ICSC 2013

Are Actor and Action Semantics Retained in Video Supervoxel Segmentation?

Chenliang Xu[1], Richard F. Doell[1], Stephen José Hanson[2], Catherine Hanson[2], and Jason J. Corso[1]

[1]SUNY at Buffalo

[2]Rutgers University

Note: images here are videos in the original slides.

Video Understanding; What?

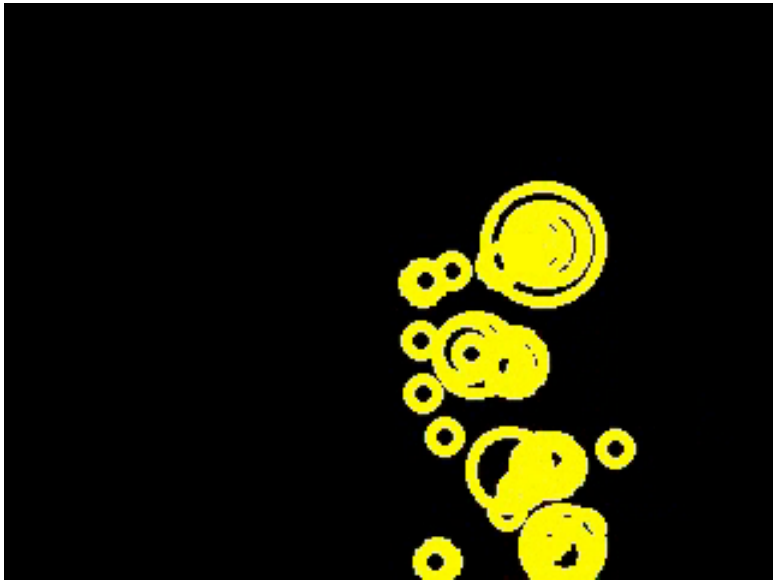
- Example human synopsis: “A person is climbing a rock-wall.”

Applications

- Real-time / Interactive
  - Human computer interaction and entertainment.
  - Healthcare monitoring and surveillance.
- Off-line
  - Video indexing and search.
  - Video to language.
  - Sports analysis.

Note: images here are videos in the original slides.

Video Understanding; What?



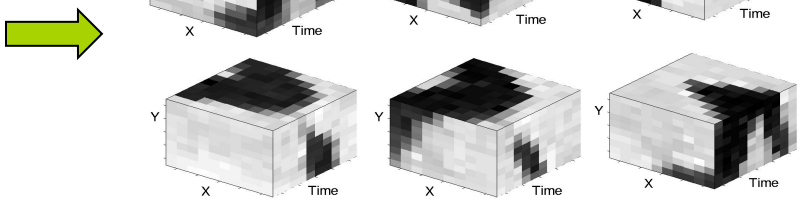
Method: Laptev. “On Space-Time Interest Points.” IJCV 2005.

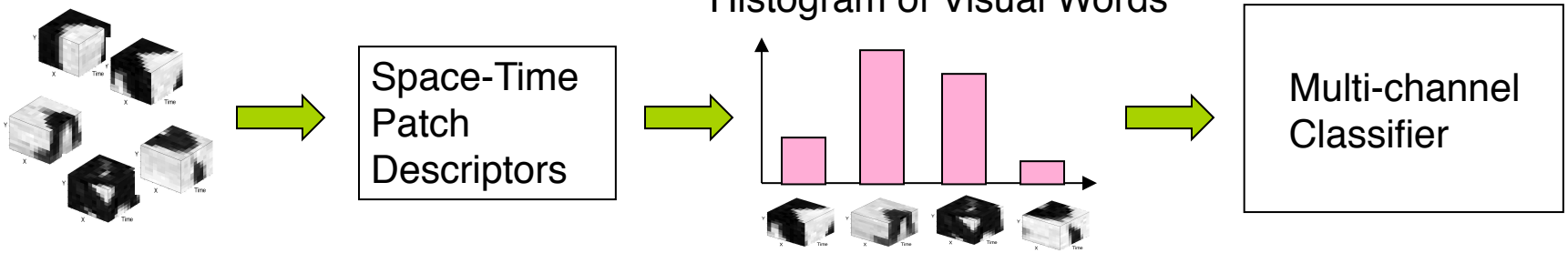
Method: Wang et al. “Action Recognition by Dense Trajectories.” CVPR 2011.

Note: images here are videos in the original slides.

The (Very Common) Bag-of-Features Pipeline

Source: materials adapted from Laptev's CVPR 2008 slides.





- Examples include Schüldt et al. ICPR 2004, Niebles et al. IJCV 2008, and many works building on this basic idea.

Note: images here are videos in the original slides.

Supervoxel Segmentation: Toward a Representation with Rich Semantics?

Note: images here are videos in the original slides.

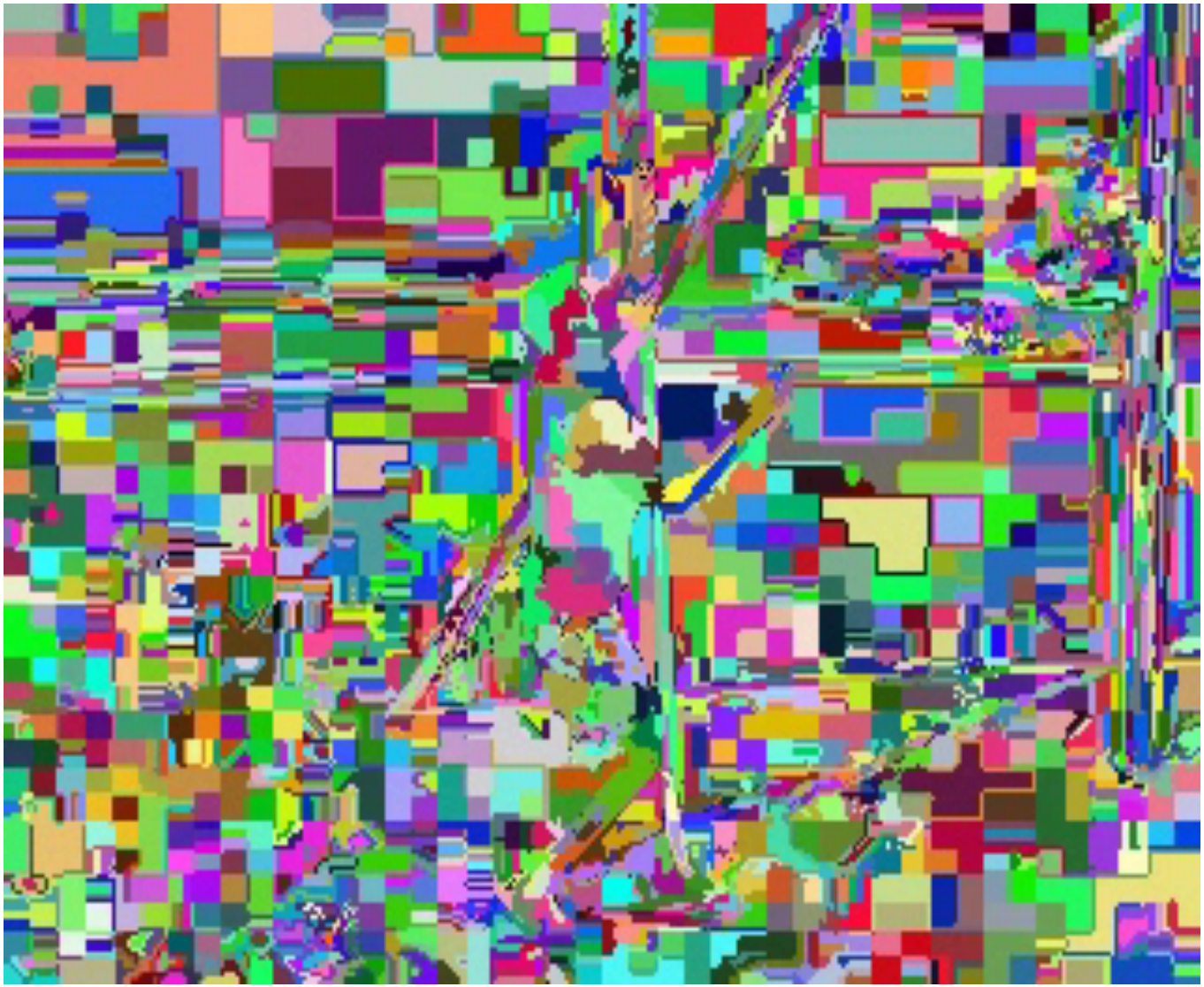
[Xu, Xiong and Corso, ECCV 2012]

Study Questions

- Primary Question:
  - Do the segmentation hierarchies retain enough information for the human perceiver to discriminate
    - Actor? (human or animal)
    - Action? (climbing, crawling, eating, flying, jumping, running, spinning, walking)
- Secondary Questions:
  - How does the semantic retention vary with
    - Density of the supervoxels?
    - Actor (human versus animal)?
    - Background (static versus moving)?
  - How does response time vary with action?

