COLLEGE OF ENGINEERING
ELECTRICAL ENGINEERING & COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN

Toward the Who and Where of Action Recognition

Jason J. Corso
Associate Professor
Electrical Engineering and Computer Science
University of Michigan

Joint work with Chenliang Xu (PhD Student)

http://web.eecs.umich.edu/~jjcorso jjcorso@eecs.umich.edu

July 21 2015 @ ERC ALLEGRO Workshop, INRIA Grenoble

What?

What?

Source: Goodfellas (copyright Columbia Pictures) used under fair use; Video trimmed from GaTech Video Segmentation data set.

What?

What? Who?

- Who’s birthday party was it?

What? Who? How many?

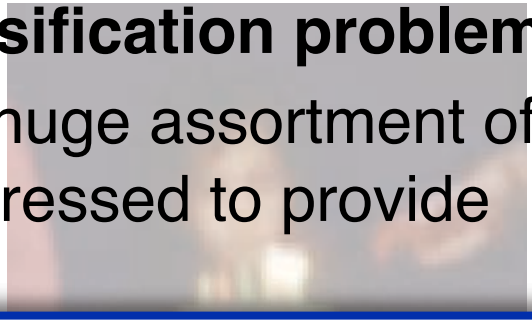
- How many kids were at the party?

What? Who? How ...

- How old was she?

Video Understanding; What? Who? How many? How old was she? Where? When? What time of year? Inside or outside? Why? Where did the boy go? Were they singing? What would happen next? ....

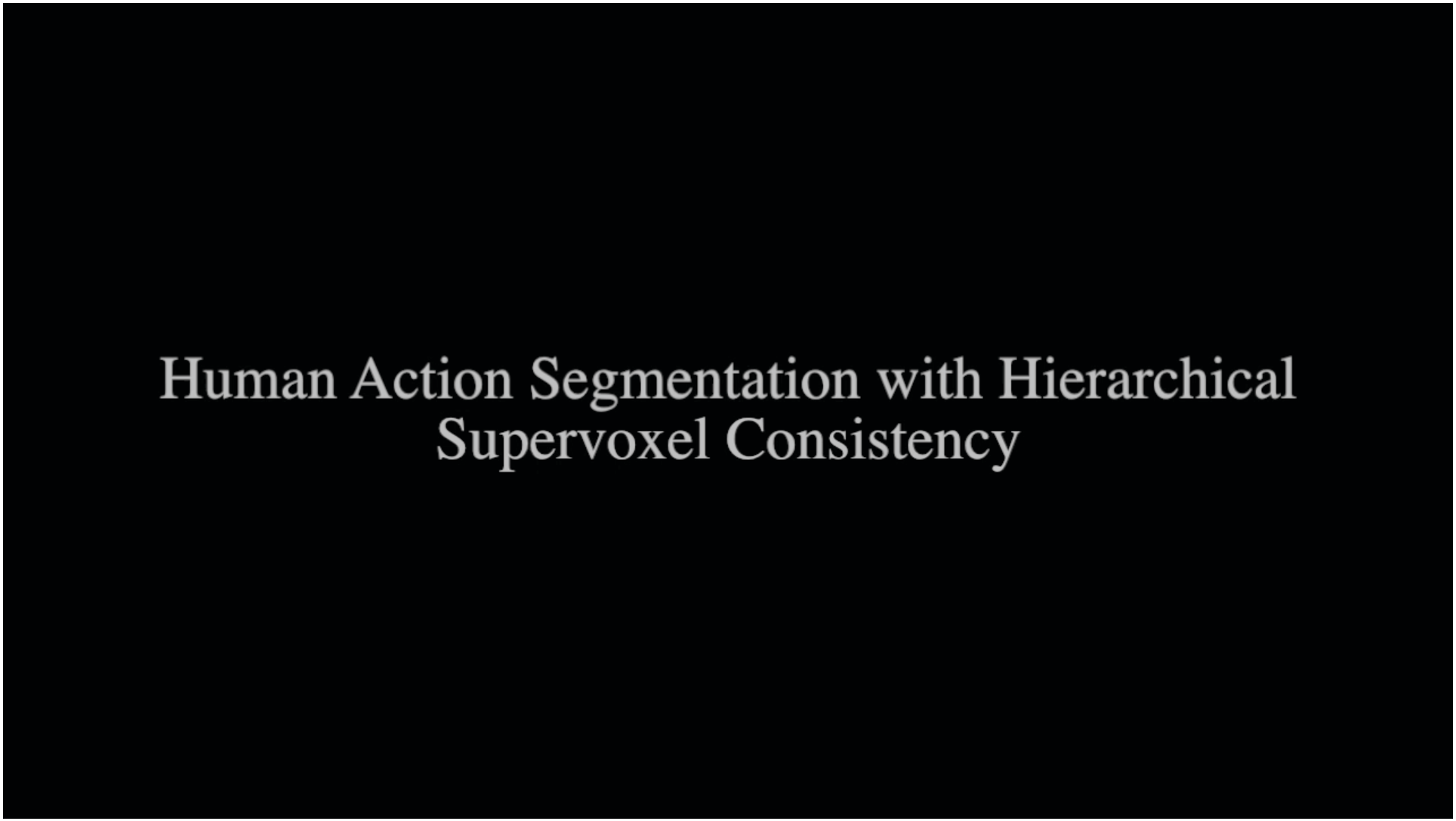
- **Video understanding is not just a classification problem.**
- The video understanding problem has a huge assortment of associated problems that need to be addressed to provide solutions to real-world applications.



Applications of Video Understanding

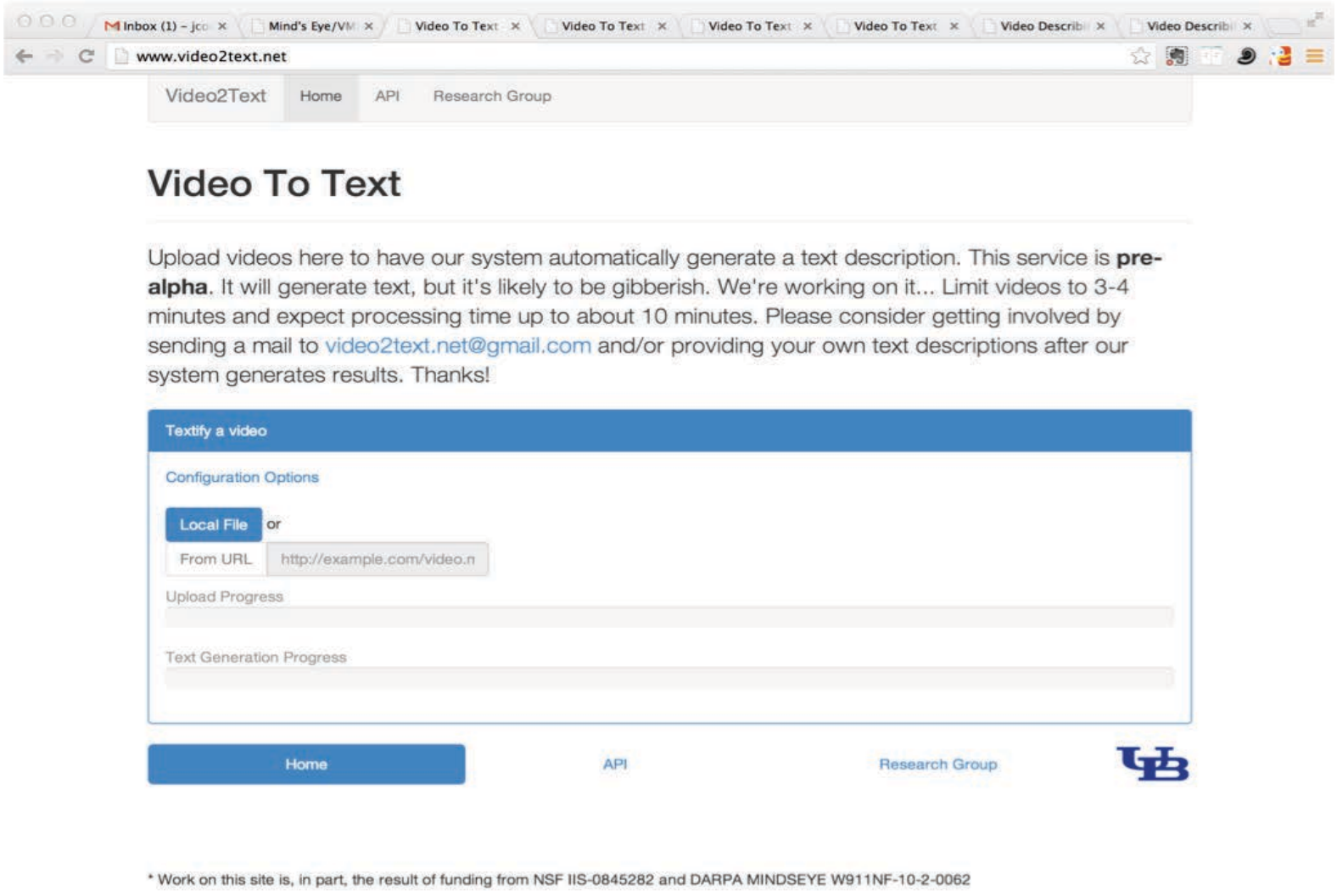
- **Real-time / Interactive**
  - Mobile robotic guidance, navigation and manipulation.
  - Human computer/machine/robot interaction and entertainment.
  - Healthcare monitoring and surveillance.
- **Off-line**
  - Video indexing and search.
  - Video to text.
  - Sports analysis.
  - Advertising analytics.

