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?


Note: images here are videos in the original slides.
The (Very Common) Bag-of-Features Pipeline

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
  \[ S = \{ S^1, S^2, \ldots, S^h \} \]
  \[ S^i = \{ s_1, s_2, \ldots \} \] such that \( s_j \subset \Gamma, \bigcup_j s_j = \Gamma, \) and \( s_i \cap s_j = \emptyset \) for pairs \( i, j \)

- Use the minimum spanning tree method.

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

Note: images here are videos in the original slides.

[Stages 1-3: Diagrams showing the stages of the algorithm, with edge weights and similarity measures.]

Note: images here are videos in the original slides.

[Stage 2: Diagram showing the graph connecting nearest voxels with edge weights computed by similarity measures.]

**Edge weight is computed by**
\[ w(e) = D[(u, v) \in e] \]
\[ D[(u, v) \in e] = \| f(u) - f(v) \| \]

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

[Xu, Xiong and Corso, ECCV 2012]
Streaming Hierarchical Video Segmentation

[Note: images here are videos in the original slides.]

Stream Video

Temporal

Temporal

Temporal

[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)
Study Results: Actor Discrimination

• 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.

Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>un</th>
<th>hu</th>
<th>an</th>
</tr>
</thead>
<tbody>
<tr>
<td>unknown</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>human</td>
<td>0.11</td>
<td>0.86</td>
<td>0.03</td>
</tr>
<tr>
<td>animal</td>
<td>0.17</td>
<td>0.05</td>
<td>0.78</td>
</tr>
</tbody>
</table>

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%.

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

Note: Images here are videos in the original slides.
Study Results: Action Discrimination

- Dominant unidirectional motion.

<table>
<thead>
<tr>
<th></th>
<th>un</th>
<th>wl</th>
<th>sp</th>
<th>rn</th>
<th>jm</th>
<th>ea</th>
<th>cl</th>
<th>cr</th>
<th>fl</th>
</tr>
</thead>
<tbody>
<tr>
<td>unknown</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>walking</td>
<td>0.11</td>
<td>0.57</td>
<td>0.12</td>
<td>0.12</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.04</td>
<td>0</td>
</tr>
<tr>
<td>spinning</td>
<td>0.15</td>
<td>0.06</td>
<td>0.65</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>running</td>
<td>0.01</td>
<td>0.07</td>
<td>0.79</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>jumping</td>
<td>0.19</td>
<td>0.01</td>
<td>0.04</td>
<td>0.09</td>
<td>0.57</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.09</td>
</tr>
<tr>
<td>eating</td>
<td>0.19</td>
<td>0.01</td>
<td>0.04</td>
<td>0.09</td>
<td>0.57</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.09</td>
</tr>
<tr>
<td>climbing</td>
<td>0.06</td>
<td>0.01</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0.90</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>crawling</td>
<td>0.20</td>
<td>0.03</td>
<td>0.06</td>
<td>0.01</td>
<td>0</td>
<td>0.01</td>
<td>0.69</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>flying</td>
<td>0.19</td>
<td>0.03</td>
<td>0.01</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Note: images here are videos in the original slides.
Study Results: Action Discrimination

- Dominant unidirectional motion.

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

Note: images here are videos in the original slides.
Study Results: Action Discrimination

- Semantic ambiguity in videos.

<table>
<thead>
<tr>
<th></th>
<th>un</th>
<th>wl</th>
<th>sp</th>
<th>rn</th>
<th>jm</th>
<th>ea</th>
<th>cl</th>
<th>cr</th>
<th>fl</th>
</tr>
</thead>
<tbody>
<tr>
<td>unknown</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>walking</td>
<td>0.1</td>
<td>0.57</td>
<td>0.12</td>
<td>0.12</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.04</td>
<td>0</td>
</tr>
<tr>
<td>spinning</td>
<td>0.15</td>
<td>0.06</td>
<td>0.65</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>running</td>
<td>0.01</td>
<td>0.07</td>
<td>0.07</td>
<td>0.79</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>jumping</td>
<td>0.19</td>
<td>0.01</td>
<td>0.04</td>
<td>0.09</td>
<td>0.57</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.09</td>
</tr>
<tr>
<td>eating</td>
<td>0.19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.76</td>
<td>0.04</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>climbing</td>
<td>0.06</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0.03</td>
<td>0</td>
<td>0.90</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>crawling</td>
<td>0.20</td>
<td>0.03</td>
<td>0</td>
<td>0.06</td>
<td>0.01</td>
<td>0</td>
<td>0.01</td>
<td>0.69</td>
<td>0</td>
</tr>
<tr>
<td>flying</td>
<td>0.19</td>
<td>0.03</td>
<td>0.01</td>
<td>0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Note: Images here are videos in the original slides.
Study Results: Action Discrimination

• Semantic ambiguity in videos.

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

<table>
<thead>
<tr>
<th>Level</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine</td>
<td>62.8%</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>72.8%</td>
<td></td>
</tr>
<tr>
<td>Coarse</td>
<td>76.7%</td>
<td></td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Level</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine</td>
<td>62.8%</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>72.8%</td>
<td></td>
</tr>
<tr>
<td>Coarse</td>
<td>76.7%</td>
<td></td>
</tr>
</tbody>
</table>

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.
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

Human
75.0%

Animal
65.9%

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


Note: images here are videos in the original slides.
Alas, what makes such a good representation?


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.
Thank you!

Note: images here are videos in the original slides.