Part I : Tracking and Data Association

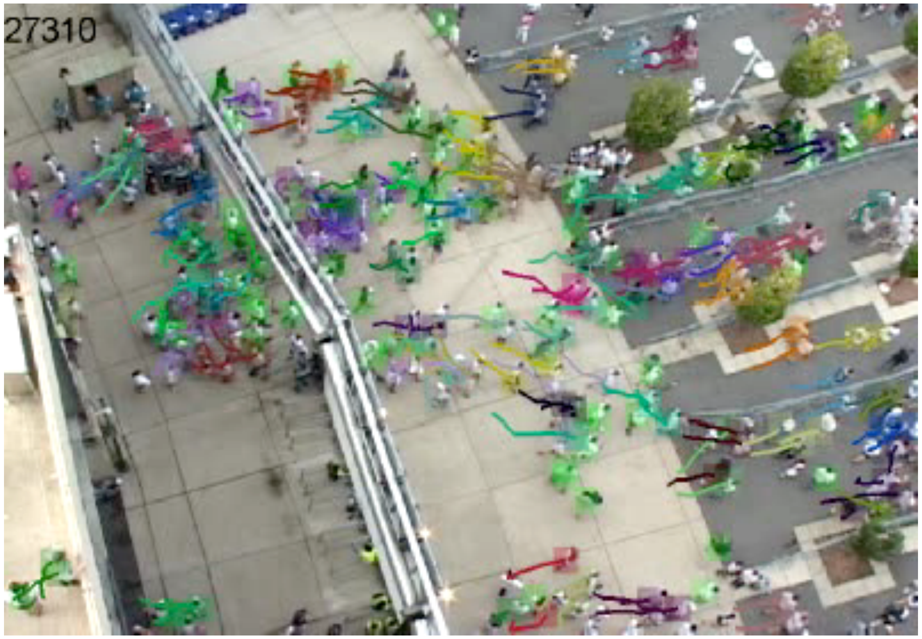
Part II : Crowd-scene Analysis

VLPR 2012, Shanghai, China
Bob Collins, July 2012

Crowd Scene Analysis

- Using computer vision tools to look at people in public places

- Real-time monitoring
  - situation awareness
  - notifications/alarms
- After-action review
  - traffic analysis



Crowd Scene Analysis

Things we might want to know:

- How many people are there?
- How to track specific individuals?
- How to determine who is with whom?

Challenges:

Crowd scenes tend to have low resolution.
You rarely see individuals in isolation.
Indeed, there are frequent partial occlusions.

Crowd Counting

FAQ: How many people participated in ...

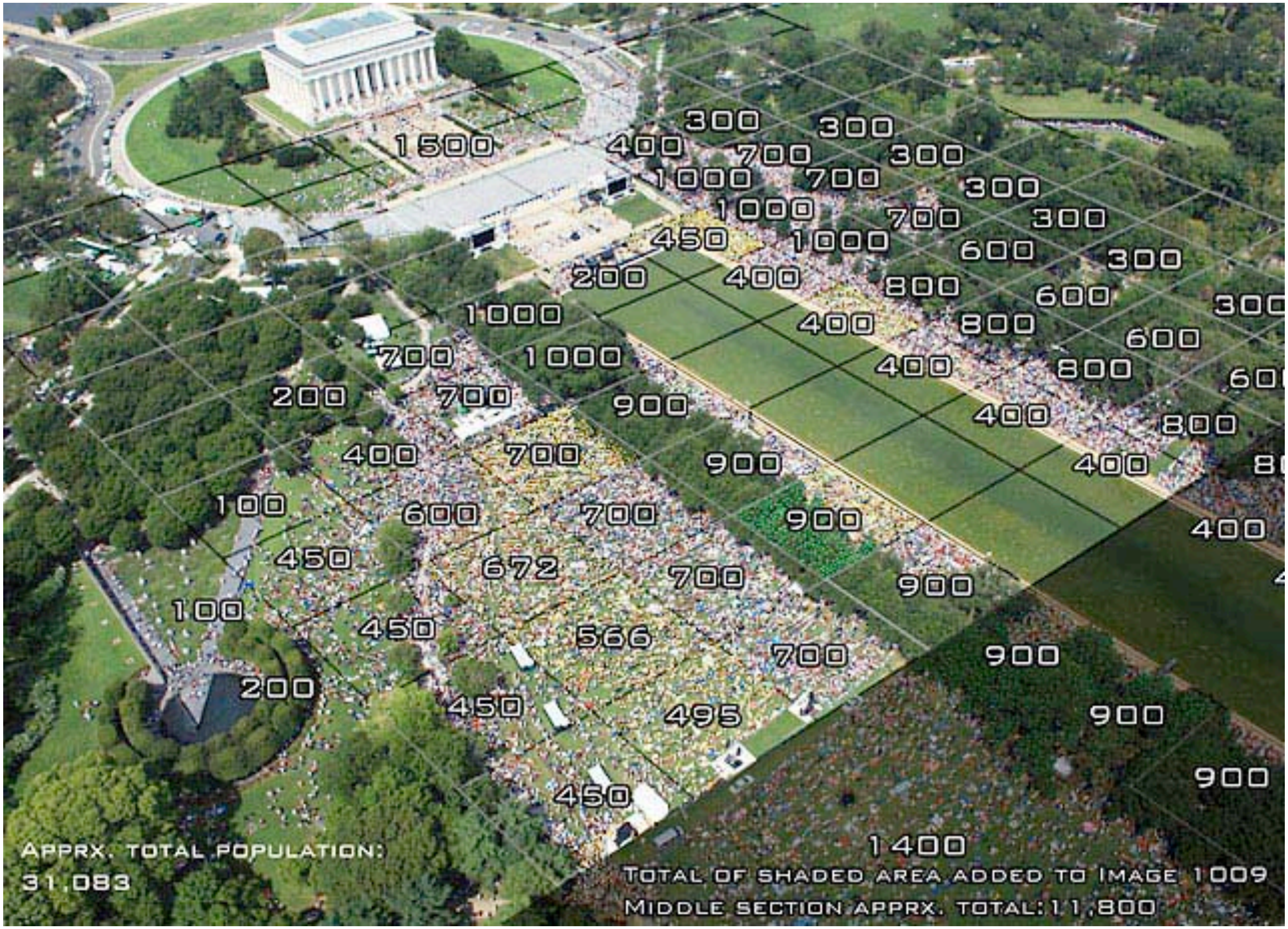
- Tahrir Square Protests
- Obama's inaguration
- Occupy Wall Street
- Kumbh Mela

Jacob’s Method

- Herbert Jacobs, Berkeley, 1960s
- count = area * density
  - 10 sqft/person – loose crowd (arm’s length from each other)
  - 4.5 sqft/person – more dense
  - 2.5 sqft/person – very dense (shoulder-to-shoulder)
- Problem: Pedestrians do not uniformly distribute over a space, but clump together into groups or clusters.
- Refinement: break area into a grid of ground patches and estimate a different density in each small patch. Accumulate these counts over whole area.

Robert Collins
Penn State

Example of Jacob's Method



source http://www.popularmechanics.com/science/the-curious-science-of-counting-a-crowd

Computer Vision Could do Better!

Cavaet: nobody really wants accurate counts

e.g. organizers of the "Million Man March" in Washington DC threatened to sue the National Park Service for estimating that only 400K people attended.

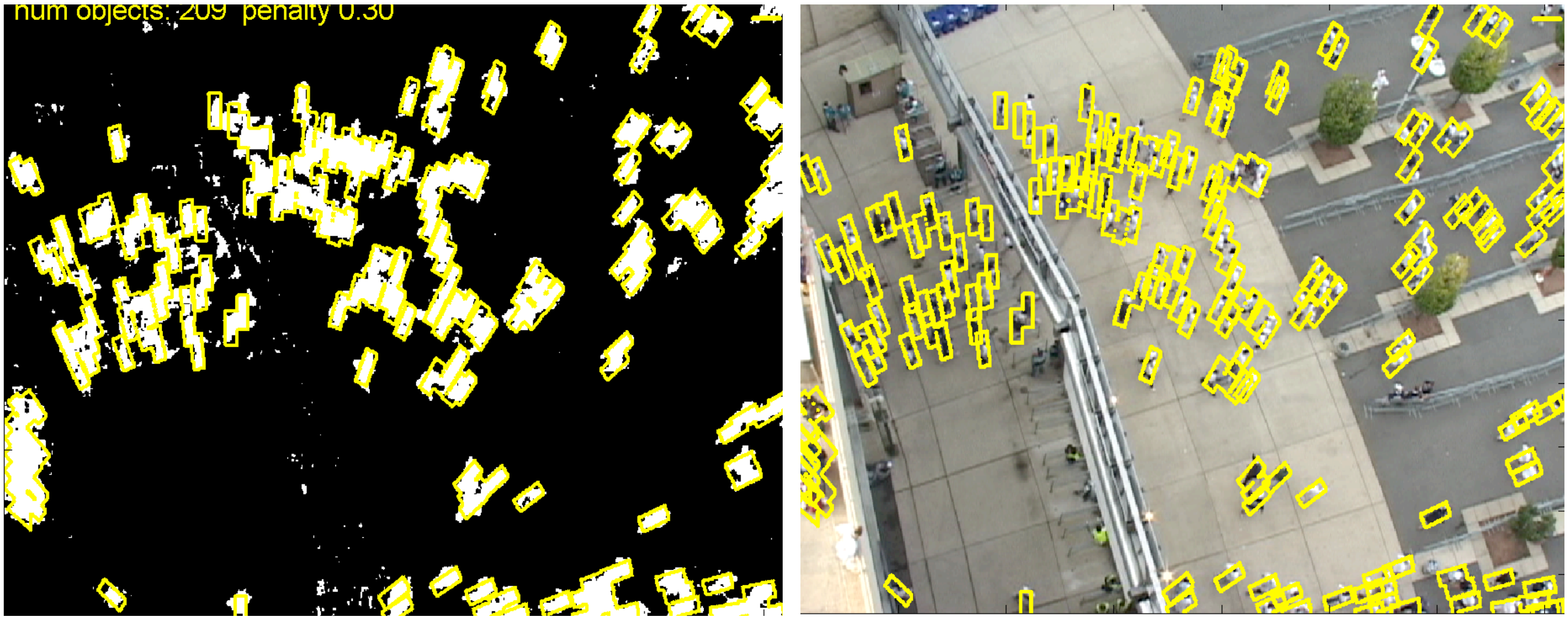
Vision-based Counting

- detection and tracking (light density)
- clustering feature trajectories that move coherently (moderate density)
- treat crowd as a dynamic texture and compute regression estimates based on measured properties (heavy density)

Robert Collins
Penn State

Detecting and Counting Individuals

Ge and Collins, "Marked Point Processes for Crowd Counting," *IEEE Computer Vision and Pattern Recognition (CVPR'09),* Miami, FL, June 2009, pp.2913-2920.

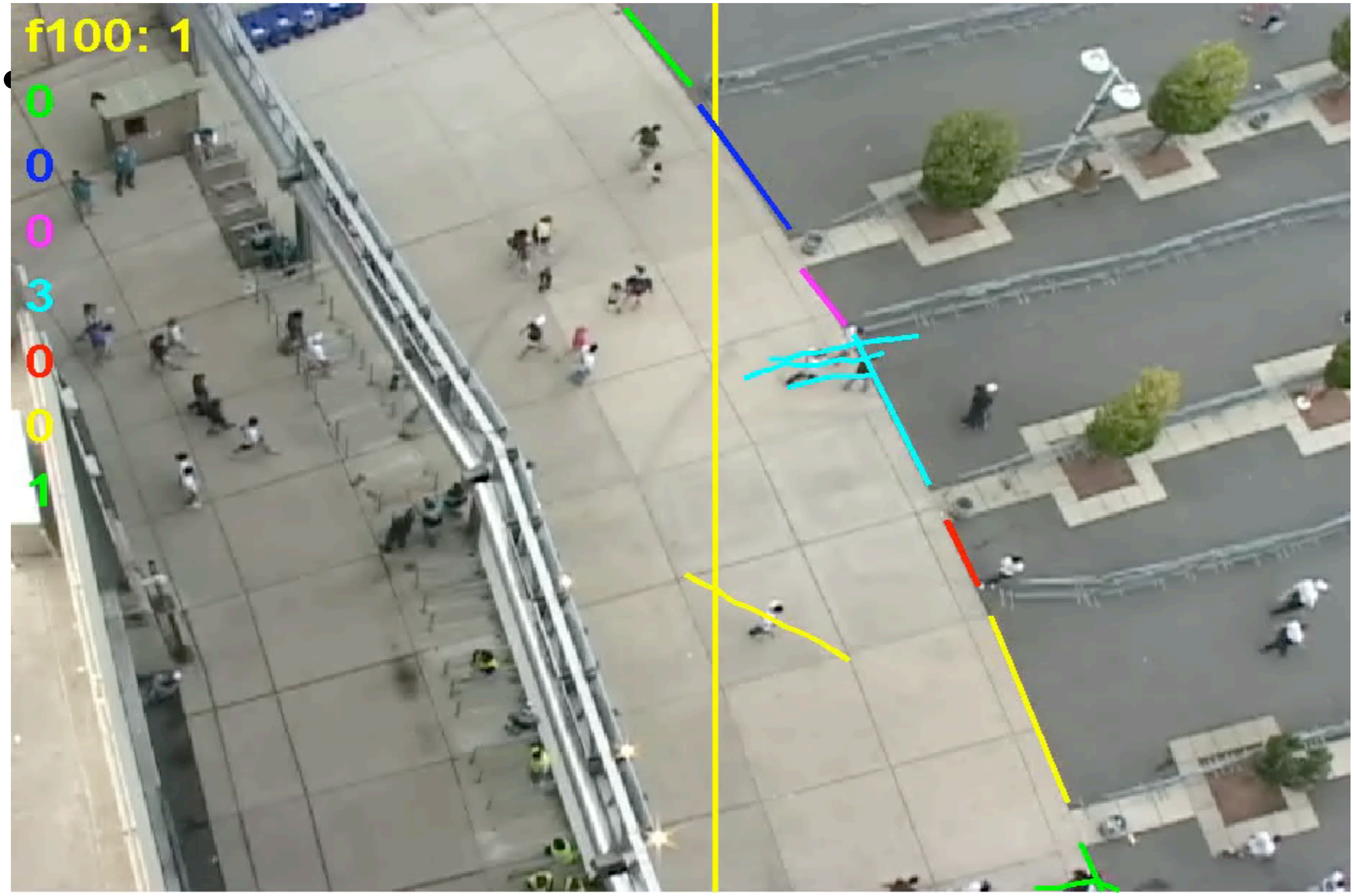


Good for low-resolution / wide-angle views.
Relies on foreground/background segmentation.
Not appropriate for very high crowd density or stationary people.