Action Segmentation



[Liu, …, Corso CVPR 2015]

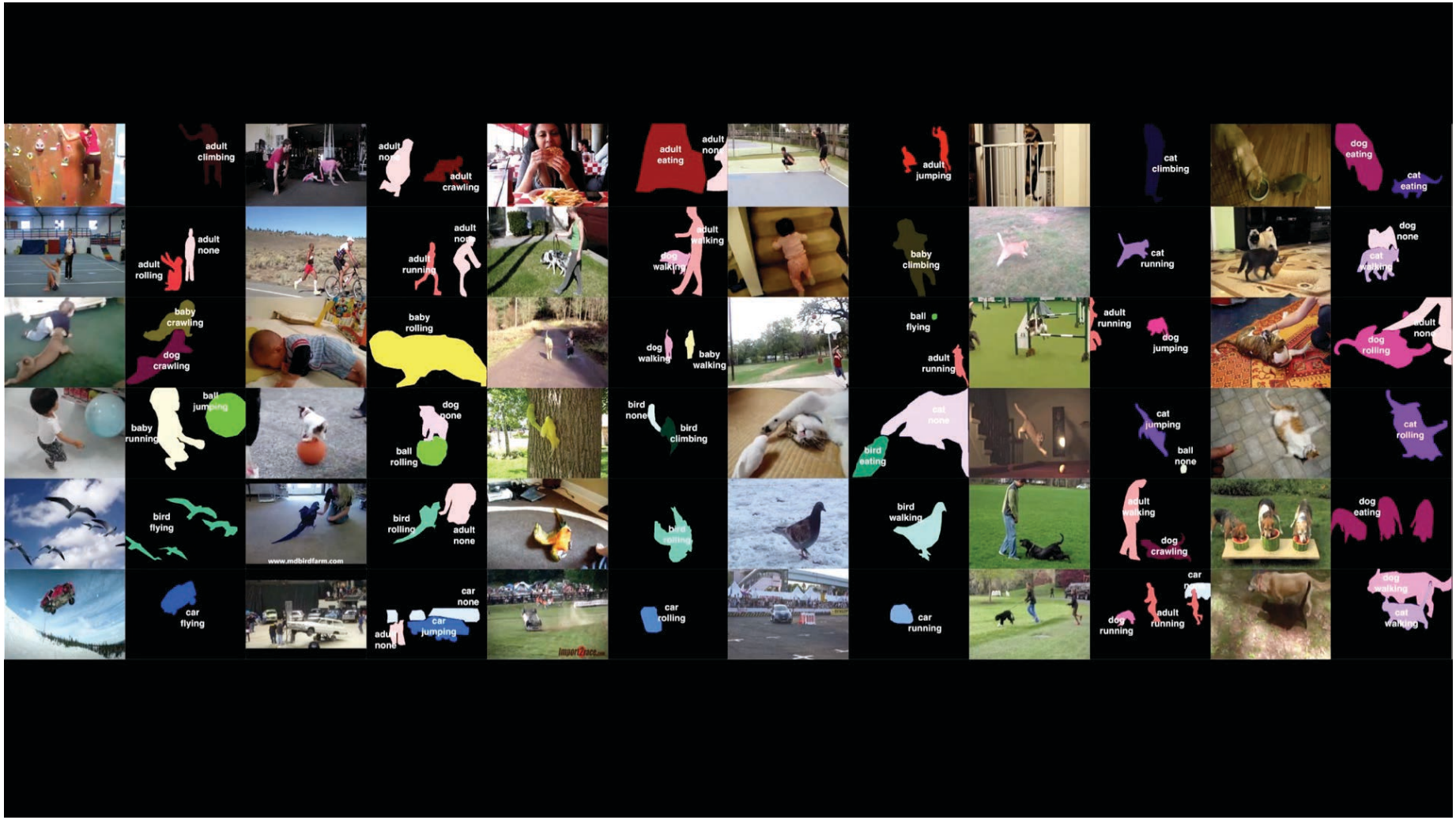
Video to Language Translation



[Das, …, Corso, WSDM 2013, CVPR 2013; Xu et al.]

