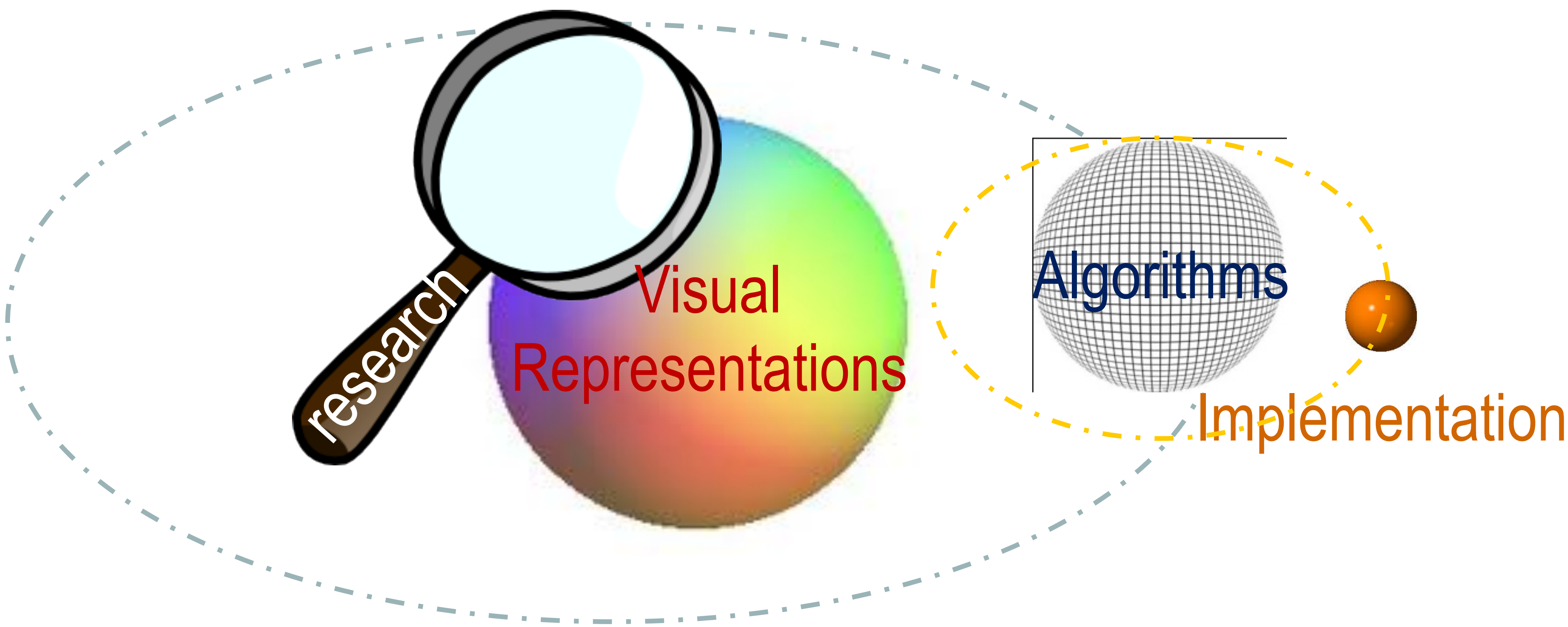
Object and Activity Recognition Grounded on Midlevel Image Representations

Sinisa Todorovic
joint work with: N. Payet, M. Amer

Oregon State University

September 20, 2011

Marr’s Vision



Slide by SC Zhu

Open Basic Problems

- Semantic gap between visual features and categories
  - General vs. task-specific representations

Open Basic Problems

- Semantic gap between visual features and categories
  - General vs. task-specific representations
  - An observation: recent research mostly task-specific

Open Basic Problems

- Semantic gap between visual features and categories
  - General vs. task-specific representations
  - An observation: recent research mostly task-specific

- What are successful general representations?
  - **Grammars and logic are back!**
  - **SIG-11 workshop at ICCV11**
  - **U Grenander, D Mumford, SC Zhu, A Yuille, L Davis, R Chellapa, N Ahuja, A Leonardis, P Felzenszwalb, S Todorovic ...**

Goal

A unified computational framework capable of:

Discovery

Detection

Segmentation

Summarization

…

of visual categories

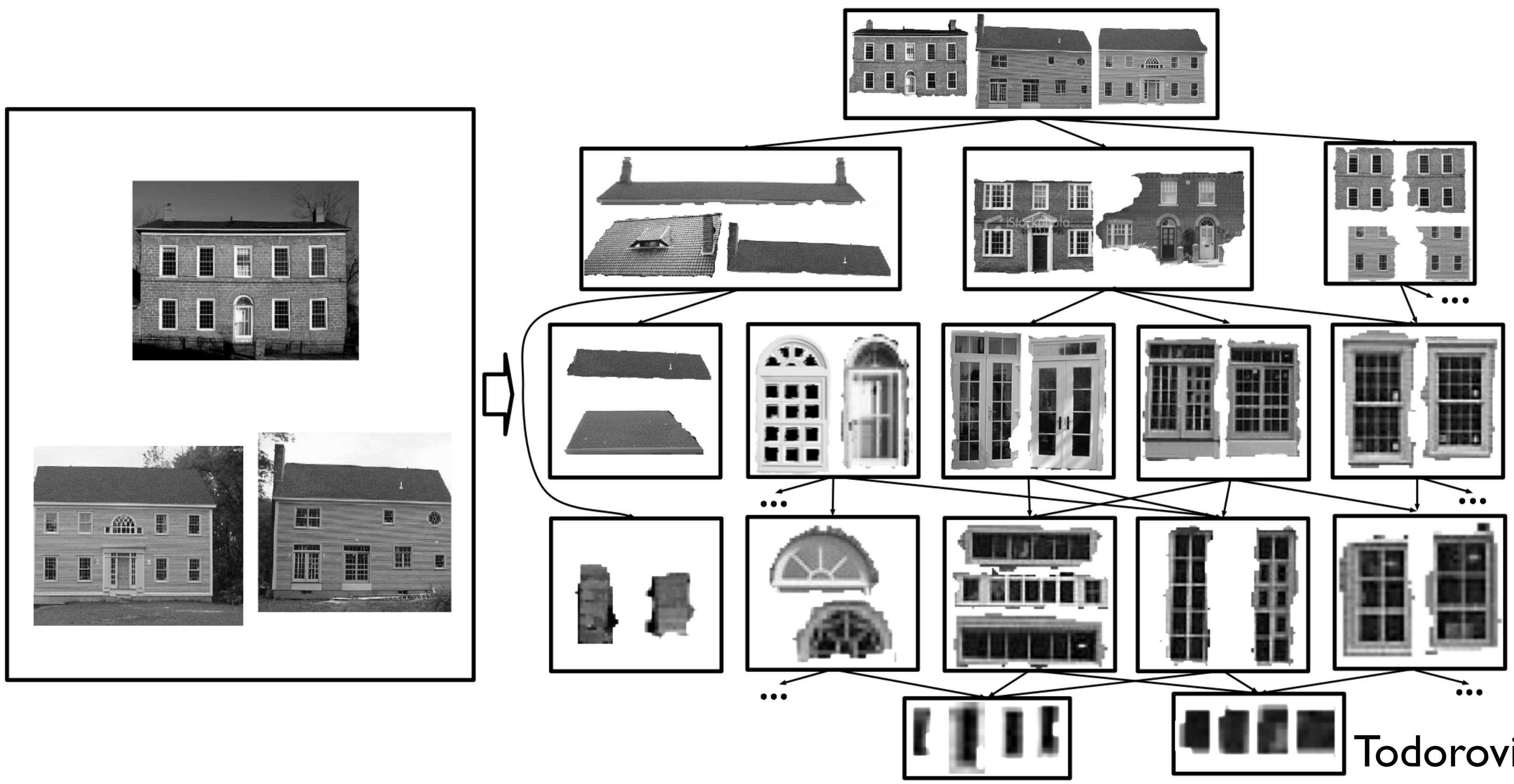
in images and video

What is an Object?

compositionality

Spatial arrangement of its parts

input images partonomy = SCFG



Todorovic & Ahuja PAMI08

What is an Object?

compositionality

Spatial arrangement of its parts

Parts = Objects in their own right

Part discovery = Suspicious coincidences

What is an Object?

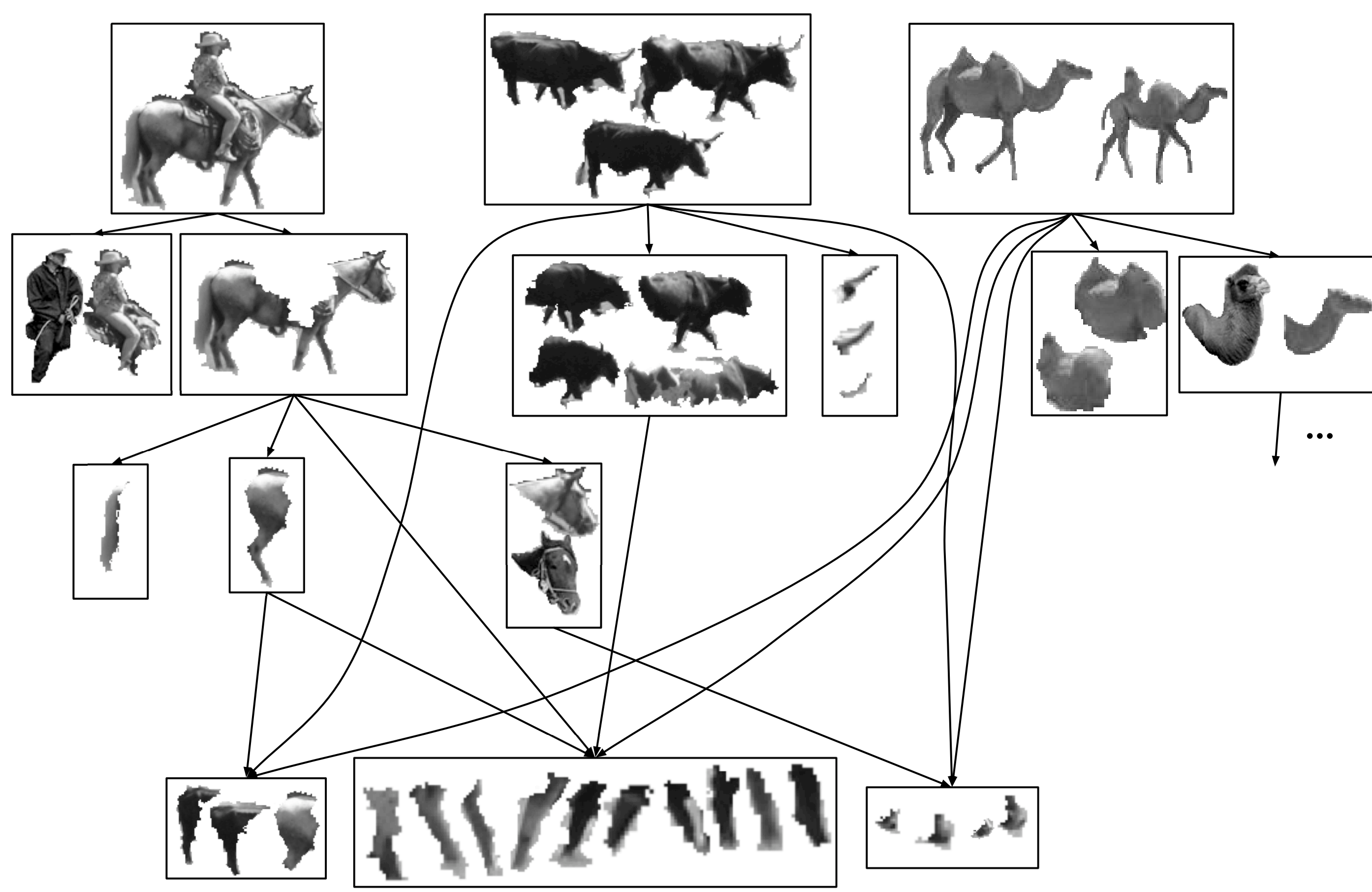
compositionality

Spatial arrangement of its parts

and

context

Spatial and semantic constraints with other objects



AND-OR graph:
a hierarchy of
random fields

Todorovic & Ahuja ICCV07

What is an Activity?

