Learning Spatiotemporal Graphs of Human Activities

William Brendel

Sinisa Todorovic

Our Goal

Long Jump

Triple Jump

- Recognize all occurrences of activities
- Identify the start and end frames
- Parse the video and find all subactivities
- Localize actors and objects involved

Weakly Supervised Setting

Weight Lifting

Large-Box Lifting

In training:

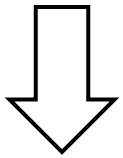
> **ONLY** class labels

Domain knowledge of temporal structure:

> **NOT AVAILABLE**

Learning What and How

Weak supervision in training

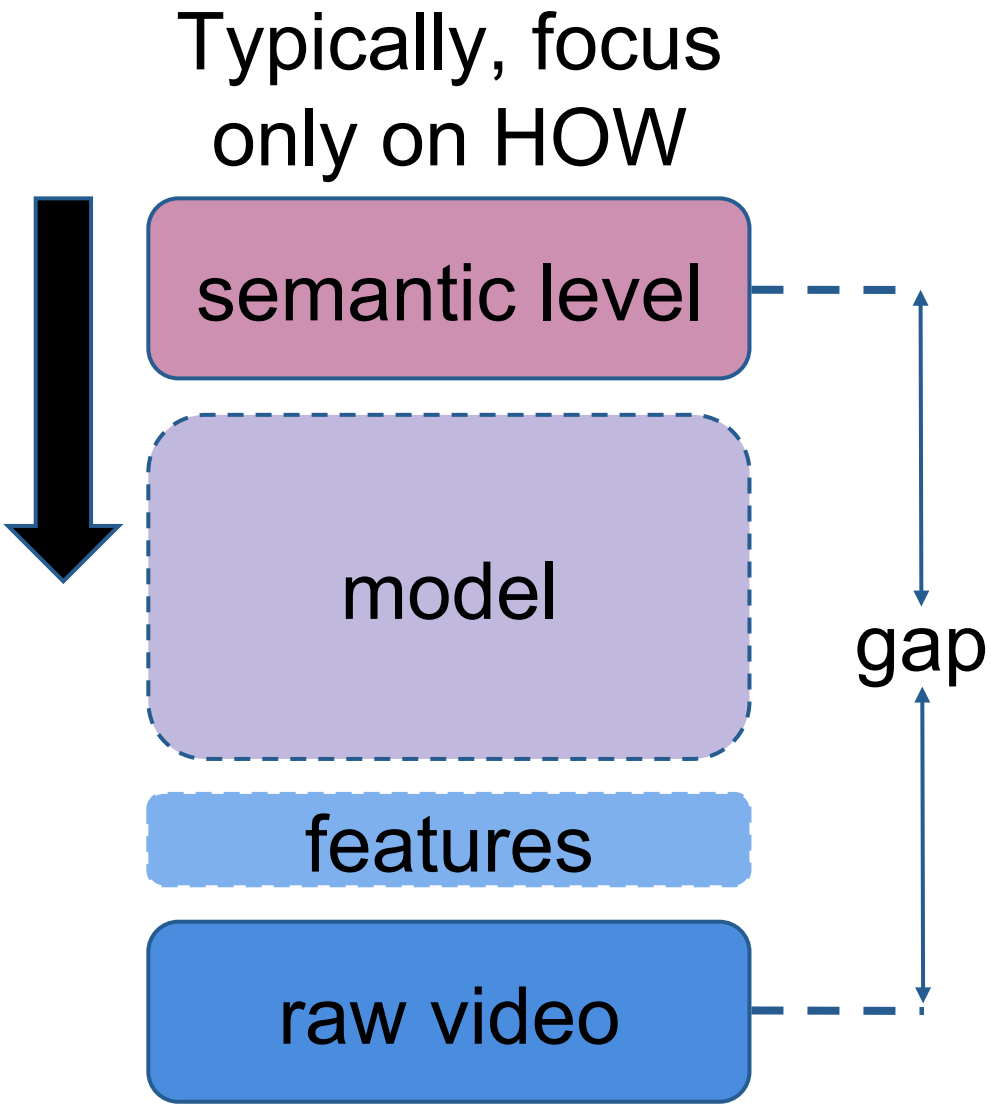


Need to learn from training videos:

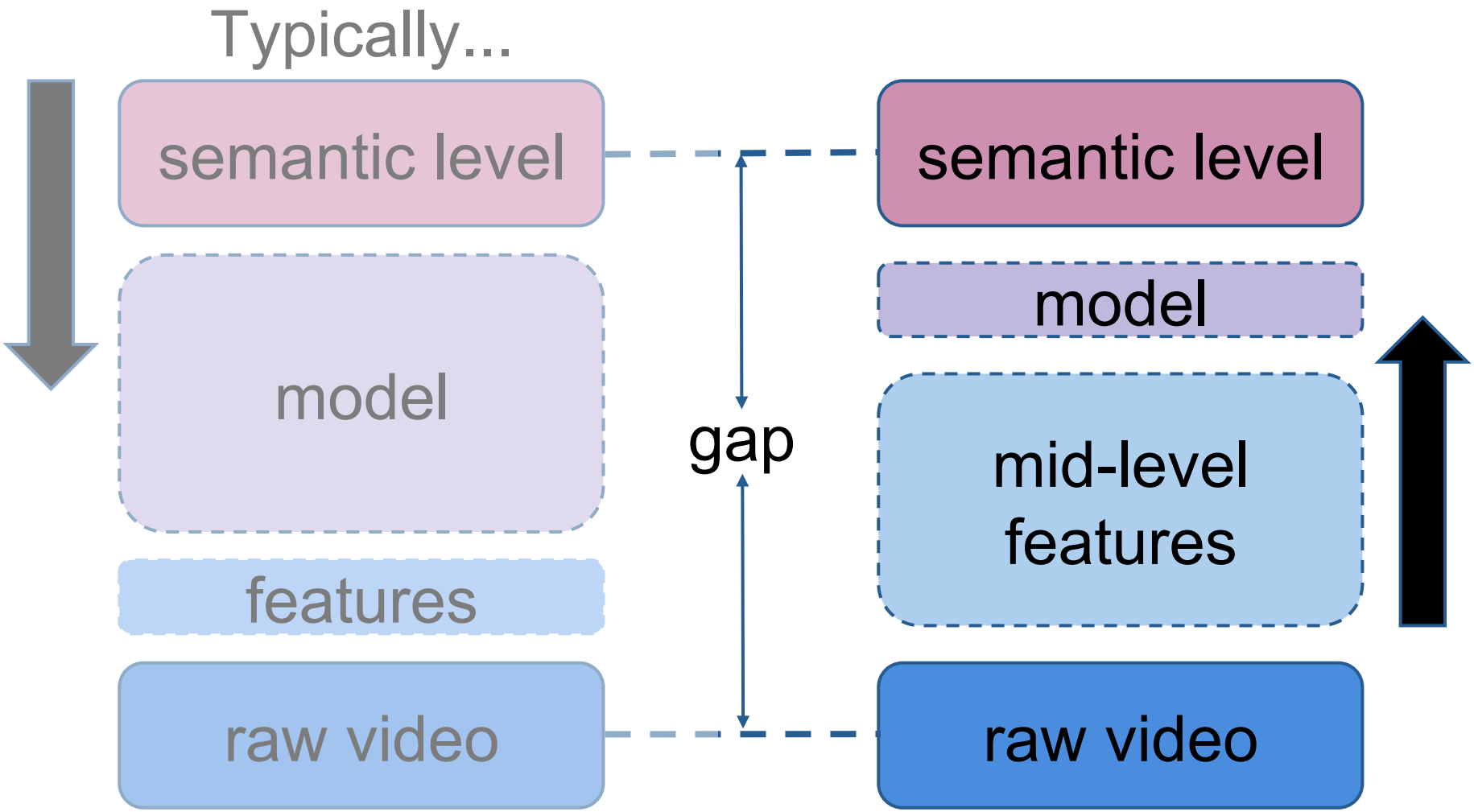
What activity parts are relevant

How relevant they are for recognition

Prior Work vs. Our Approach

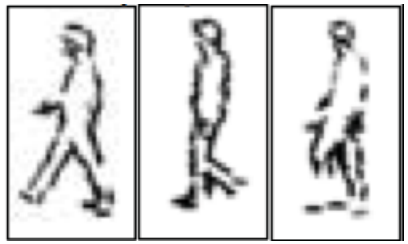


Prior Work vs. Our Approach



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, …



Our Features: 2D+t Tubes

- Allow simpler:
  - Modeling
  - Learning (few examples)
  - Inference

Sukthankar & Hebert 07,
Gorelick & Irani 08,
Pritch & Peleg 08, ...