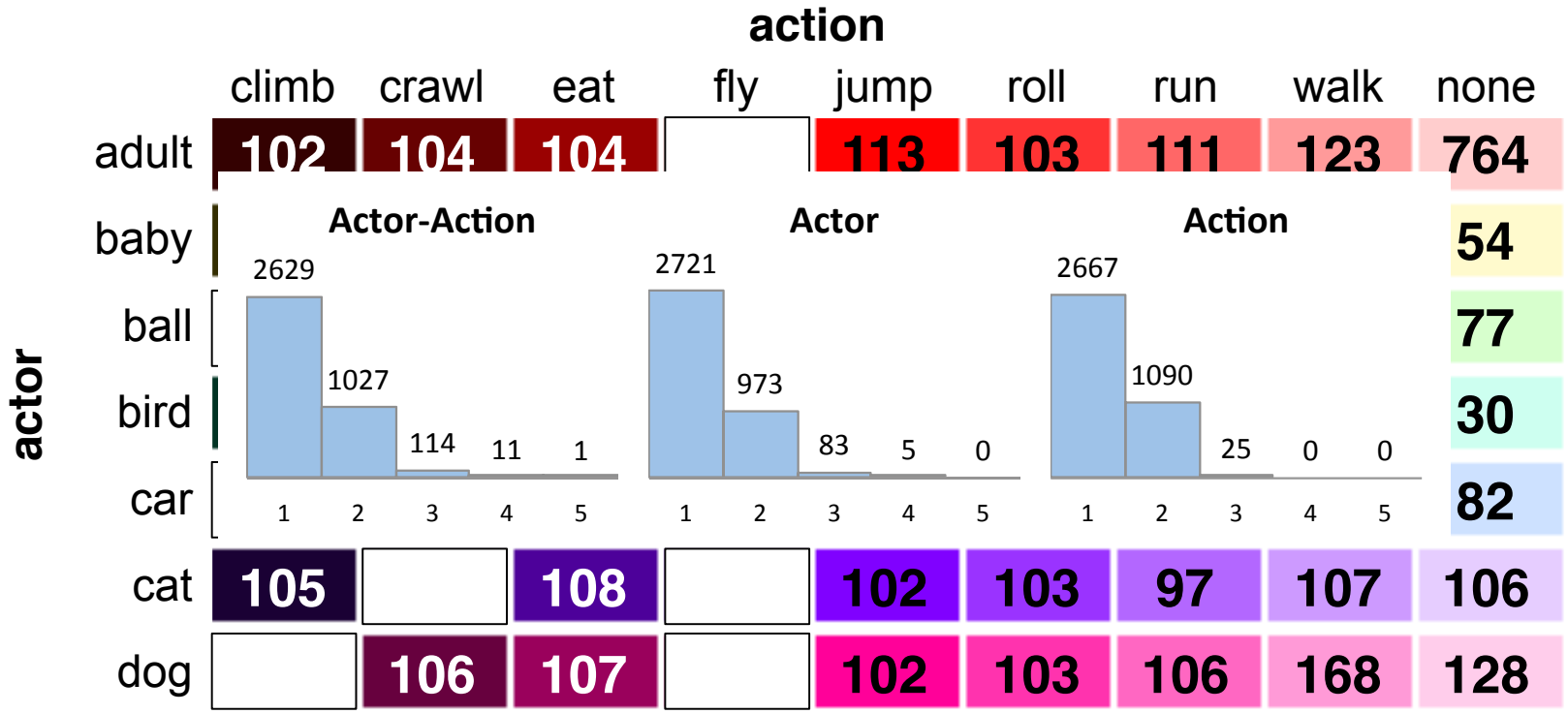
Toward the Who and Where

Joint Actor-Action Understanding



[Xu, …, Corso CVPR 2015]

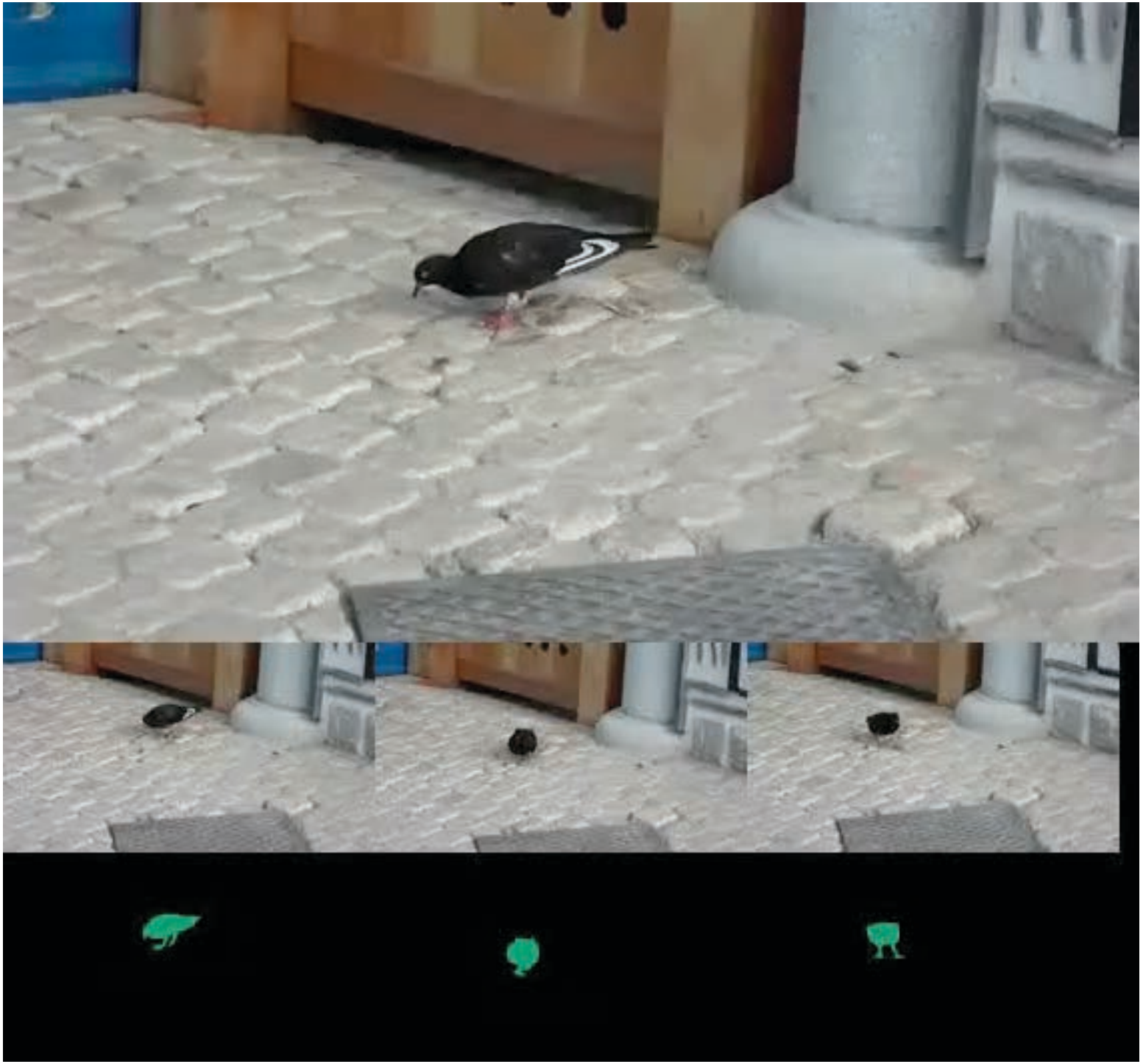
A2D: Actor-Action Dataset Statistics



- We have collected a dataset of 3782 videos.
  - Average length: 136F; Minimum: 24F; Maximum: 332F.
  - One-third have more than one actor performing different actions.
  - We split the dataset into 80% training and 20% testing divided evenly over all actor-action tuples.
- This is the **first actor+action large dataset** in vision.

[Xu, …, Corso CVPR 2015]


adult-climbing
adult-crawling
adult-eating
adult-jumping
adult-rolling
adult-running
adult-walking
baby-climbing
baby-crawling
baby-rolling
baby-walking
ball-flying
ball-jumping
ball-rolling
bird-climbing
bird-eating
bird-flying
bird-jumping
bird-rolling
bird-walking
car-flying
car-jumping
car-rolling
car-running
cat-climbing
cat-eating
cat-jumping
cat-rolling
cat-running
cat-walking
dog-crawling
dog-eating
dog-jumping
dog-rolling
dog-running
dog-walking



Actor-Action Problem Statement

- Let $\mathcal{V}$ denote a video.
  - $\mathcal{X}$: set of actor labels {adult, baby, ball, bird, car, cat, dog}.
  - $\mathcal{Y}$: set of action labels {climbing, crawling, eating, flying, jumping, rolling, running, walking, none}.

	climb	crawl	eat	fly	jump	roll	run	walk	none
adult	102	104	104		113	103	111	123	764
baby	104	103				106		109	54
ball				91	93	105			77
bird	99		102	103	102	107		109	30
car				103	108	102	111		82
cat	105		108		102	103	97	107	106
dog		106	107		102	103	106	168	128

- The general actor-action understanding problem is a posterior maximization:

$$(\mathbf{x}^*, \mathbf{y}^*) = \text{argmax}_{\mathbf{x},\mathbf{y}} P(\mathbf{x}, \mathbf{y}|\mathcal{V})$$

  - $\mathbf{x}$ and $\mathbf{y}$ are sets of random variables for actor and action. The various instantiations of the actor-action problem define their specific dimensionalities.

[Xu, …, Corso CVPR 2015]

Three Instantiations

- Single-Label Actor-Action Recognition.
  - $\mathbf{x}$ and $\mathbf{y}$ are simply scalars x and y .
  - Depict the single actor and action label for a give video $\mathcal{V}$.

[Xu, …, Corso CVPR 2015]

Actor-Action Recognition

- A standard action recognition pipeline with dense trajectory features. [Wang et al. IJCV'13]
  - Train one versus all SVM models for video classification.