Note: images here are videos in the original slides.

Can Humans Perceive Actor/Action from Supervoxels?



Note: images here are videos in the original slides.

Can Humans Perceive Actor/Action from Supervoxels?

Note: images here are videos in the original slides.

Can Humans Perceive Actor/Action from Supervoxels?

Note: images here are videos in the original slides.

Can Humans Perceive Actor/Action from Supervoxels?

Note: images here are videos in the original slides.

Can Humans Perceive Actor/Action from Supervoxels?

Note: images here are videos in the original slides.

Can Humans Perceive Actor/Action from Supervoxels?

Note: images here are videos in the original slides.

Video Supervoxel Segmentation

Note: images here are videos in the original slides.

Hierarchical Video Supervoxel Segmentation

- Basic problem statement:
- Segmentation hierarchy

Segmentation — Video Input

$$\mathcal{S}^* = \operatorname*{argmin}_{\mathcal{S}} E(\mathcal{S}|\mathcal{V})$$

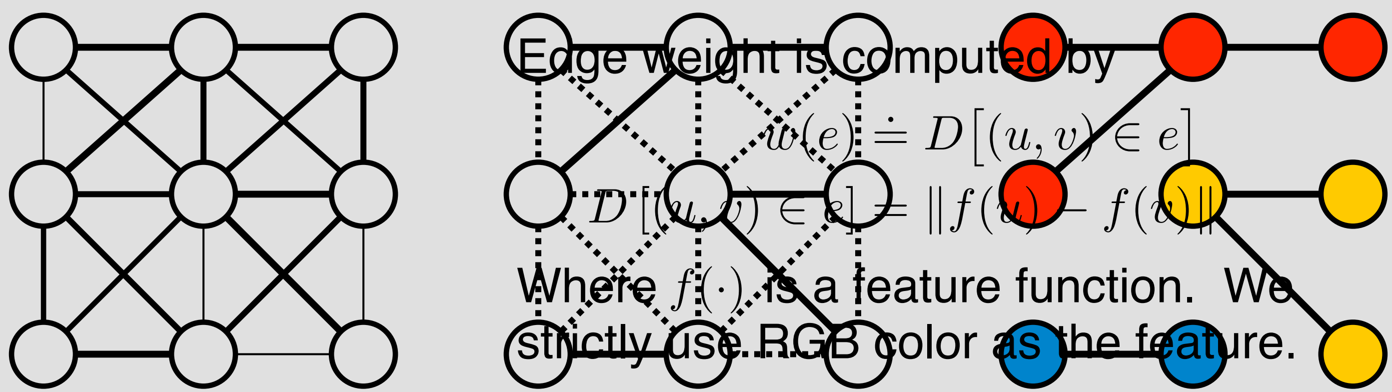
$$\mathcal{S} \doteq \{S^1, S^2, \dots, S^h\}$$

$S^i \doteq \{s_1, s_2, \dots\}$ such that $s_j \subset \Gamma$, $\cup_j s_j = \Gamma$, and $s_i \cap s_j = \varnothing$ for pairs i, j

- Use the minimum spanning tree method.

$$E(S^1|\mathcal{V}) = \tau \sum_{s \in S^1} \sum_{e \in \mathrm{MST(s)}} w(e) + \sum_{s,t \in S^1} \min_{e \in \langle s,t \rangle} w(e)$$

Stage 3: [illegible] $E(S^1|\mathcal{V})$.



Edge weight is computed by

$$w(e) \doteq D\big[(u,v) \in e\big]$$

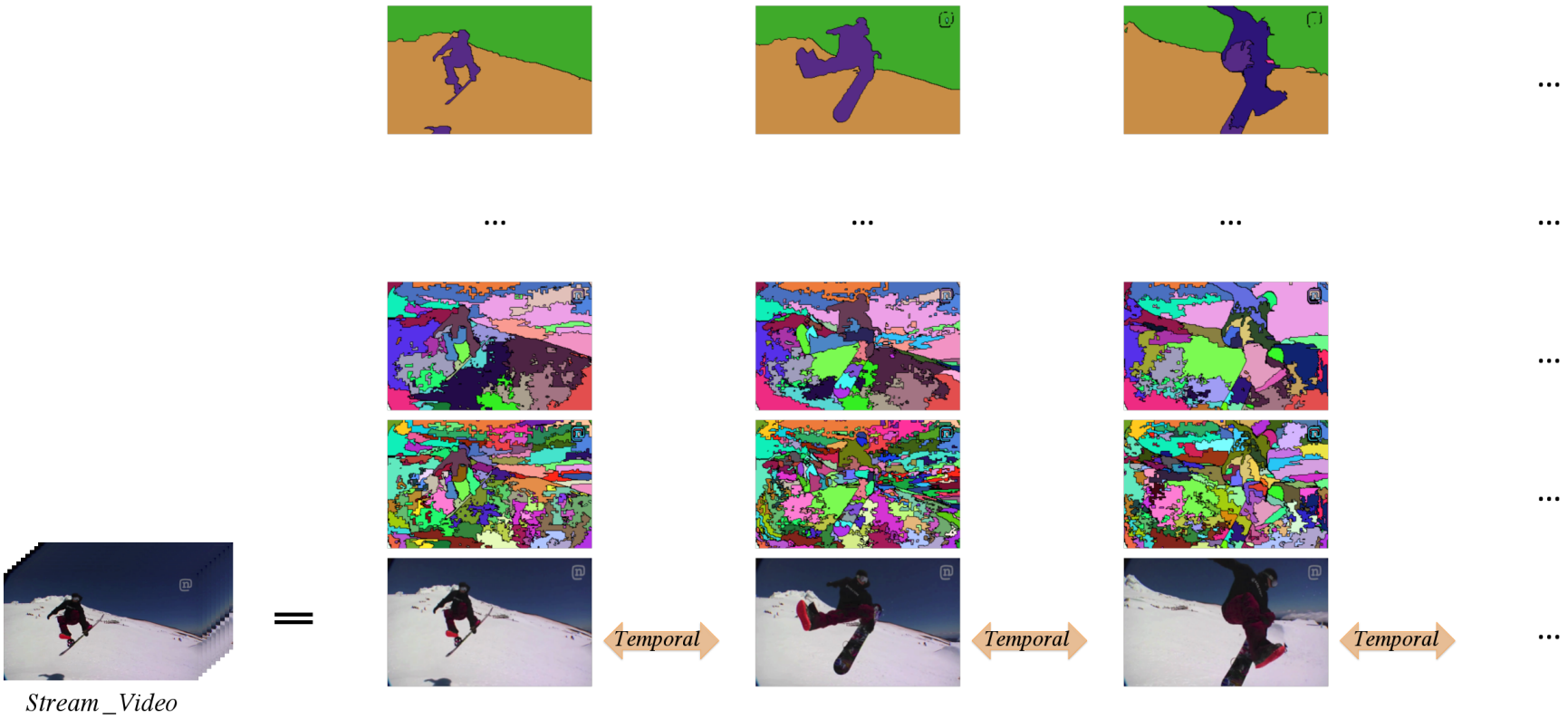
$$D\big[(u,v) \in e\big] = \|f(u) - f(v)\|$$

Where $f(\cdot)$ is a feature function. We strictly use RGB color as the feature.

