4th International Workshop on
Shape Perception in Human and Computer Vision

Shape of Human Activities

Sinisa Todorovic

joint work with William Brendel

Oregon State University

Activity Recognition

Activities with:

- Rich temporal structure
- Shared subactivities

Goal: Recognition and Segmentation

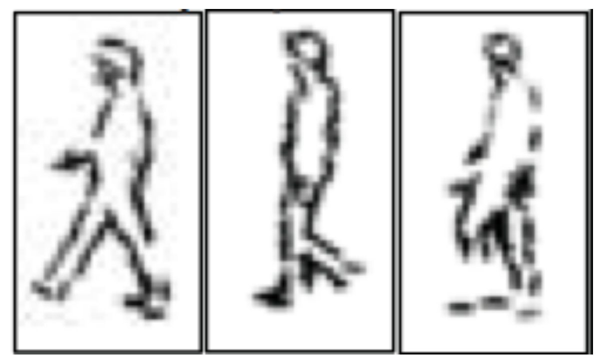
long jump

high jump

- Recognize activities
- Identify the start and end frames
- Explain recognition: space-time structure
- Segment people and objects

Prior Work – Video Representation

- Space-time points
  - Laptev & Schmid 08, Niebles & Fei-Fei 08,...
- Still human postures
  - Soatto 07, Ning & Huang 08,...
- Action Templates
  - Yao & Zhu 09,...
- Point tracks
  - Sukthankar & Hebert 10,...
- Motion segments
  - Gorelick & Irani 08, Pritch & Peleg 08,...



Prior Work – Video Representation

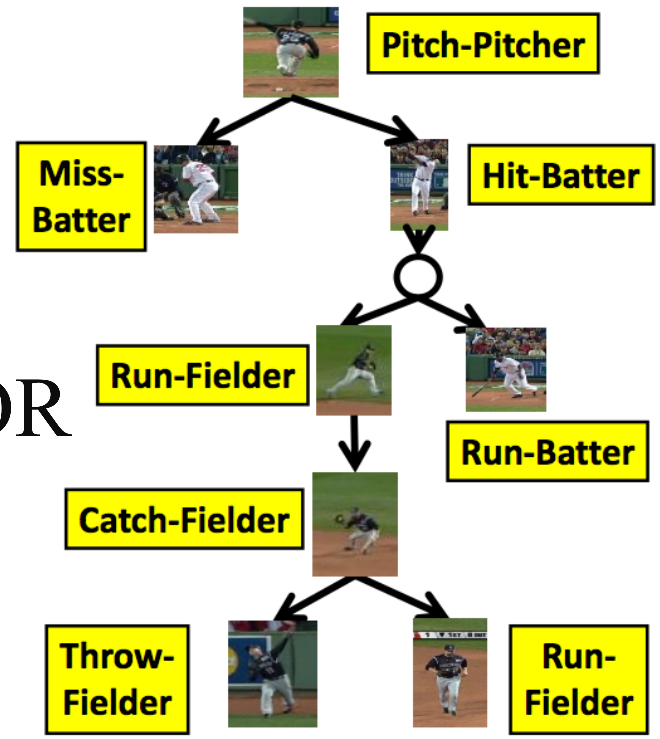
- Space-time points
  - Laptev & Schmid 08, Niebles & Fei-Fei 08,...
- Still human postures
  - Soatto 07, Ning & Huang 08,...
- Action Templates
  - Yao & Zhu 09,...
- Point tracks
  - Sukthankar & Hebert 10,...
- Motion segments
  - Gorelick & Irani 08, Pritch & Peleg 08,...

Too local

Do not capture long-term spatiotemporal structure

Prior Work – Activity Representation

- Classifiers, e.g., Bag-of-Words
  - Ke, Herbert ICCV'05
  - Hamid, Essa ICCV07
  - Laptev, Schmid CVPR'08
  - ...



- Graphical models, e.g., AND-OR
  - Ivanov, Bobick PAMI00
  - Xiang, Gong IJCV'06
  - Ryoo, Aggarwal ICCV'09
  - Gupta, Davis CVPR09
  - Liu, Zhu CVPR09
  - ...