Our Features: 2D+t Tubes

- Allow simpler:
  - Modeling
  - Learning (few examples)
  - Inference

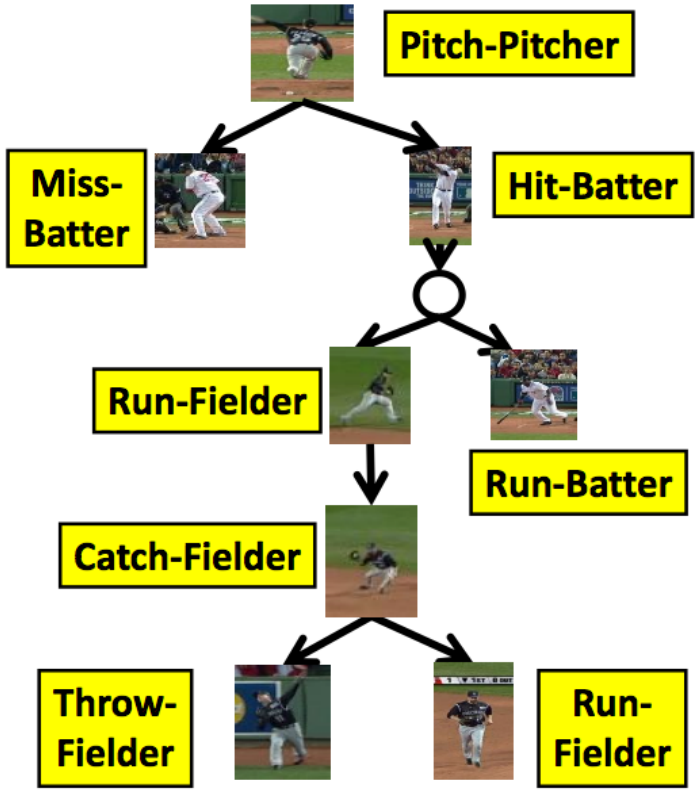
Sukthankar & Hebert 07,
Gorelick & Irani 08,
Pritch & Peleg 08, ...

- We use 2D+t tubes for building a statistical generative model of activities

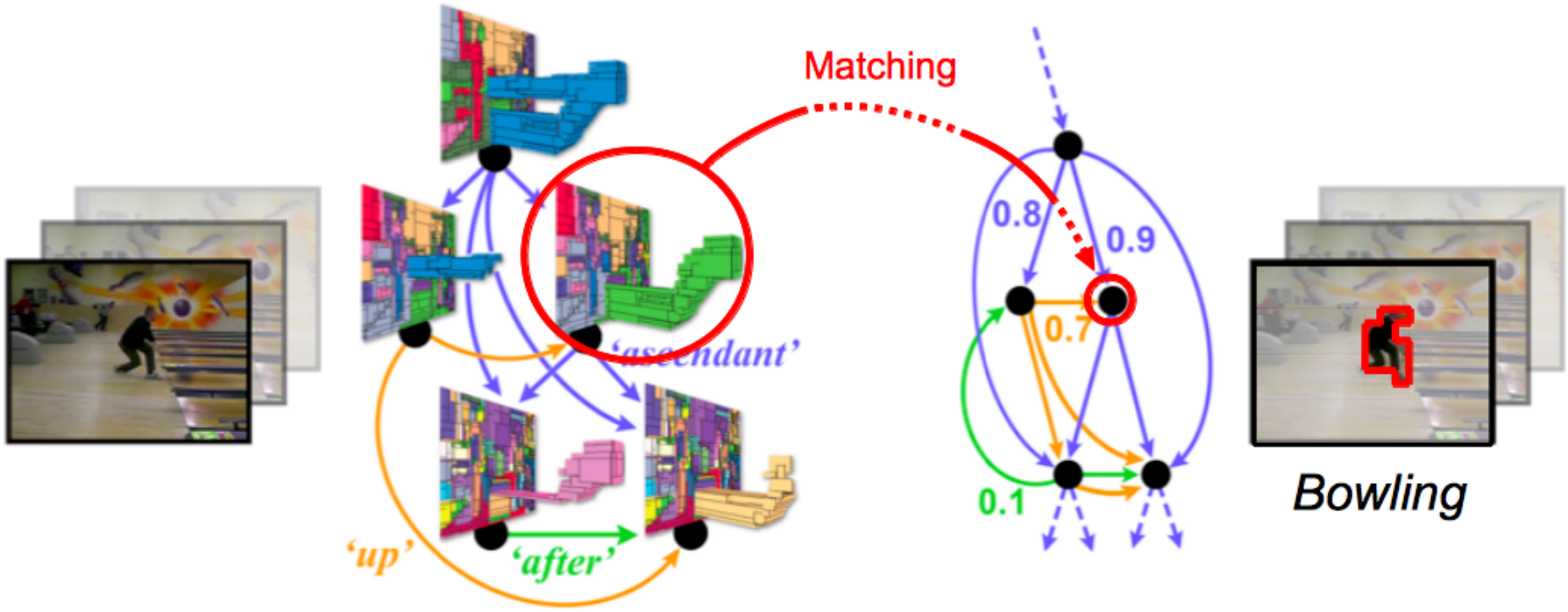
Prior Work – Activity Representation

- Graphical models, Grammars
  - Ivanov & Bobick 00
  - Xiang & Gong 06
  - Ryoo & Aggawal 09
  - Gupta & Davis 09
  - Liu & Zhu 09
  - Niebles & Fei-Fei 10
  - Lan et al. 11

- Probabilistic first-order logic
  - Tran & Davis 08
  - Albanese et al. 10
  - Morariu & Davis 11
  - Brendel et al. 11...