Note: images here are videos in the original slides.

[Xu, Xiong and Corso, ECCV 2012]

Streaming Hierarchical Video Segmentation



Note: images here are videos in the original slides

[Xu, Xiong and Corso, ECCV 2012]

Main Study

Note: images here are videos in the original slides.

Study Setup: Data Set

- Video Time (Action starts immediately after play.)
  - About 4 Seconds / shown at half-frame-rate
- Stratified according to
  - **Actors:** human or animal
  - **Background:** static or moving
  - **Actions:** climbing, crawling, eating, flying, jumping, running, spinning, walking
- 3 Levels of the segmentation hierarchy
  - Fine: 8th level / Medium: 16th level / Coarse: 24th level
  - Q: a best level in the hierarchy?
- In total, we have 96 videos
  - 2 actors * 2 backgrounds * 8 acts * 3 levels

Note: images here are videos in the original slides.

Study Setup: Data Collection

- Study cohort of 20 college-age participants.
  - No student is studying segmentation.
  - 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.

Note: images here are videos in the original slides.

Discriminate Actor? (human or animal)

Note: images here are videos in the original slides.

Study Results: Actor Discrimination

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

Confusion Matrix

- Overall actor discrimination rate: 82.4%.
- Unknown was chosen when less confident.
- Suspects
  - Performance is so high due to one dominant actor.
    - Locate by svx motion, then determine by svx shape.
  - Performance on human is better than animal due to more variation of animal location and orientation.

Note: images here are videos in the original slides.

Discriminate Action? (one of eight)

Note: images here are videos in the original slides.

Study Results: Action Discrimination

- Overall action discrimination rate: 70.4%.

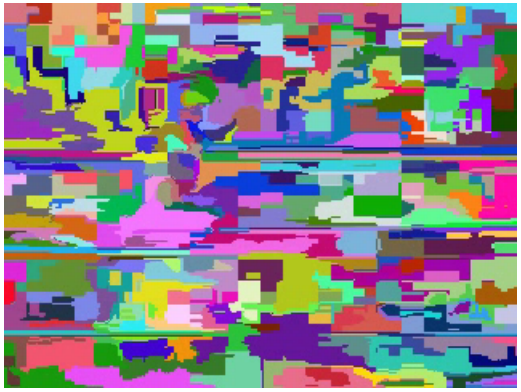
	un	wl	sp	rn	jm	ea	cl	cr	fl
unknown	0	0	0	0	0	0	0	0	0
walking	0.11	0.57	0.12	0.12	0	0.01	0.01	0.04	0
spinning	0.15	0.06	0.65	0.03	0	0	0.01	0.04	0.06
running	0.01	0.07	0.07	0.79	0.04	0	0	0.01	0
jumping	0.19	0.01	0.04	0.09	0.57	0	0	0.01	0.09
eating	0.19	0	0	0	0	0.76	0.04	0	0.01
climbing	0.06	0.01	0	0	0.03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

Note: images here are videos in the original slides.

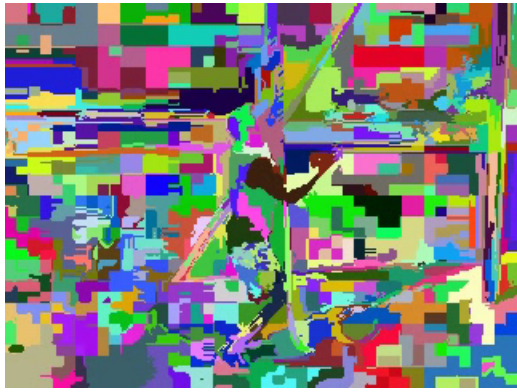
Study Results: Action Discrimination

- 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	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

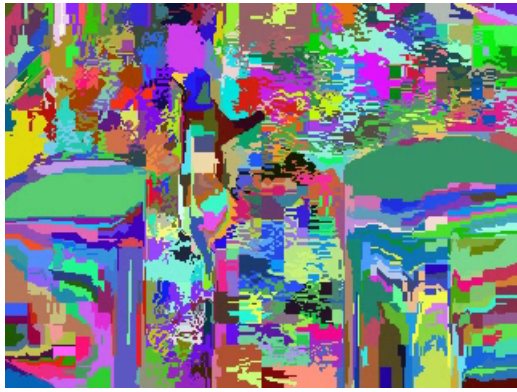


Human_Running



Human_Climbing

Animal_Running



Animal_Climbing

Note: images here are videos in the original slides.

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	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

Human_Running

Human_Climbing

Animal_Running

Animal_Climbing

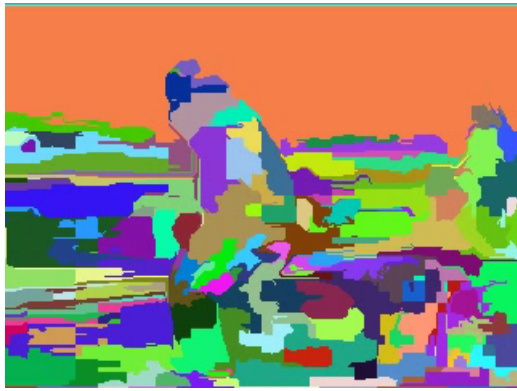
Note: images here are videos in the original slides.

Study Results: Action Discrimination

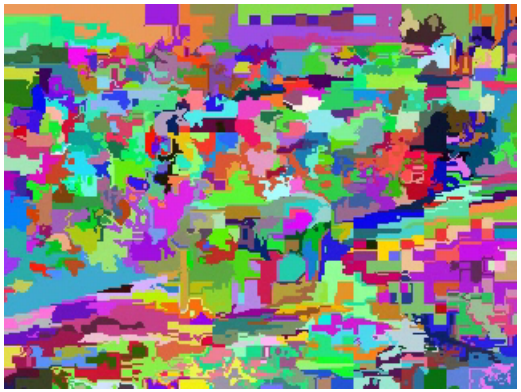
- Semantic ambiguity in videos.

	un	wl	sp	rn	jm	ea	cl	cr	fl
unknown	0	0	0	0	0	0	0	0	0
walking	0.1	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

Human_Walking



Human_Jumping



Animal_Walking

Animal_Jumping

Note: images here are videos in the original slides.

Study Results: Action Discrimination

- Semantic ambiguity in videos.