Prior Work – Activity Representation

- Classifiers, e.g., Bag-of-Words
  - Ke, Herbert ICCV'05
  - Hamid, Essa ICCV07
  - Laptev, Schmid CVPR'08

- **Require many examples**
- **Narrow goal: classification**

- Graphical models, e.g., AND-OR
  - Ivanov, Bobick PAMI00
  - Xiang, Gong IJCV'06
  - Ryoo, Aggarwal ICCV'09
  - Gupta, Davis CVPR09
  - Liu, Zhu CVPR09

- **Pre-fixed model structure**
- **Hard to learn**
- **Hard to infer**

Hypothesis

- Point-based features provide poor cues
- More expressive models are needed

Hypothesis

- Point-based features provide poor cues
- More expressive models are needed

⇩ To bridge the semantic gap

- **Use mid-level features: Activity shape**
  - **Less training examples**
  - **Allow simpler learning and inference**

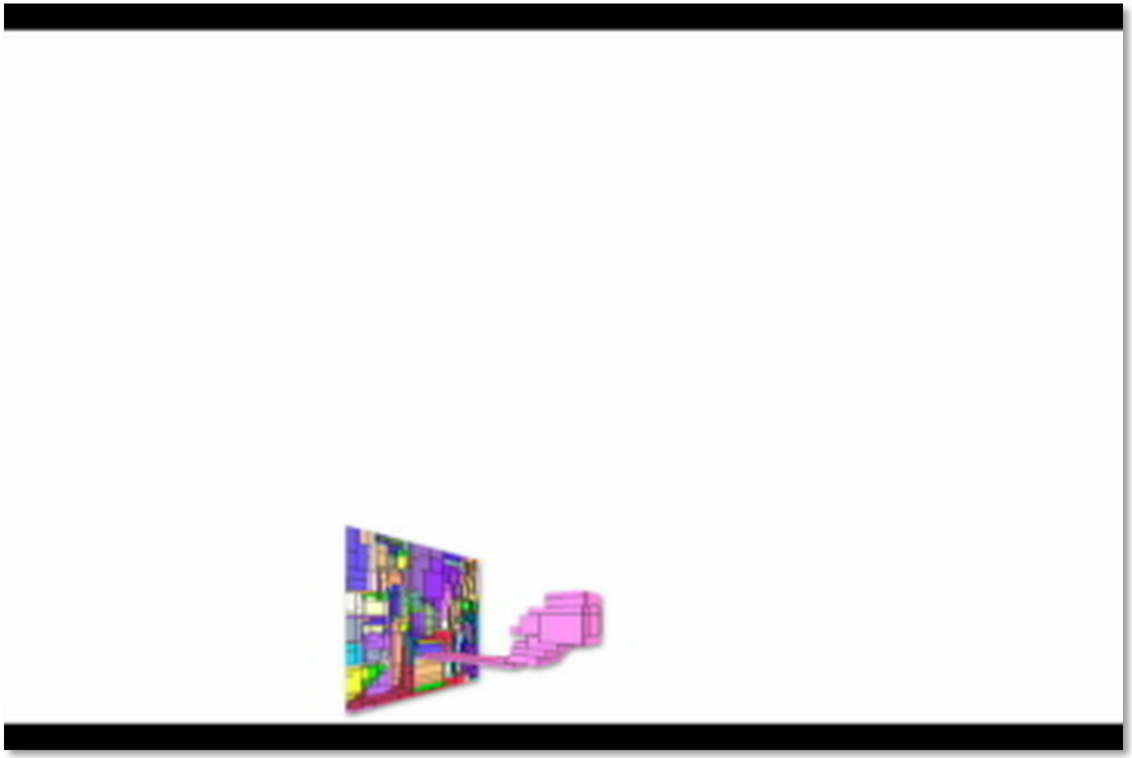
Spatiotemporal Segmentation

Irani & Peleg 94, Weiss 97, Shi & Malik 98, DeMenthon 02, Cohen 04, Greenspan et al. 02, Ahuja 05, Medioni 05, Todorovic 09, Essa 10,...