Robert Collins
Penn State

GateA Path Counts

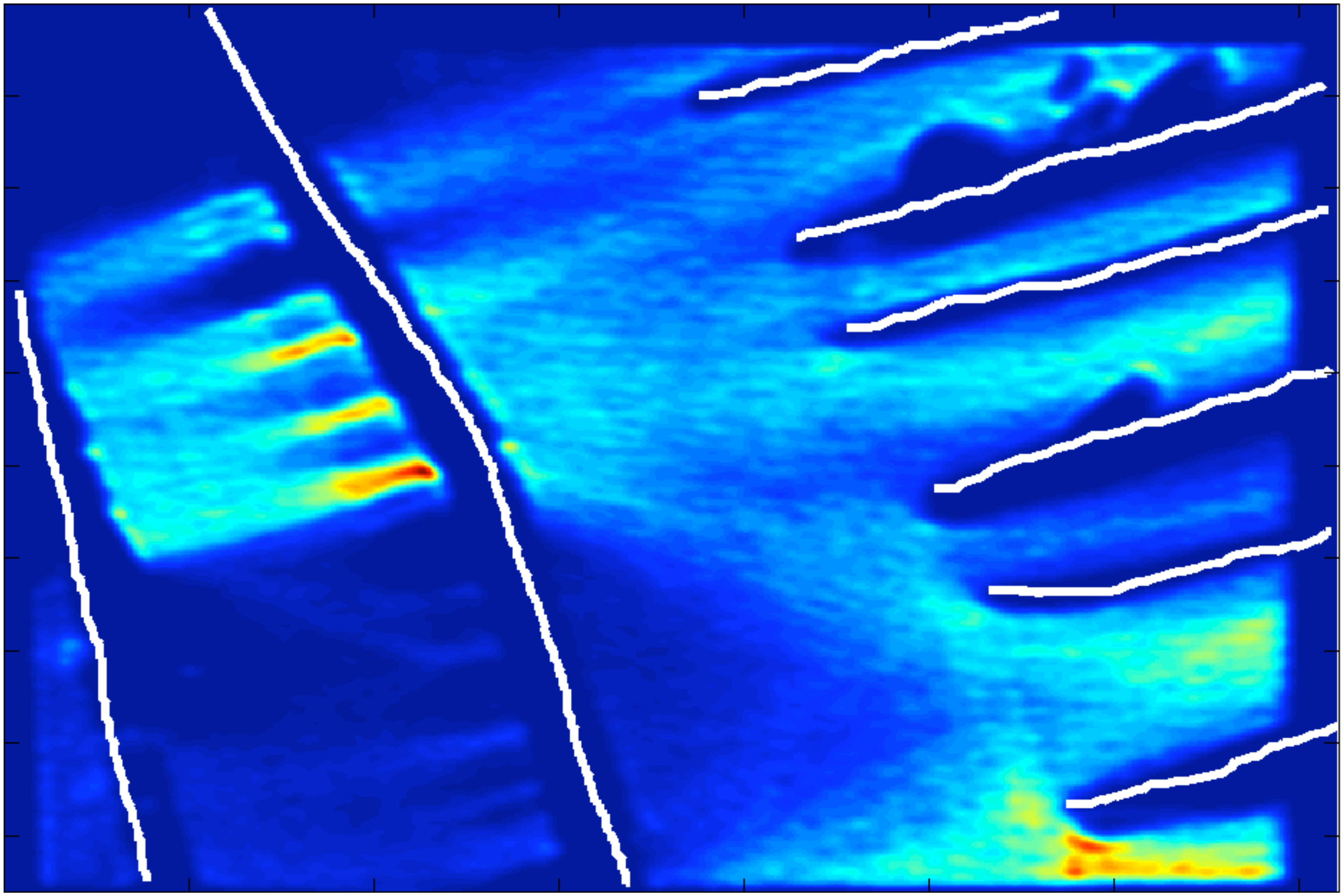
movie



Maintain a running count of number of people whose trajectories cross a set of user-specified lines (color-coded).

Robert Collins
Penn State

Crowd Flow/Density

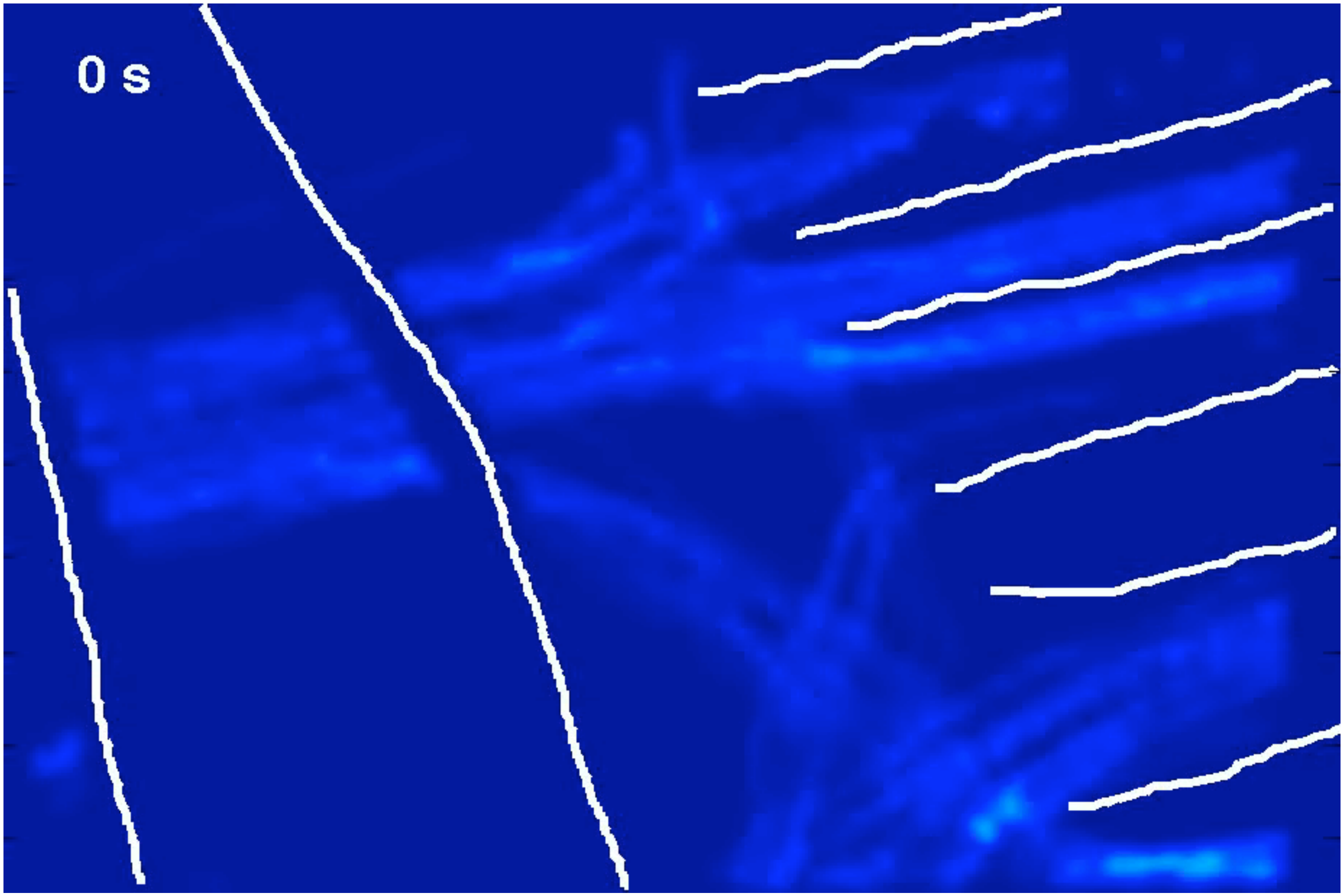


30 minute period

Robert Collins
Penn State

Crowd Flow/Density

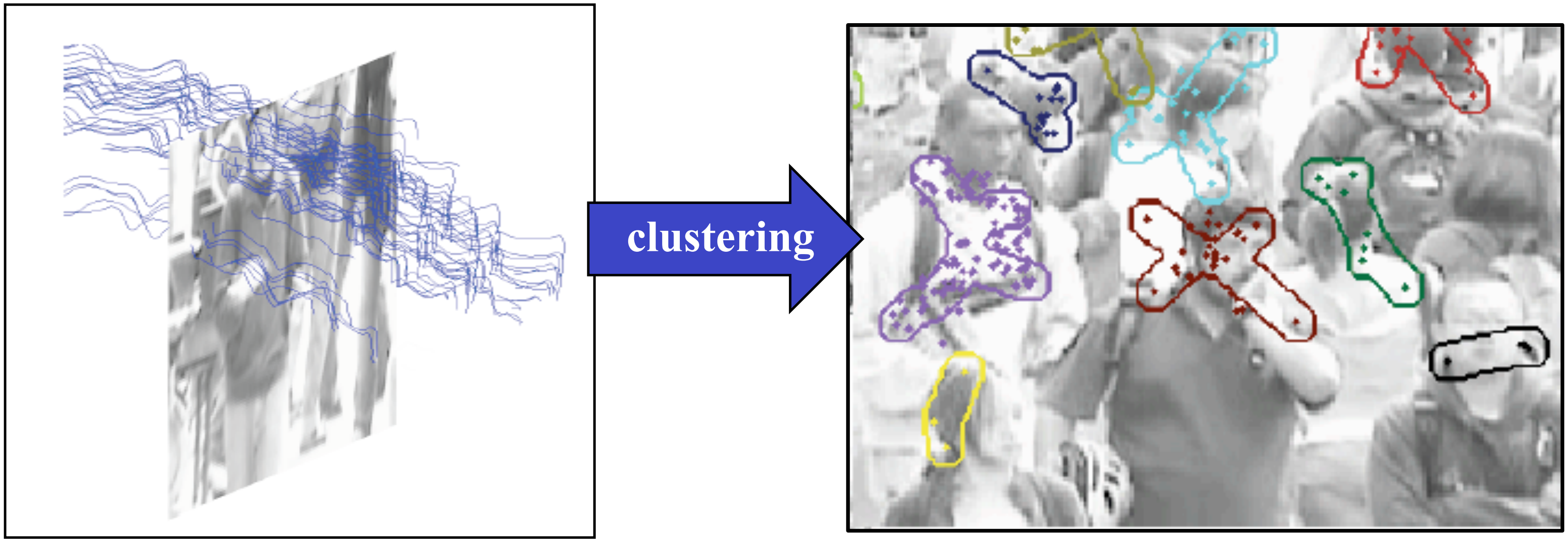
movie



Time Lapse. Integrated over spatial/temporal windows.

Motion Segmentation

Idea: track many small features (e.g. corners) over time and cluster sets of features that have similar motion.



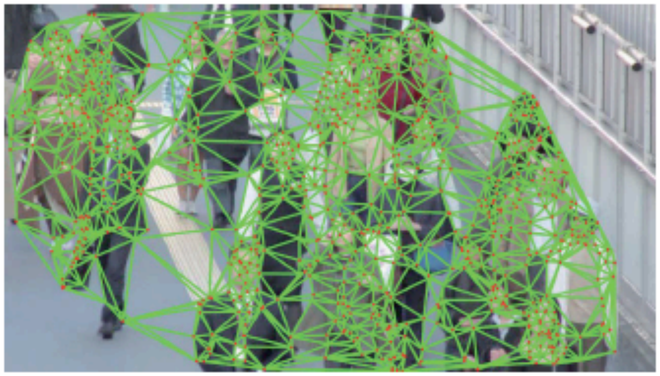
corner trajectories **independently moving objects**

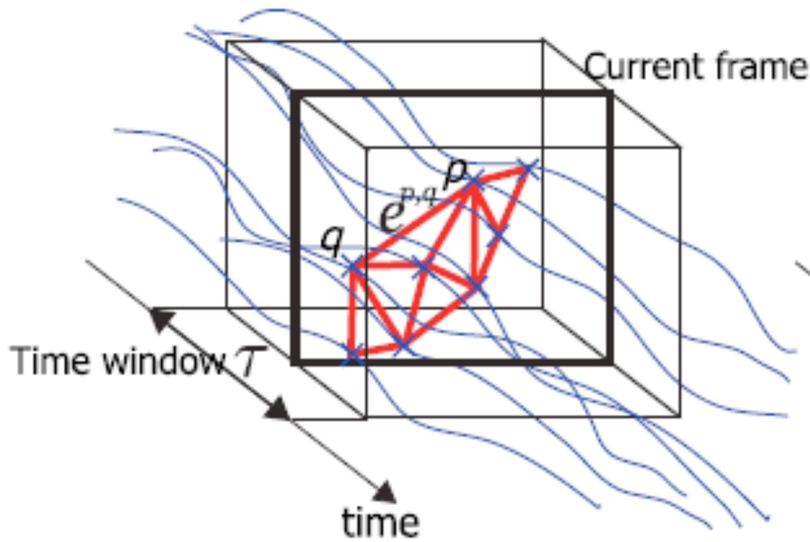
- G. J. Brostow and R. Cipolla, "Unsupervised bayesian detection of independent motion in crowds," in IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 594–601.
- V. Rabaud and S. Belongie, "Counting crowded moving objects," in IEEE Computer Vision and Pattern Recognition, New York City, 2006, pp. 705–711.
- D. Sugimura, K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Using individuality to track individuals: Clustering individual trajectories in crowds using local appearance and frequency trait," in International Conference on Computer Vision, 2009, pp. 1467–1474.