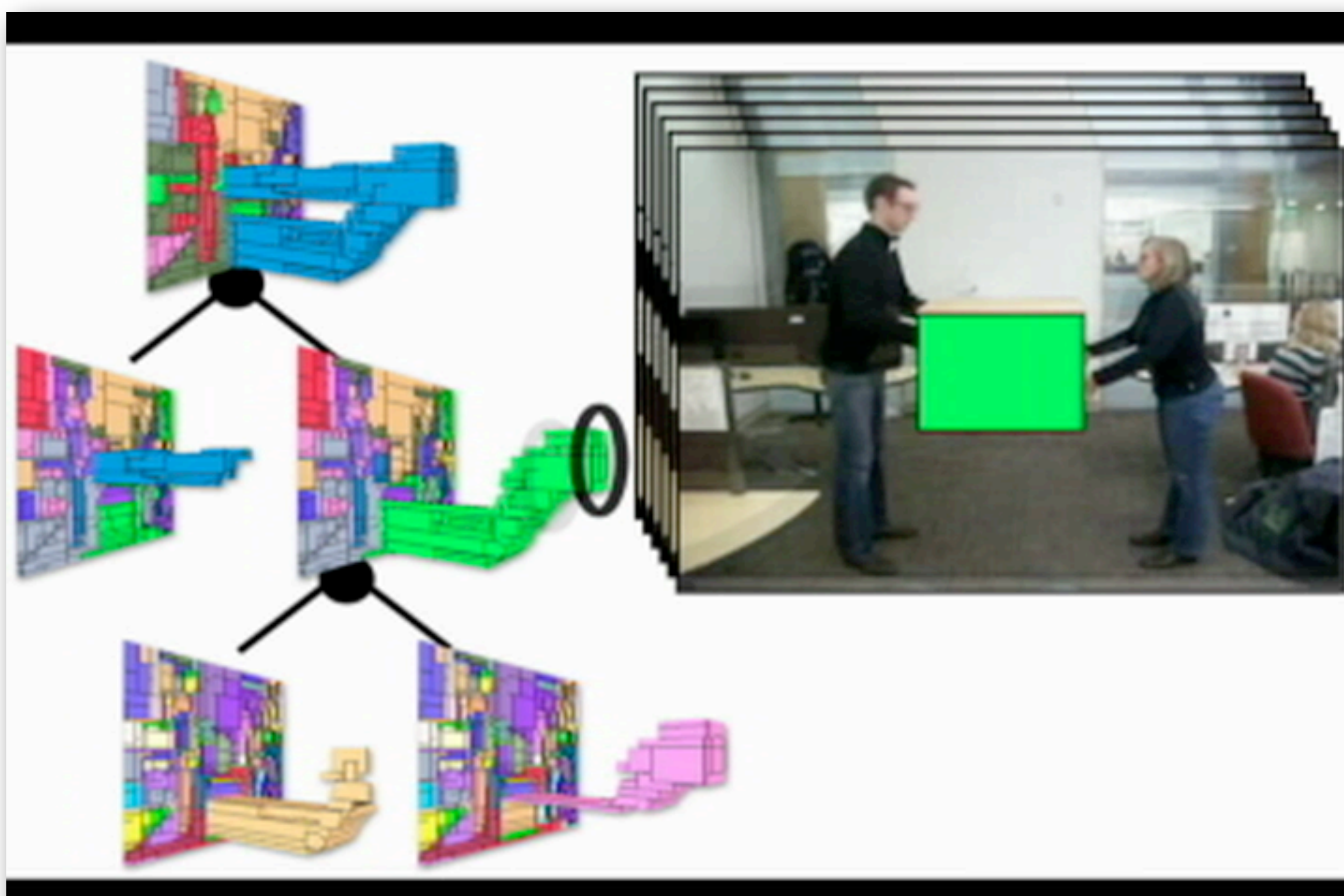
compositionality

Spatiotemporal arrangement of its parts

and

context

Spatiotemporal constraints with other activities



What is an Activity?

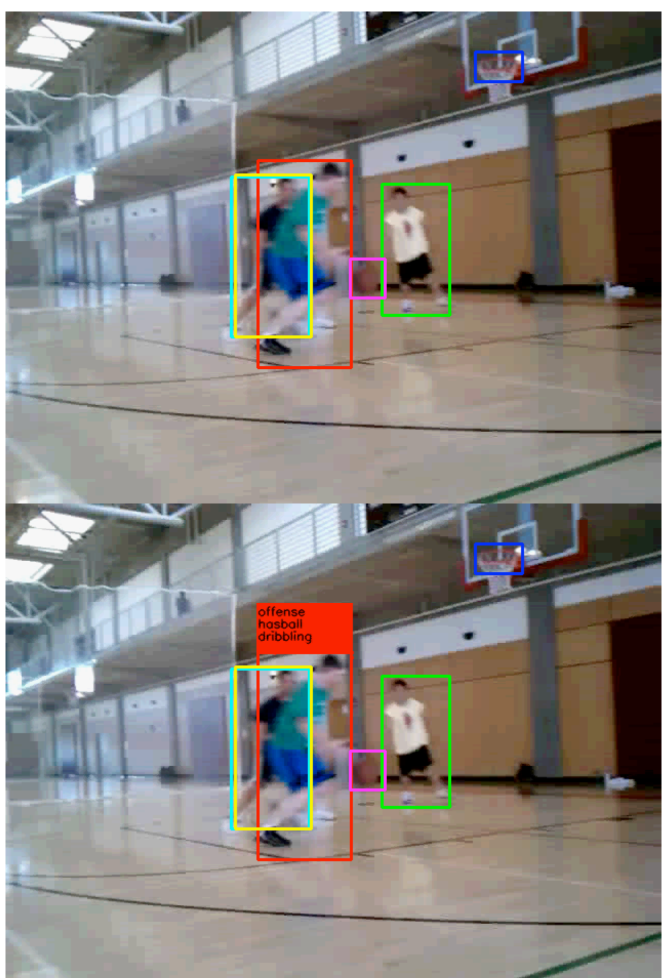
compositionality

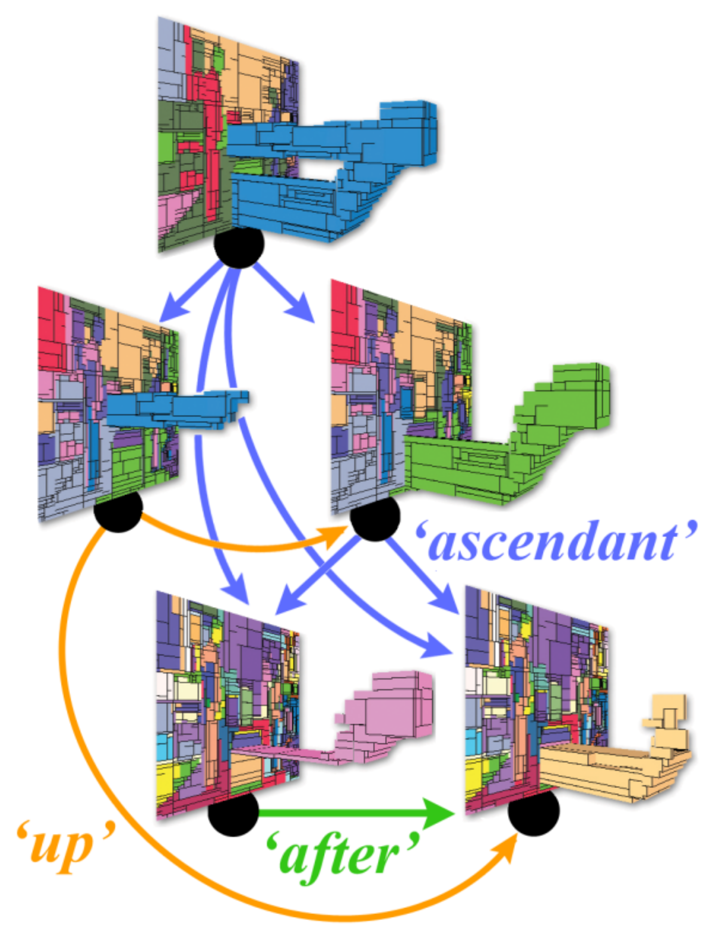
Spatiotemporal arrangement of its parts

and

context

Spatiotemporal constraints with other activities



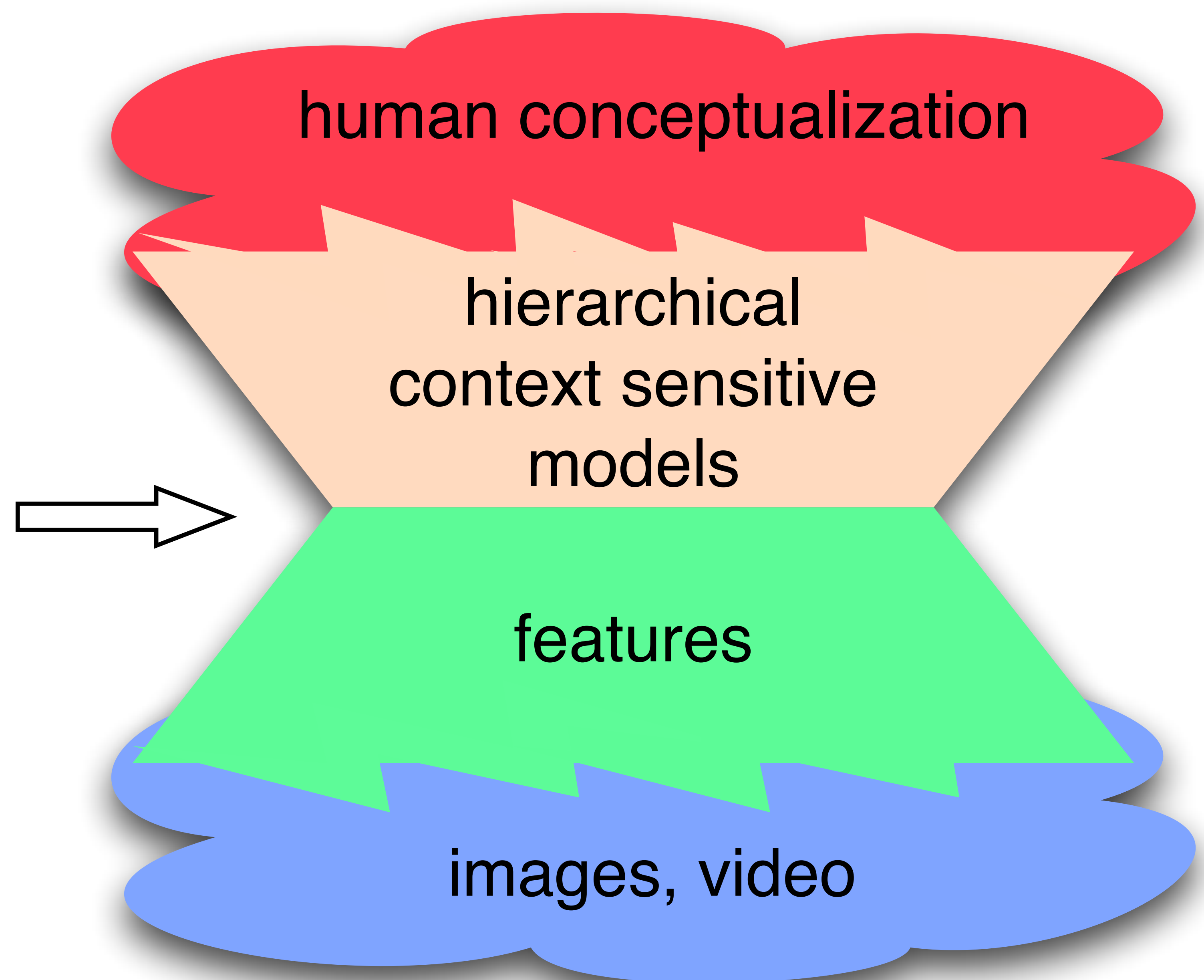


AND-OR graph:
a hierarchy of
random fields

Brendel & Todorovic CVPR11, ICCV11

Issues

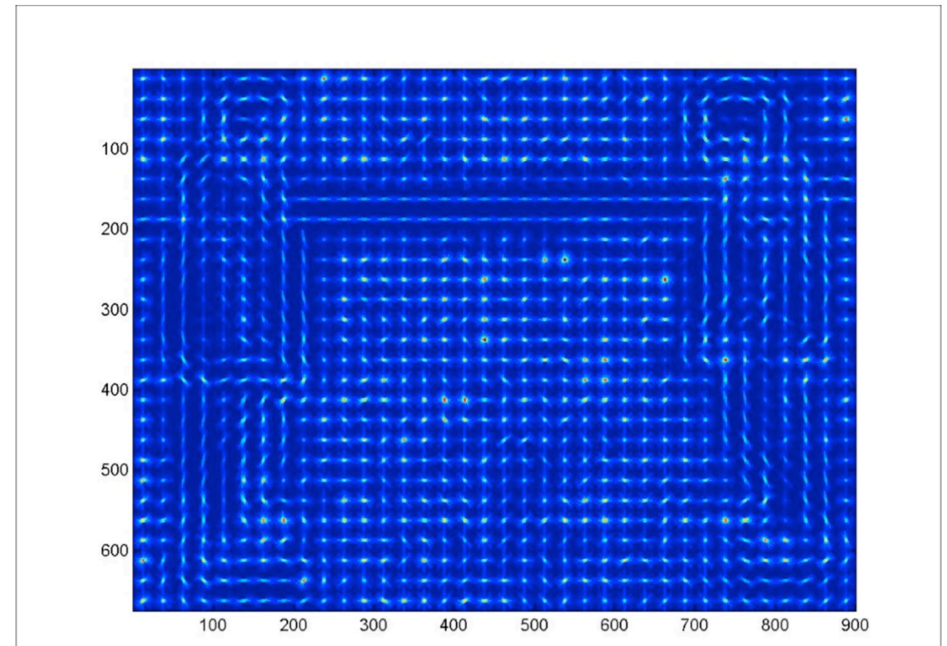
Bottleneck



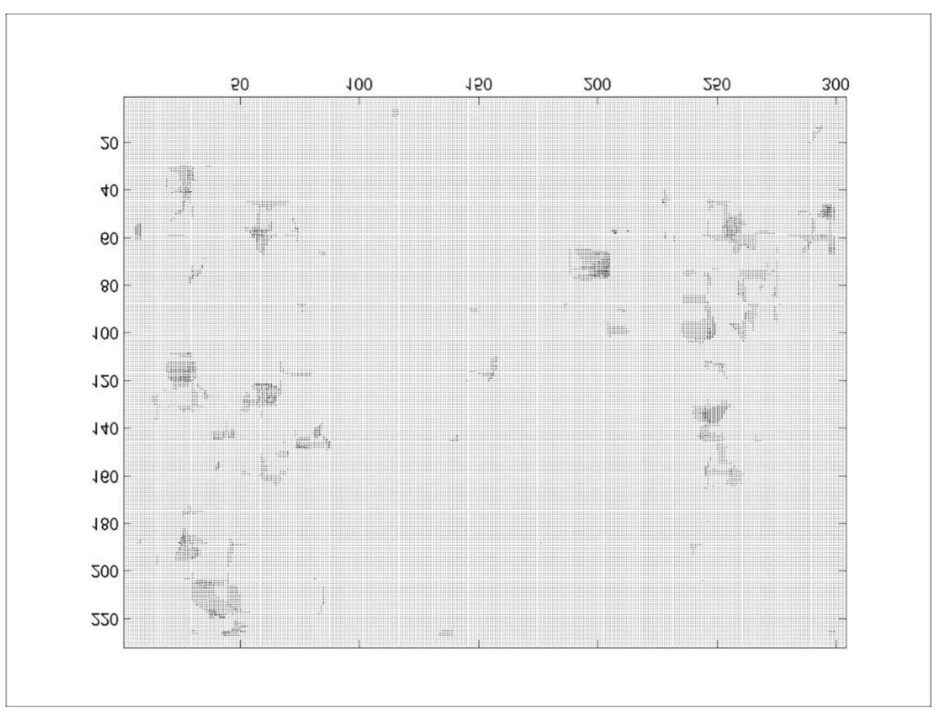
Grounding grammars on
pre-selected low-level features