	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

Human_Walking

Human_Jumping

Animal_Walking

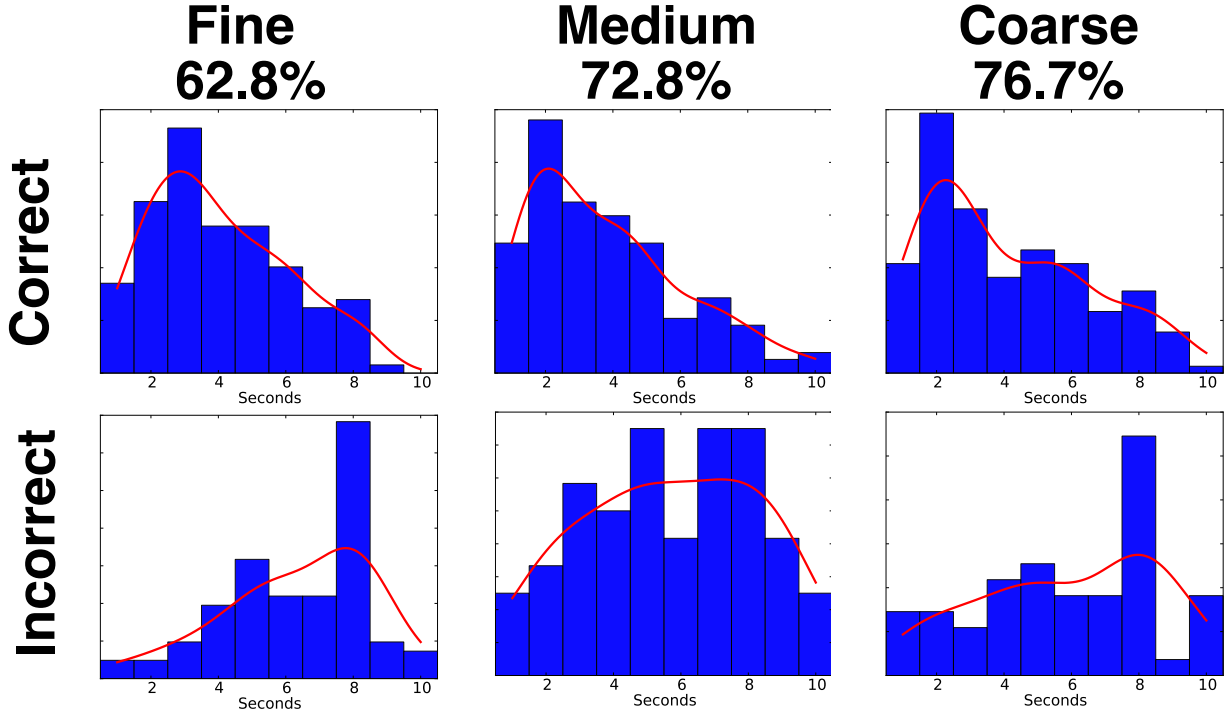
Animal_Jumping

Note: images here are videos in the original slides.

How does the performance vary with density of the supervoxels?

Note: images here are videos in the original slides.

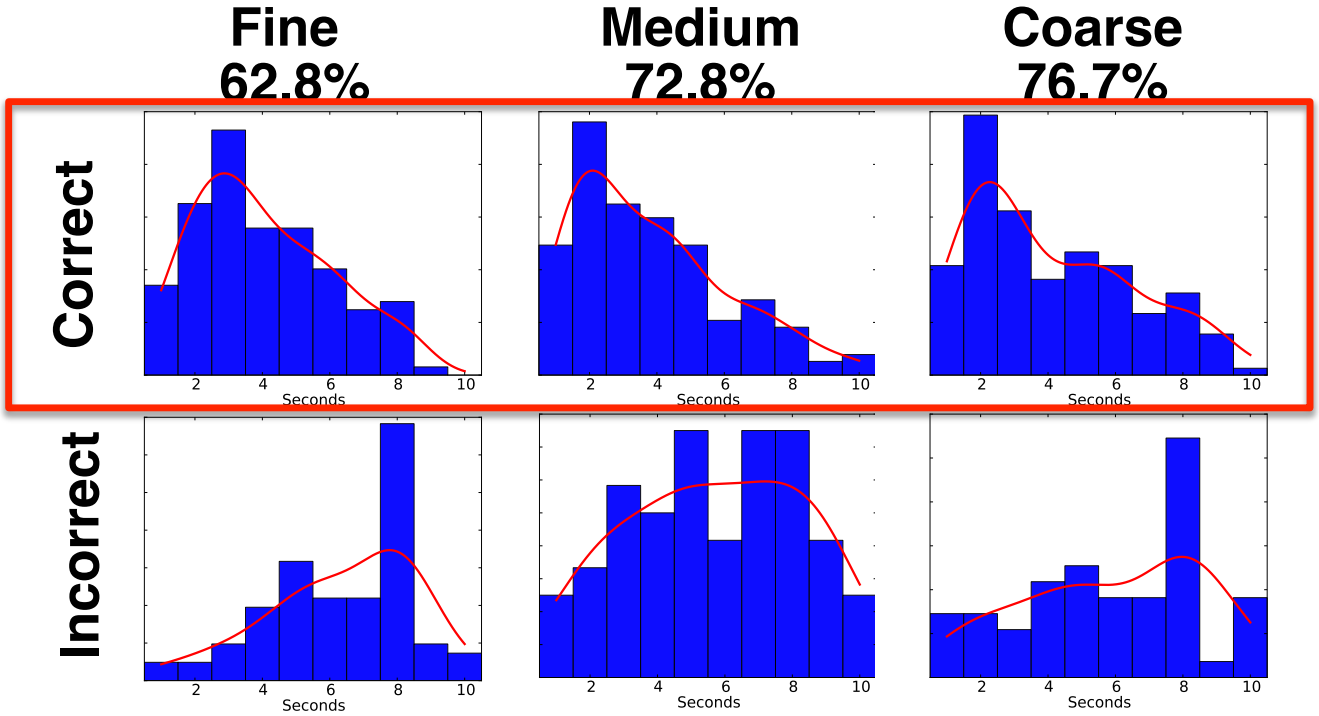
Study Results: Performance by Level



- 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.

Note: images here are videos in the original slides.

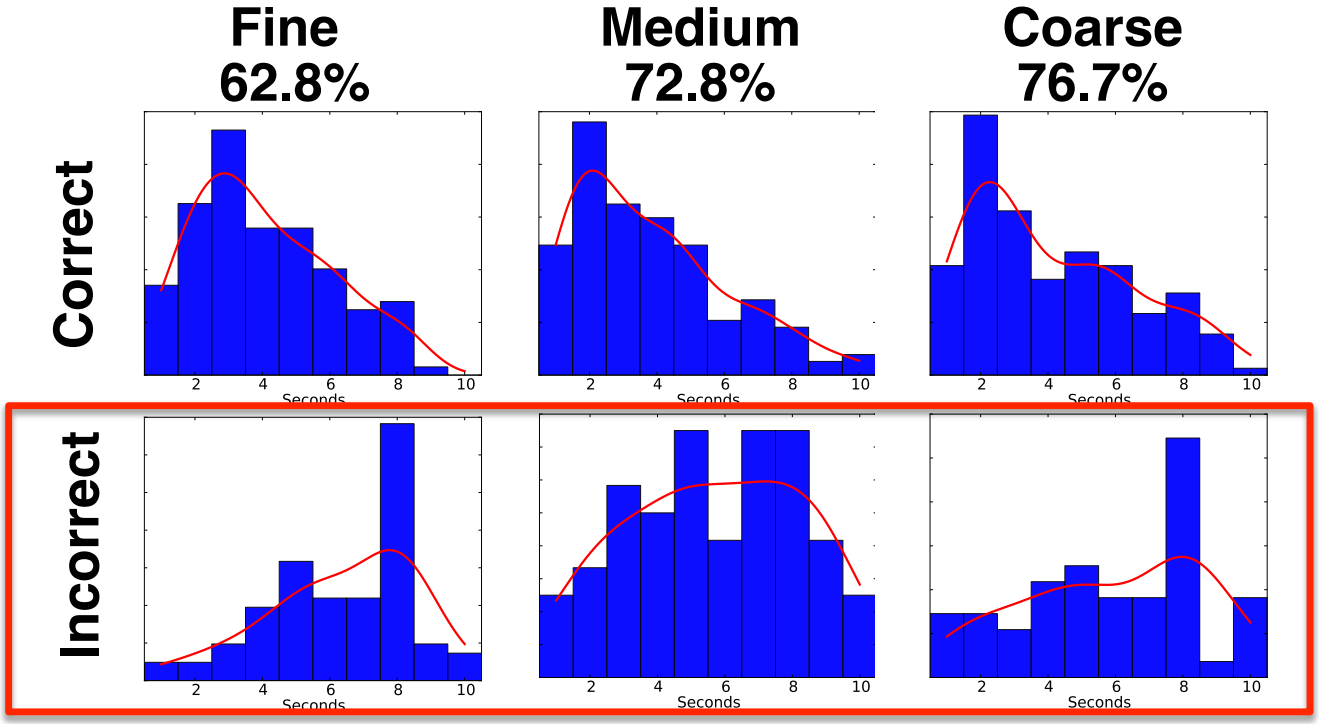
Study Results: Performance by Level



- 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.