Approach



Input Video

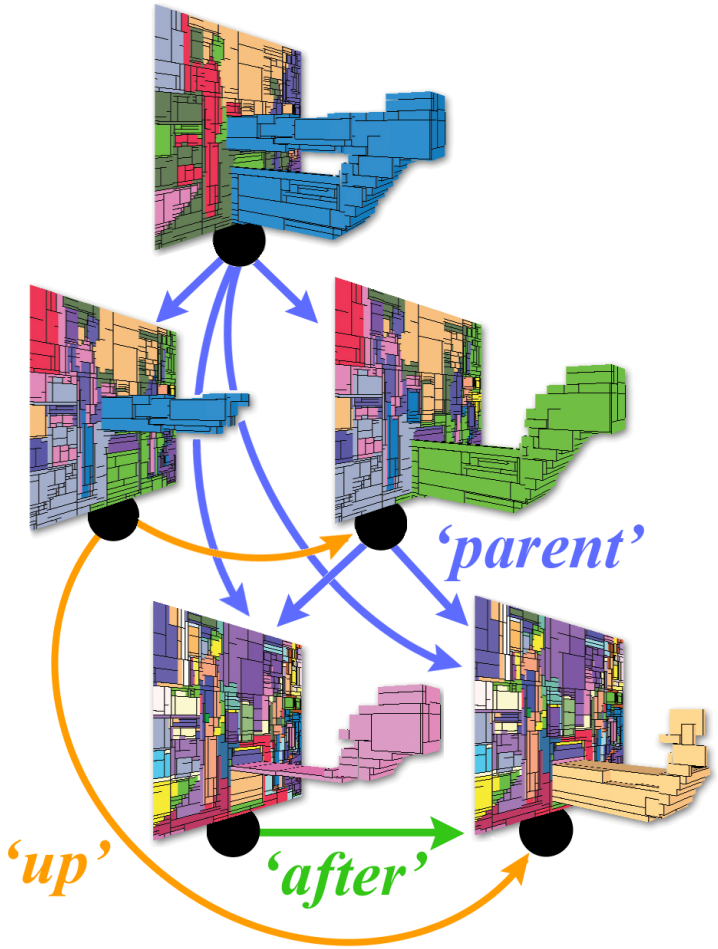
Spatiotemporal Graph

Activity Model

Recognition Localization

Blocky Video Segmentation

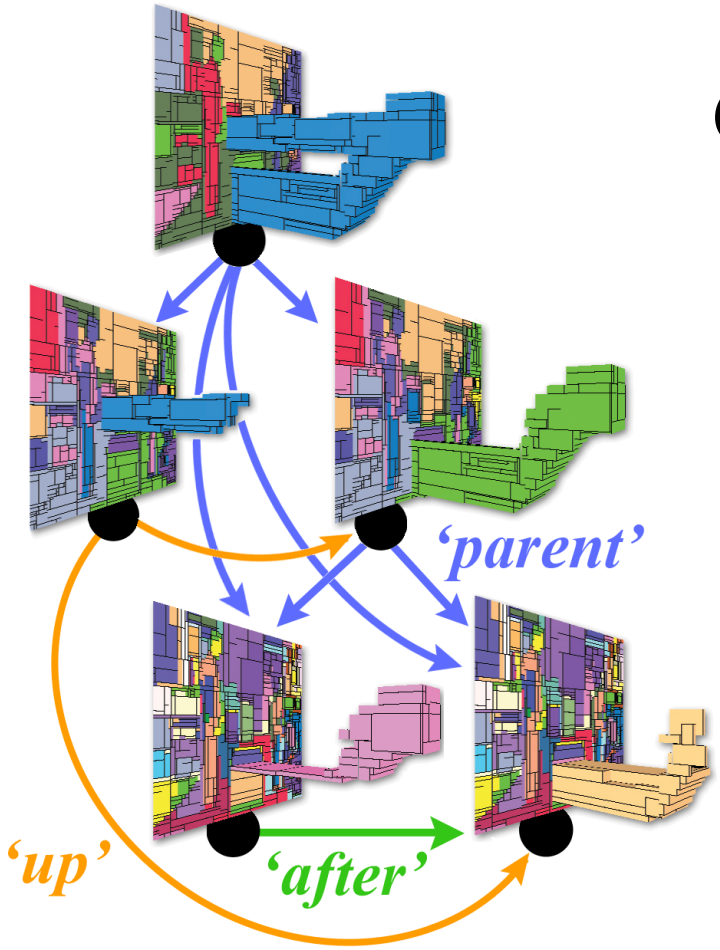
Activity as a Spatiotemporal Graph



Descriptors of nodes and edges:

- Node descriptors: F
  - Motion
  - Object shape
- Adjacency Matrices: $\{A_i\}$
  - Allen temporal relations
  - Spatial relations
  - Compositional relations