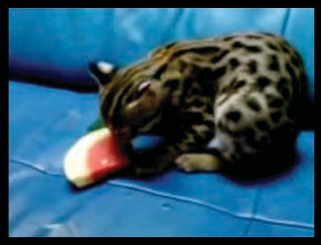
- NaÏve Bayes.
  - Separate classifiers over $\mathcal{X}$ and $\mathcal{Y}$ then score them together.
- Joint Product Space.
  - Train classifier for each actor-action tuple in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.
- Trilayer Model.
  - Learn classifiers over the actor space $\mathcal{X}$, the action space $\mathcal{Y}$ and the joint actor-action space $\mathcal{Z}$.
  - During inference, it separately infers the naÏve Bayes terms and the joint product space terms and then takes a linear combination of them to yield the final score.

[Xu, …, Corso CVPR 2015]

Experiments

GroundTruth: car-flying
Naïve Bayes: adult-flying
JointPS: car-flying
Trilayer: car-flying

GroundTruth: bird-walking
Naïve Bayes: bird-crawling
JointPS: bird-walking
Trilayer: bird-walking



GroundTruth: cat-eating
Naïve Bayes: baby-eating
JointPS: cat-eating
Trilayer: cat-eating

GroundTruth: adult-crawling
Naïve Bayes: dog-crawling
JointPS: adult-crawling
Trilayer: adult-crawling

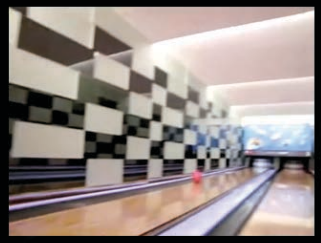
The naïve Bayes model applies actor and action classifiers independently to test videos, and thus it results labels that are outside of the 36 valid actor-action combinations. Wrong predictions are marked in red above.

When we consider both actor and action in the evaluation, it is clearly beneficial to jointly model them.
The joint produce space model and the trilayer model correctly predict the above videos.

[Xu, …, Corso CVPR 2015]

Experiments

GroundTruth: baby-walking
Naïve Bayes: baby-rolling
JointPS: baby-crawling
Trilayer: baby-walking



GroundTruth: ball-rolling
Naïve Bayes: car-rolling
JointPS: car-running
Trilayer: ball-rolling

GroundTruth: bird-eating
Naïve Bayes: cat-eating
JointPS: baby-climbing
Trilayer: bird-eating

GroundTruth: adult-crawling
Naïve Bayes: adult-walking
JointPS: bird-walking
Trilayer: adult-crawling

The trilayer model outperforms the other two models in terms of both individual actor or action tasks as well as the joint actor-action task (as shown in above videos). This implies that the side information of the actor when doing action recognition provides useful information to improve the inference task.

The visualization is only for single-label actor-action recognition and the multiple-label one supports the same conclusion.

[Xu, …, Corso CVPR 2015]

Experiments

	Single-Label			Multiple-Label		
	Classification Accuracy			Mean Average Precision		
Model	Actor	Action	<A, A>	Actor	Action	<A, A>
Naïve Bayes	63.53	68.63	48.39	69.87	72.94	51.76
JointPS	63.24	64.34	52.54	70.49	70.77	55.24
Trilayer	**65.95**	**68.77**	**53.89**	**73.04**	**74.27**	**58.85**

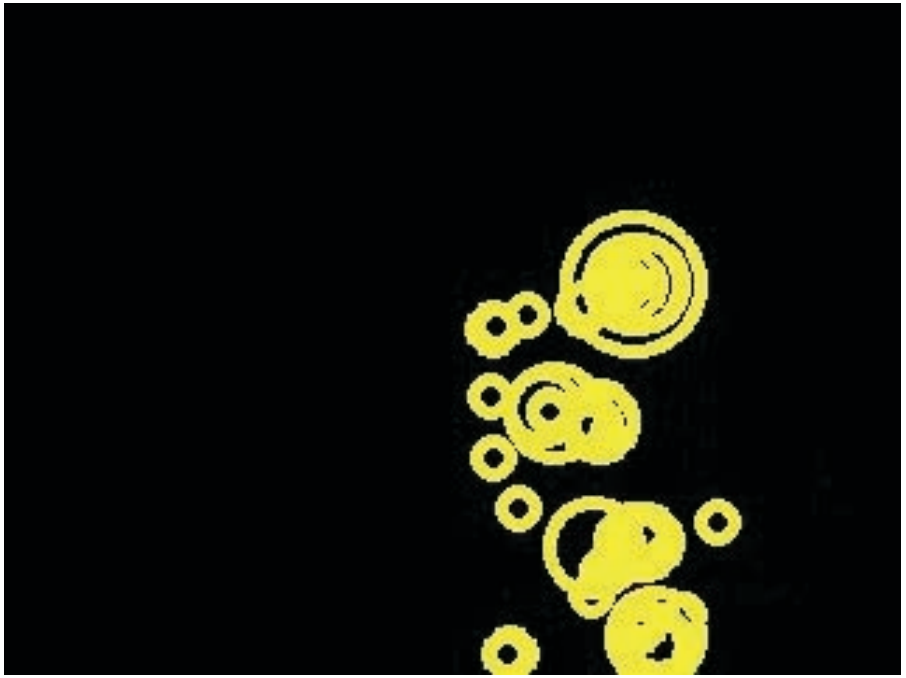
- Metrics
  - Classification Accuracy for single-label.
  - Mean Average Precision for multiple-label.
- Observation 1:
  - The independent model for action outperforms the joint product space model for action.
  - When consider both <A,A>, it is beneficial to jointly model them.
- Observation 2:
  - Trilayer model outperforms the other two methods.
- Observation 3:
  - Even if we only care about actor or action, it is better to jointly model them.

[Xu, …, Corso CVPR 2015]

Actor-Action Semantic Segmentation

[Xu, …, Corso CVPR 2015]

What do we need to represent to segment actions?



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

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

What do we need to represent to segment actions?

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

What do we need to represent to segment actions?

Video Segmentation: A Complementary Feature?

[Xu, Xiong, Corso ECCV 2012]

Space-Time Video Segmentation

- Represent video $\mathcal{V}$ on lattice Γ.
- Basic problem statement:

Segmentation Video Input

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} E(\mathcal{S}|\mathcal{V})$$