Note: images here are videos in the original slides.

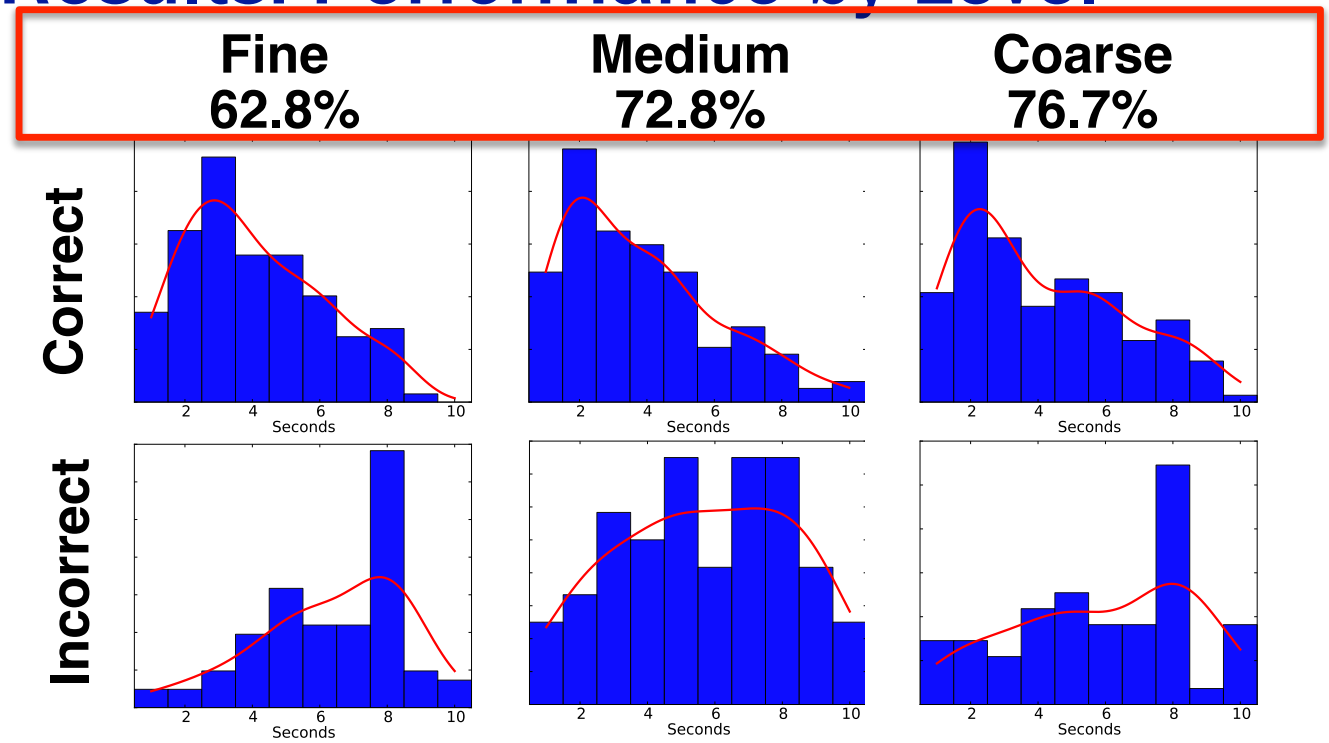
Study Results: Performance by Level



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

Note: images here are videos in the original slides.

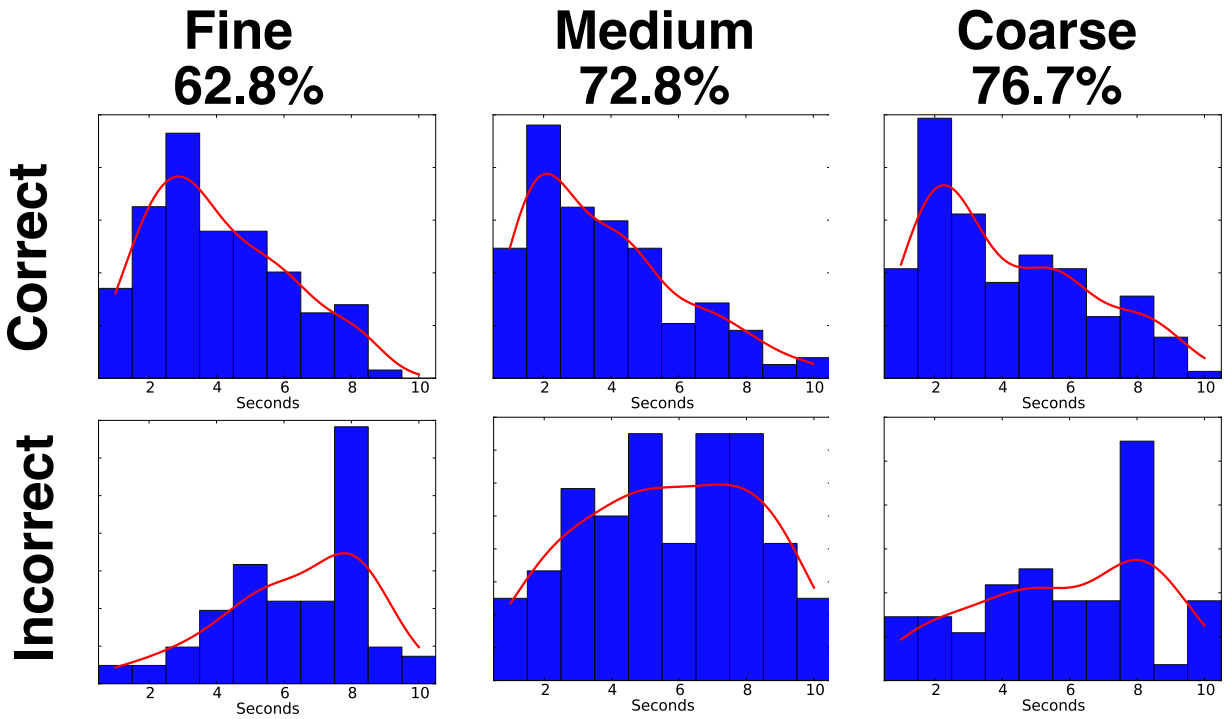
Study Results: Performance by Level



- Information in finer details are unlikely to be needed when performing the task.

Note: images here are videos in the original slides.

Study Results: Performance by Level



Fine

Medium

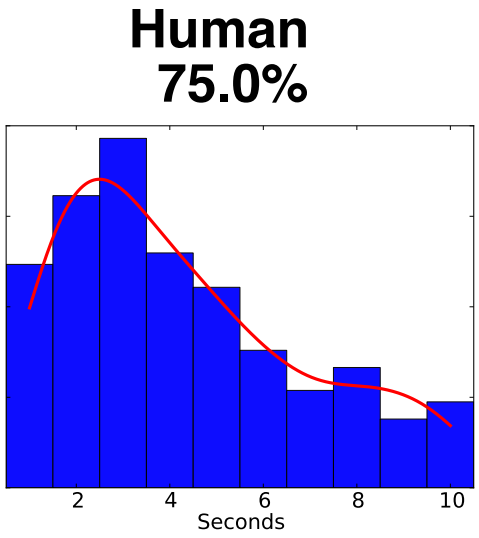
Coarse

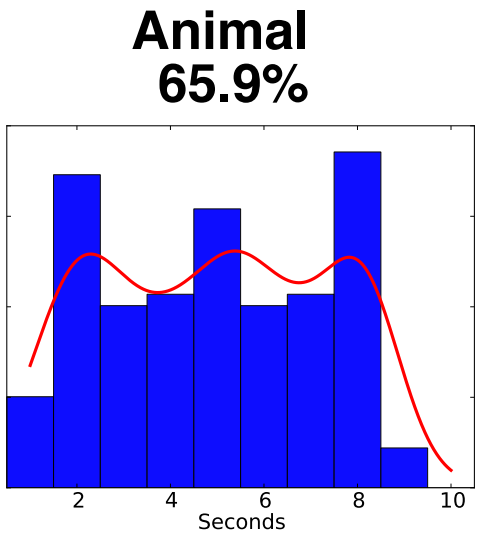
Note: images here are videos in the original slides.