Example Low-Level Features

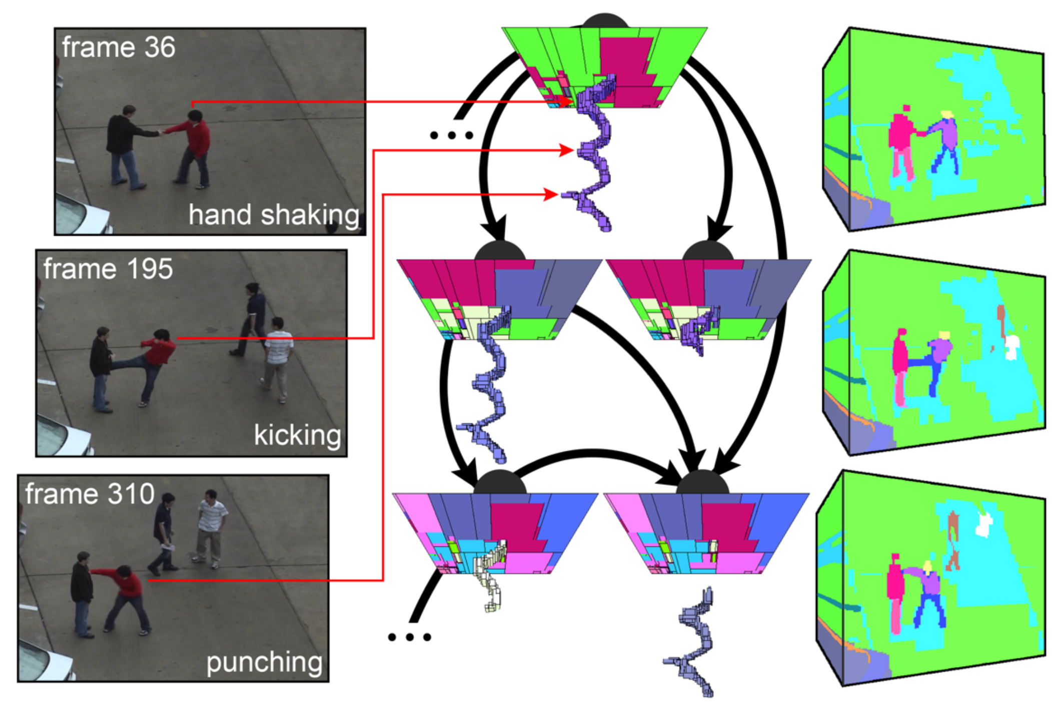
STIPs



HOGs



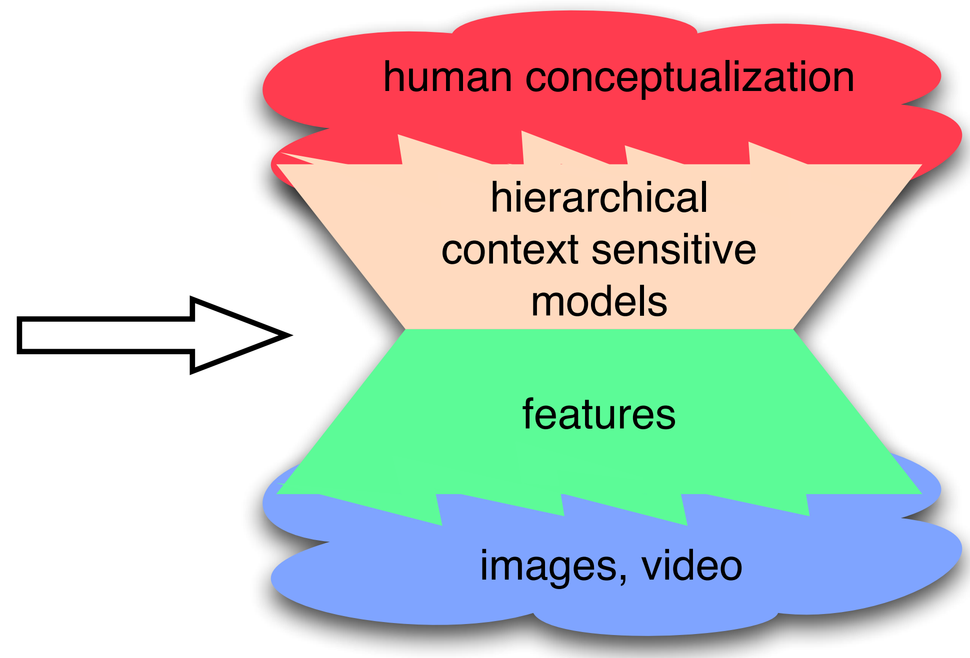
optical flow



video segmentation

Issues

Bottleneck



Grounding requires (probabilistic) repeatability of:

features and their spatiotemporal placement

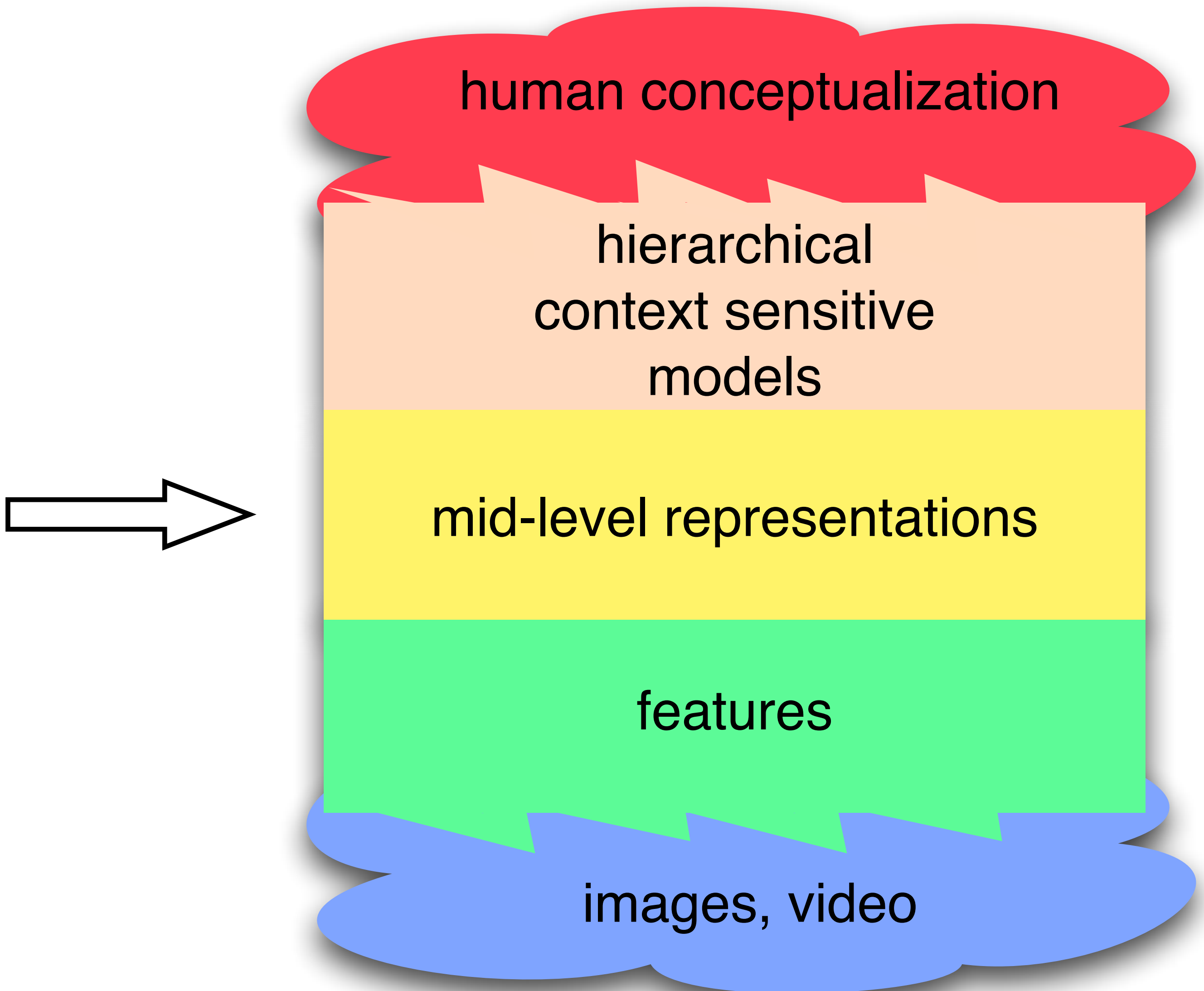
Satisfied only in particular settings/tasks
(e.g., single-view object recognition)

Example Issues

low-level ⟶ mid-level

point-based features	regions, contours
stable, repeatable	unstable, poor repeatability
poor informativeness	informative

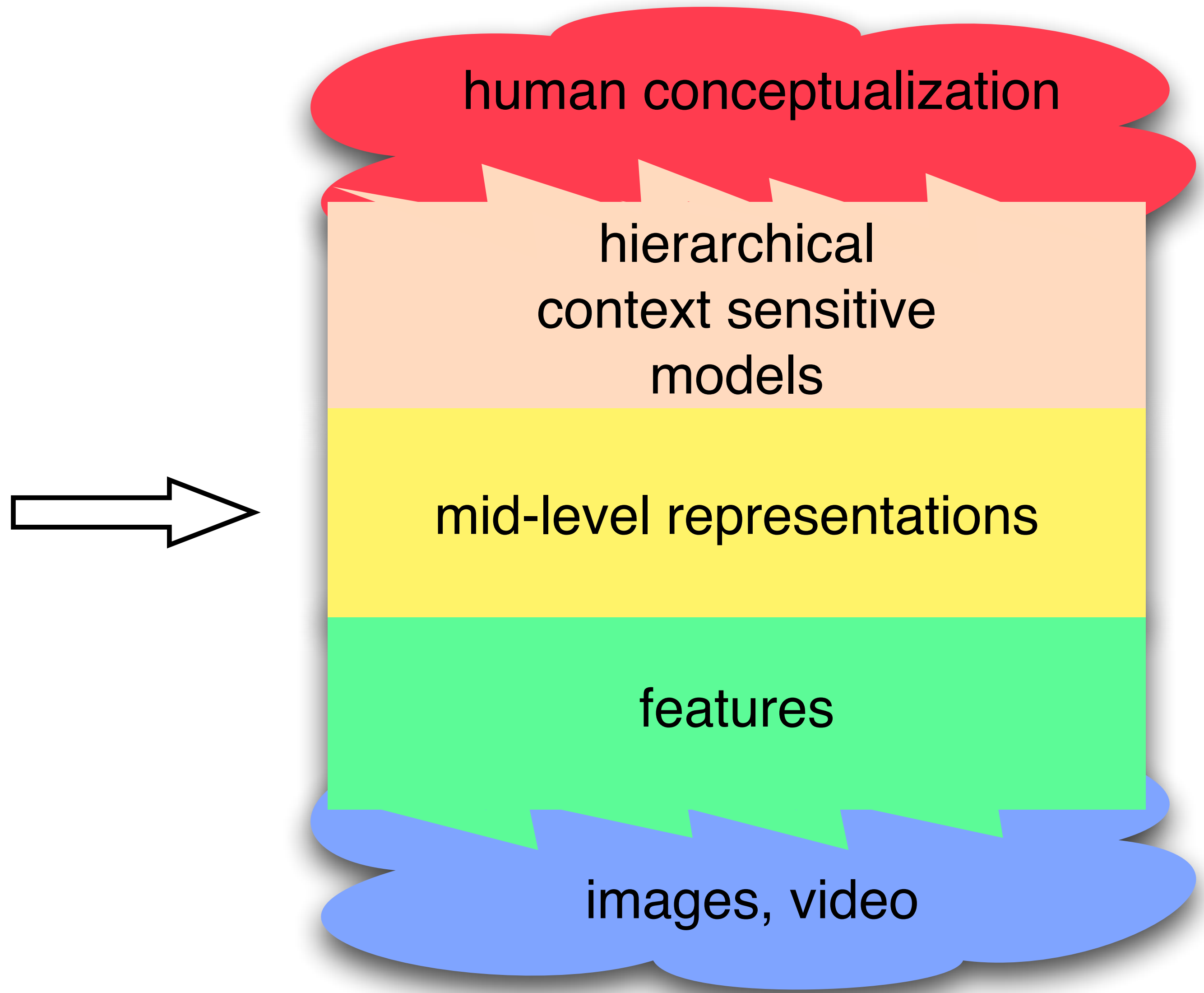
Hypothesis



Grounding grammars on

summaries of low-level features

Hypothesis



Grounding grammars on

summaries of low-level features

where the summarization is guided top-down

Objective

Formalize a mid-level feature that will be:

informative like regions (e.g., encode structure)

and

repeatable like points (e.g., view-invariant)

