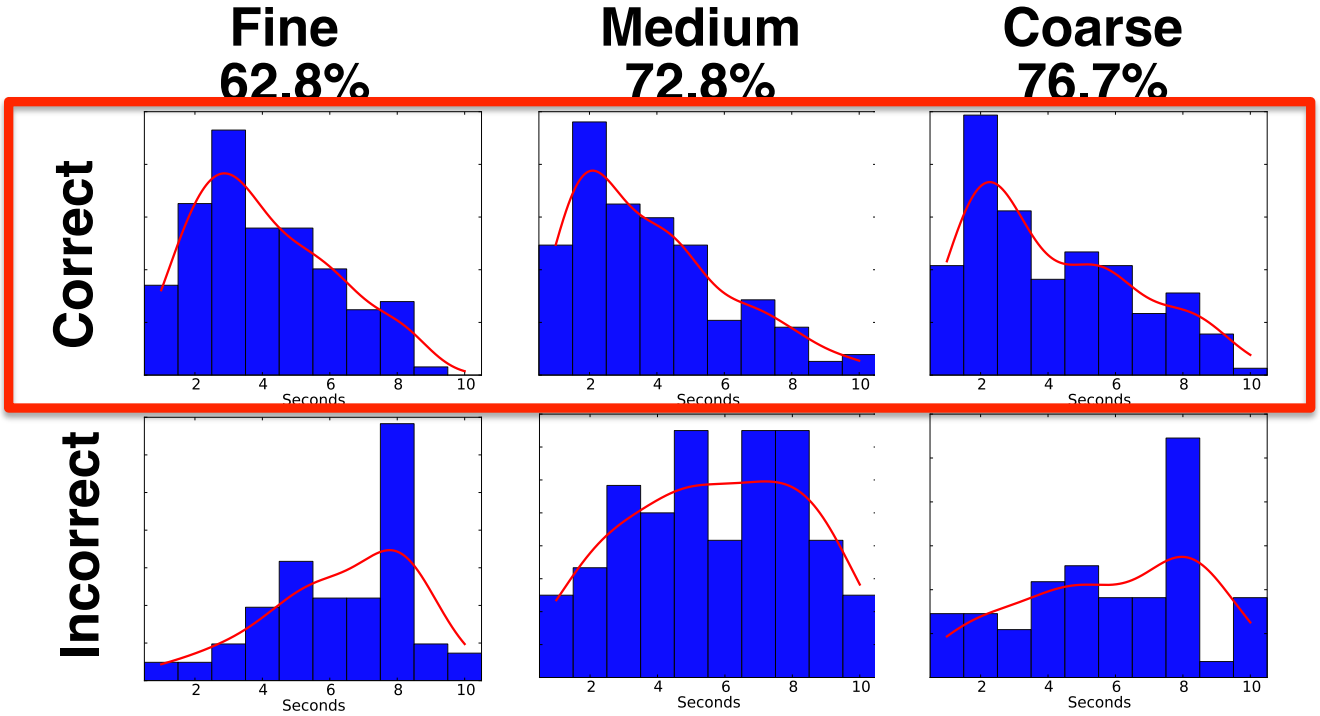
Study Results: Performance by Level



- Bar figures are the response time.
  - X-axis: Time at the half-frame-rate.
  - Y-axis: density of responses.
  - Blue bars: simple histogram.
  - Red curve: Gaussian kernel density estimate.

[Xu, Doell,… Corso ICSC 2013 and IJSC In Press]

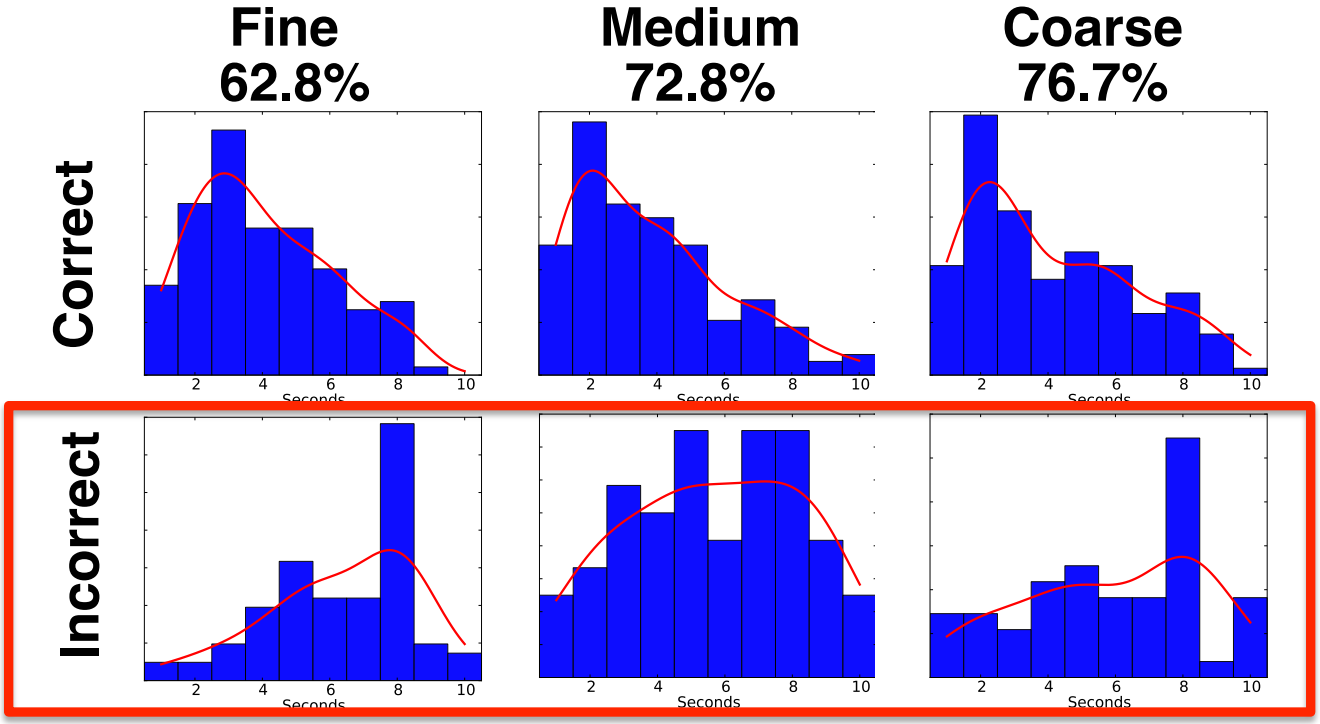
Study Results: Performance by Level



- Correct action matches:
  - Response distributions are early equivalent.
  - Heavily weighted toward the shorter end of X-axis.
- If the participant knows the answer then typically knows it quickly.

[Xu, Doell,… Corso ICSC 2013 and IJSC In Press]

Study Results: Performance by Level



- Incorrect action matches:
  - Different patterns.
  - Fine videos peaked at about eight seconds.
- Participant watched the whole video and still got the wrong action perception.

[Xu, Doell,… Corso ICSC 2013 and IJSC In Press]

Summary of Study

- Segmentation hierarchies generate rich decompositions of the video content.
- They compress the signal significantly, but does enough semantic information remain to discriminative actor and act?

- Yes! 82% accuracy on actor and 70% on act.
- Act discrimination increases with coarseness of the signal.
- Act discrimination for human actors is better than animals.
- Act discrimination for a static background is better than a moving background.

- Limitation: 20 participants on 32 input videos. Moving to 64 input videos and Mechanical Turk.