Activity as Segmentation Graph



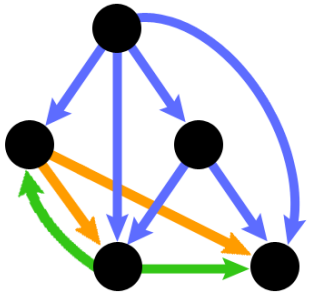
$$G = (V, E, \text{"descriptors"})$$

$$= (F, \{A_1, ..., A_n\})$$

node descriptors

adjacency matrices of distinct relations between the tubes

Activity Graph Model



Probabilistic Graph Mixture

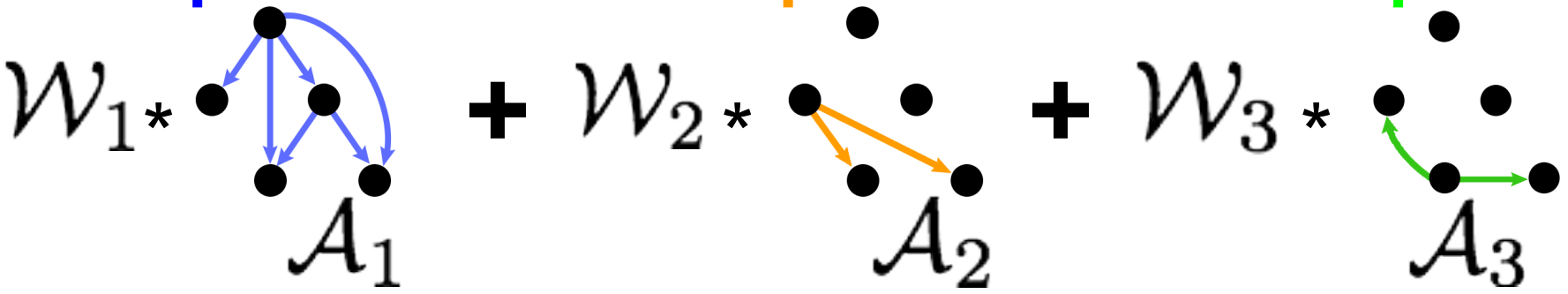
$$\mathcal{G} = (\mathcal{F}, \{\mathcal{A}_i, \mathcal{W}_i\})$$

model node descriptors

mixture weights

model adjacency matrices

compositional **spatial** **temporal**



Activity Model

An activity instance: $\boldsymbol{G = (F, \{A_1,..., A_n\})}$

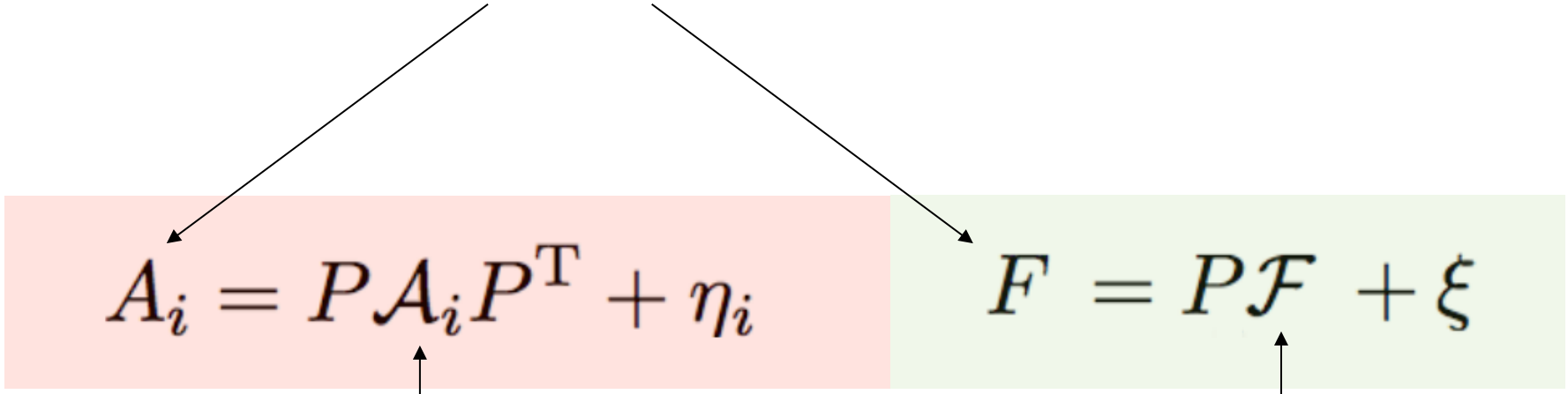
$$A_i = P\mathcal{A}_i P^{\mathrm{T}} + \eta_i$$