Robert Collins
Penn State

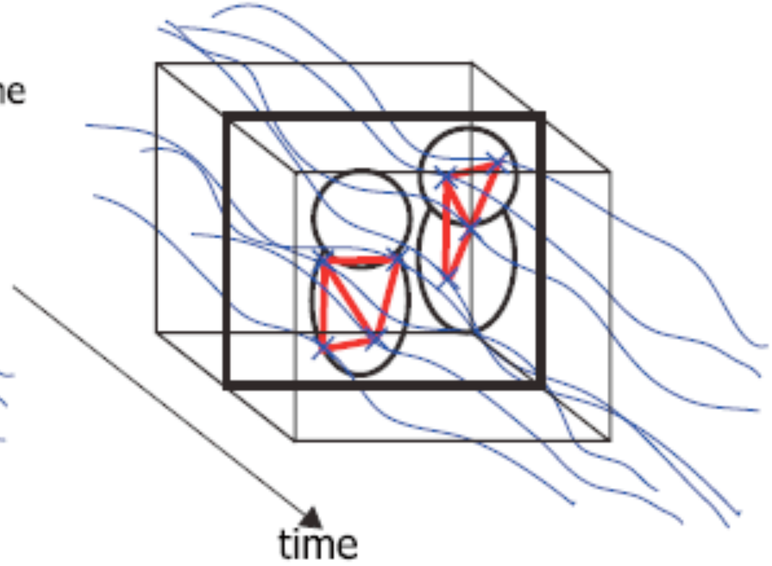
Motion Segmentation

Basic steps: Form a corner connectivity graph. Assign each edge a dissimilarity score based on distance and motion coherence of trajectories. Prune edges with high scores. The remaining connected components are the independent objects.





connectivity graph

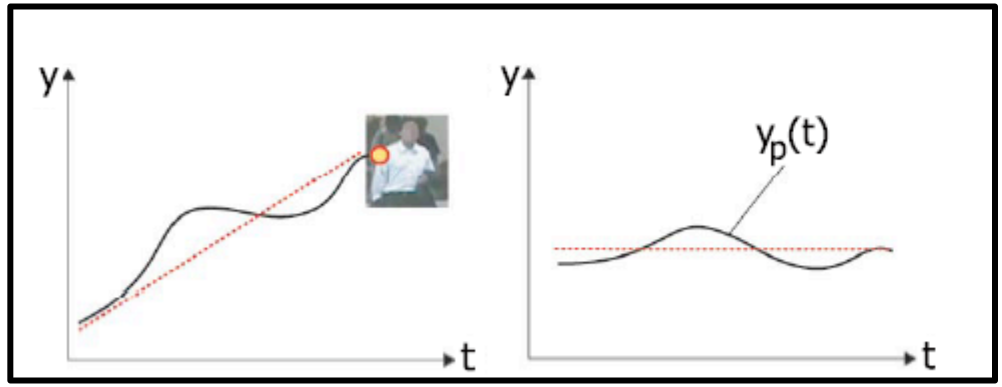


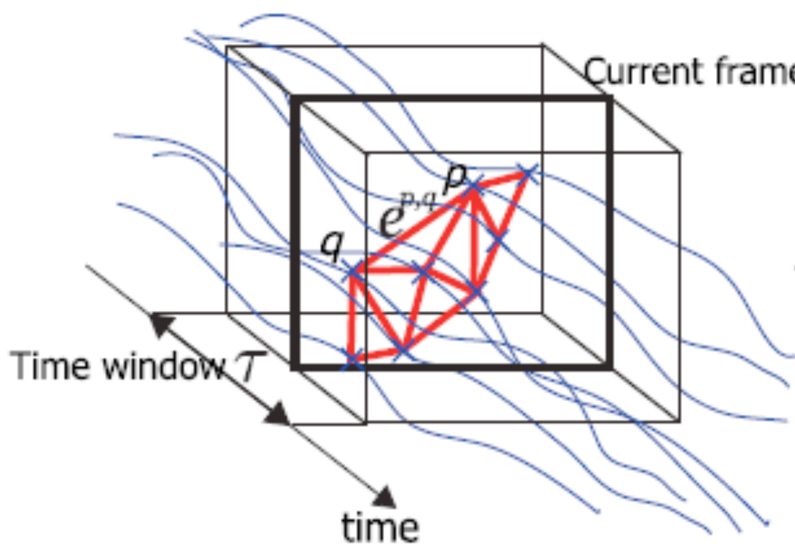
connected components after pruning

Robert Collins
Penn State

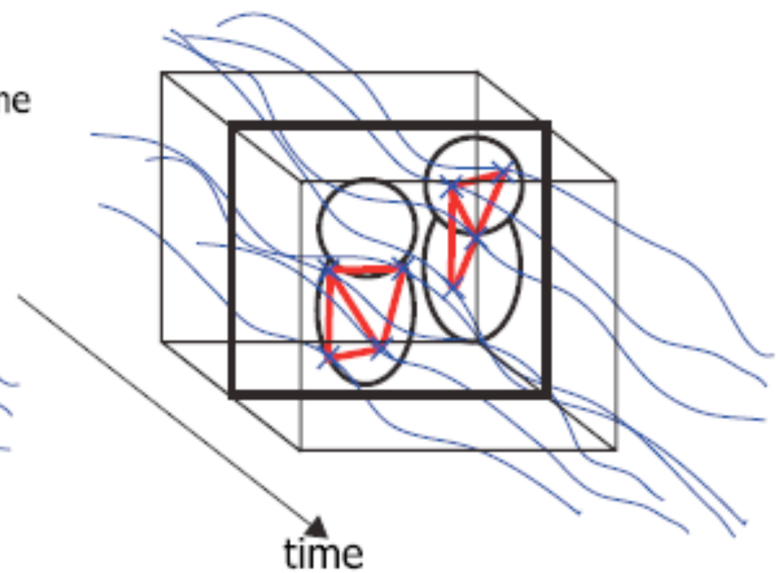
Motion Segmentation

Note: Sugimara et.al. add a feature based on gait periodicity to help disambiguate nearby people.





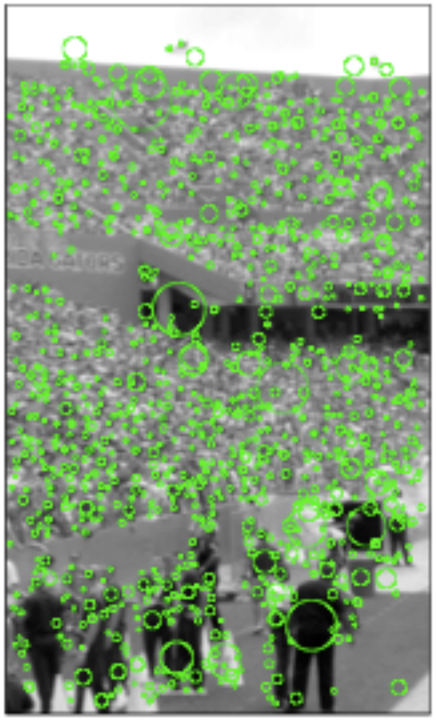
connectivity graph

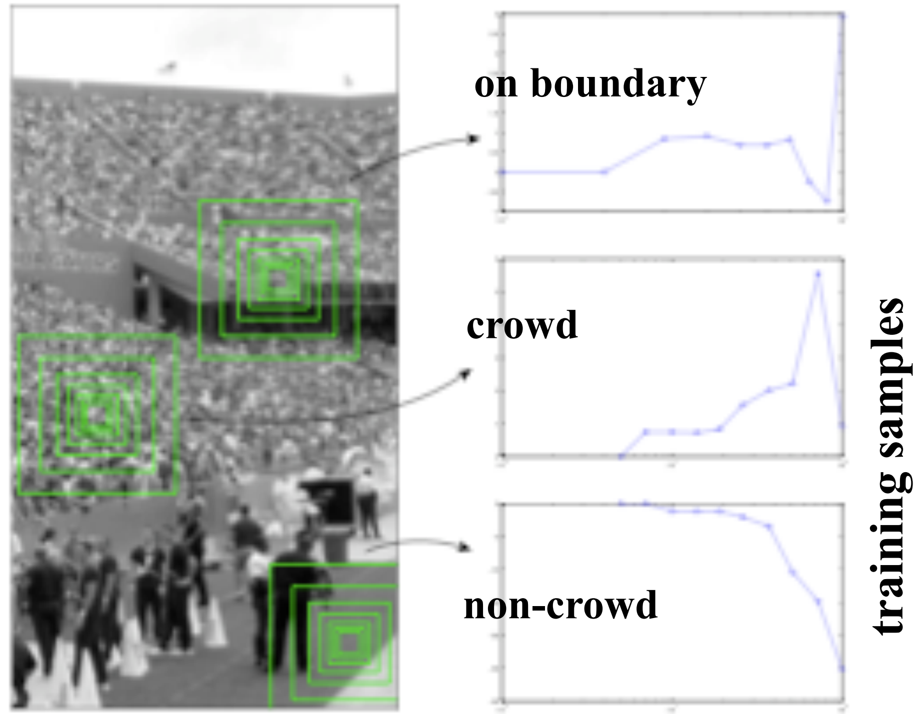


connected components after pruning

Texture-based Crowd Detection

Arandjelovic, “Crowd Detection from Still Images,” BMVC 2008



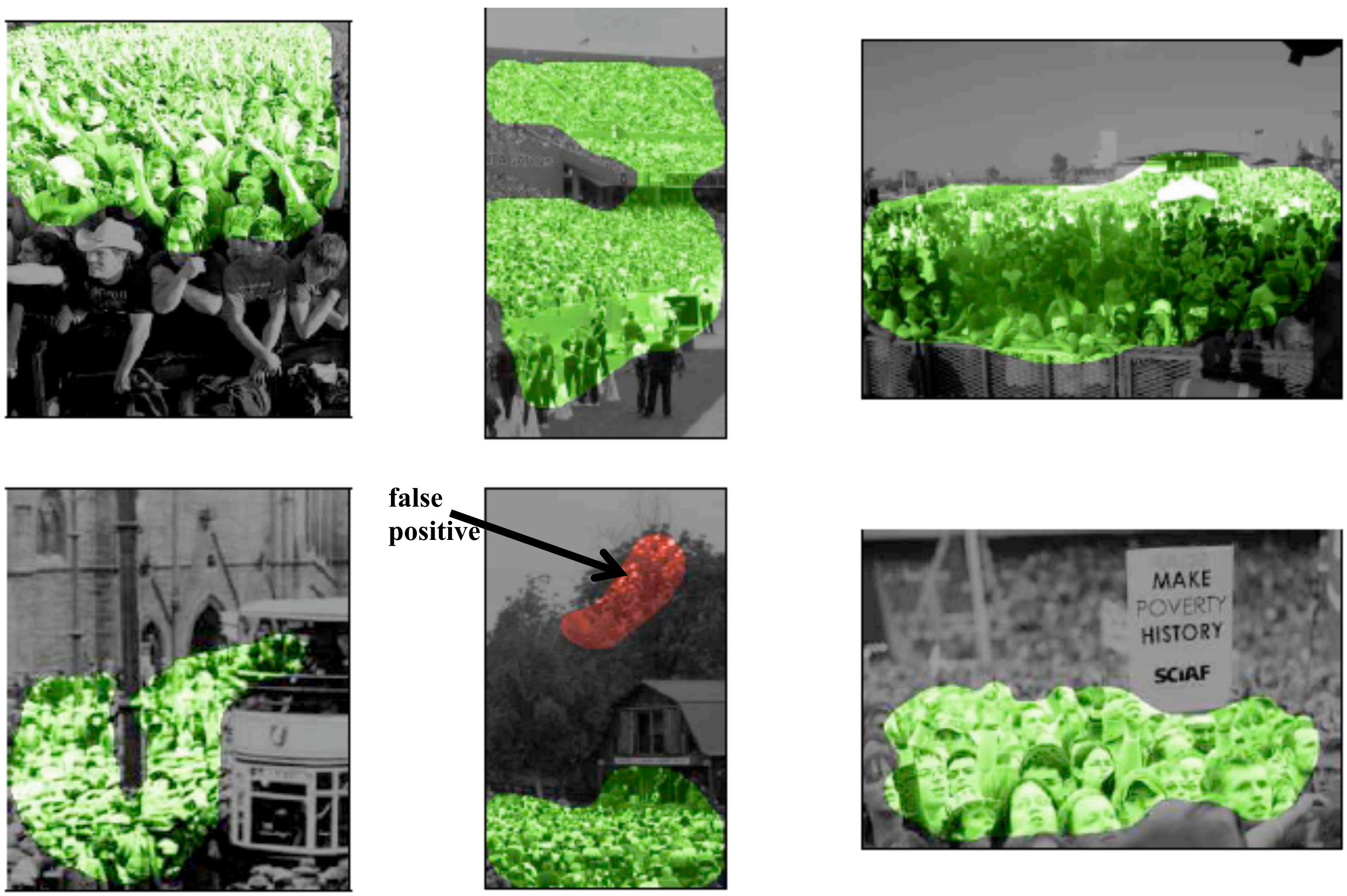


- SIFT descriptors
- K-means clustering to form “SIFT-Words”

- Likelihood ratio of distributions of word counts over 10 patch sizes yields 10-D feature vector
- Radial basis SVM for classification into crowd / non-crowd

Texture-based Crowd Detection

Sparse classifications turned into dense segmentation using graph cuts. Unary costs based on SVM output and pairwise costs based on magnitude of patch likelihood scores (small magnitudes indicate interclass boundaries).