Bags of Right Features/Detections

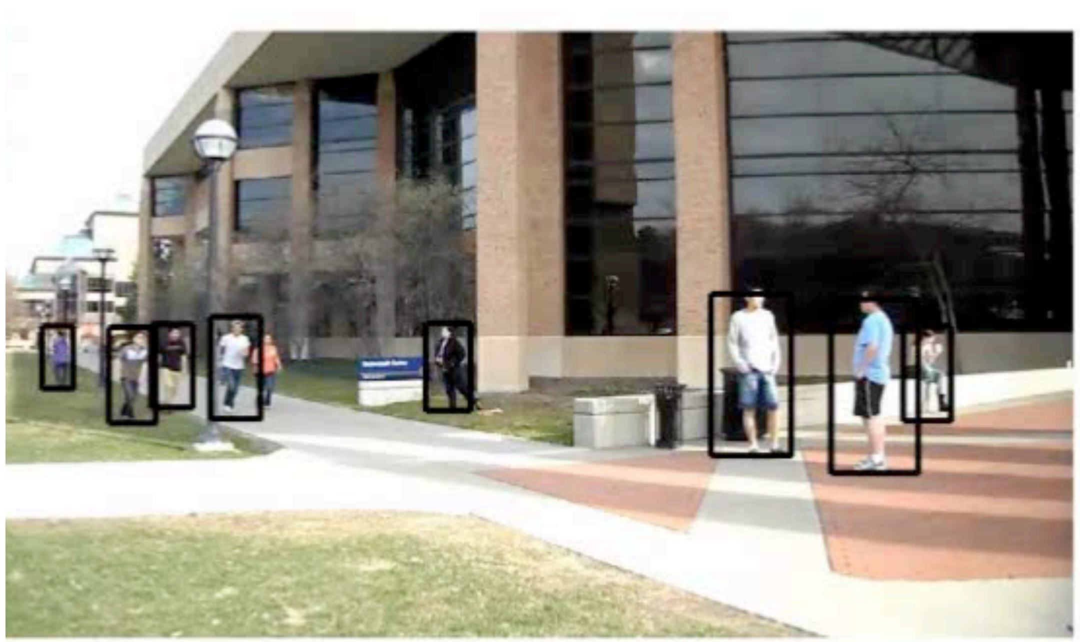
If the category occurs,

it has to be in the spotlight of many BORDs

so they can jointly support the occurrence hypothesis

Perina & Jojic CVPR11

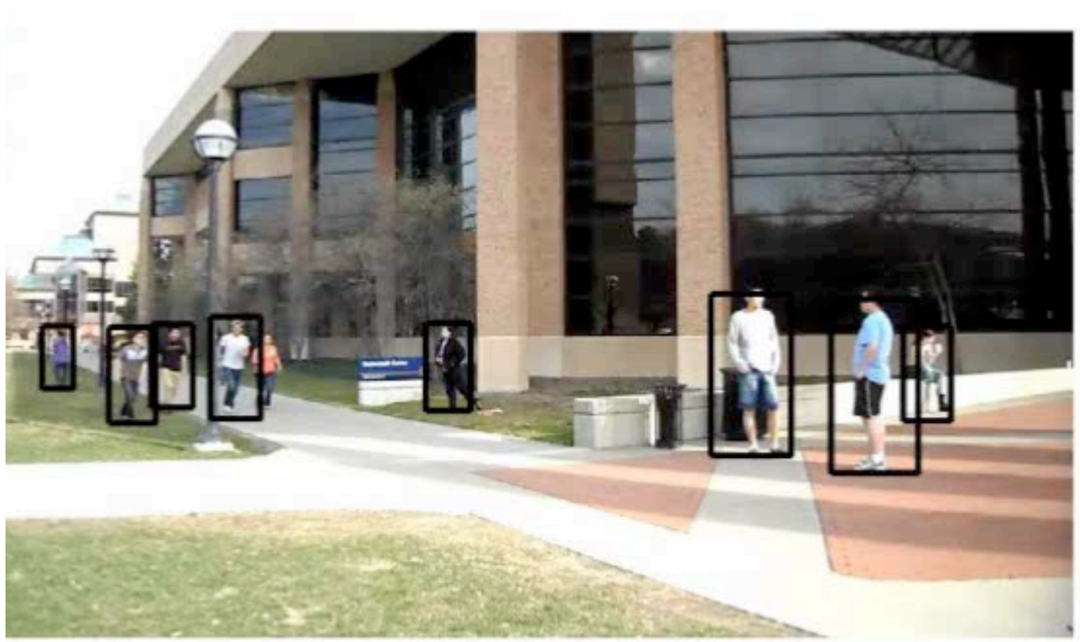
Example Problem: Activity Recognition



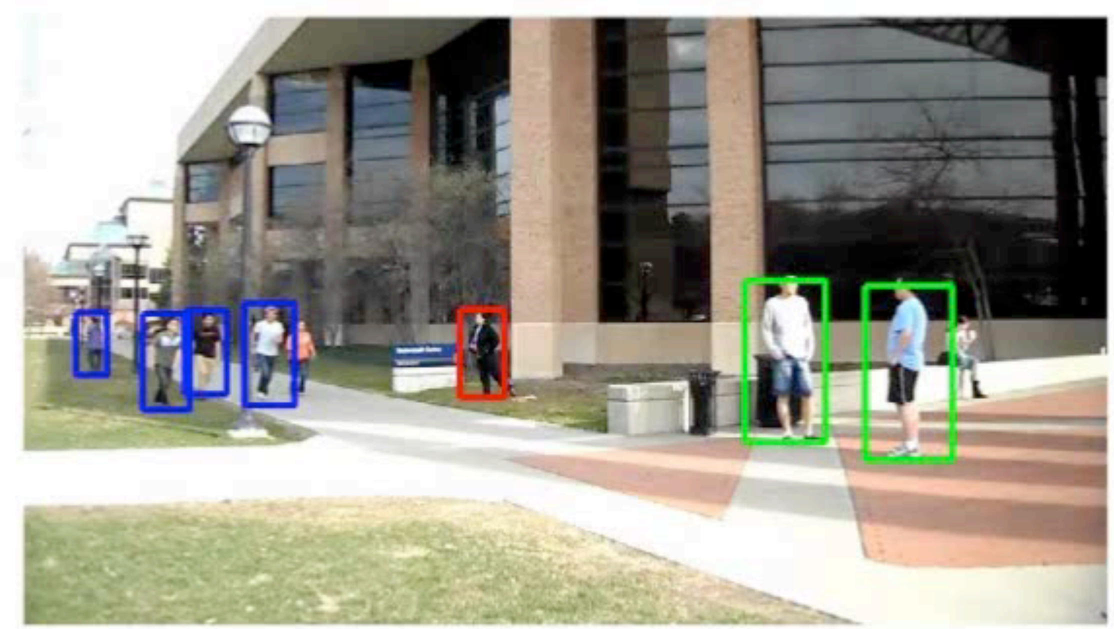
Given a video with noisy people detections

Amer & Todorovic ICCV11

Example Problem



Given a video with noisy people detections



Detect and localize: all activity instances & actors

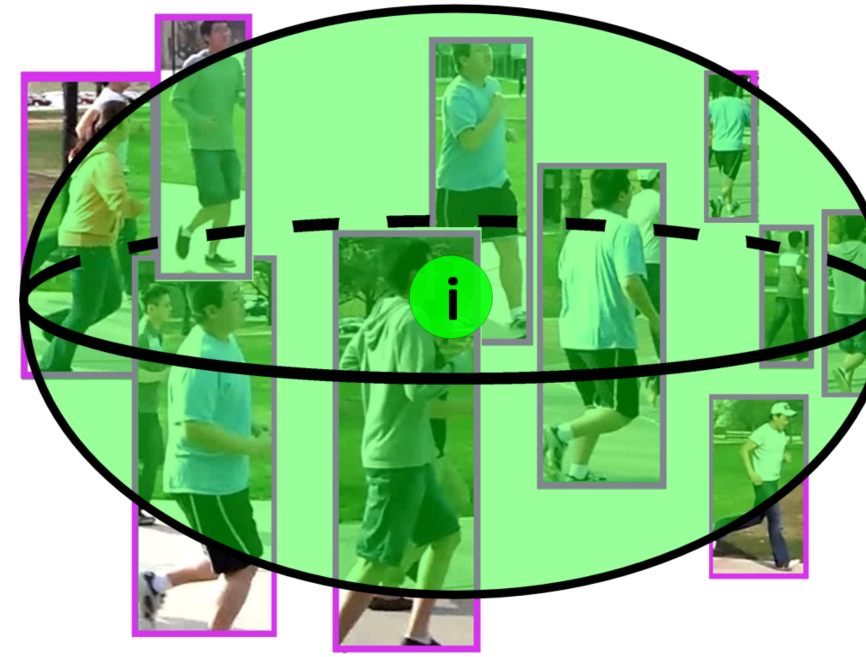
Amer & Todorovic ICCV11

Example Problem

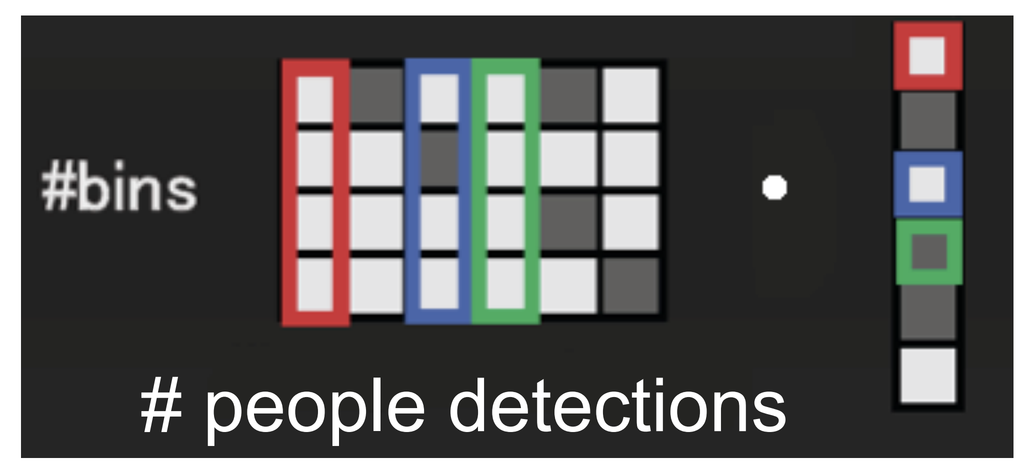
Grounding the grammar on a space-time grid of Bags of Right Detections (BORDs)