Model adjacency matrices

Edge type: $\boldsymbol{i = 1, 2,..., n}$

Activity Model

An activity instance: $\boldsymbol{G = (F, \{A_1,..., A_n\})}$



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

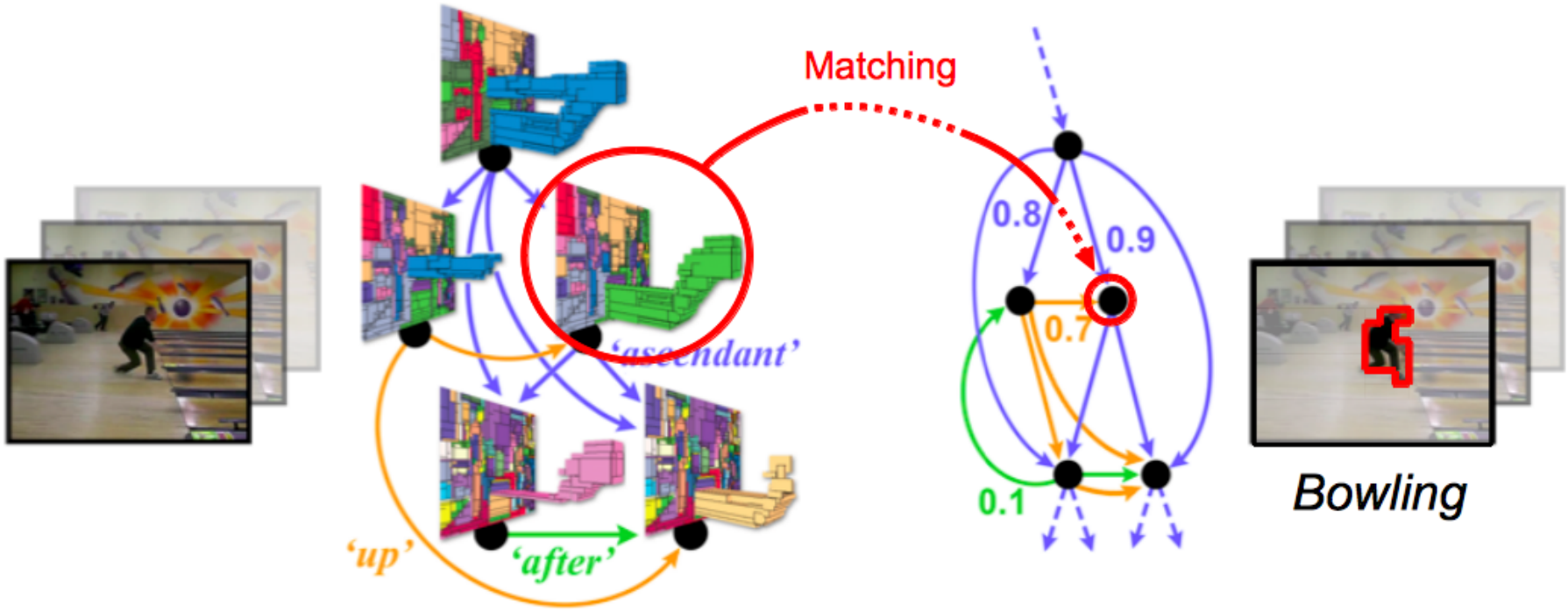
$$F = P\mathcal{F} + \xi$$

Model adjacency matrices

Edge type: $\boldsymbol{i = 1, 2,..., n}$

Model matrix of node descriptors

Inference



Input Video

Spatiotemporal Graph

Activity Model

Recognition Localization

Inference = Robust Least Squares

Goal:

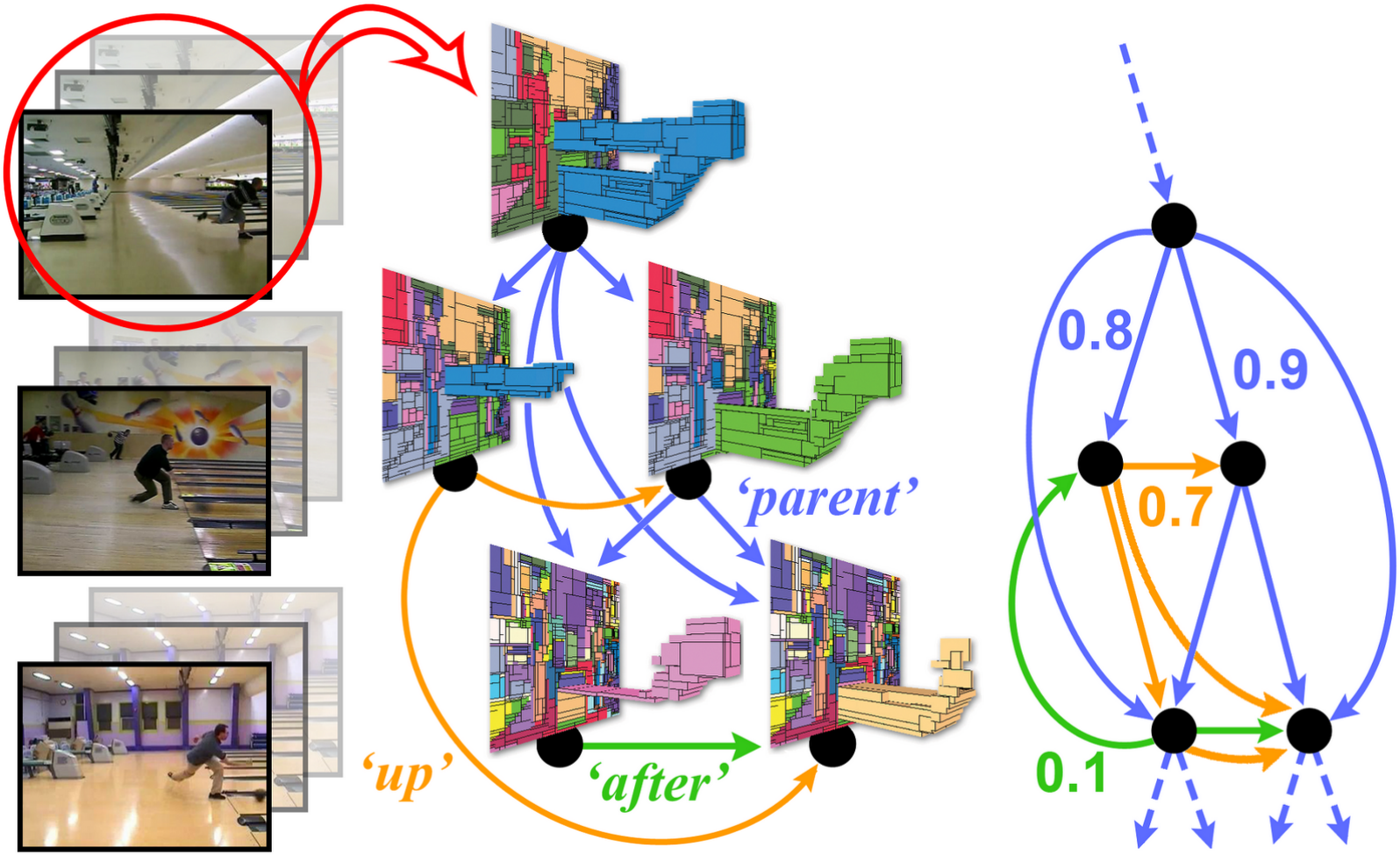
- For every activity model
- Estimate the permutation matrix

$$\min_{P} \sum_{i} \|A_i - P\mathcal{A}_i P^{\mathrm{T}}\|^2 + \|F - P\mathcal{F}\|^2$$

subject to

$$PP^{\mathrm{T}} = I, \quad P \in \{0,1\}^{m \times m}$$

Learning the Activity Graph Model



Training videos → Training graphs → Graph model

Learning

Given K training graphs $\{G_k : k = 1, ..., K\}$

Adjacency matrix

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

Node descriptor

$$F_k = P_k \mathcal{F} + \xi$$

Edge type: $i = 1, 2, ..., n$

Learning

Given K training graphs,

ESTIMATE

Adjacency matrix

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

Node descriptor

$$F_k = P_k \mathcal{F} + \xi$$

Model parameters

Learning

Given K training graphs,

ESTIMATE

Adjacency matrix

$$A_{ki} = P_k A_i P_k^{\mathrm{T}} + \eta_i$$

Node descriptor

$$F_k = P_k \mathcal{F} + \xi$$

Permutation matrix

Learning = Robust Least Squares

Given K Training graphs: $\{A_{ki}\},\ F_k,\ k = 1, \ldots, K$

Estimate: $\{\mathcal{A}_i, \mathcal{W}_i\},\ \mathcal{F}$ and $\{P_k\}$

$$\min \sum_{k,i} \|P_k^{\mathrm{T}} A_{ki} P_k - \mathcal{A}_i\|^2 + \|P_k^{\mathrm{T}} F_k - \mathcal{F}\|^2$$

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

Learning = Structural EM

E-step → expected model structure

M-step → matching of the training graphs and model

$$\forall i,\ \mathcal{A}_i = \frac{1}{K}\sum_{k=1}^{K} \mathbf{P}_k^{\mathrm{T}} \mathbf{A}_{ki} \mathbf{P}_k,$$

$$\mathcal{F} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{P}_k^{\mathrm{T}} \mathbf{F}_k.$$

$$\mathcal{W} = \mathrm{Cov}^{-1}(\mathcal{F})$$

$$\min \sum_{k,i} \|P_k^{\mathrm{T}} A_{ki} P_k - \mathcal{A}_i\|^2 + \|P_k^{\mathrm{T}} F_k - \mathcal{F}\|^2$$