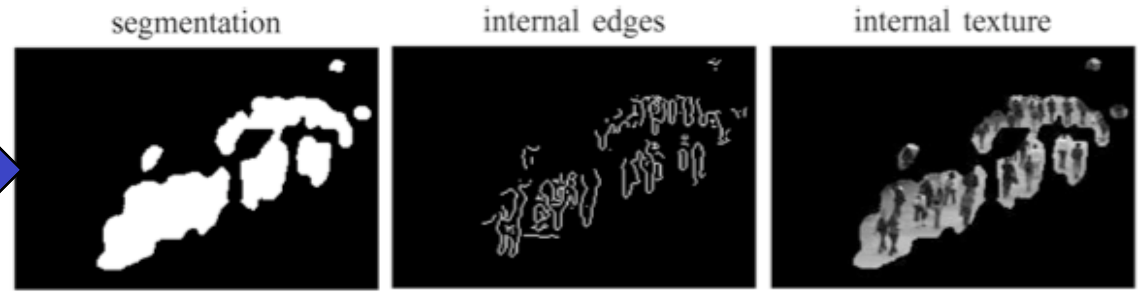
Robert Collins
Penn State

Texture-based Counting

Chan and Vasconcelos, "Counting People with Low-level Features and Bayesian Regression",
IEEE Transactions on Image Processing, **Vol 21 (4), 2160-2177, April 2012**

short video clip

segment



motion segmentation
using dynamic textures

Extract feature vector for each frame:

- **region features**
  e.g. area, perimeter, num connected components...
- **internal edge features**
  e.g. num edges, histogram of orientations
- **grey-level texture features**
  e.g. homogeneity, energy, entropy

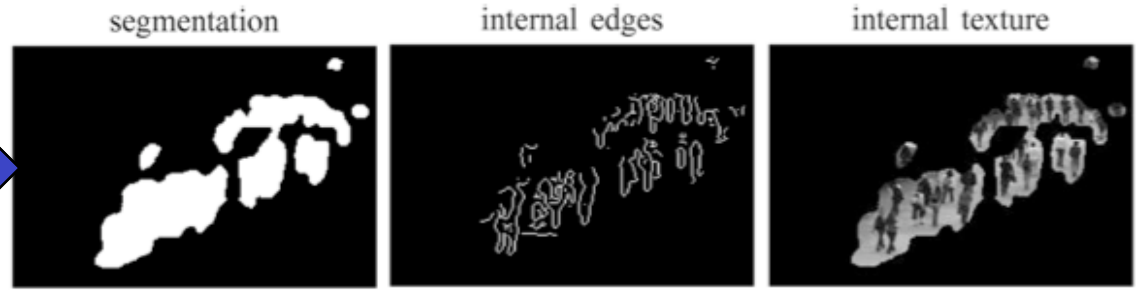
Robert Collins
Penn State

Texture-based Counting

Chan and Vasconcelos, “Counting People with Low-level Features and Bayesian Regression”, *IEEE Transactions on Image Processing,* Vol 21 (4), 2160-2177, April 2012

short video clip

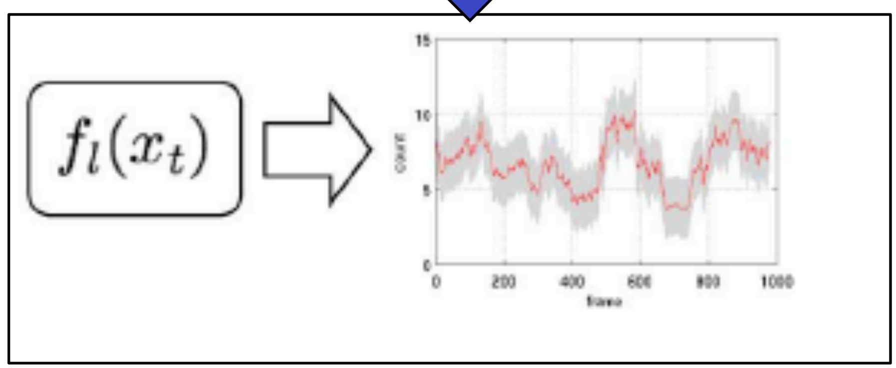
segment



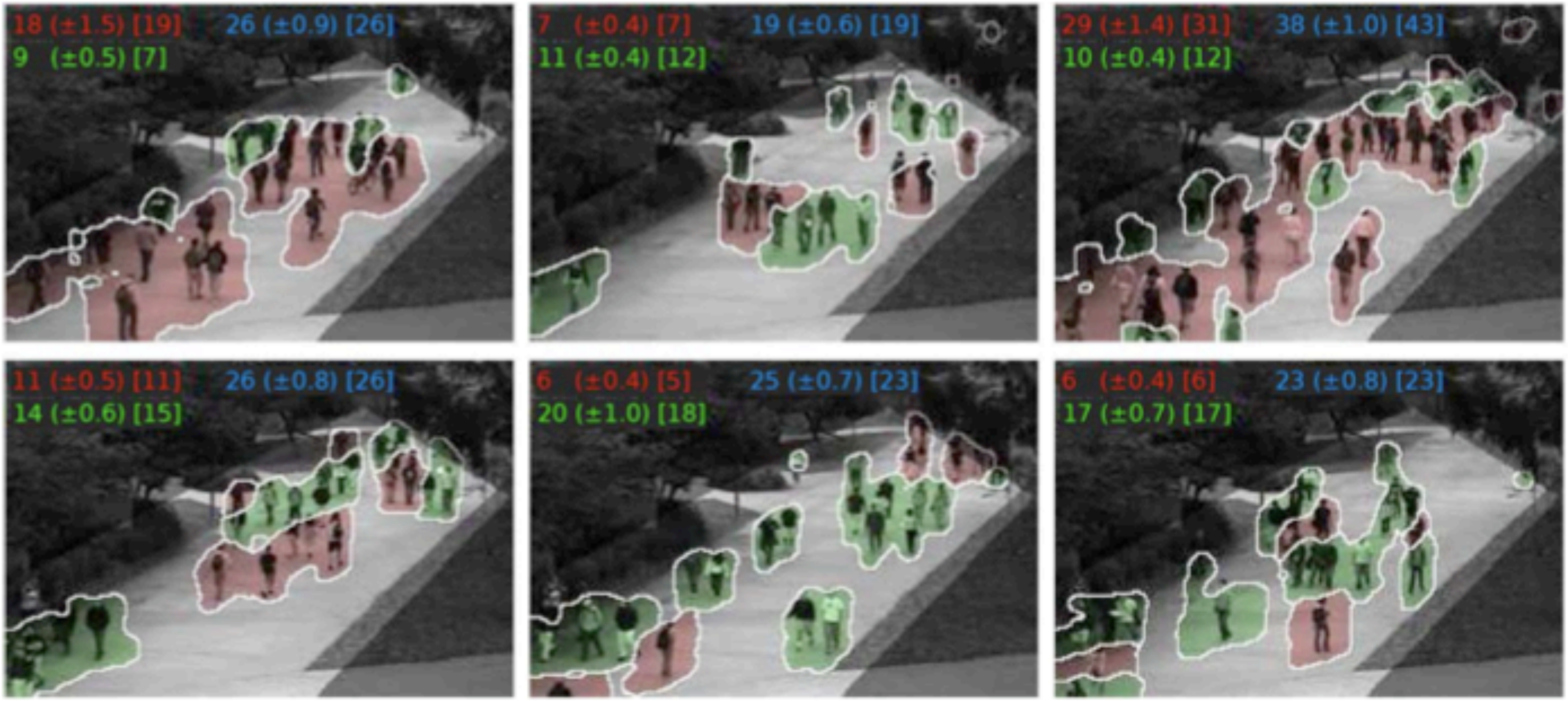
Extract feature vector for each frame:

estimate

estimate counts using
learned regression function



Texture-based Crowd Detection



green/red = crowd walking towards/away   blue = total
numeric results formatted as: estimated count (uncertainty) [true count]

Robert Collins
Penn State

Texture-based Crowd Detection

green/red = crowd walking towards/away
numeric results formatted as: estimated count (uncertainty)

Tracking in Dense Crowds

Goal: Track targets in high-density crowd scenes.

Challenges: lots of occlusion; small object sizes; appearances are similar

Idea: Model typical crowd behavior to provide better motion priors.

Point of View: Macro vs Micro

- Macroscopic level: modeling dynamic behavior of the whole crowd; holistic
  - density, flow, mean speed of a traffic stream
  - analogy to fluid streams; particle flow
  - behavior is reactive, a function of environment and density

  Crowd Flow

- Microscopic level: models decision makers, their goals, and interactions; individualistic
  - intelligent agents make decisions based on goals and social rules
  - simulating realistic interactions

  Social Force Models

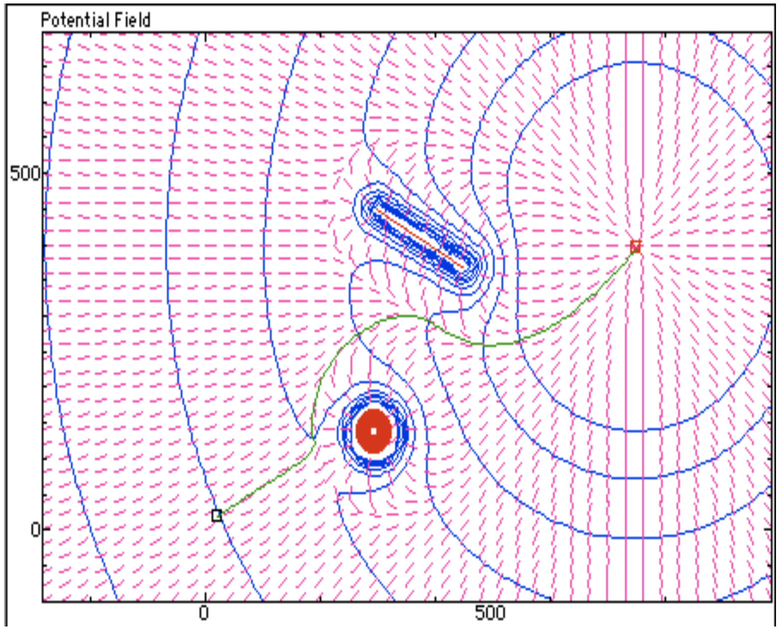
Crowd Flow: Floor Fields

Saad Ali and Mubarak Shah, Floor Fields for Tracking in High Density Crowd Scenes, The 10th European Conference on Computer Vision (ECCV), 2008.

Inspired by particle flow evacuation models.

Represents how global scene structure affects local pedestrian motion decisions.

Long-range goals/influences transformed into local forces (similar to potential fields for robotic path planning).



potential field

Floor Fields

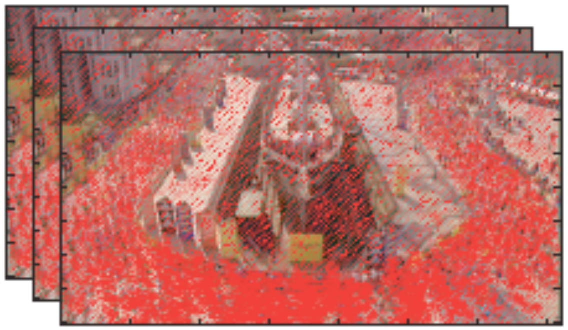
- **Static Floor Field (SFF)**
  attraction field; represents typical crowd motion towards interesting locations, dominant paths, exits

- **Boundary Floor Field (BFF)**
  repulsive forces; boundaries, walls, obstacles

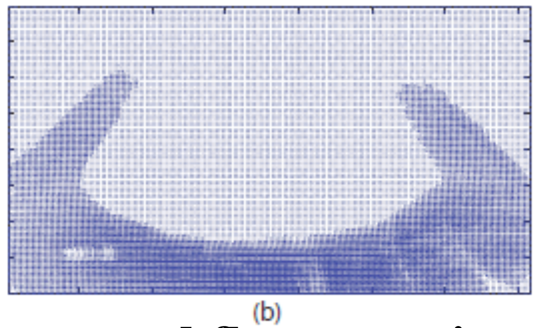
- **Dynamic Floor Field (DFF)**
  current motion of neighboring individuals computed in temporal sliding window

Static Floor Field

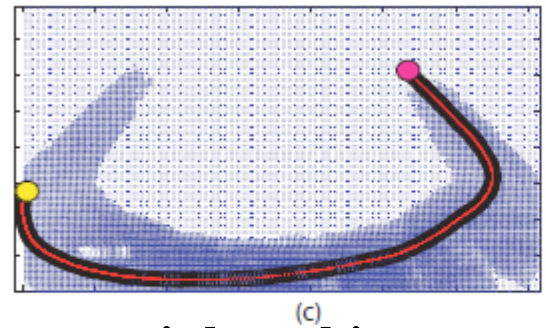
example: marathon runners turning a corner