Structure is encoded in the overlap of BORDs

Bags of Right Detections



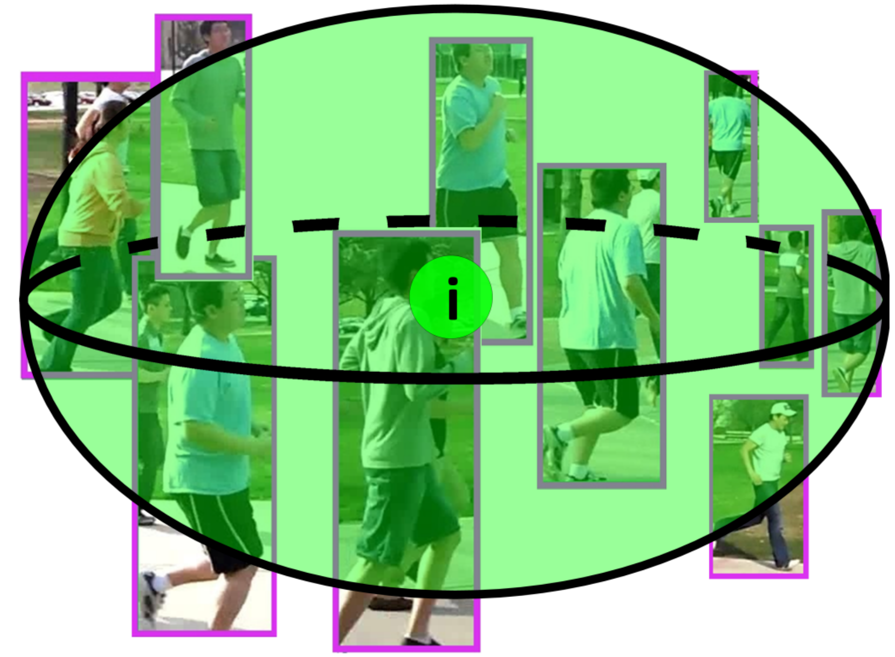
BORD i



$$S_i \cdot \mathbf{x}$$

shape context indicator

Bags of Right Detections



BORD i

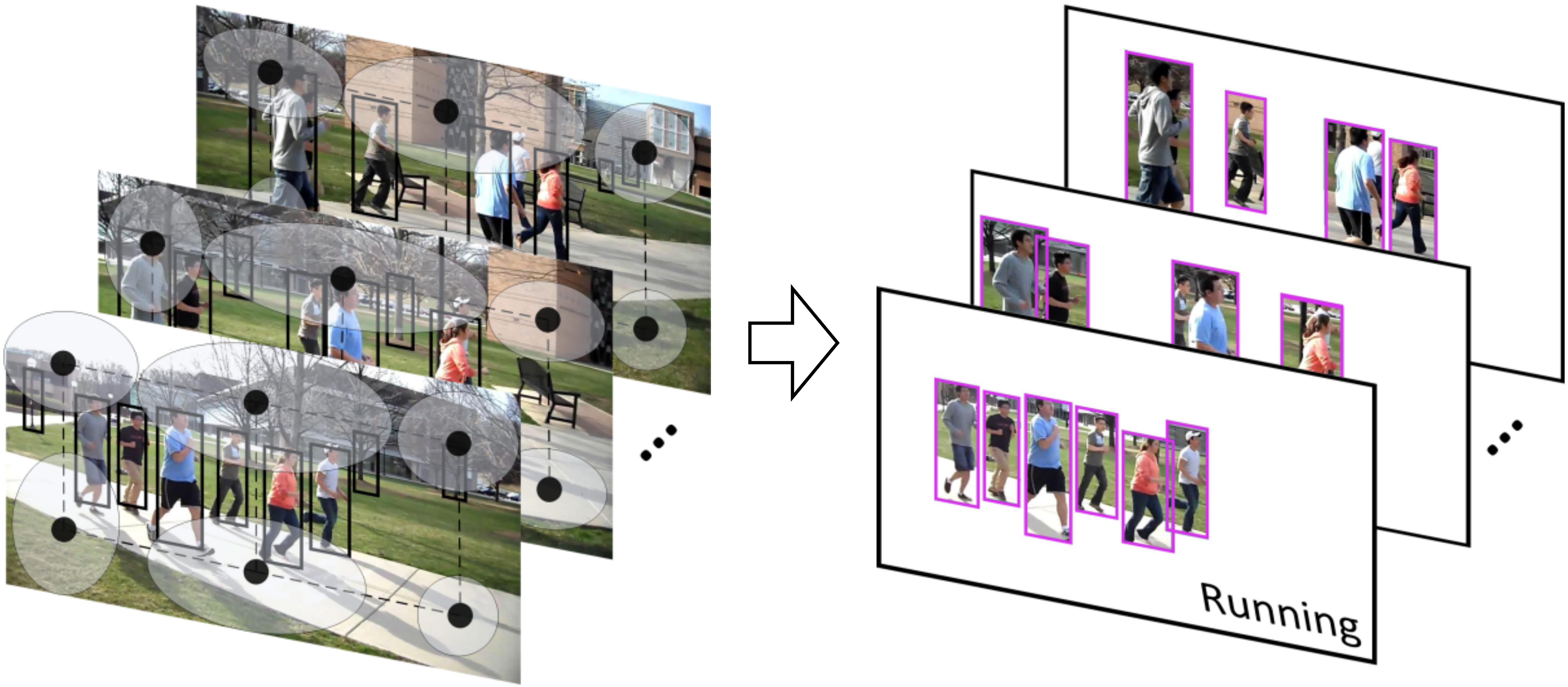
$\mathbf{x}$ - latent variable

constrained by all BORDs

$$S_i \cdot \mathbf{x}$$

shape context   indicator

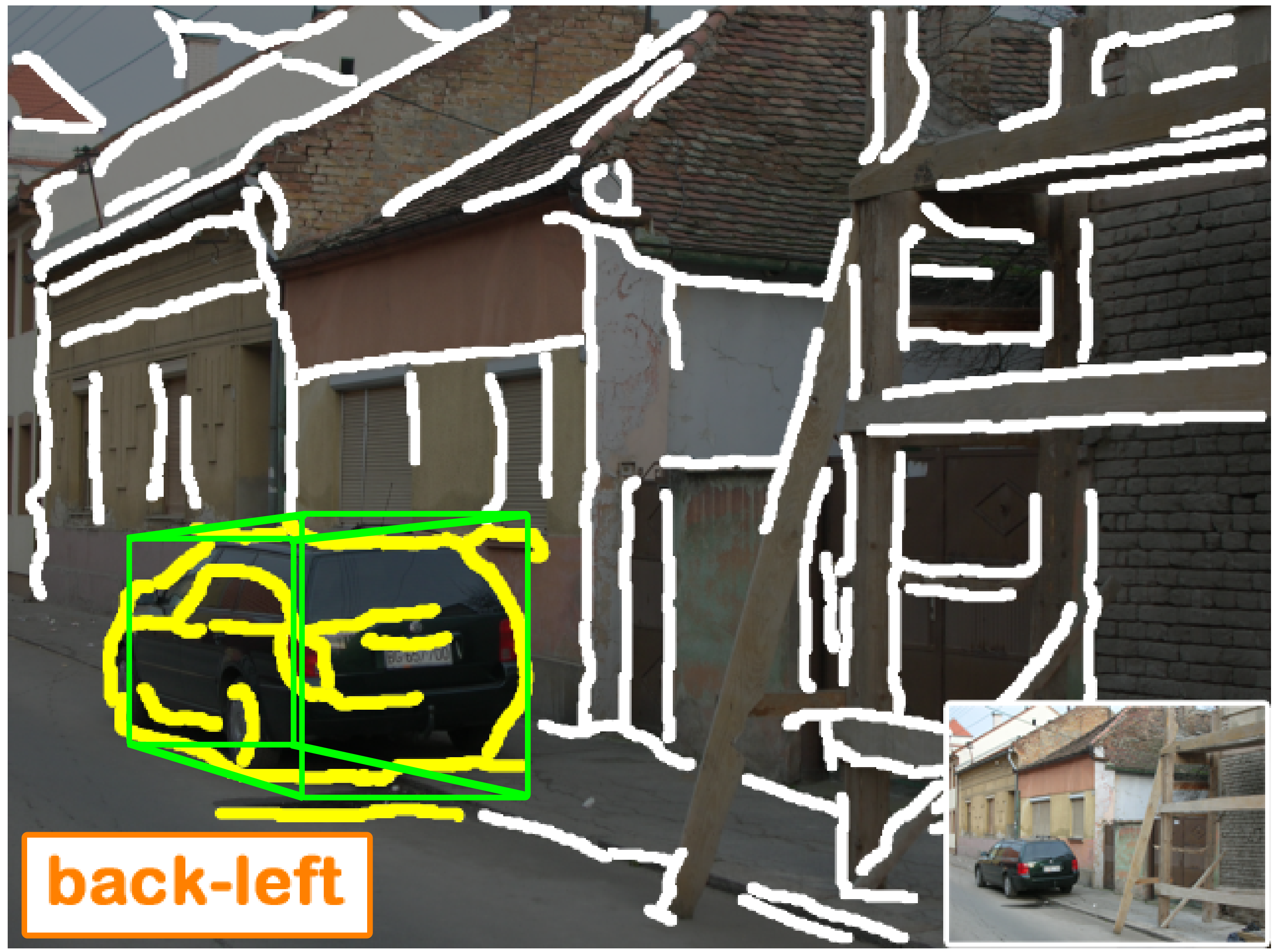
Detection and Localization



BORDs jointly constrain the solution

Amer & Todorovic ICCV11

Example Problem: Object Recognition



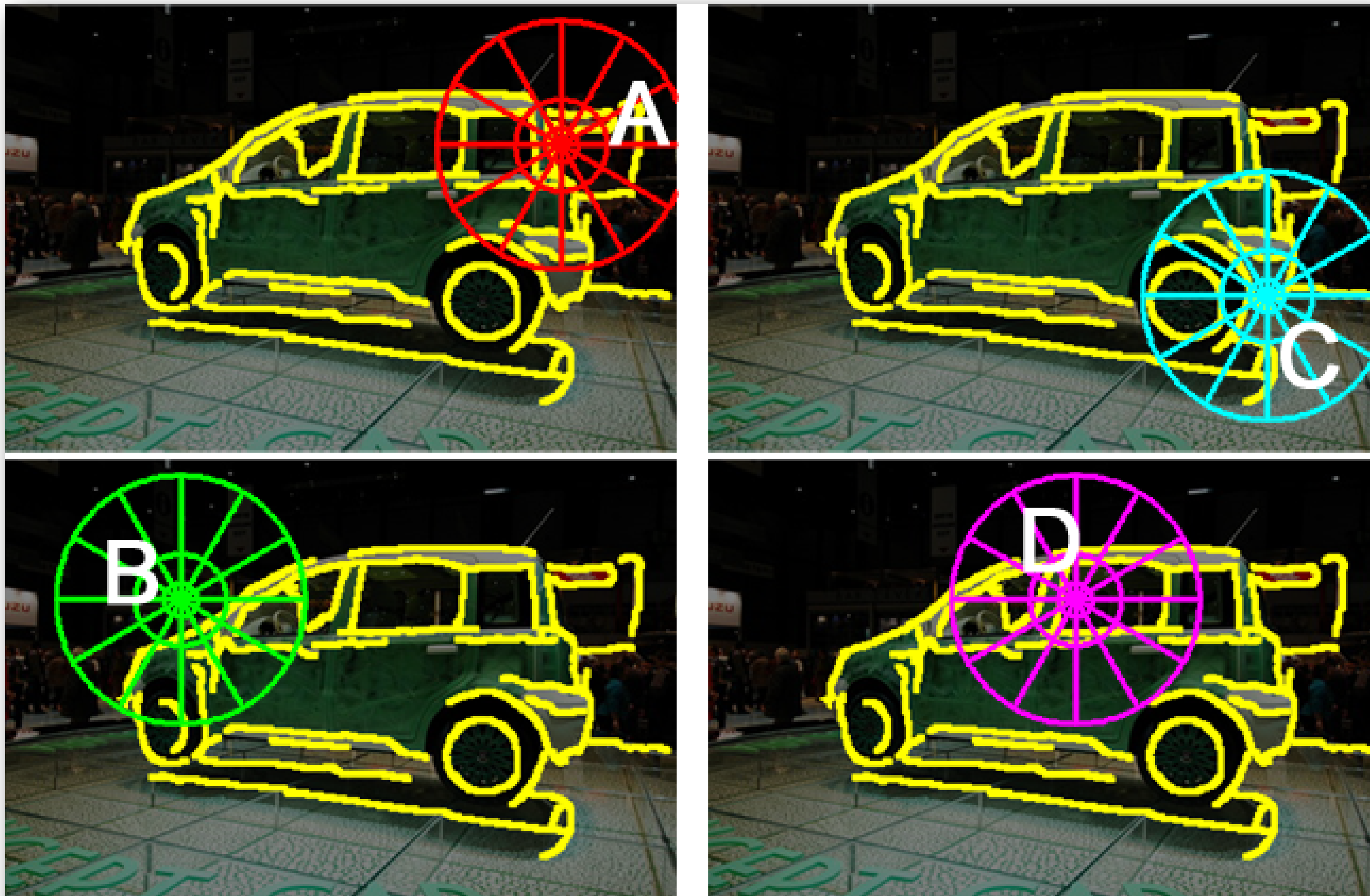
Given a set of edges in the image

detect and localize all object instances

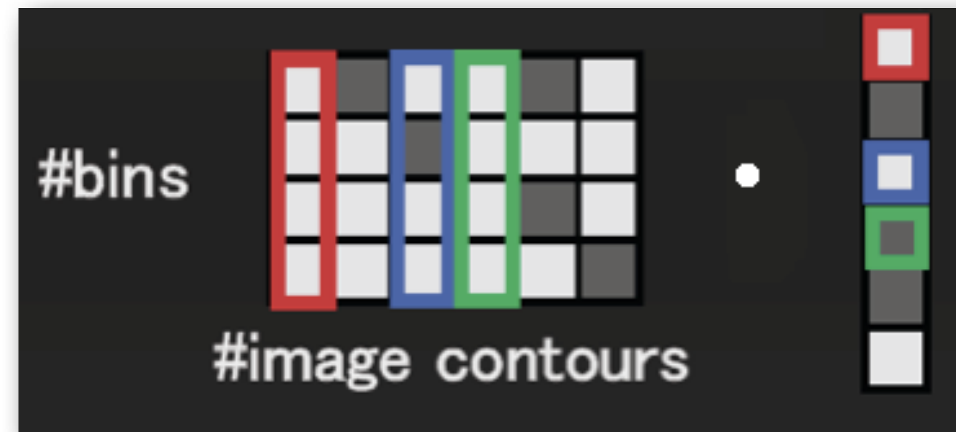
and estimate their 3D pose

Payet & Todorovic ICCV11

Bags of Right Detections



BORD i



$$S_i \cdot \mathbf{x}$$

shape context

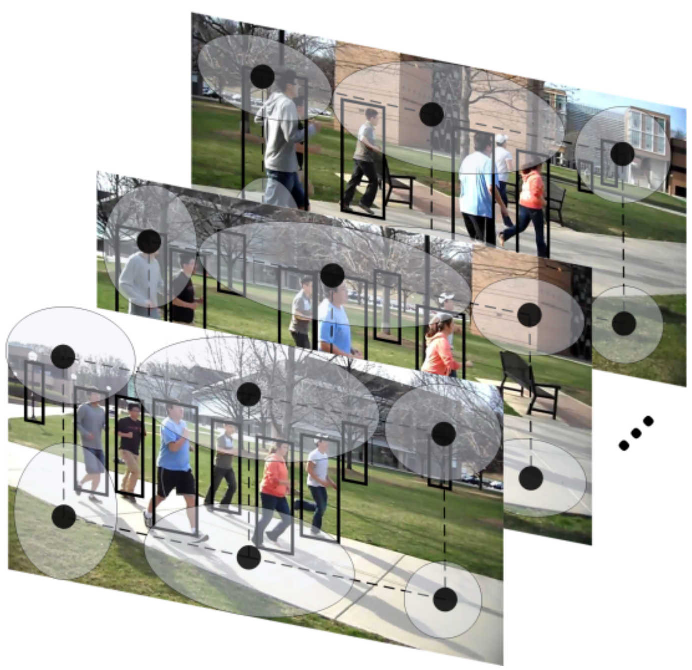
latent indicator

BORDs jointly constrain the solution

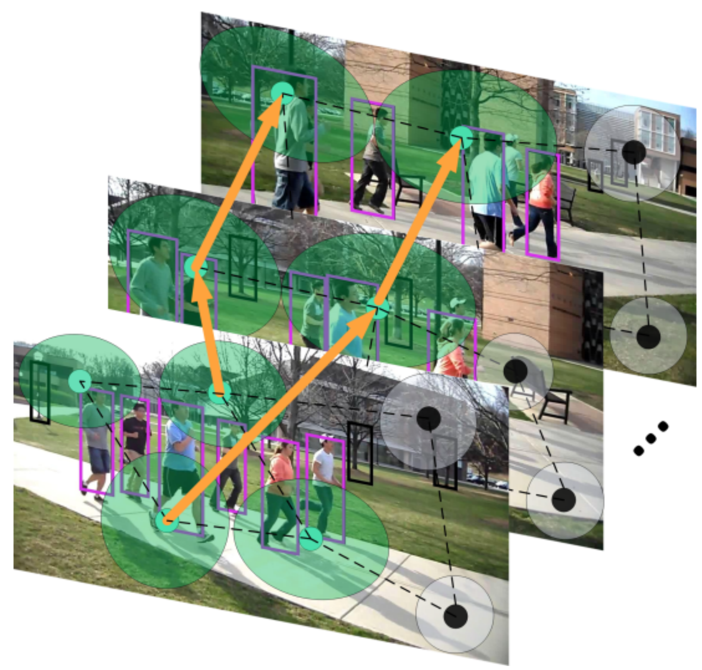
Payet & Todorovic ICCV11

Our Approach

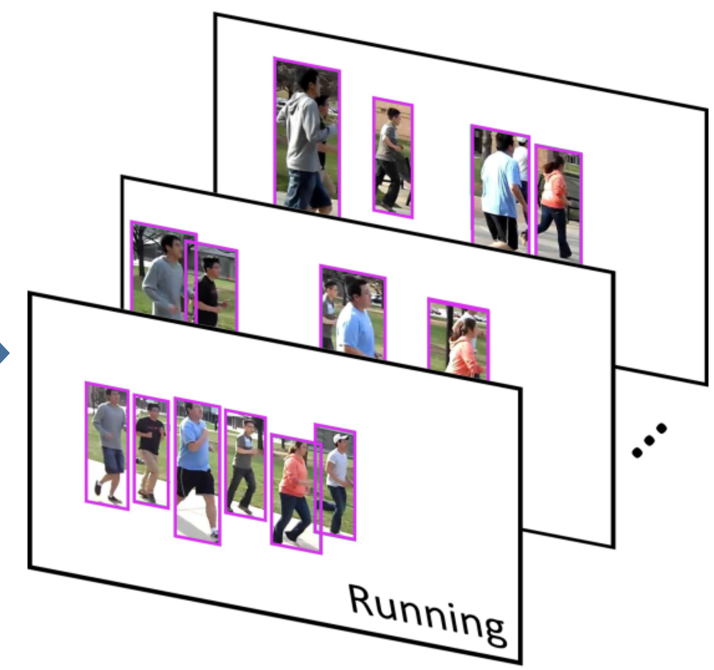
Initial placement of BORDs on a regular grid