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

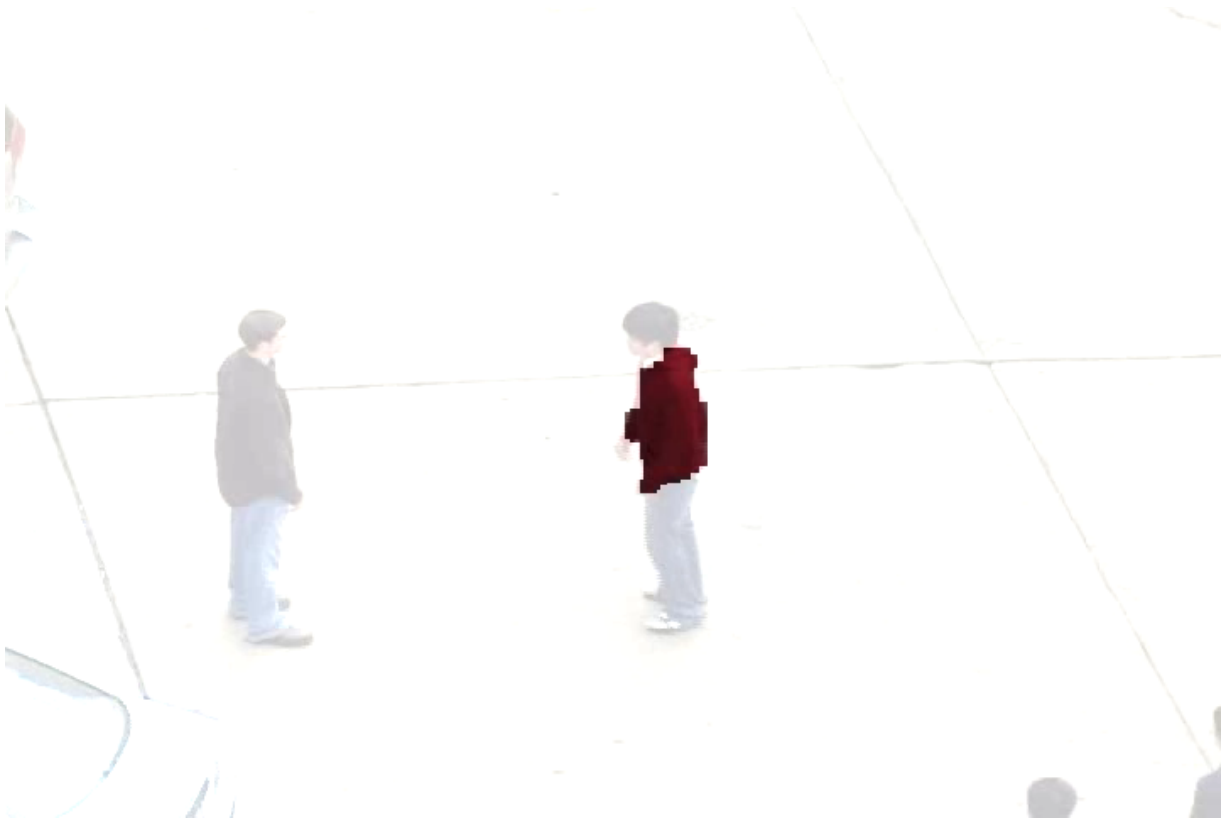
Estimatation of model parametrs

Estimation of permutation matrices

Learning Results

Correctly learned activity-characteristic tubes

Recognition and Segmentation



Activity “handshaking”

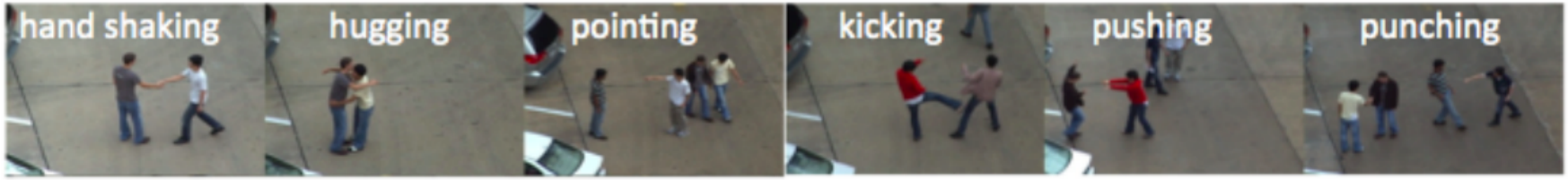
Detected and segmented characteristic tube

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%**
[18]	75%	87.5%	62.5%	50%	75%	75%

Human interaction activities

[18] Ryoo et al. ’10

Conclusion

- Fast spatiotemporal segmentation
- New activity representation = graph model
- Unified learning and inference = Least squares
- Learning under weak supervision:

    - WHAT activity parts are relevant and

    - HOW relevant they are for recognition