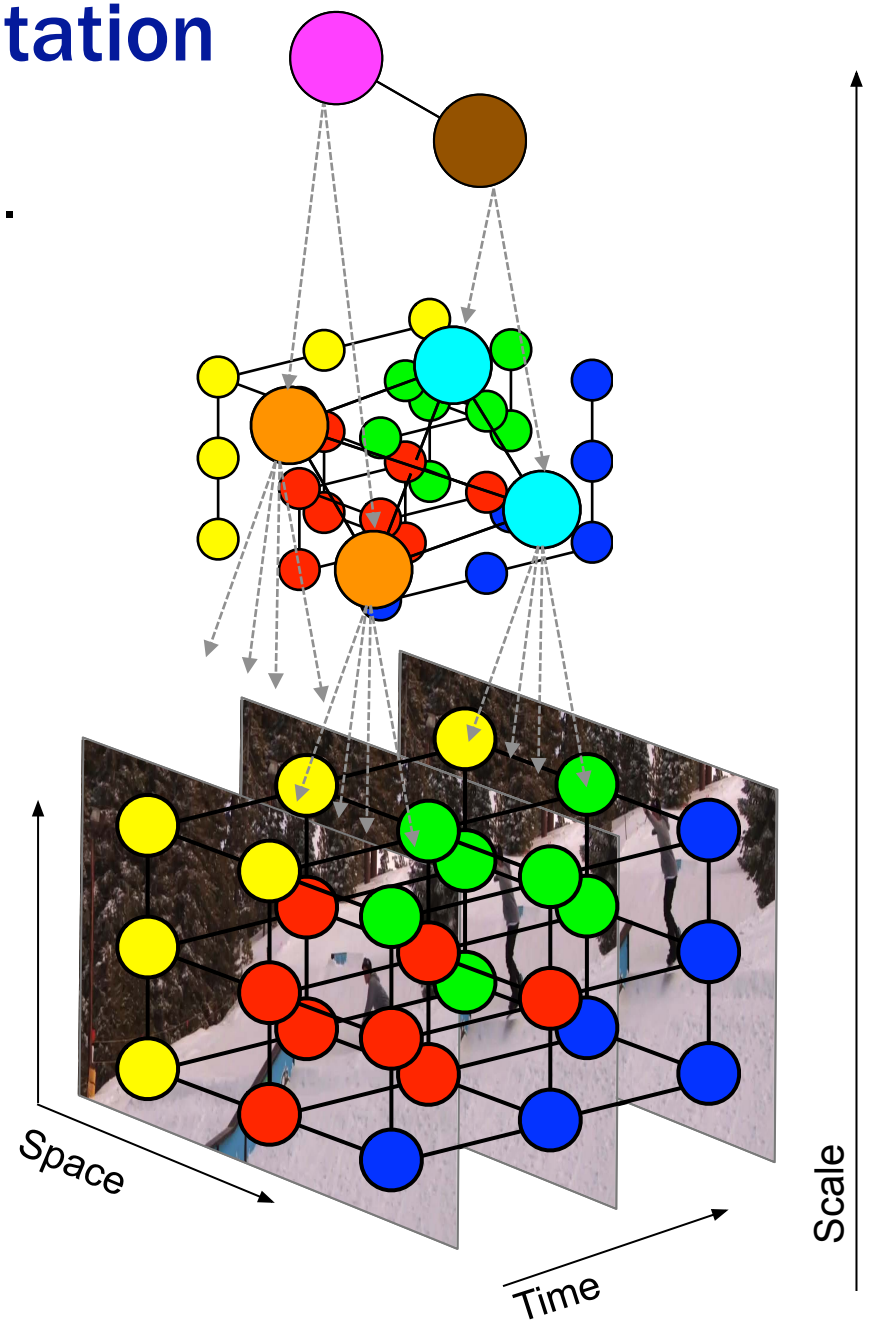
- Segmentation hierarchy

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

$$S^i \doteq \{s_1, s_2, \ldots\}$$

such that $s_j \subset \Gamma, \cup_j s_j = \Gamma,$

and $s_i \cap s_j = \varnothing$ for pairs i, j



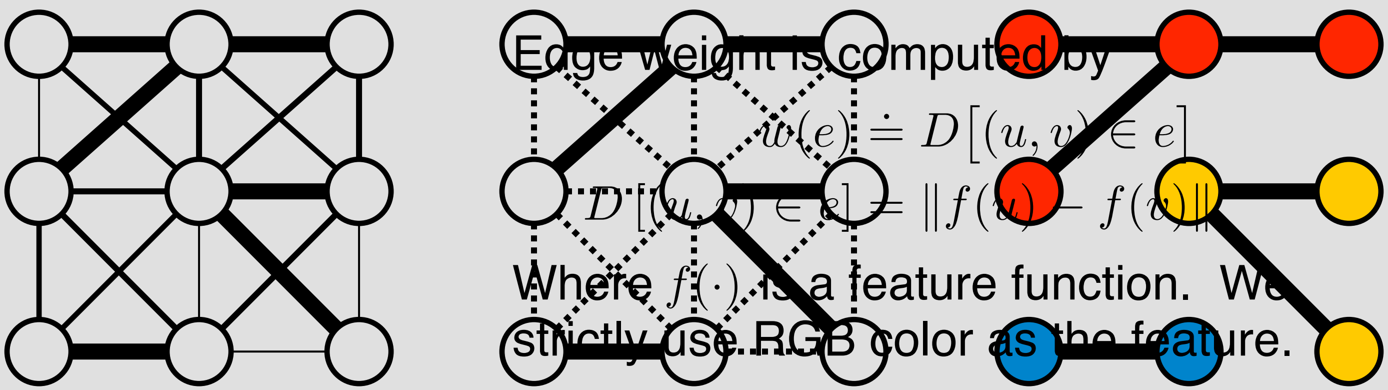
[Xu, Corso CVPR 2012]

Choosing the Energy Function

- Mumford-Shah (Discretized)
  - Explained variation and boundary complexity.
  - Let δS^1 be the boundary of S^1.

- Minimum Spanning Forest
  - Intra-segment agreement plus inter-segment difference.

$$E(S^1|\mathcal{V}) = \tau \sum \sum w(e)$$



Edge weight is computed by

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

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

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

$$E(S^1|\mathcal{V}) = \frac{\sum_{s_i \cup s_j} w_{s_i s_j}}{\sum_{s_i \cap s_j} w_{s_i s_j}}$$

$$E(S^1|\mathcal{V}) = \sum_i \phi(s_i|\mathcal{V}) + \sum_i \sum_{j \in \mathcal{N}(i)} \psi(s_i, s_j|\mathcal{V})$$

Corso, Sharon and Yuille. MICCAI 2006, TMI 2008

Ref: Geman and Geman. PAMI 1984.

Corso et al. MICCAI 2007, CVPR 2008

[Xu, Corso CVPR 2012]

Video Segmentation: A Complementary Feature?

Video Segmentation: A Complementary Feature?

Video Segmentation: A Complementary Feature?

Video Segmentation: A Complementary Feature?

What semantics are retained in video segmentation?

- We have systematically studied this question.
- 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?

[Xu, …, Corso IJSC 2014]

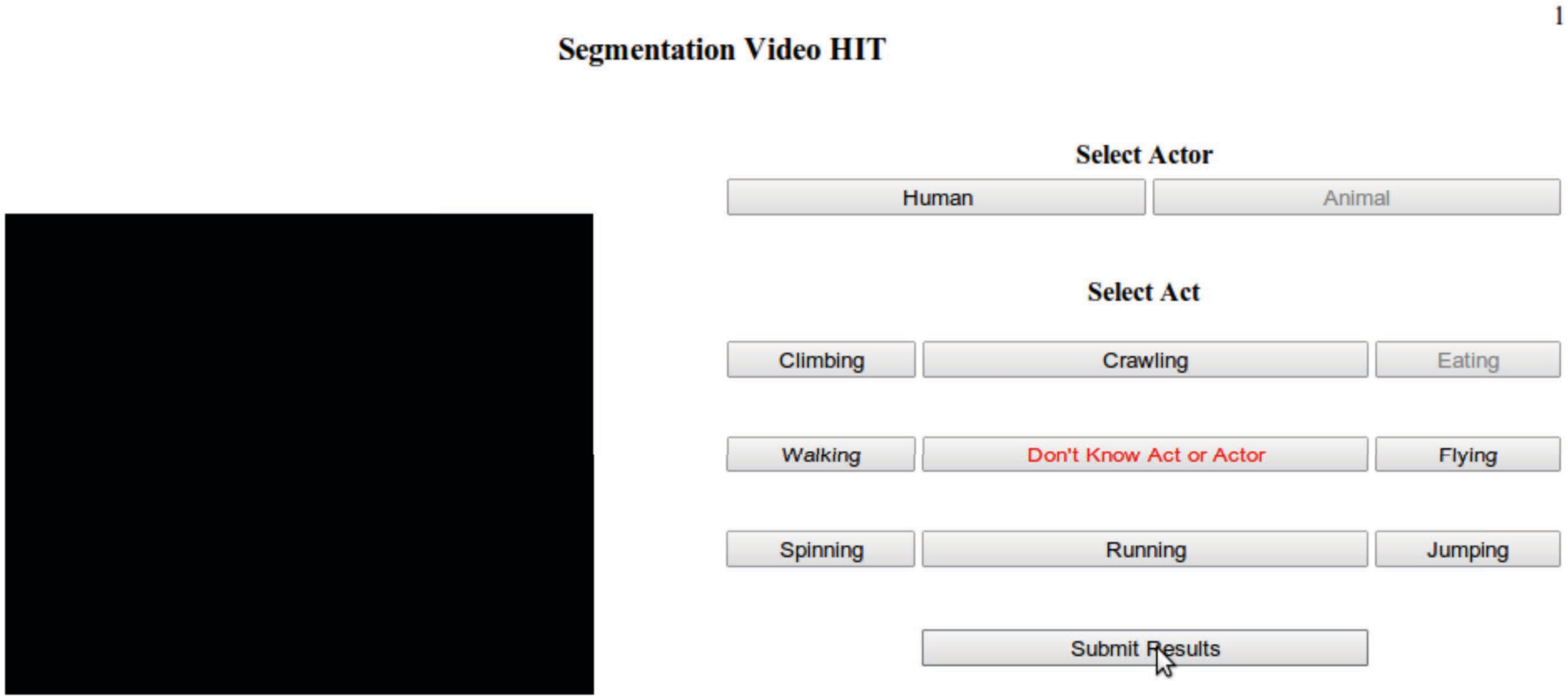
Study Setup: Data Set

- 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

[Xu, …, Corso IJSC 2014]

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.