Search for optimal features by warping the grid



Detection & Localization



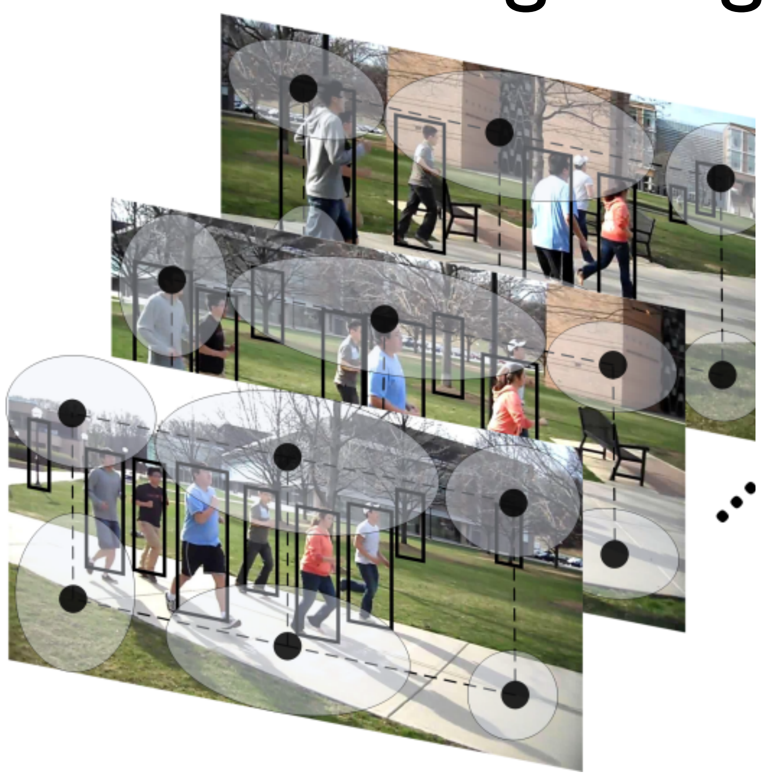
guided by

chains graphical model

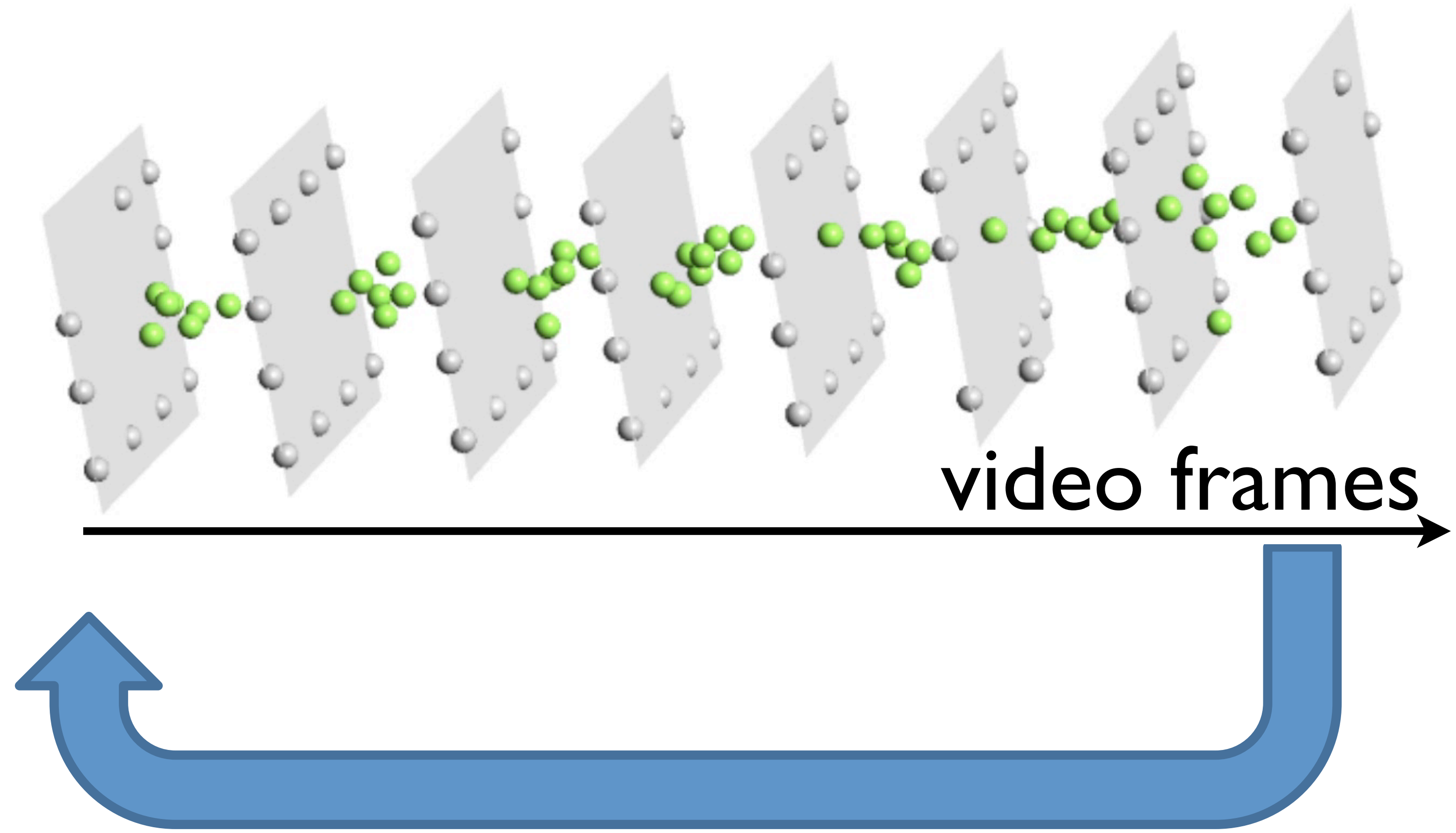
Amer & Todorovic ICCV11

Our Approach

Initial placement of BORDs on a regular grid



Search for optimal features by warping the grid



guided by

chains graphical model

MAP Inference:

1. Warp the grid to expected locations
2. Select MAP BORDs

The Chains Model

$$P(M, O, L_S, L_E, F)$$

Features → Observables

End frame of activity

Start frame of activity

Latent variables

Ordering of BORDs

Number of BORDs

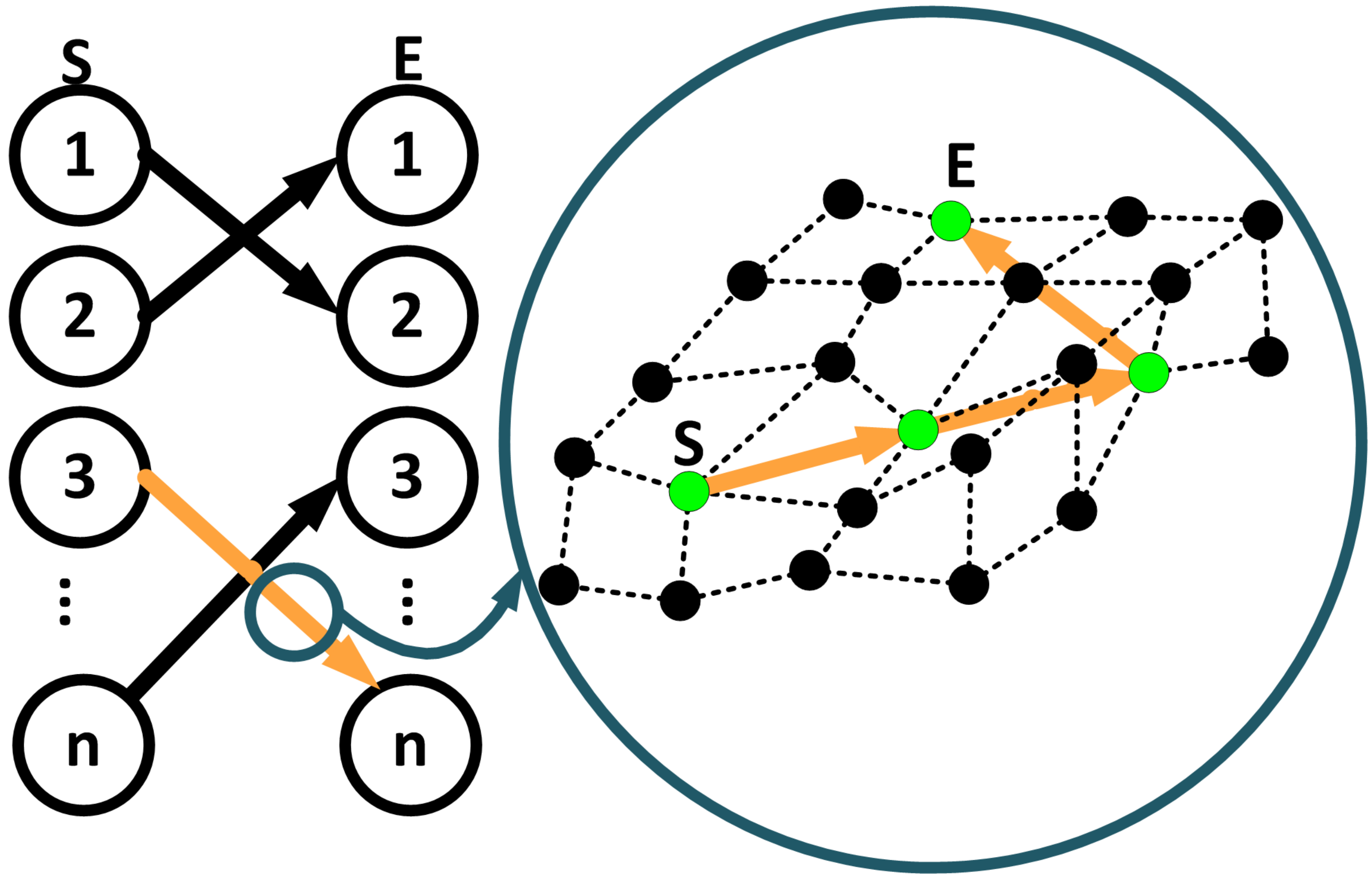
The Chains Model

$$P(M, O, L_S, L_E, F)$$

$$= P(M, O)P(L_S|M, O, F)P(L_E|M, O, F)$$

$$\cdot \prod_i P(F_{O(i+1)}|F_{O(i)}) \prod_{i \in F-O} P_{\mathrm{bgd}}(F_i)$$

MAP Inference



$$P(M, O, L_S, L_E | F)$$

MAP Inference

$$P(L_S, L_E | F) \propto \sum_{M,O} P(M, O, L_S, L_E, F)$$

MAP Inference

$$P(L_S, L_E|F) \propto \sum_{M,O} P(M, O, L_S, L_E, F)$$

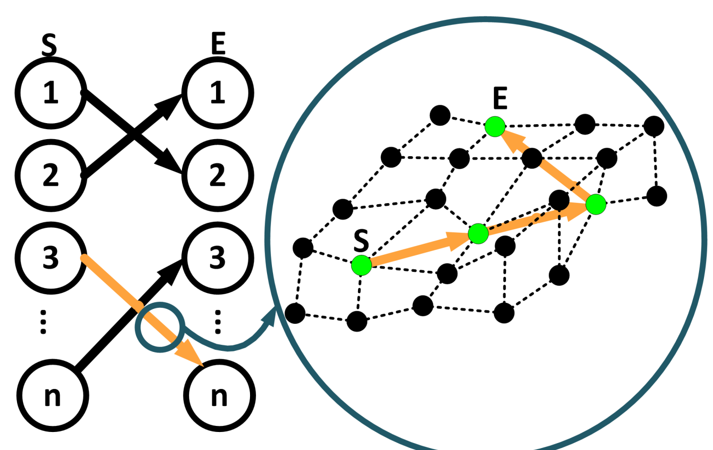
$$= \sum_{M,O} P(M) \left[\prod_{i,j} P(F_j|F_i) \right] P(L_S, L_E|M, O, F)$$