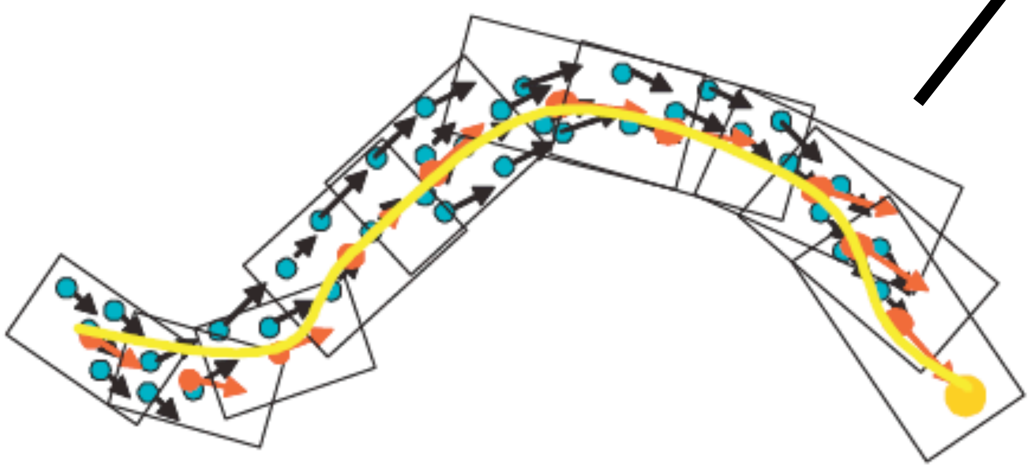
optic flow



averaged flow over time



sink-seeking

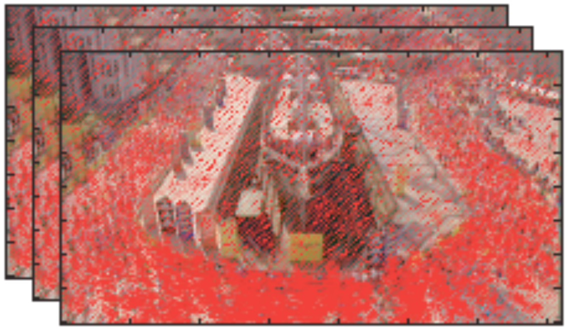


mean-shift-like procedure to determine particle flow (path, distance) to nearest goal location.

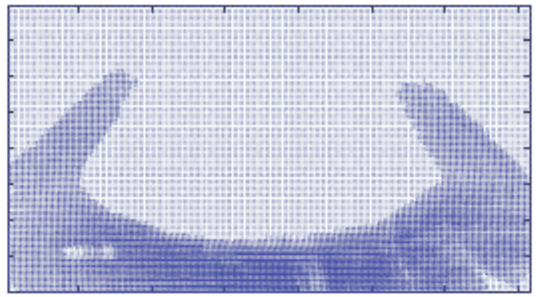
Robert Collins
Penn State

Static Floor Field

example: marathon runners turning a corner

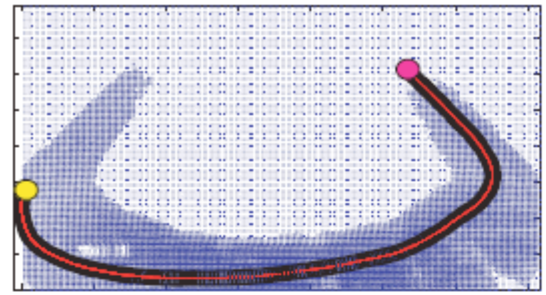


optic flow



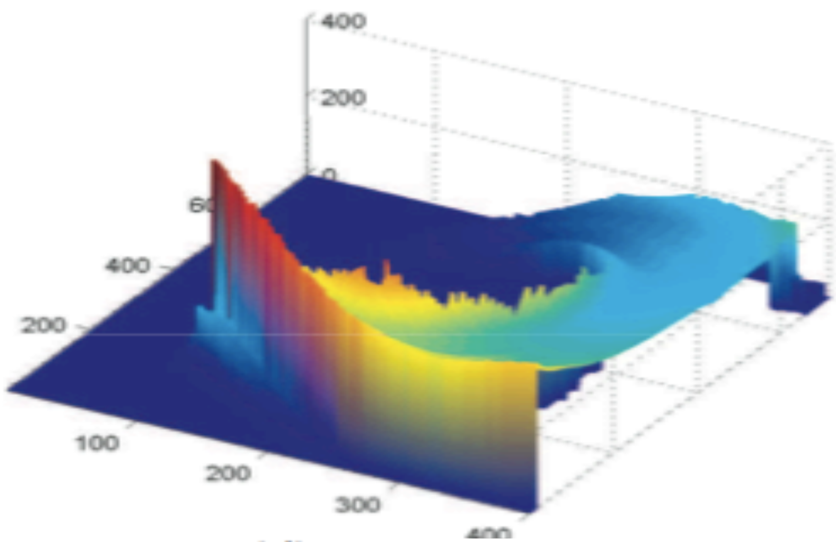
(b)

averaged flow over time



(c)

sink-seeking

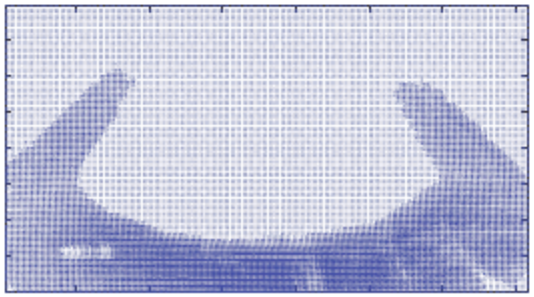


(d)

SFF = path length surface. Low values are "better". Intuition: drop a ball on surface and it rolls towards nearest sink.

Robert Collins
Penn State

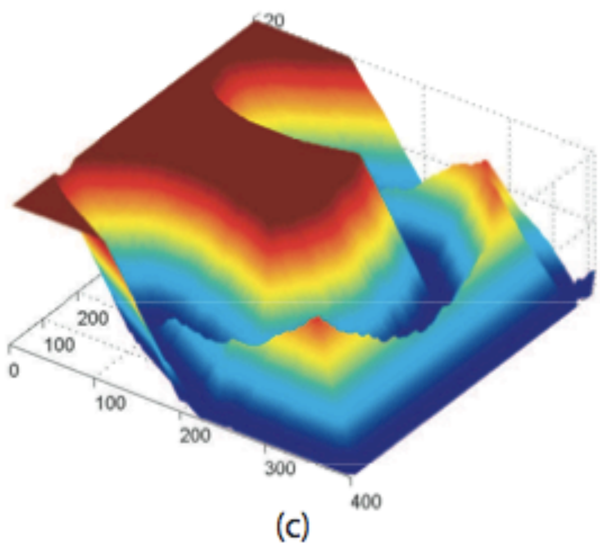
Boundary Floor Field



averaged flow

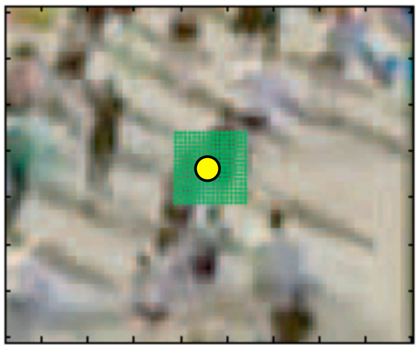
segmented flow

edge map
(real+virtual boundaries)

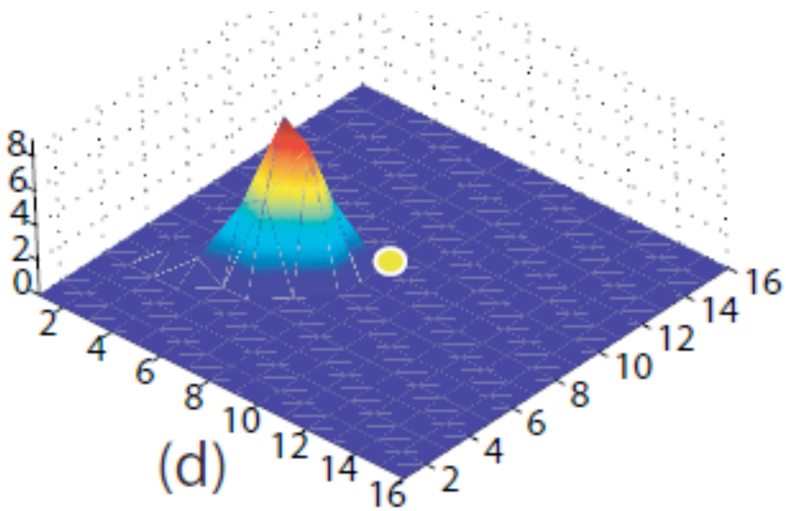


BFF = truncated distance transform.
High values are "better".
Intuition: go/no-go surface with deep valleys forming the barriers.

Dynamic Floor Field



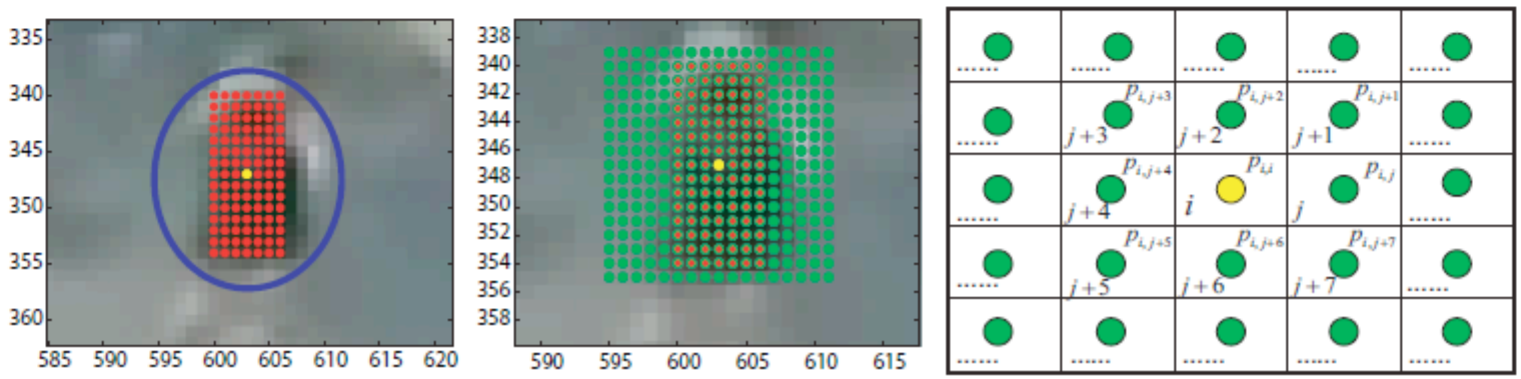
local neighborhood around target location (yellow dot)



DFF = current local motion likelihood computed from flow in a narrow temporal window.

Intuition: this is how nearby particles are currently moving.

How Floor Fields are Used



(a) (b) (c)

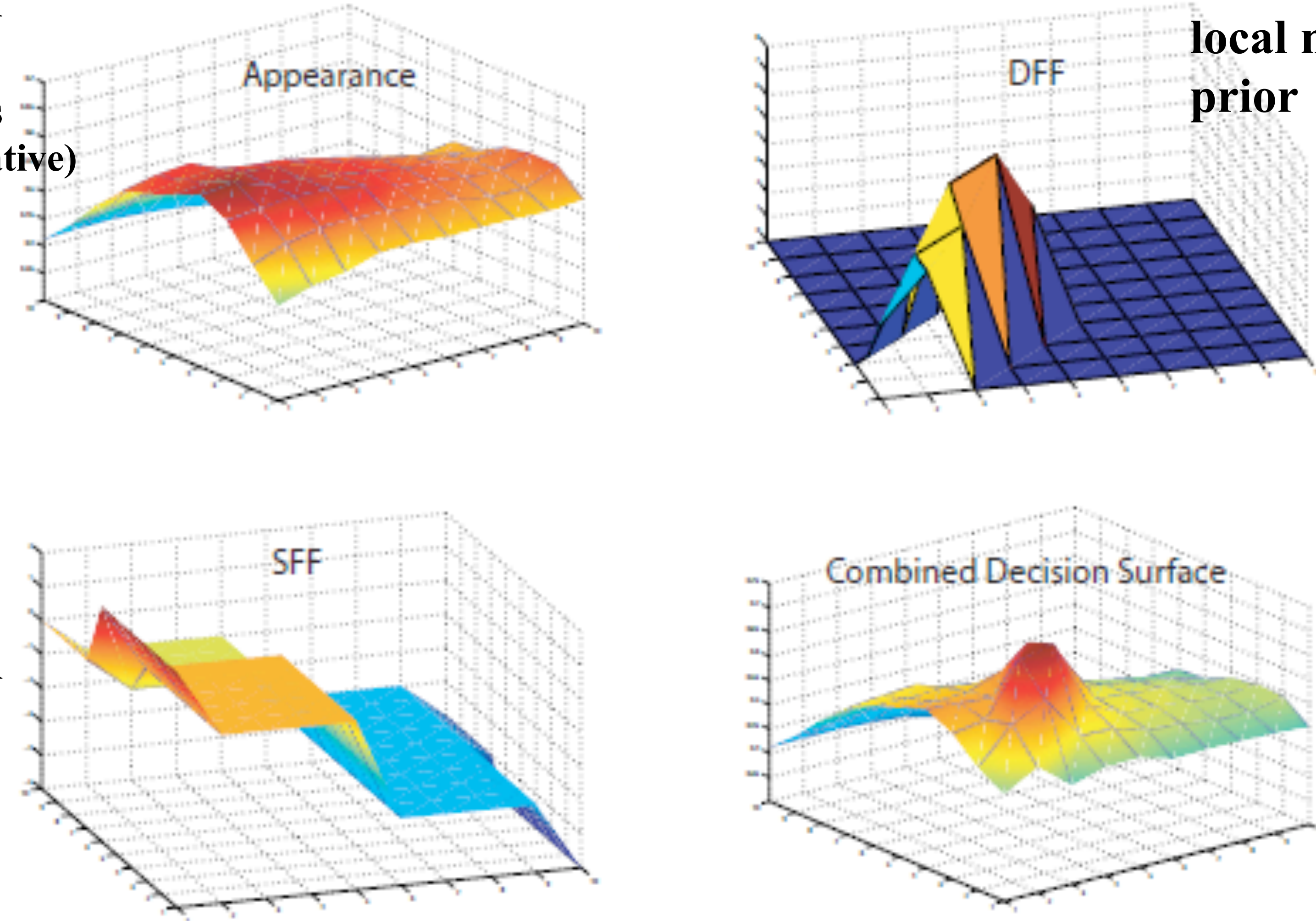
For current target location, compute matrix of local transition probabilities combining appearance and floor field terms.