Activity Shape

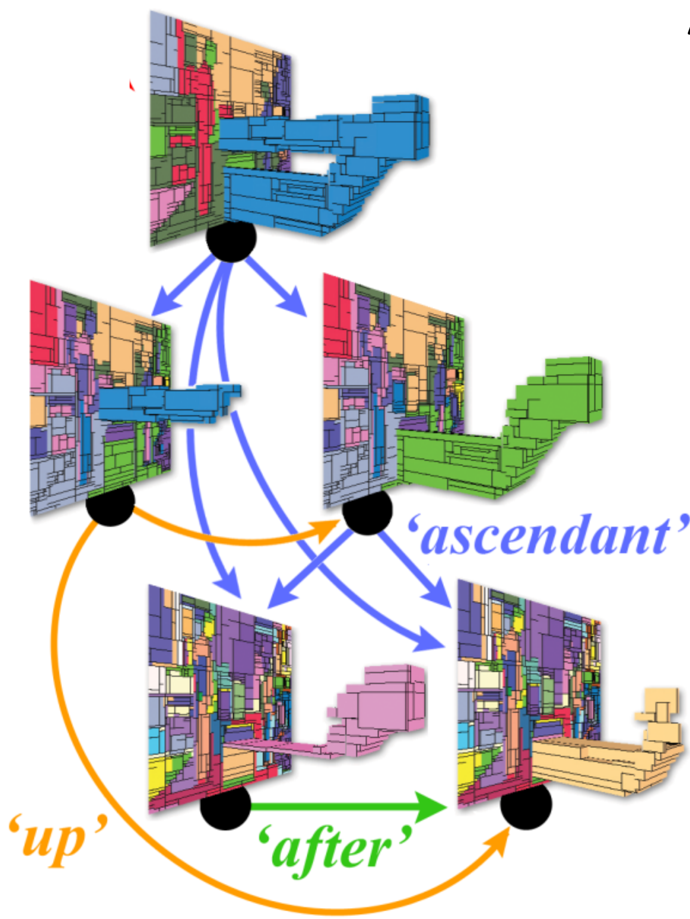
- Objects occupy space-time tubes
- Because they
  - are cohesive in space
  - have locally smooth trajectories in time

Activity Shape = Segmentation Graph



- As the right scale is unknown...
- The graph captures spatiotemporal structure

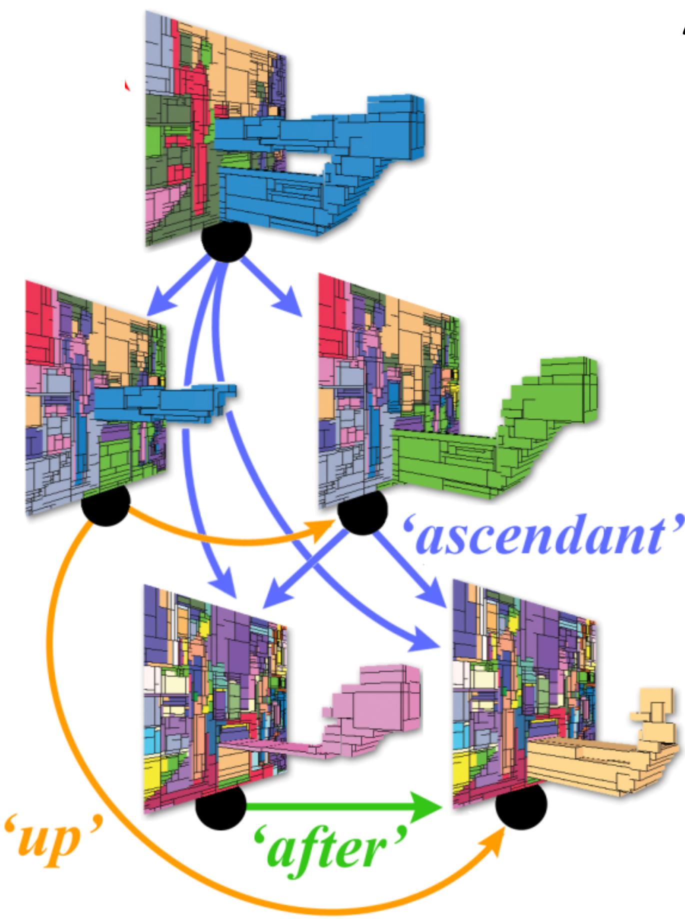
Activity Shape = Segmentation Graph



Attributes of nodes and edges:

– Intrinsic properties: F
  - Motion
  - Object shape

Activity Shape = Segmentation Graph

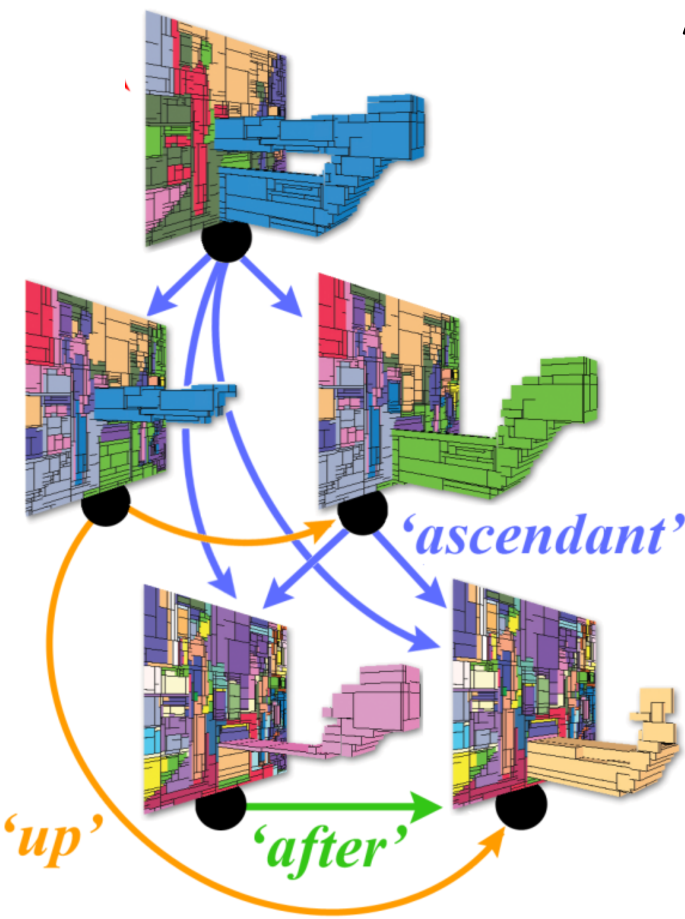


Attributes of nodes and edges:

– Intrinsic properties: F
  - Motion
  - Object shape
– Adjacency matrices: A
  - Allen temporal relations
  - Spatial relations
  - Compositional relations

$$G = (V, E) = \{(A_1, F_1), ..., (A_L, F_L)\}$$

Activity Shape = Segmentation Graph

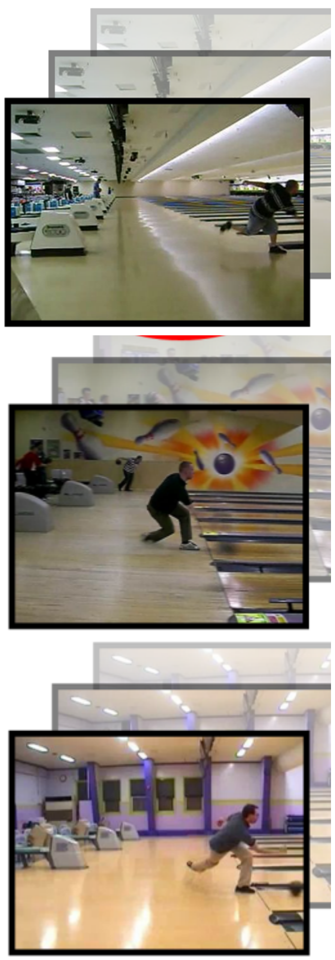


Attributes of nodes and edges:

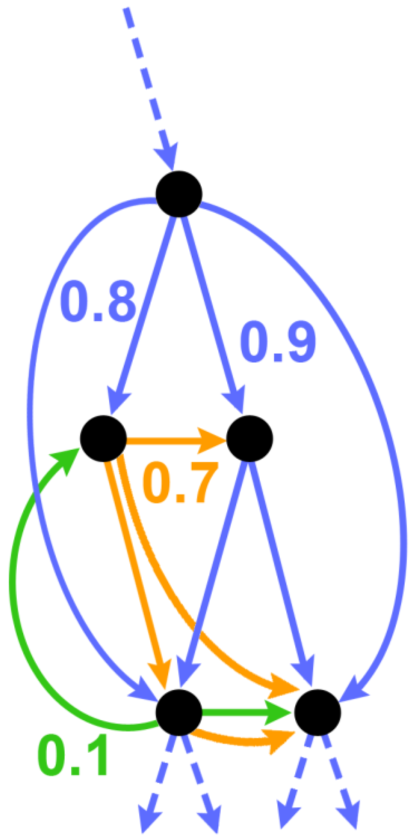
- Intrinsic properties: F
  - Motion
  - Object shape
- Adjacency matrices: A
  - Allen temporal relations
  - Spatial relations
  - Compositional relations