How does the performance vary with actor (human versus animal)?

Note: images here are videos in the original slides.

Study Results: Performance by Actor





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

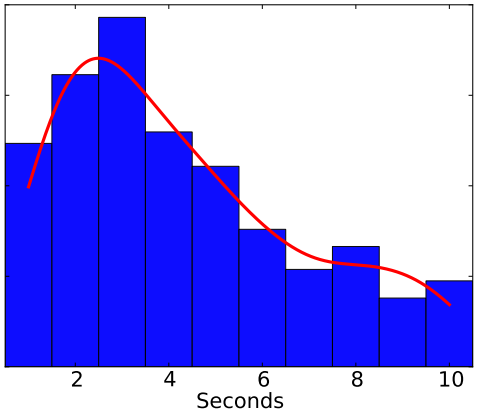
Actor Discrimination

- In general, human action has better match.
- For speed (one peak vs. multiple peaks)
  - Greater variation in appearance of animals.
- Human activity is easier to perceive than animal.
  - A correlation between knowing the actor and recognizing the action correctly.

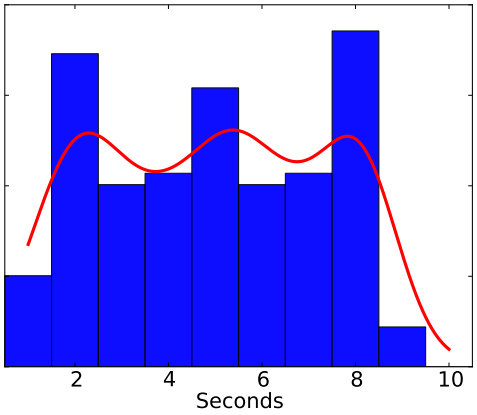
Note: images here are videos in the original slides.

Study Results: Performance by Actor

Human
75.0%

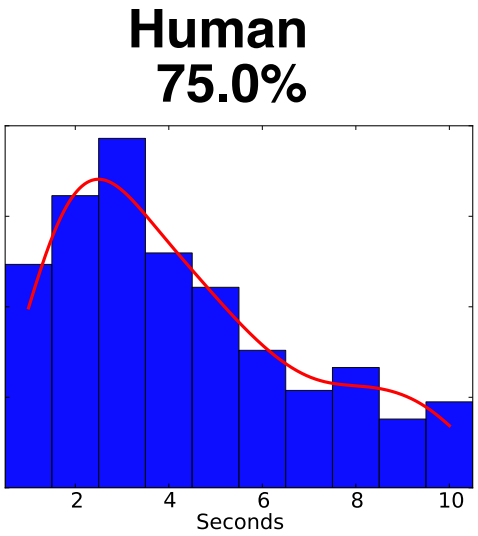


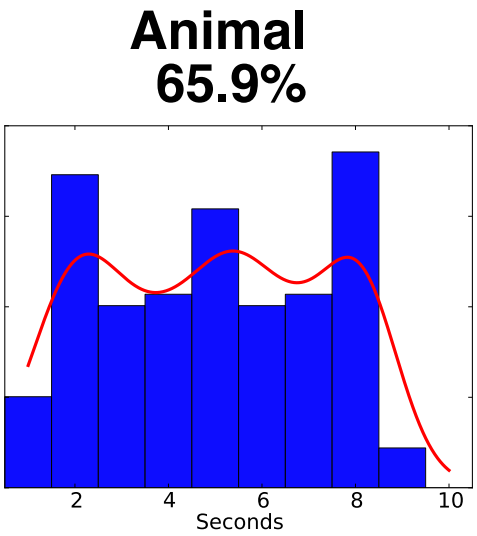
Animal
65.9%



Note: images here are videos in the original slides.

Study Results: Performance by Actor



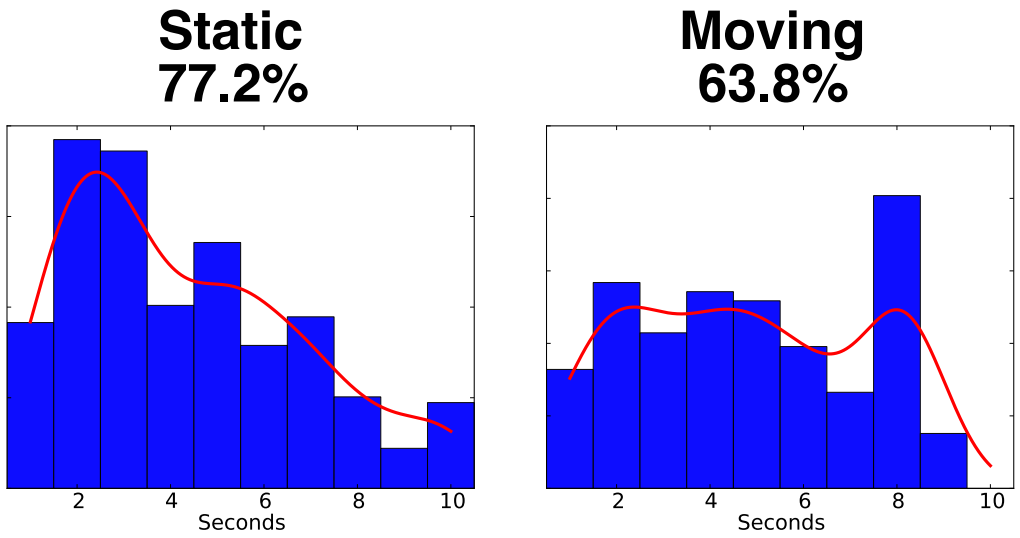


Note: images here are videos in the original slides.

How does the performance vary with background (static versus moving)?

Note: images here are videos in the original slides.

Study Results: Performance by Background

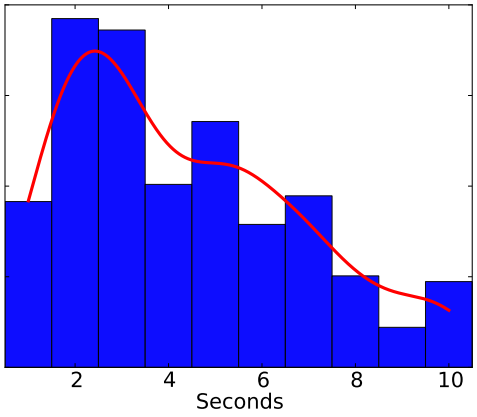


- Static Background:
  - The dominant actor is more easily picked out.
- Moving Background:
  - The flat curve suggests the response time for a single video highly depends on the specific background within that video.

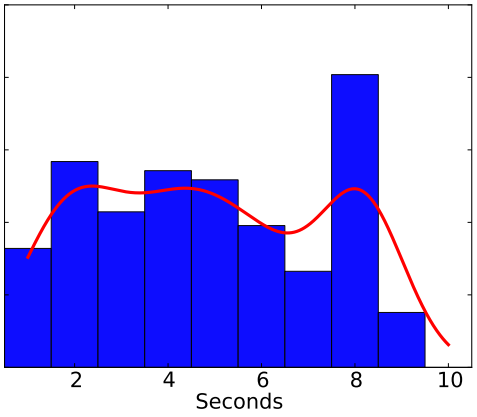
Note: images here are videos in the original slides.