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.
- Suspected Reasons:
  - 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.

[Xu, …, Corso IJSC 2014]

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

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	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, …, Corso IJSC 2014]

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	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, …, Corso IJSC 2014]

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?

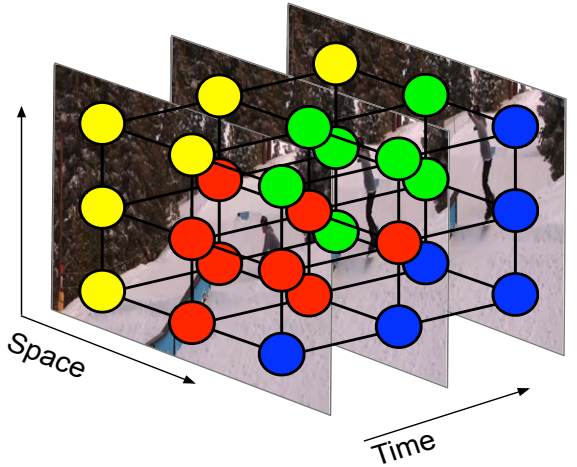
- Yes! 82% accuracy on actor and 70% on action.
- Performance increases with coarseness of the signal.
- Performance for human actors is better than animals.
- Performance for a static background is better than a moving background.

[Xu, …, Corso IJSC 2014]

Actor-Action Semantic Seg. Problem Statement

- Let $\mathcal{V} = \{v_1, \ldots, v_n\}$ denote a video with n supervoxels in a video segmentation represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
  - $\mathcal{X}$: set of actor labels {adult, baby, ball, bird, car, cat, dog}.
  - $\mathcal{Y}$: set of action labels {climbing, crawling, eating, flying, jumping, rolling, running, walking, none}.

$$(\mathbf{x}^*, \mathbf{y}^*) = \text{argmax}_{\mathbf{x},\mathbf{y}} P(\mathbf{x}, \mathbf{y} | \mathcal{V})$$

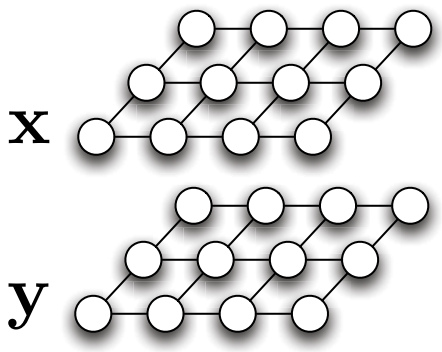


- Actor-Action Semantic Segmentation.
  - $\mathbf{x} = \{x_1, \ldots, x_n\}$ and $\mathbf{y} = \{y_1, \ldots, y_n\}$ are two sets of random variables defined on the set of supervoxels $\mathcal{V} = \{v_1, \ldots, v_n\}$ of a video, and assign each $x_i \in \mathcal{X}$ and each $y_i \in \mathcal{Y}$.

[Xu, …, Corso CVPR 2015]

Naïve Bayes Model

- Assume independent over actor and action.



$$P(\mathbf{x}, \mathbf{y}|\mathcal{V}) = P(\mathbf{x}|\mathcal{V})P(\mathbf{y}|\mathcal{V})$$
$$= \prod_{i \in \mathcal{V}} P(x_i)P(y_i) \prod_{i \in \mathcal{V}} \prod_{j \in \mathcal{E}(i)} P(x_i, x_j)P(y_i, y_j)$$
$$\propto \prod_{i \in \mathcal{V}} \phi_i(x_i)\psi_i(y_i) \prod_{i \in \mathcal{V}} \prod_{j \in \mathcal{E}(i)} \phi_{ij}(x_i, x_j)\psi_{ij}(y_i, y_j)$$

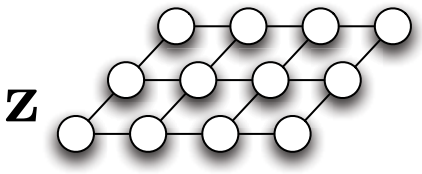
Separate potential functions defined on actor and action nodes alone.

Pairwise potential functions within sets of actor nodes and sets of action nodes (contrast-sensitive Potts model) to encourage smoothness.

[Xu, …, Corso CVPR 2015]

Joint Product Space Model

- Each actor-action pair is considered as one class.



$$\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$$

$$P(\mathbf{x}, \mathbf{y}|\mathcal{V}) \doteq P(\mathbf{z}|\mathcal{V}) = \prod_{i \in \mathcal{V}} P(\mathbf{z}_i) \prod_{i \in \mathcal{V}} \prod_{j \in \mathcal{E}(i)} P(z_i, z_j)$$

$$\propto \prod_{i \in \mathcal{V}} \varphi_i(z_i) \prod_{i \in \mathcal{V}} \prod_{j \in \mathcal{E}(i)} \varphi_{ij}(z_i, z_j)$$

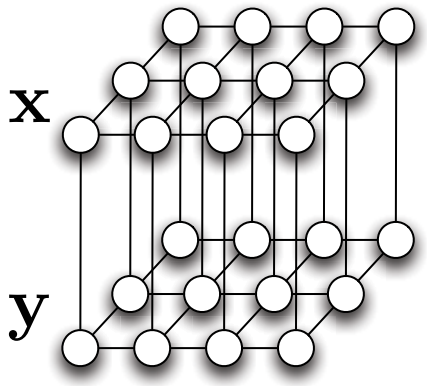
$$= \prod_{i \in \mathcal{V}} \varphi_i([x_i, y_i]) \prod_{i \in \mathcal{V}} \prod_{j \in \mathcal{E}(i)} \varphi_{ij}([x_i, y_i], [x_j, y_j])$$

Potential function for joint actor-action product space label.

[Xu, …, Corso CVPR 2015]

Bilayer Model

- Capture the covariance across the actor and action labels.



$$P(\mathbf{x}, \mathbf{y}|\mathcal{V}) = \prod_{i \in \mathcal{V}} P(x_i, y_i) \prod_{i \in \mathcal{V}} \prod_{j \in \mathcal{E}(i)} P(x_i, x_j) P(y_i, y_j)$$

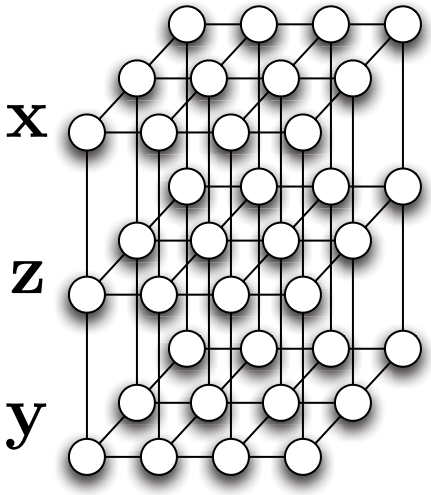
$$\propto \prod_{i \in \mathcal{V}} \phi_i(x_i) \psi_i(y_i) \xi_i(x_i, y_i) \prod_{i \in \mathcal{V}} \prod_{j \in \mathcal{E}(i)} \phi_{ij}(x_i, x_j) \psi_{ij}(y_i, y_j)$$