MAP Inference

$$P(L_S, L_E | F) \propto \sum_{M,O} P(M, O, L_S, L_E, F)$$

$$= \sum_{M,O} P(M) \left[\prod_{i,j} P(F_j | F_i) \right] P(L_S, L_E | M, O, F)$$

$$= \boldsymbol{\pi}_{\text{start}}^T \left[\sum_m P(M = m) X^m \right] \boldsymbol{\pi}_{\text{end}}$$



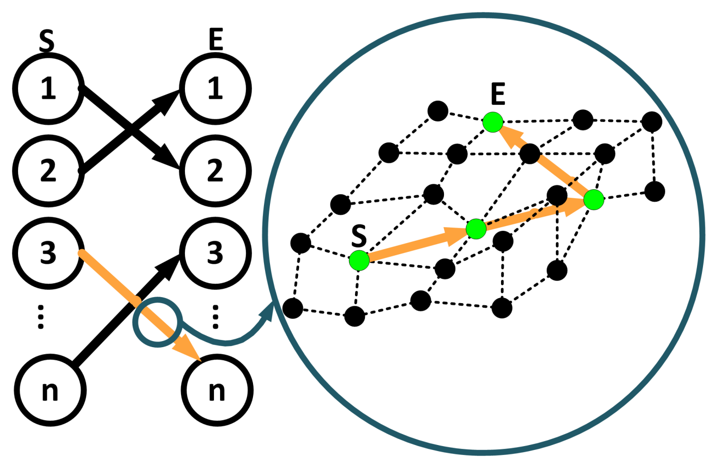
MAP Inference = LP

$$\begin{aligned}\text{minimize} \quad & \mathrm{tr}\{\boldsymbol{C}_{\boldsymbol{b}}^{\mathrm{T}}\boldsymbol{X}\} + \alpha\|(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}\boldsymbol{Q}\|_1 + \beta\|(\boldsymbol{I}-\boldsymbol{X})\boldsymbol{Q}\|_1 \\ \text{subject to} \quad & \boldsymbol{X} \geq 0, \;\; \boldsymbol{X}\mathbf{1}_n = \mathbf{1}_n, \;\; \boldsymbol{b} \geq 0, \;\; \|\boldsymbol{b}\|_2^2 = 1.\end{aligned}$$

Searching for optimal features

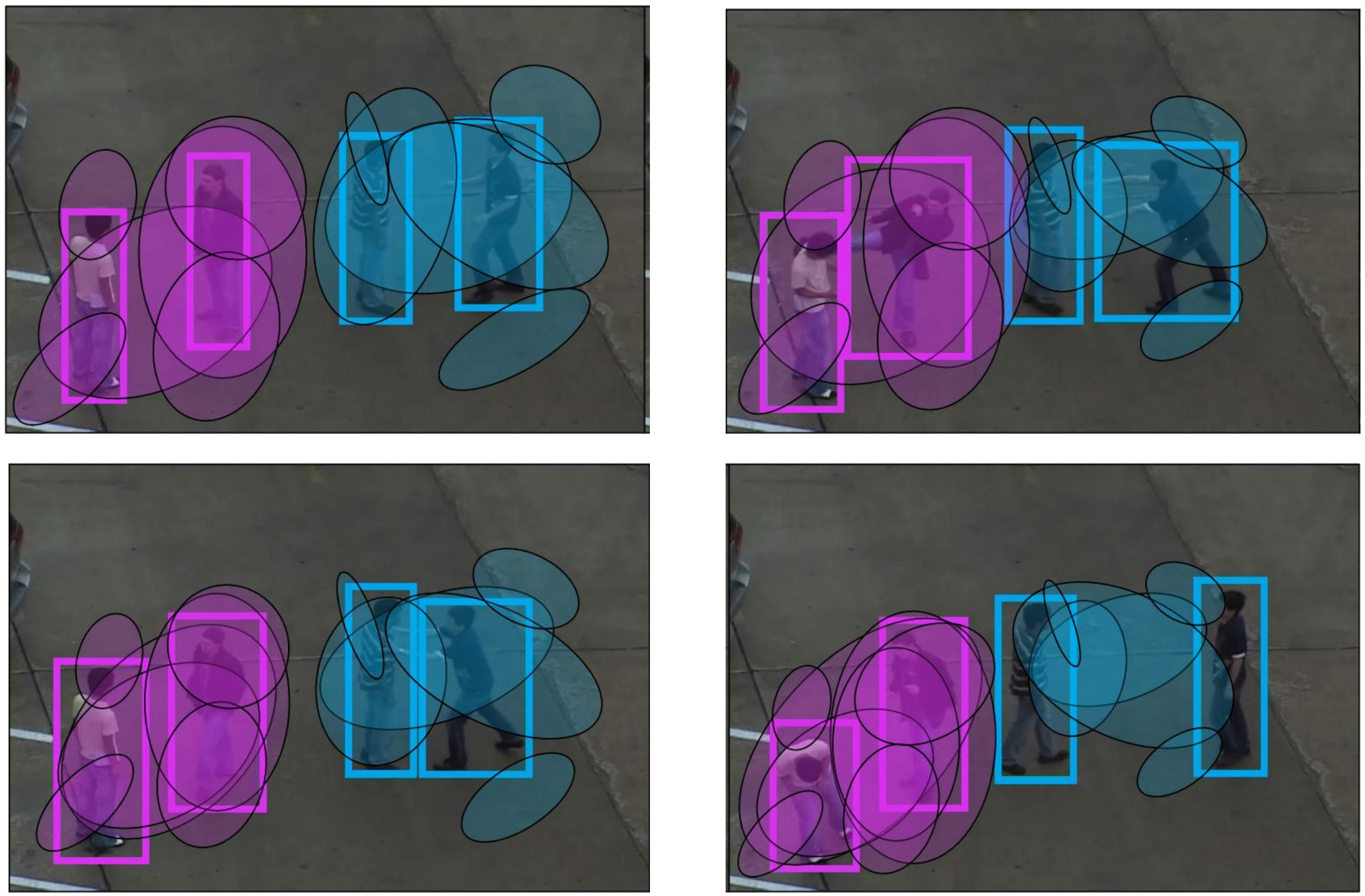
under non-rigid shape deformations

of the grid of BORDs



Results

video frames



Correct detection and localization of:
kicking and pushing
and actors involved

Results

video frames

Correct detection and localization of:
handshaking and hugging
and actors involved