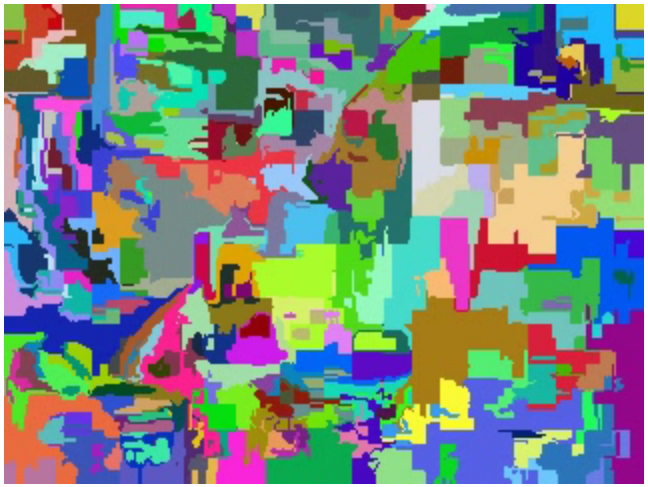
Study Results: Performance by Background

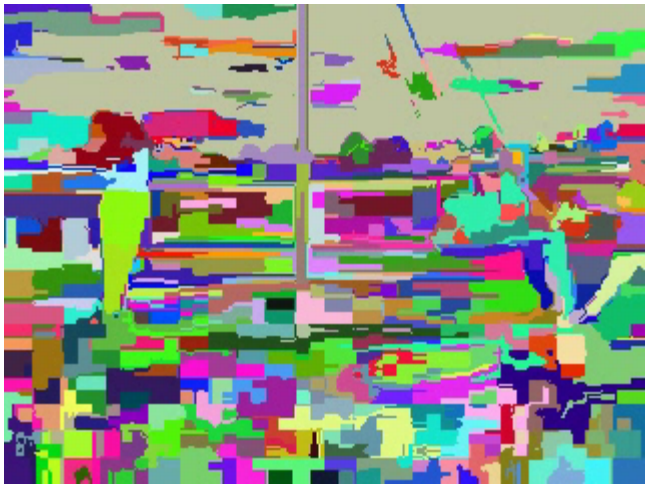
Static
77.2%



Moving
63.8%



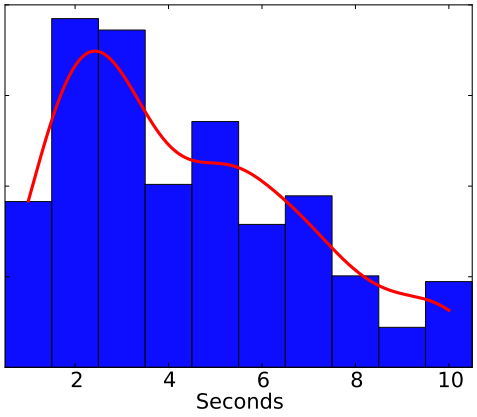




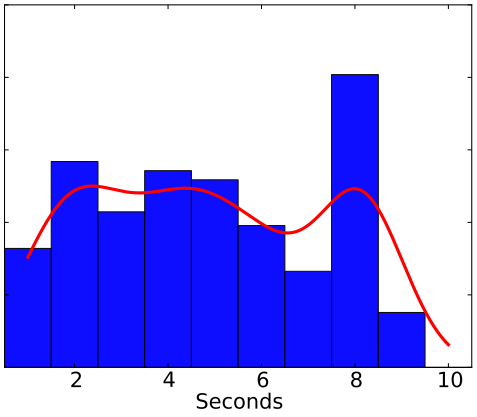
Note: images here are videos in the original slides.

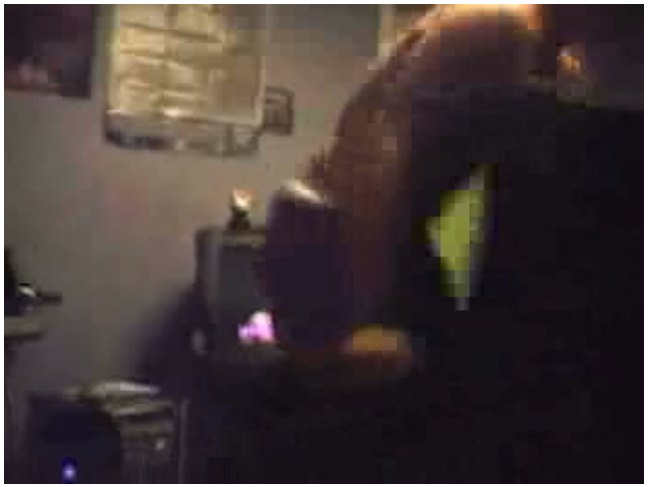
Study Results: Performance by Background

Static
77.2%



Moving
63.8%



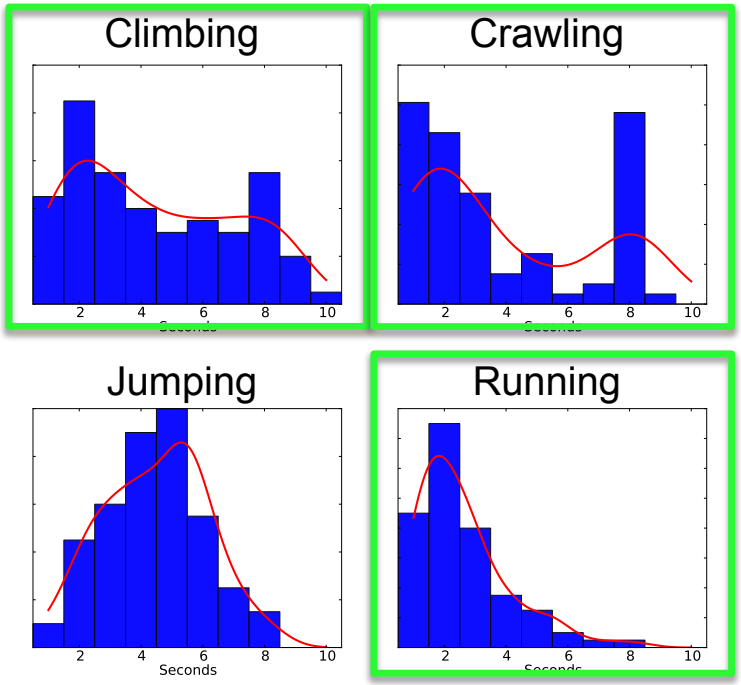


Note: images here are videos in the original slides.

How does response time vary with action?

Note: images here are videos in the original slides.