$$p_{ij} = C \underbrace{e^{k_D D_{ij}} e^{k_S S_{ij}} e^{k_B B_{ij}}}_{\text{SFF/BFF/DFF influence terms (priors)}} \underbrace{R_{ij}}_{\text{appearance term (likelihood)}}$$

How Floor Fields are Used

multimodal likelihood
(appearance is not discriminative)

local motion prior

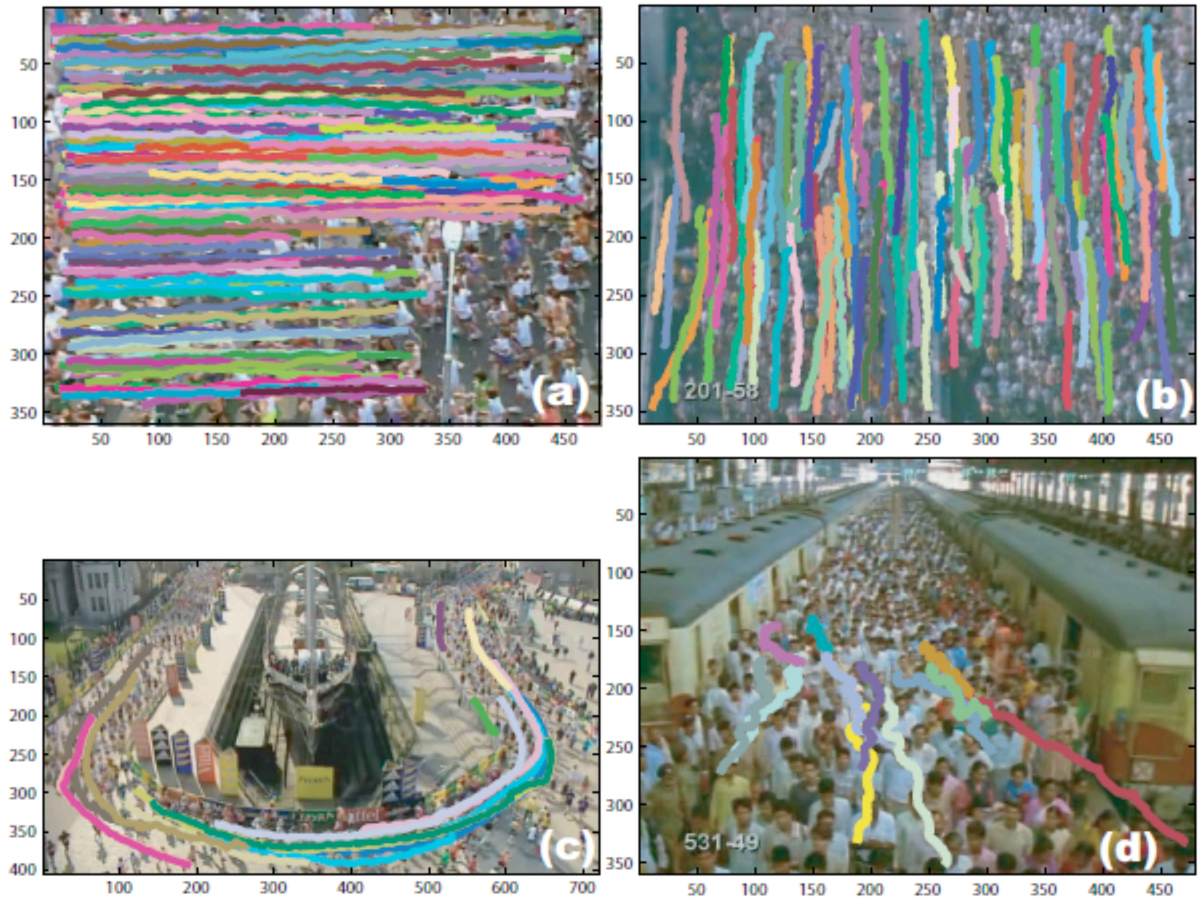


scene goal prior

much more reliable (unimodal) posterior

Robert Collins
Penn State

Tracking Examples



Tracking Examples

Floor Field Drawbacks

• SFF can’t represent multimodal goals / motion at single point in the scene

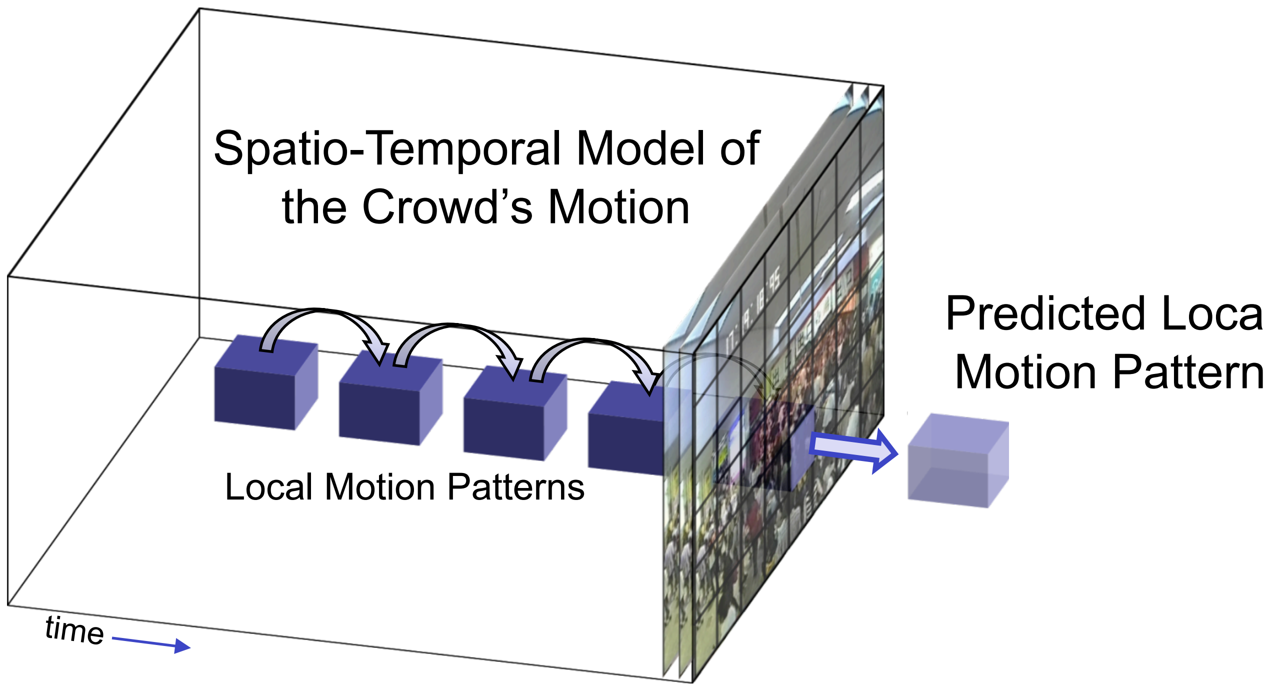
• DFF allows some local temporal adaptation, but only correct when target moves similar to neighbors

• Hard to track outlier behaviors (moving against traffic)

HMM-based Flow Model

Kratz and Nishino, Tracking with Local Spatio-Temporal Motion Patterns in Extremely Crowded Scenes, IEEE Trans Pattern Analysis and Machine Intelligence, 2012.

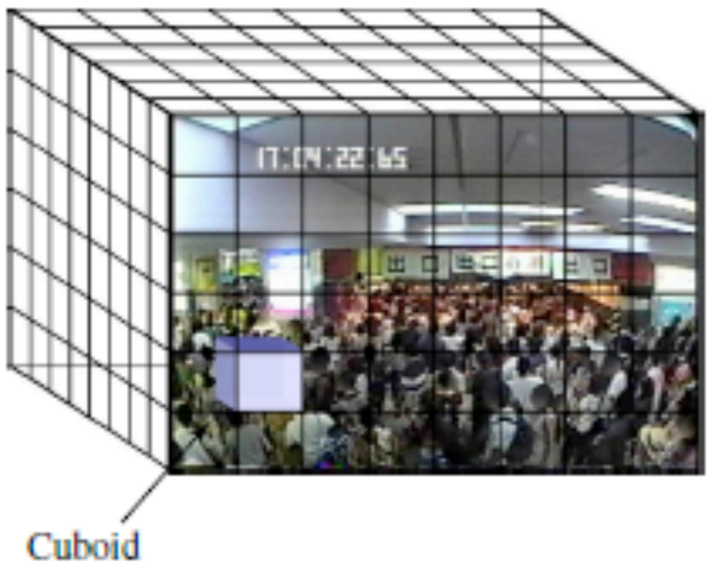
Intuition: model multi-modal, time-varying flow by training an HMM at each scene location.

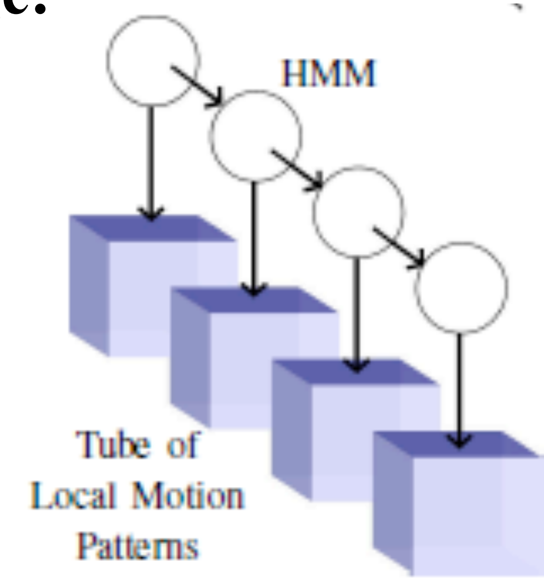


Robert Collins
Penn State

HMM-based Flow Model

Training stage:



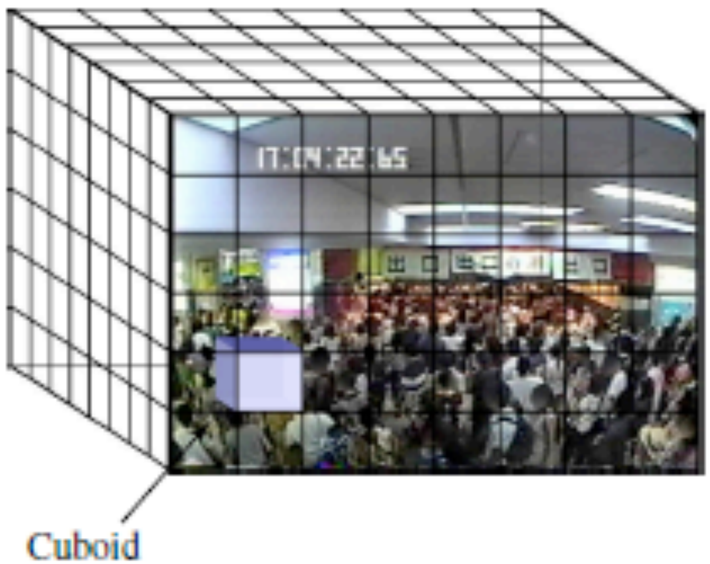


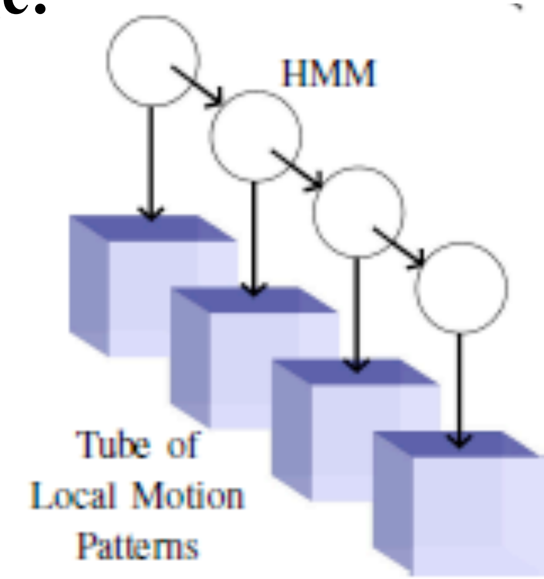
- dice training video into space-time cuboids
- estimate 3D Gaussian motion pattern in each cuboid (space-time gradients)

- in each time-tube of cuboids
  - discretize motion patterns by online clustering
  - train an HMM

HMM-based Flow Model

Training stage:





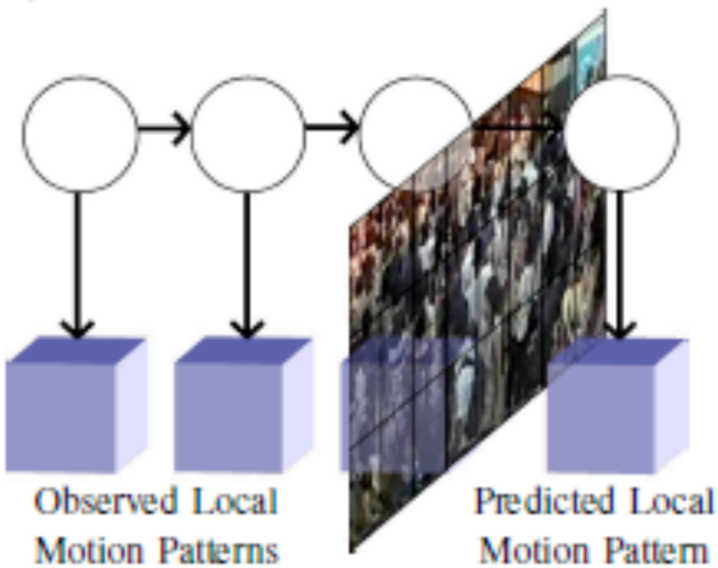
The HMMs can model time-dependencies between multiple motions at a single spatial location.

e.g. "this location has two dominant flow directions that tend to be interleaved"
"this location exhibits many rapidly-changing flow directions"
"this location has a single dominant flow"

Robert Collins
Penn State

HMM-based Flow Model

Tracking stage:



project into 2D

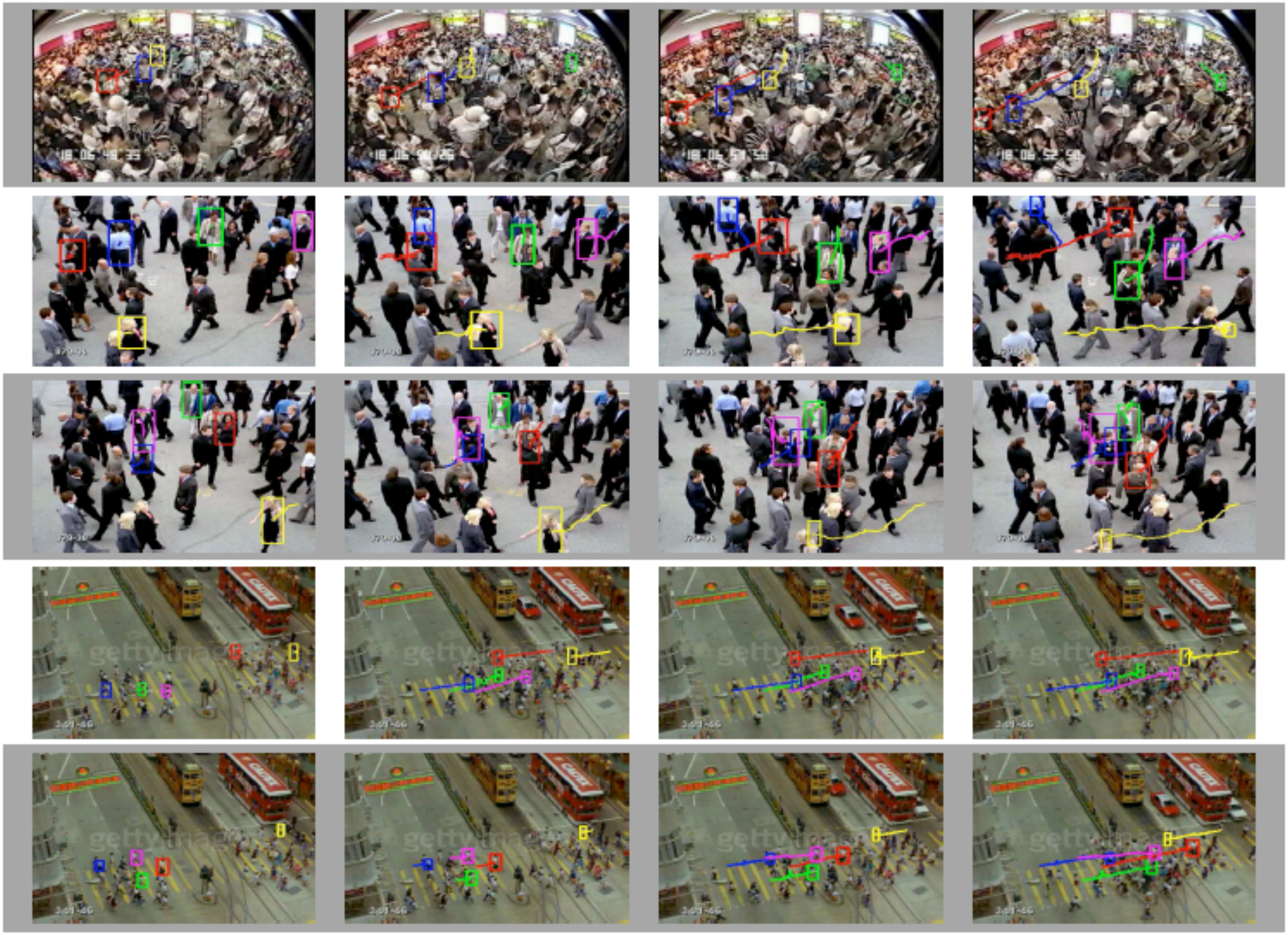
Predicted
2D Motion

• at runtime, use observed motion patterns up to time t-1 to compute expected motion at at target's center at time t.

• project this 3D motion pattern into 2D to get predicted image flow distribution

• use this distribution as a motion prior for particle filter tracking

Sample Results



play video outside ppt

Data Driven Flow Modeling

- Floor fields and HMM-based flow are scene-centric models (must be trained previously on video from the same scene viewpoint)

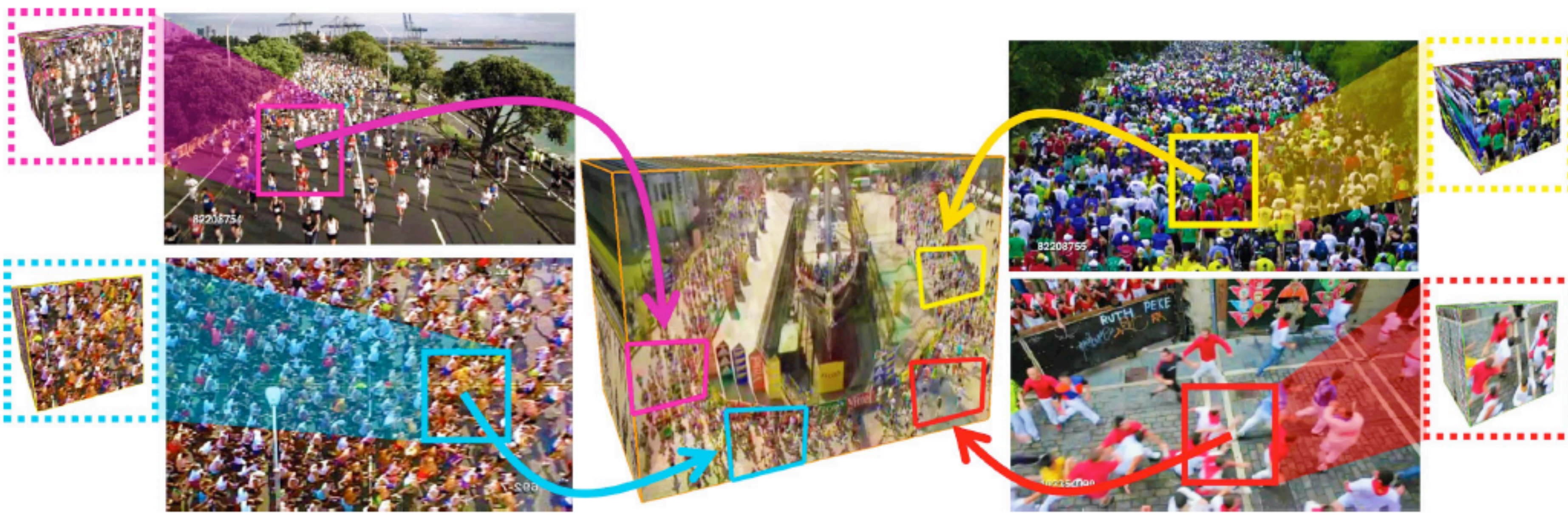
- They also have trouble tracking "rare" motions because they accumulate distributions of typical scene behavior

- Idea: try non-parametric data-driven approaches that have been very successful in texture synthesis and inpainting.

Robert Collins
Penn State

Data-Driven Flow

Rodriguez, Sivic, Laptev, and Audibert, "Data-driven Crowd Analysis in Videos", ICCV 2011.



Insight: Any given crowd video can be viewed as a composite mixture of patches taken from a large dataset of previously viewed videos.

Two-Stage Matching

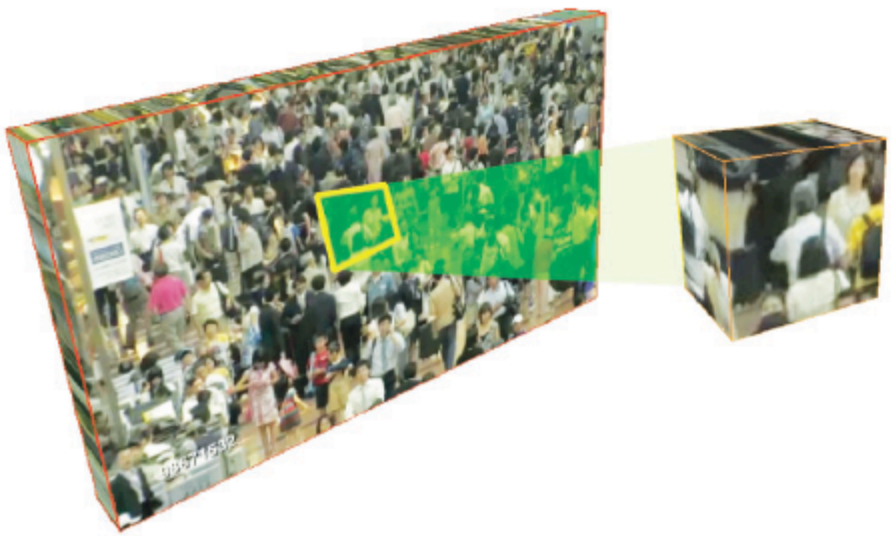
- First stage: Global matching using GIST descriptor of first frame to find videos roughly matching orientation and scale (viewpoint) of input video.

input video

matches from database

Two-Stage Matching

- Second stage: Local patch matching based on HOG3D descriptors (histograms of spatio-temporal gradients) to find patches with similar structure and motion as neighborhood around target.



spatio-temporal patch centered on target

k-nearest neighbor matches from pool of stage 1 videos

Motion Transfer

- Motion information is averaged over the matching patches and incorporated into a motion prior during Kalman filter tracking.
- This data-driven prior, using different videos, does better than averaging scene flow over the actual input sequence.

red = ground truth; green = data-driven flow, yellow = averaged scene flow

Performance on Rare Events

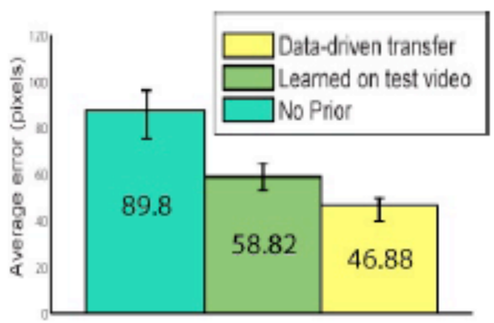


Figure 9. Comparison of average tracking errors when tracking people in rare crowd events based on 21 tracks and $k = 3$.