A learned potential function over the product space of labels.
E.g. $\varphi_i([x_i, y_i])$

[Xu, …, Corso CVPR 2015]

Trilayer Model

- Same actor performs different actions.
- Same actions being performed by different actors.



$$P(\mathbf{x},\mathbf{y},\mathbf{z}|\mathcal{V}) = P(\mathbf{x}|\mathcal{V})P(\mathbf{y}|\mathcal{V})P(\mathbf{z}|\mathcal{V}) \prod_{i\in\mathcal{V}} P(x_i,z_i)P(y_i,z_i)$$

$$\propto \prod_{i\in\mathcal{V}} \phi_i(x_i)\psi_i(y_i)\varphi_i(z_i)\mu_i(x_i,z_i)\nu_i(y_i,z_i)\cdot$$

$$\prod_{i\in\mathcal{V}} \prod_{j\in\mathcal{E}(i)} \phi_{ij}(x_i,x_j)\psi_{ij}(y_i,y_j)\varphi_{ij}(z_i,z_j)$$

$$\mu_i(x_i,z_i) = \begin{cases} w(y_i'|x_i) & \text{if } x_i = x_i' \text{ for } z_i = [x_i', y_i'] \\ 0 & \text{otherwise} \end{cases}$$

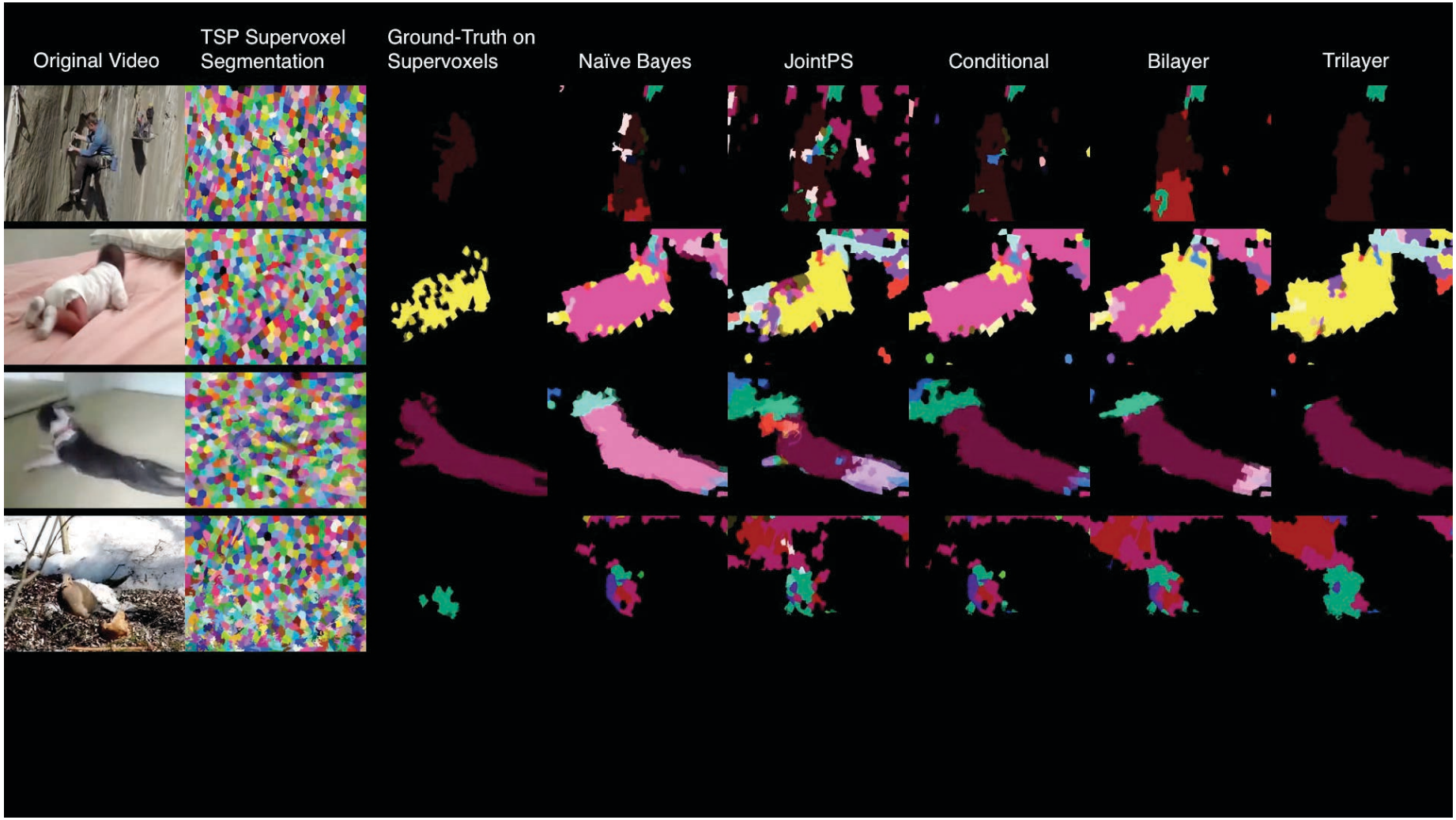
$$\nu_i(y_i,z_i) = \begin{cases} w(x_i'|y_i) & \text{if } y_i = y_i' \text{ for } z_i = [x_i', y_i'] \\ 0 & \text{otherwise} \end{cases}$$

Scores of conditional classifiers.

[Xu, …, Corso CVPR 2015]

Experiments

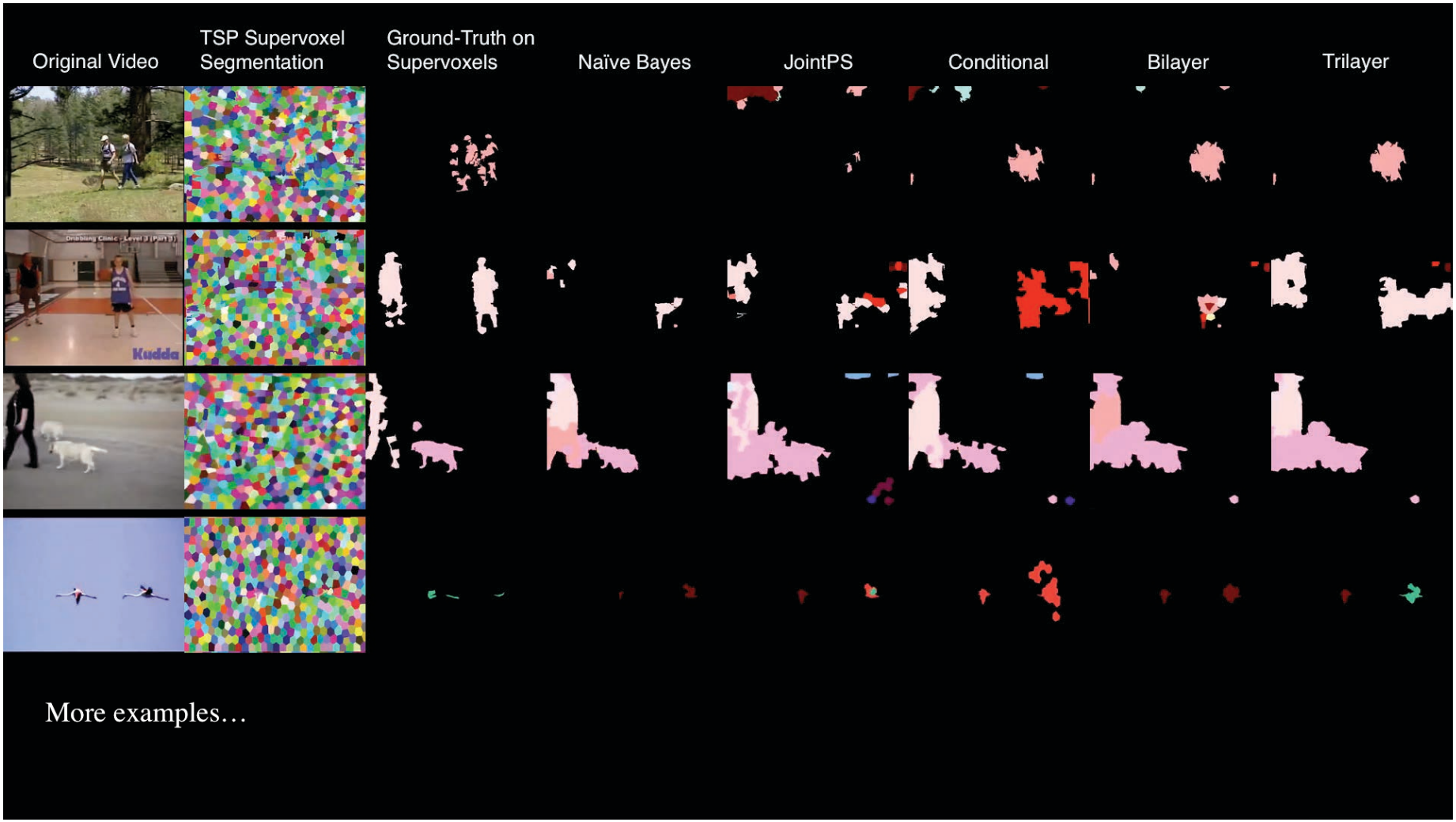
	climb	crawl	eat	fly	jump	roll	run	walk	none
adult	102	104	104		113	103	111	123	764
baby	104	103				106		109	54
ball				91	93	105			77
bird	99		102	103	102	107		109	30
car				103	108	102	111		82
cat	105		108		102	103	97	107	106
dog		106	107		102	103	106	168	128