Study Results: Speed by Action



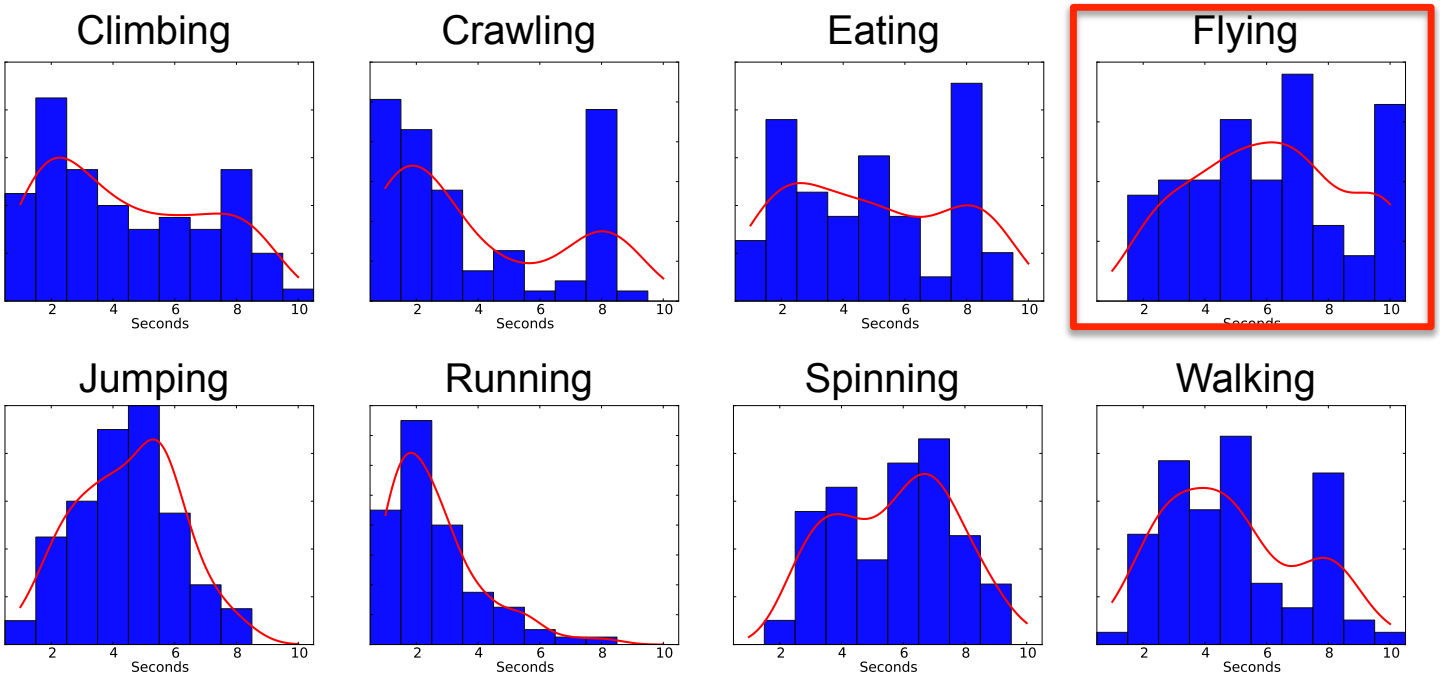
	un	wl	sp	rn	jm	ea	cl	cr	fl
unknown	0	0	0	0	0	0	0	0	0
walking	0.11	0.57	0.12	0.12	0	0.01	0.01	0.04	0
spinning	0.15	0.06	0.65	0.03	0	0	0.01	0.04	0.06
running	0.01	0.07	0.07	0.79	0.04	0	0	0.01	0
jumping	0.19	0.01	0.04	0.09	0.57	0	0	0.01	0.09
eating	0.19	0	0	0	0	0.76	0.04	0	0.01
climbing	0.06	0.01	0	0	0.03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

Action Discrimination

- Actions whose semantics have been strongly retained are generally responded to more quickly.

Note: images here are videos in the original slides.

Study Results: Speed by Action



- Actions whose semantics have been strongly retained are generally responded to more quickly.
- Unusual actions take more time to get a response.

Note: images here are videos in the original slides.

Summary of Study

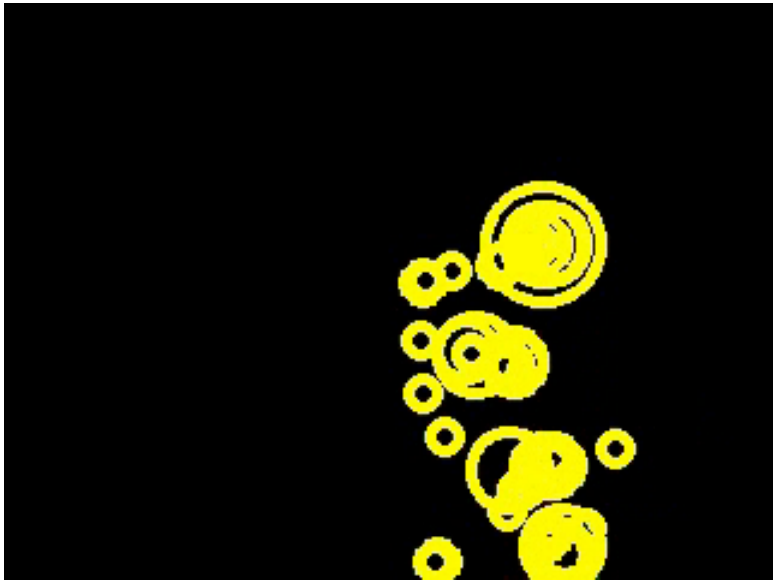
- Segmentation hierarchies generate rich decompositions of the video content.
- They compress the signal significantly, but does enough semantic information retained to discriminate actor and action?

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

- Future Work:
  - Semantic ambiguity and large study cohort.
  - More Supervoxel Algorithms: SWA etc.
  - Action recognition based on Supervoxels.

Note: images here are videos in the original slides.

Video Understanding; What?

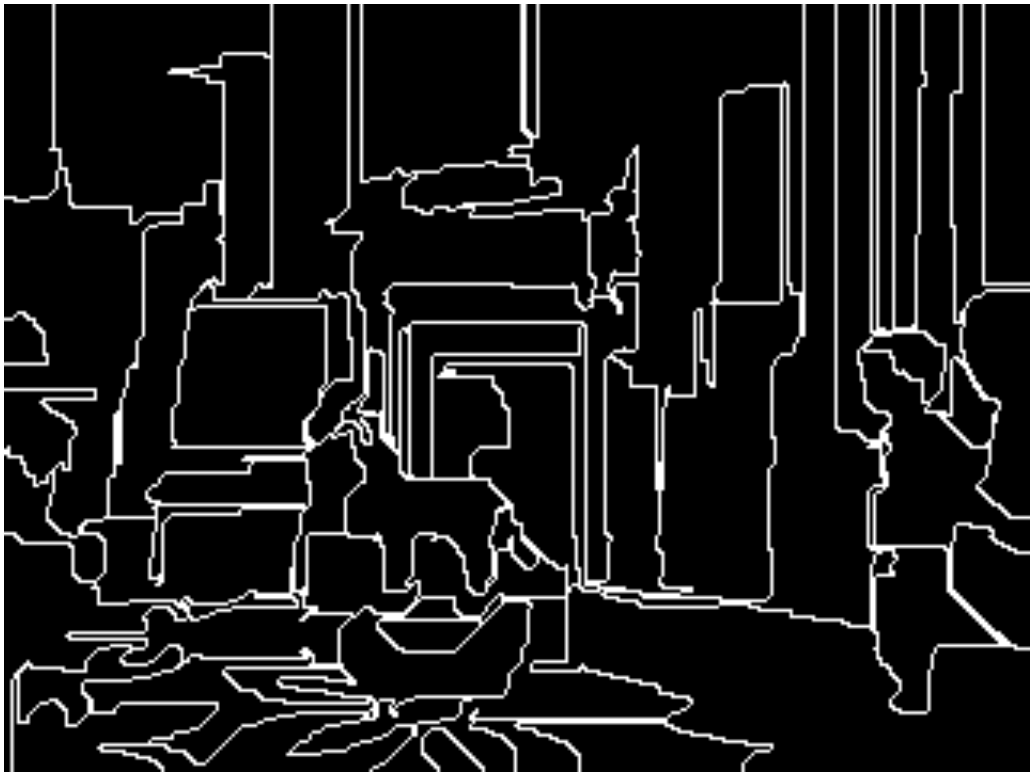


Method: Laptev. “On Space-Time Interest Points.” IJCV 2005.

Method: Wang et al. “Action Recognition by Dense Trajectories.” CVPR 2011.

Note: images here are videos in the original slides.

Alas, what makes such a good representation?



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

Note: images here are videos in the original slides.

Segmentation: Toward a Representation with Rich Semantics?

Note: images here are videos in the original slides.

JAMES S.
MCDONNELL
FOUNDATION

Thank you!

Note: images here are videos in the original slides.