Our Approach

training videos

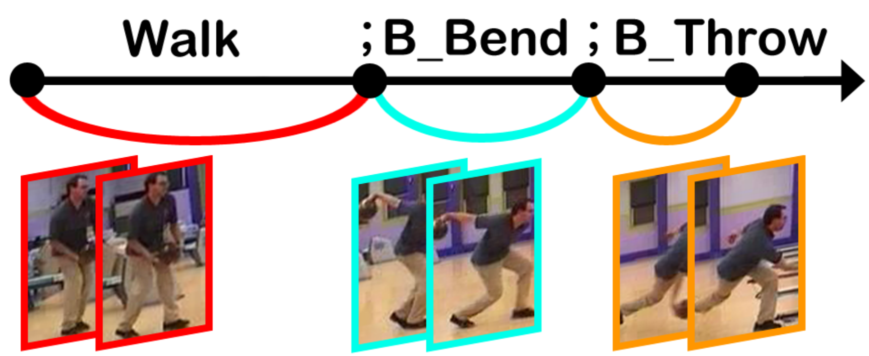


activity model

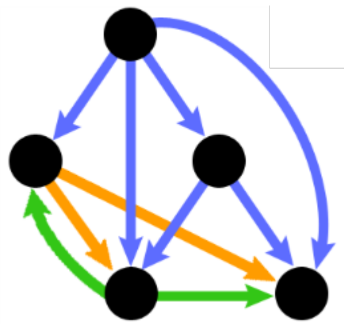


In a new video:

- Recognize
- Segment
- Explain



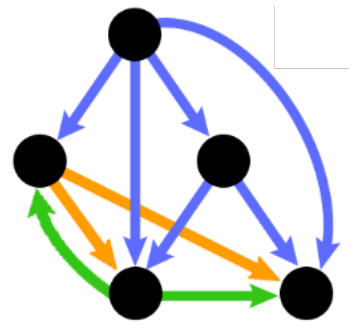
Activity-Shape Model



Video = Graph instance

sampled from the model

Activity-Shape Model

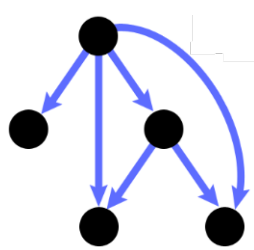
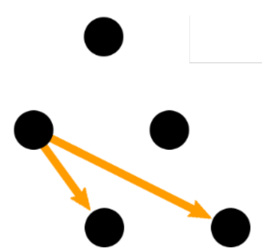
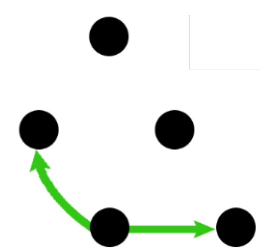


Video = Graph instance

sampled from the model

Model = Probabilistic Graph Mixture

compositional **spatial** **temporal**

$\mathcal{W}_1 \circ$  + $\mathcal{W}_2 \circ$  +...+ $\mathcal{W}_L \circ$ 

$(\mathcal{A}_1, \mathcal{F}_1)$ $(\mathcal{A}_2, \mathcal{F}_2)$ $(\mathcal{A}_L, \mathcal{F}_L)$

Generative Process

video: $G = \{(A_1, F_1), ..., (A_L, F_L)\}$

adjacency matrix

$$A_i = P\mathcal{A}_i P^{\mathrm{T}} + \eta_i$$

node descriptor

$$F_i = P\mathcal{F}_i + \xi_i$$

model parameters

$$i = 1, 2, ..., L$$

Activity-Shape Model

adjacency matrix

node descriptor

$$A_i = P\mathcal{A}_i P^{\mathrm{T}} + \eta_i \qquad F_i = P\mathcal{F}_i + \xi_i$$

permutation matrix

noise

$$i = 1, 2, ..., L$$

Learning

GIVEN K training videos $\{G_k : k = 1, ..., K\}$

$$A_{ki} = P_k \mathcal{A}_i P_k^{\mathrm{T}} + \eta_i \qquad F_{ki} = P_k \mathcal{F}_i + \xi_i$$

permutation matrices

$$i = 1, 2, ..., L$$

Learning

GIVEN K training videos

ESTIMATE

adjacency matrix

$$A_{ki} = P_k \mathcal{A}_i P_k^{\mathrm{T}} + \eta_i$$

node descriptor

$$F_{ki} = P_k \mathcal{F}_i + \xi_i$$

permutation matrices

$$i = 1, 2, \ldots, L$$

Learning

GIVEN K training videos

ESTIMATE

adjacency matrix

node descriptor

$$A_{ki} = P_k \mathcal{A}_i P_k^{\mathrm{T}} + \eta_i \qquad F_{ki} = P_k \mathcal{F}_i + \xi_i$$

permutation matrices

noise

$$i = 1, 2, ..., L$$

Learning and Inference

constraint on permutation matrices

$$\forall\ k,\ \ P_k P_k^{\mathrm{T}} = I,\ \ P_k \in \{0, 1\}^{m \times n}$$

$$\left.\begin{array}{l}\text{Learning}\\ \\ \text{Inference}\end{array}\right\} = \text{Quadratic Integer Program}$$