Results

video frames

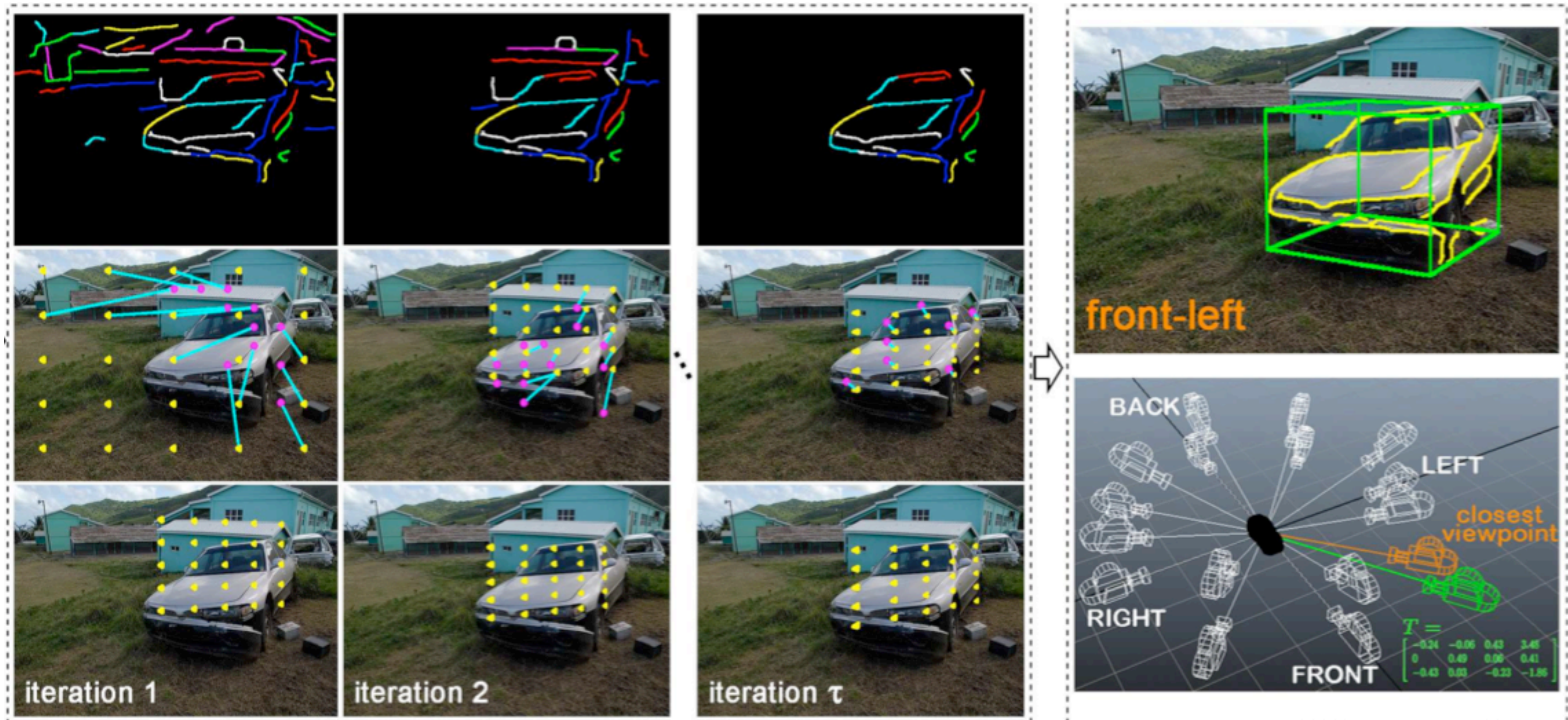
Failure example

correct: handshaking and hugging

wrong: actors involved

Example Problem: Object Recognition

MAP inference = LP



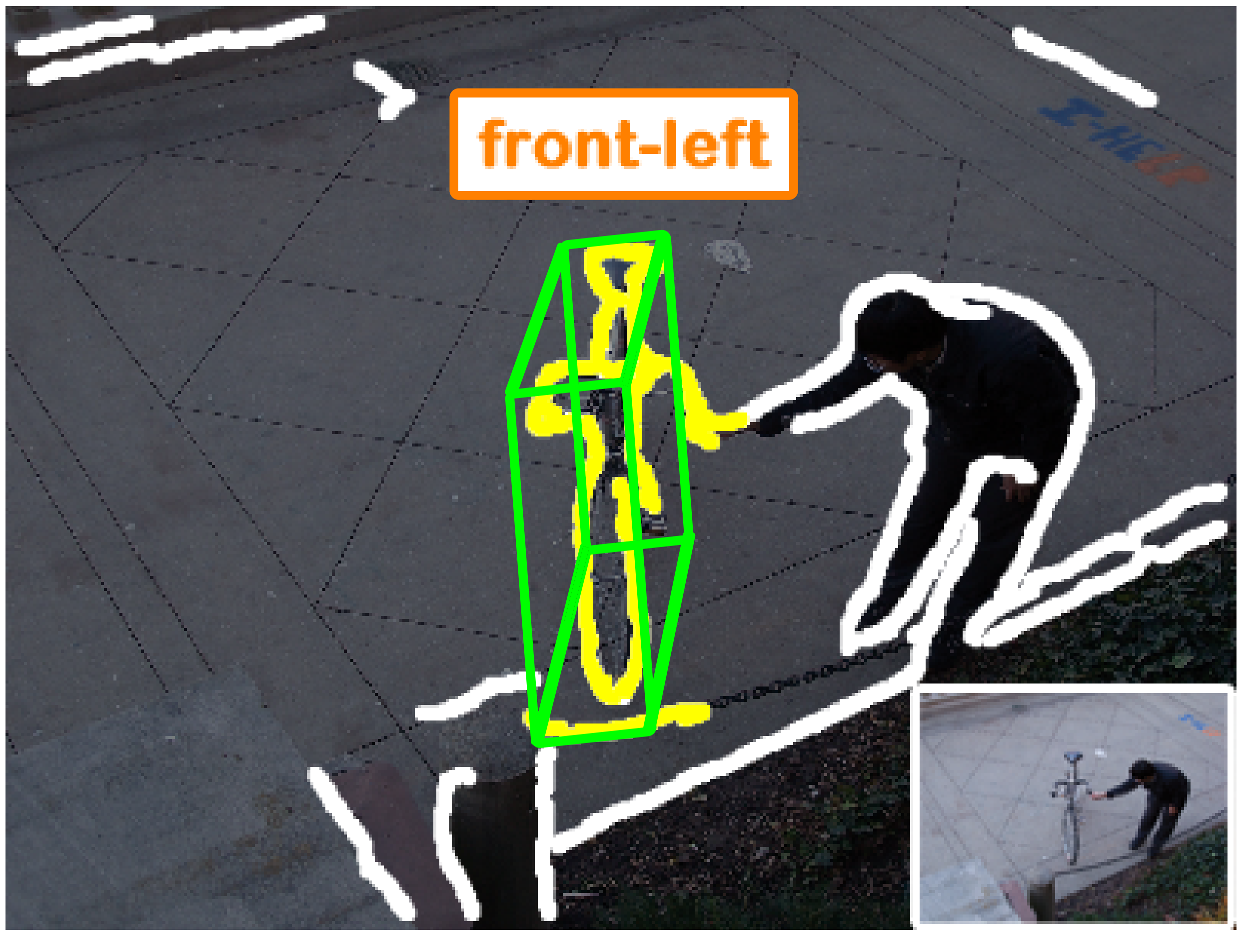
Searching for optimal features

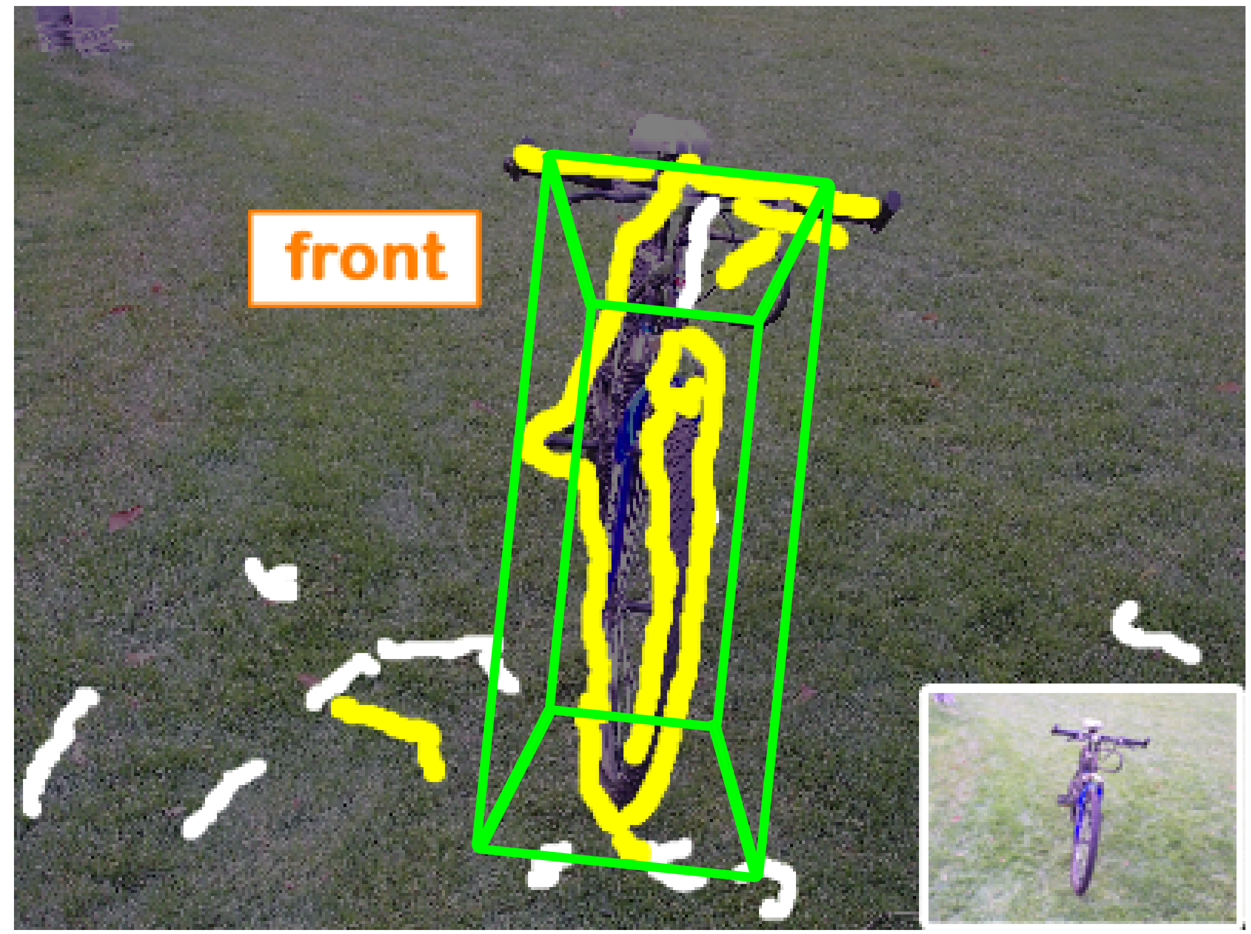
under non-rigid shape deformations

of the grid of BORDs

Payet & Todorovic ICCV11

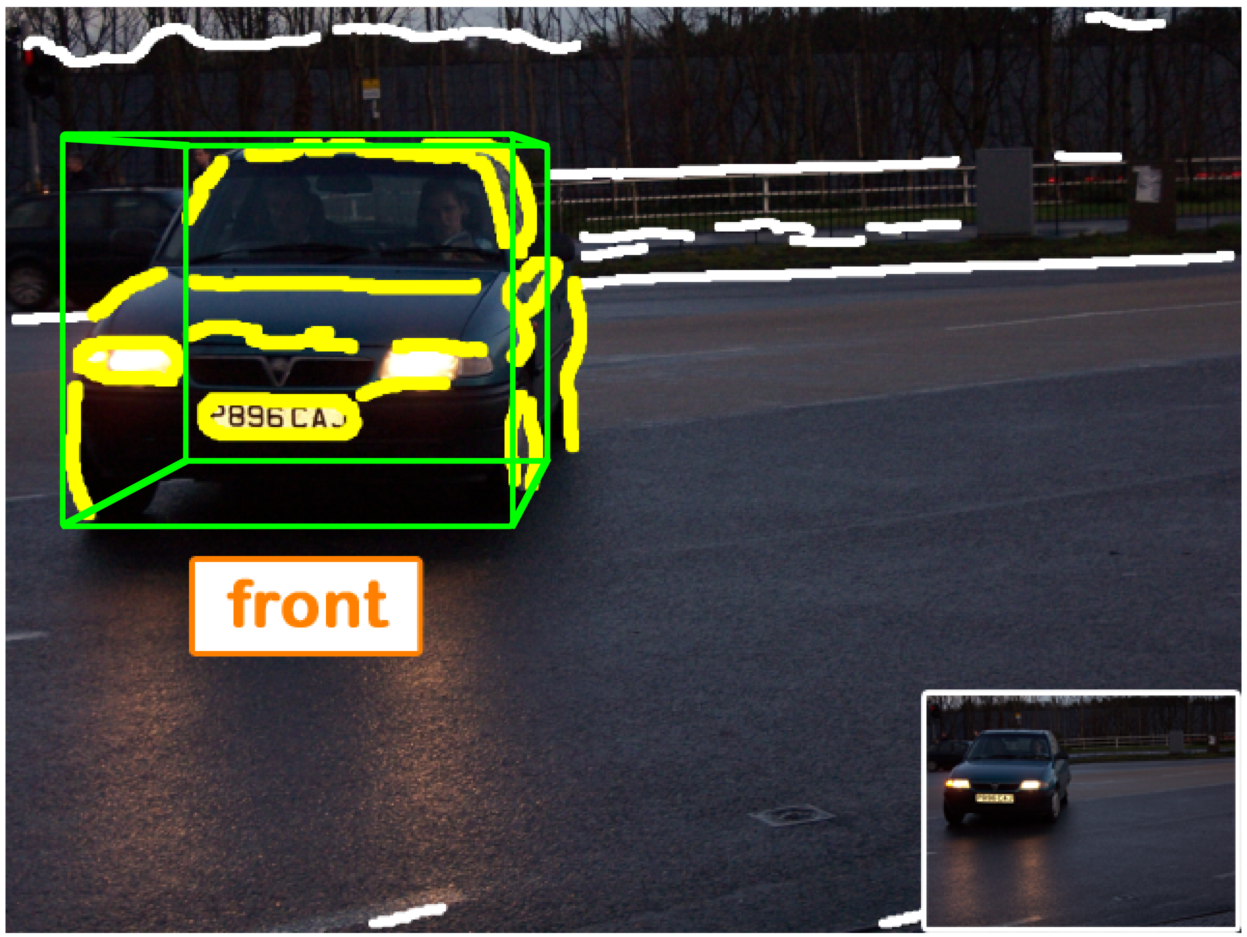
Results

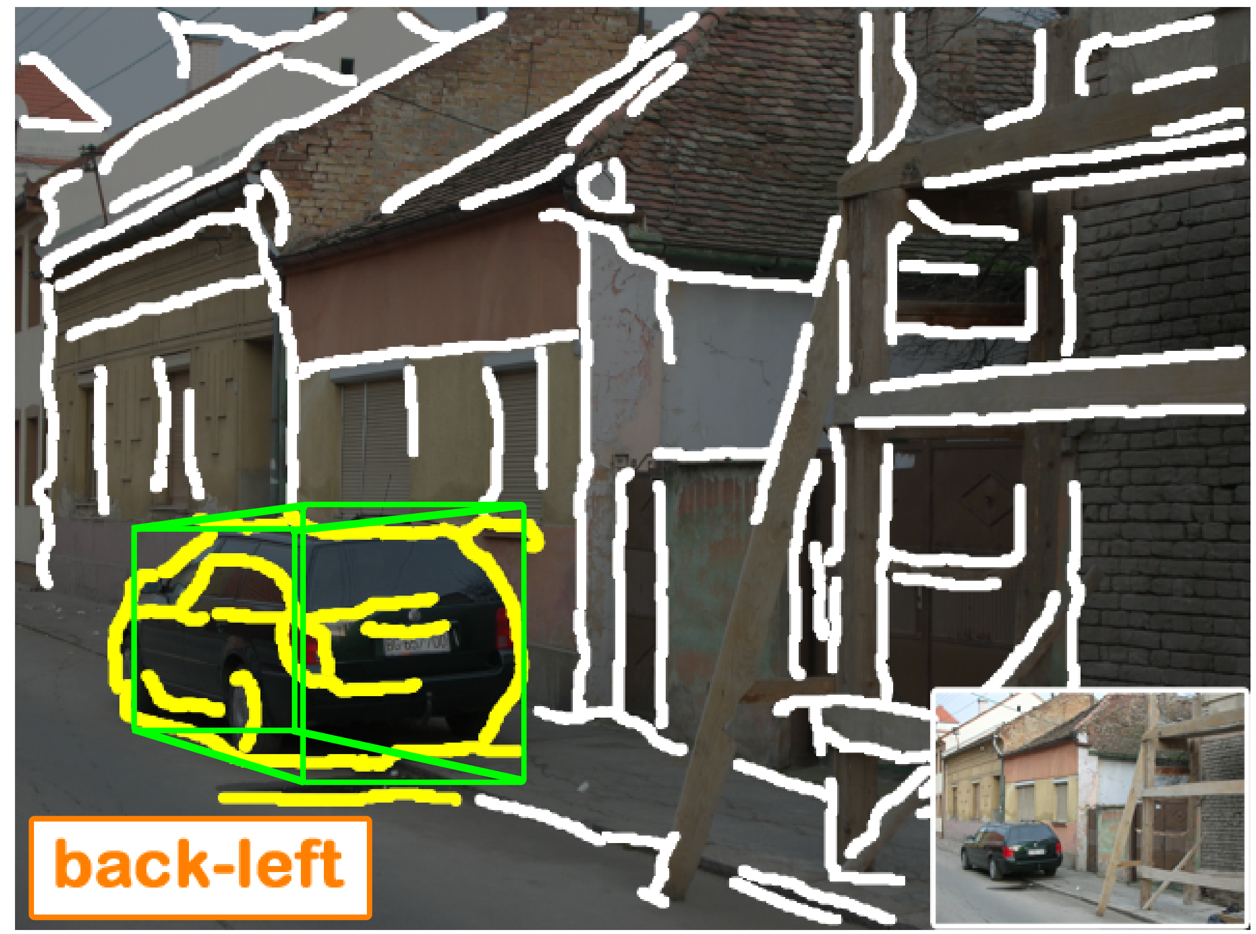




Correct detection, localization, and pose estimation

Results



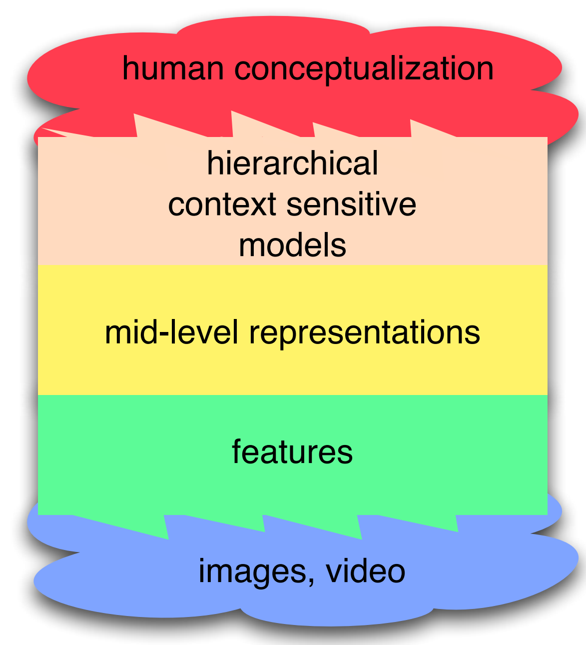


Correct detection, localization, and pose estimation

Conclusion

- Prior work: pre-selected features, typically low-level for repeatability

- Proposed mid-level features:
  - Allow abstraction of low-level features
  - Reduce the semantic gap
  - Enable addressing multiple tasks
  - Repeatable, and jointly encode structure



Acknowledgment

NSF IIS 1018490