[Xu, …, Corso CVPR 2015]

Experiments

	climb	crawl	eat	fly	jump	roll	run	walk	none
adult	102	104	104		113	103	111	123	764
baby	104	103				106		109	54
ball				91	93	105			77
bird	99		102	103	102	107		109	30
car				103	108	102	111		82
cat	105		108		102	103	97	107	106
dog		106	107		102	103	106	168	128



[Xu, …, Corso CVPR 2015]

Experiments

- Metric: average per-class accuracy.
- Conditional model is a simplified bilayer model with different aggregation.
- Unary (nodes), Full (nodes & edges).

Unary Term Only

Model	Average Per Class Accuracy		
	Actor	Action	<A, A>
Naïve Bayes	43.54	39.35	15.59
JointPS	41.23	37.26	19.93
Conditional	43.54	39.98	21.25
Bilayer	43.54	39.35	15.59
Trilayer	43.41	40.14	21.80

Full Model

Model	Average Per Class Accuracy		
	Actor	Action	<A, A>
Naïve Bayes	45.35	42.11	18.65
JointPS	41.87	38.24	20.71
Conditional	45.35	40.87	23.05
Bilayer	44.74	42.74	23.13
Trilayer	**45.63**	**44.22**	**25.36**

[Xu, …, Corso CVPR 2015]

Recognition–segmentation performance gap?

- Trilayer model
  - Segmentation: 26.46% per-class accuracy (at pixel level).
  - Recognition: 66.86% per-class accuracy (at video level).
- Our observations:
  - All action recognition classifiers have high per-class performance, but low per-pixel performance.
  - Performance is related to segment size
    - Smaller segments, worse performance; whole video is best.

- Full-video action recognition may help semantic segmentation as it has good per-class performance.
  - **But how?**

Recognition as Label Cost

- We use video-level recognition as label cost for pixel-level models. The penalty are related with the recognition confidence.
  - Encourage compact labeling.
  - Enforce the video-level recognition confidence globally in a video.

$$E(\mathbf{z}) = \sum_{i \in \mathcal{V}} \phi_i(z_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{E}(i)} \phi_{i,j}(z_i, z_j) + \phi_{\mathcal{L}}(\mathbf{z})$$

$$\phi_{\mathcal{L}}(\mathbf{z}) = w_{\mathcal{L}} \sum_{l} \psi_l \cdot \delta(l, \mathbf{z})$$

$$\delta(l, \mathbf{z}) = \begin{cases} 1, & \text{if } l \in \mathbf{z} \\ 0, & \text{otherwise} \end{cases}$$

  - Note that we are using only the product-space notation for simplicity.

Recognition as Label Cost

	Per-Class
Trilayer	26.46
Svx Unary	20.11
Label Cost	**28.36**

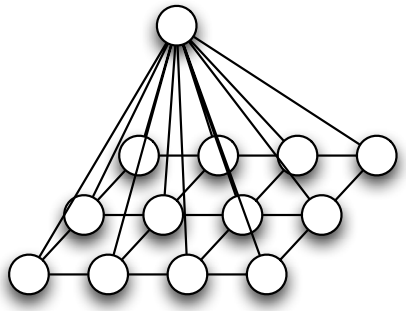
- Per-class segmentic segmentation scores are improved.
  - Trilayer is the best practice in our previous pixel-level models.
  - Here, we use joint product space supervoxel unary and the model improves the performance dramatically.
  - We expect higher performance with trilayer unary.

Recognition as One-Node

- Model the video-level recognition as one node that is connected to all nodes in a video.
- The formulation is similar to a P^N model.

$$E(\mathbf{z}, h) = \sum_{i \in \mathcal{V}} \phi_i(z_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{E}(i)} \phi_{i,j}(z_i, z_j) + \phi_h(h) + \phi_c(\mathbf{z}, h)$$

$\mathbf{h}$



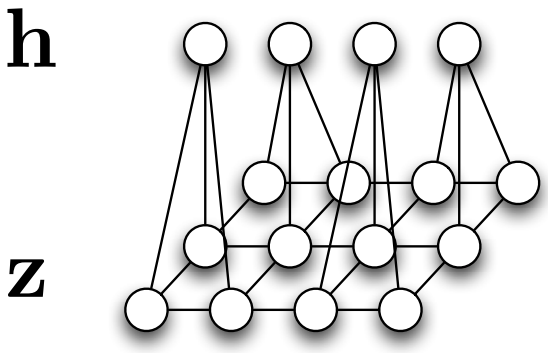
$\mathbf{z}$

$$\phi_c(\mathbf{z}, h) = w_c \sum_{i \in \mathcal{V}} \delta(z_i \neq h)$$

	Per-Class
Trilayer	26.46
Svx Unary	20.11
Label Cost	28.36
One-Node	**31.81**

Recognition on Single-level Supervoxels

- We use coarse-level supervoxel as a video decomposition and extract action Fisher vectors from each supervoxel.
- We obtain the recognition scores of coarse-level supervoxels and use them to supervise the underlying semantic segmentation.



$$E(\mathbf{z}, h) = \sum_{i \in \mathcal{V}} \phi_i(z_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{E}(i)} \phi_{i,j}(z_i, z_j) + \sum_p \phi_h(h_p) + \sum_p \phi_c(\mathbf{z}_p, h_p)$$

	Per-Class
Trilayer	26.46
Svx Unary	20.11
Label Cost	28.36
One-Node	31.81
Single-Level	**34.84**

Conclusion, Acknowledgements, and Thanks!

- **A call to action! And its many constituent problems like localization, segmentation, detection, description, etc.**
- Our contributions:
  - New actor-action understanding problem and dataset.
  - Thorough evaluation demonstrate a benefit for jointly modeling actors and actions *even when you only compare about one of them*.
  - A new trilayer approach to recognition and semantic segmentation that combines both the independent actor and action variations and product-space interactions.
  - New investigation into how global video-level recognition can guide local video semantic segmentation.

- We are grateful for our funding support from NSF, DARPA, IARPA, ARO, NIH, CIA, Google & FHWA.

- Code is available for the methods discussed today.
  - http://web.eecs.umich.edu/~jjcorso/r/code.html

Questions?