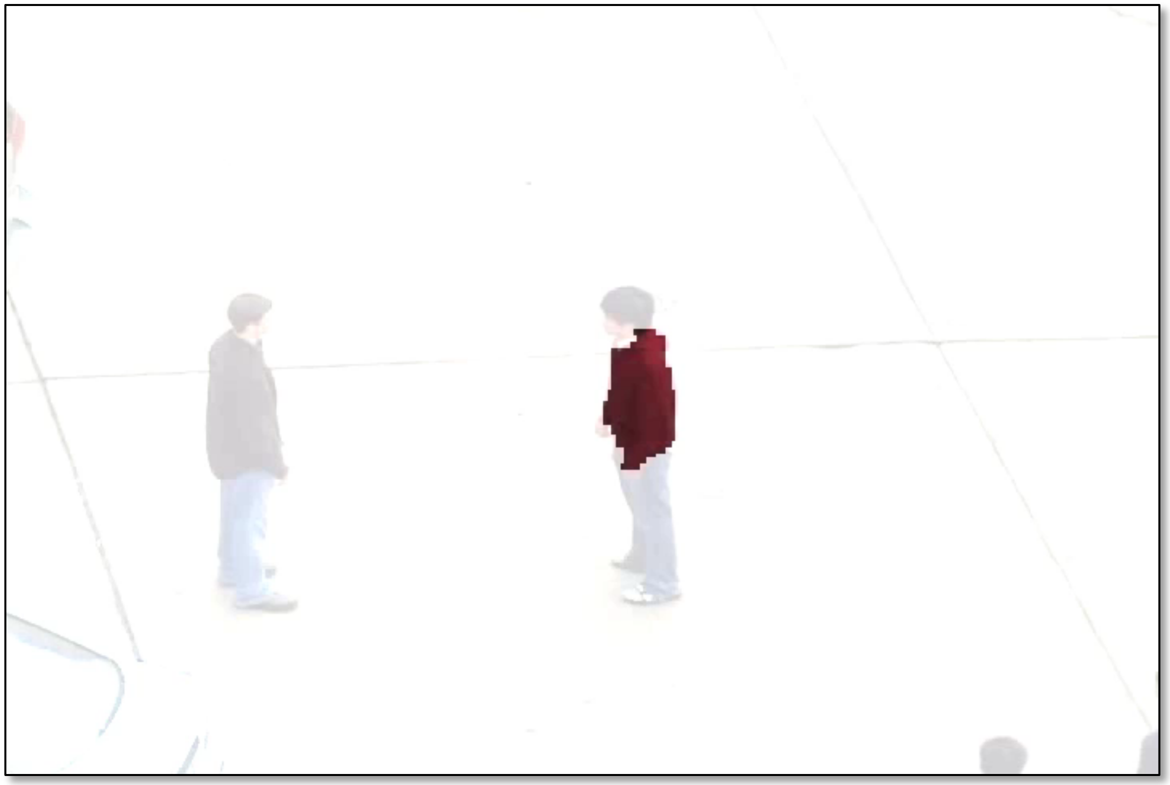
Learning Results

correctly learned activity-characteristic tubes

Learning Results

correctly learned activity-characteristic tubes

Recognition and Segmentation



activity "handshaking"
detected and segmented characteristic tube

Recognition and Segmentation

activity "kicking"
detected and segmented characteristic tube

Classification on UTexas Dataset

	hand shaking	hugging	kicking	pointing	punching	pushing
Our	**81.7%**	**89.6%**	**68.6%**	**66.4%**	**84.5%**	**82.7%**
[17]	75%	87.5%	62.5%	50%	75%	75%

human interaction activities

Conclusion

- Shape-based video representation enables:
  - Simpler activity models, learning, inference...
  - Richer interpretation: recognition + segmentation
- Difficulties
  - Correspondence between model and data features