Social Force Model

Helbing and Molnár (1995). "Social force model for pedestrian dynamics". Physical Review E 51 (5): 4282–4286

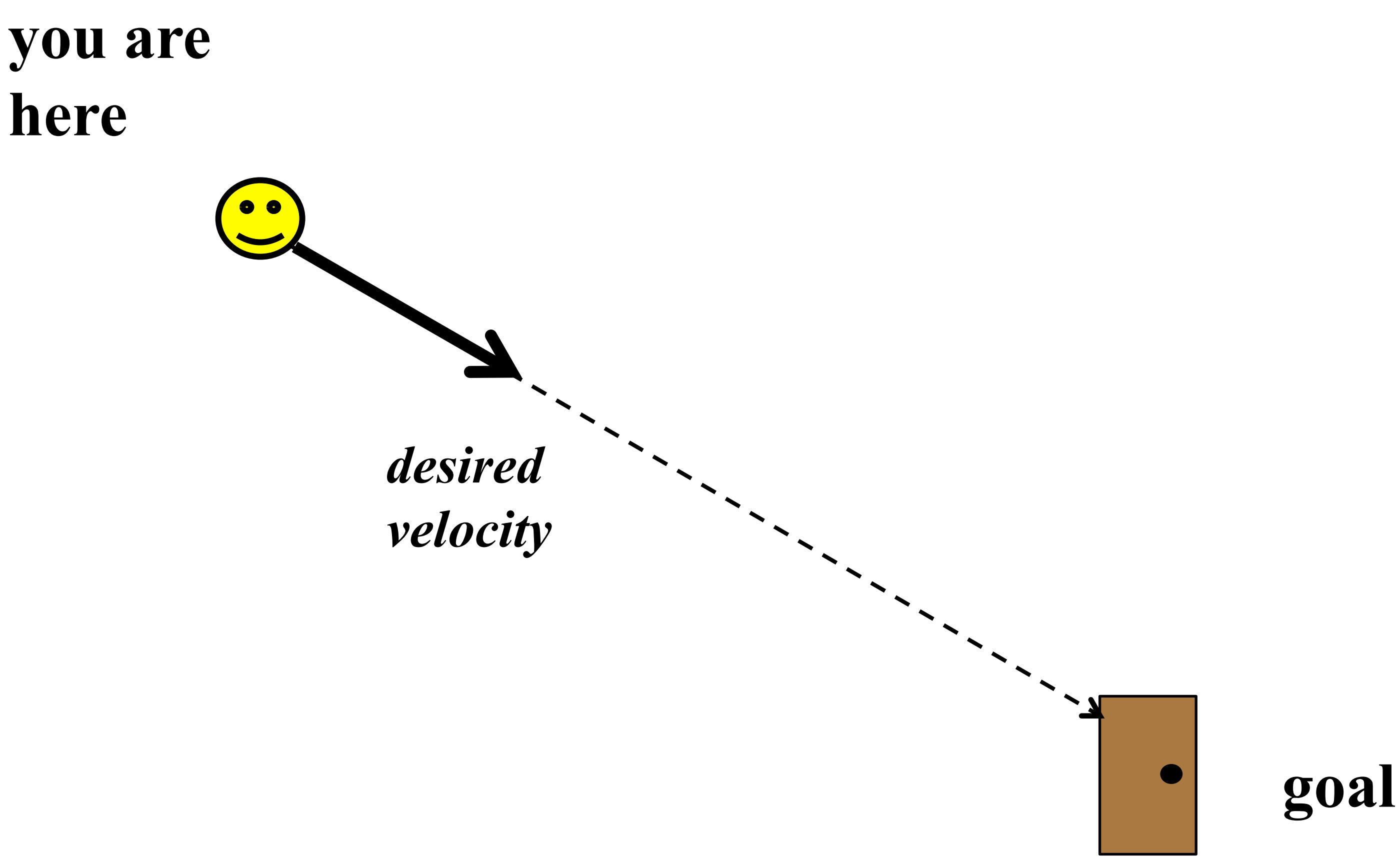
Social forces represent similar information as floor fields.

But one important distinction: working in an agent-centered point of view rather than a scene-centered one.

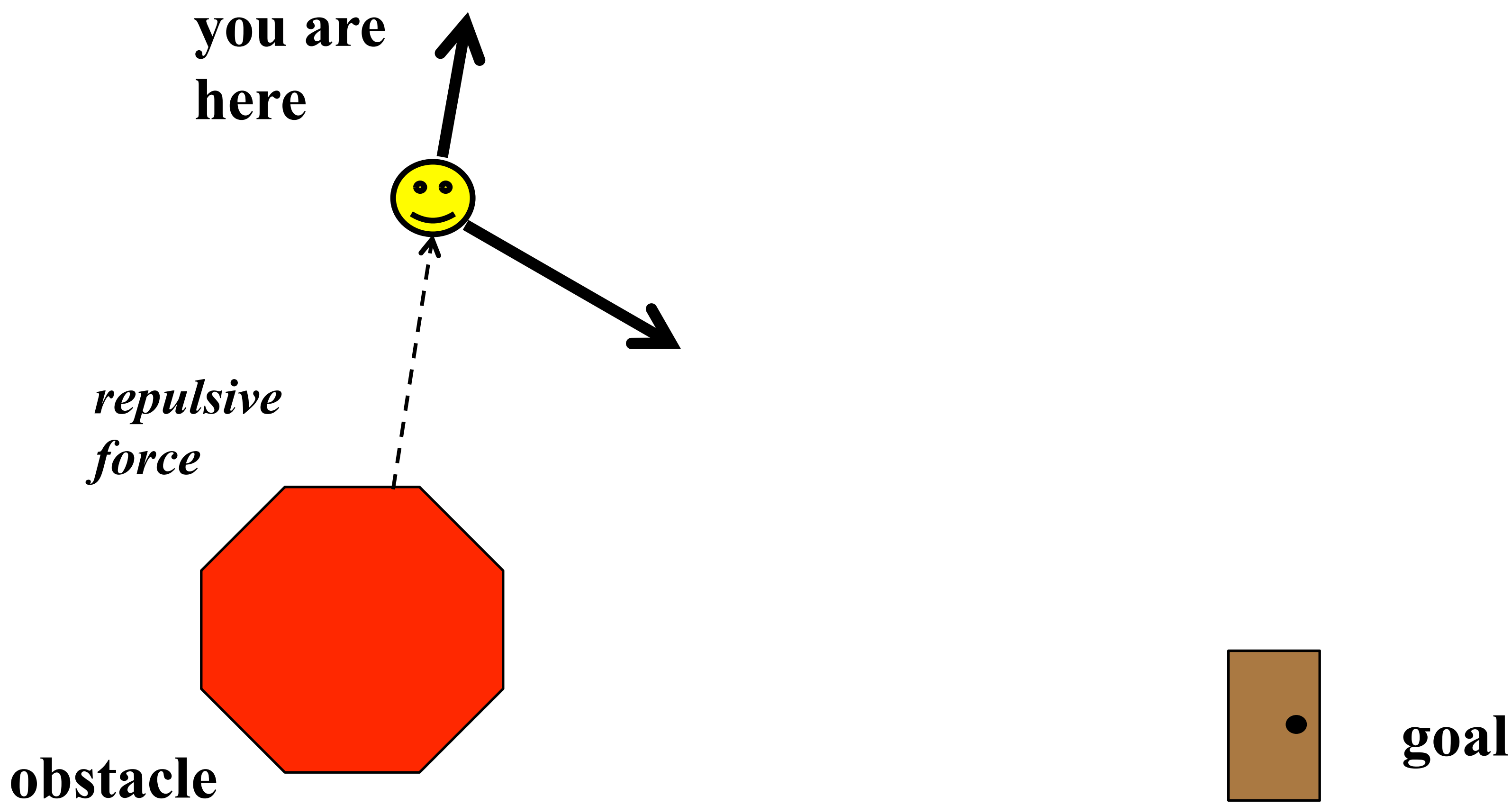
In other words, microscopic rather than macroscopic.

Robert Collins
Penn State

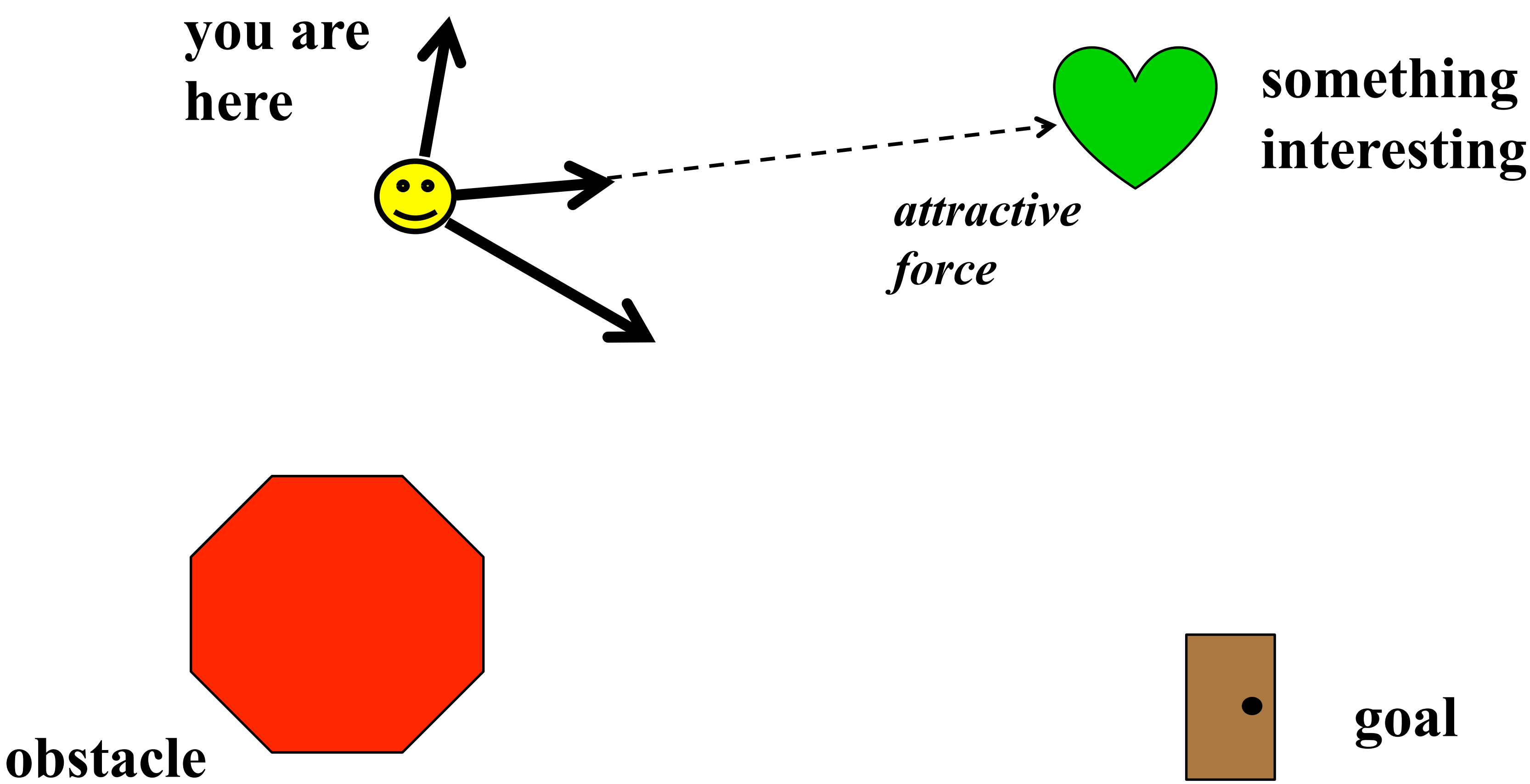
Social Force Model



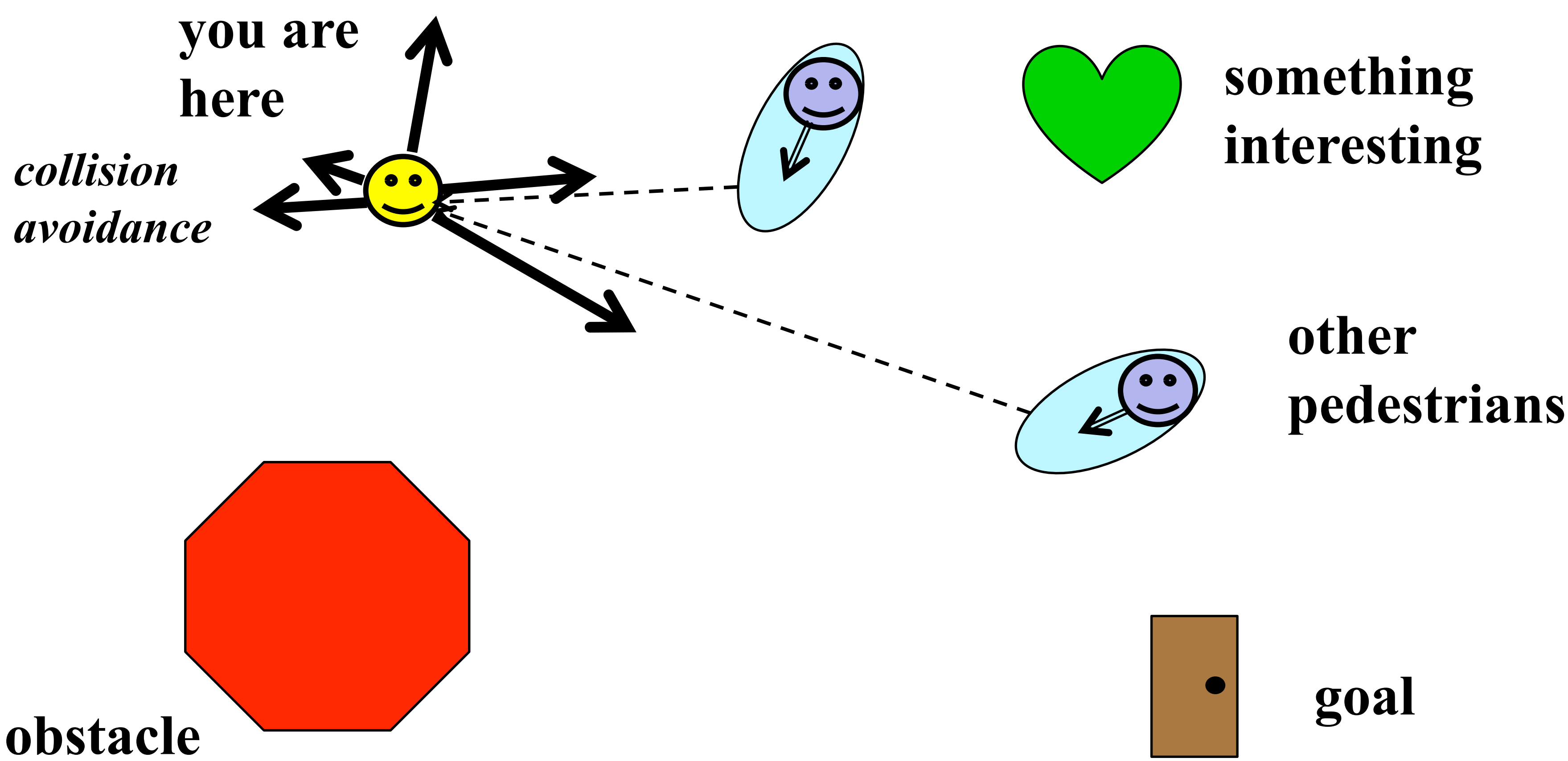
Social Force Model



Social Force Model

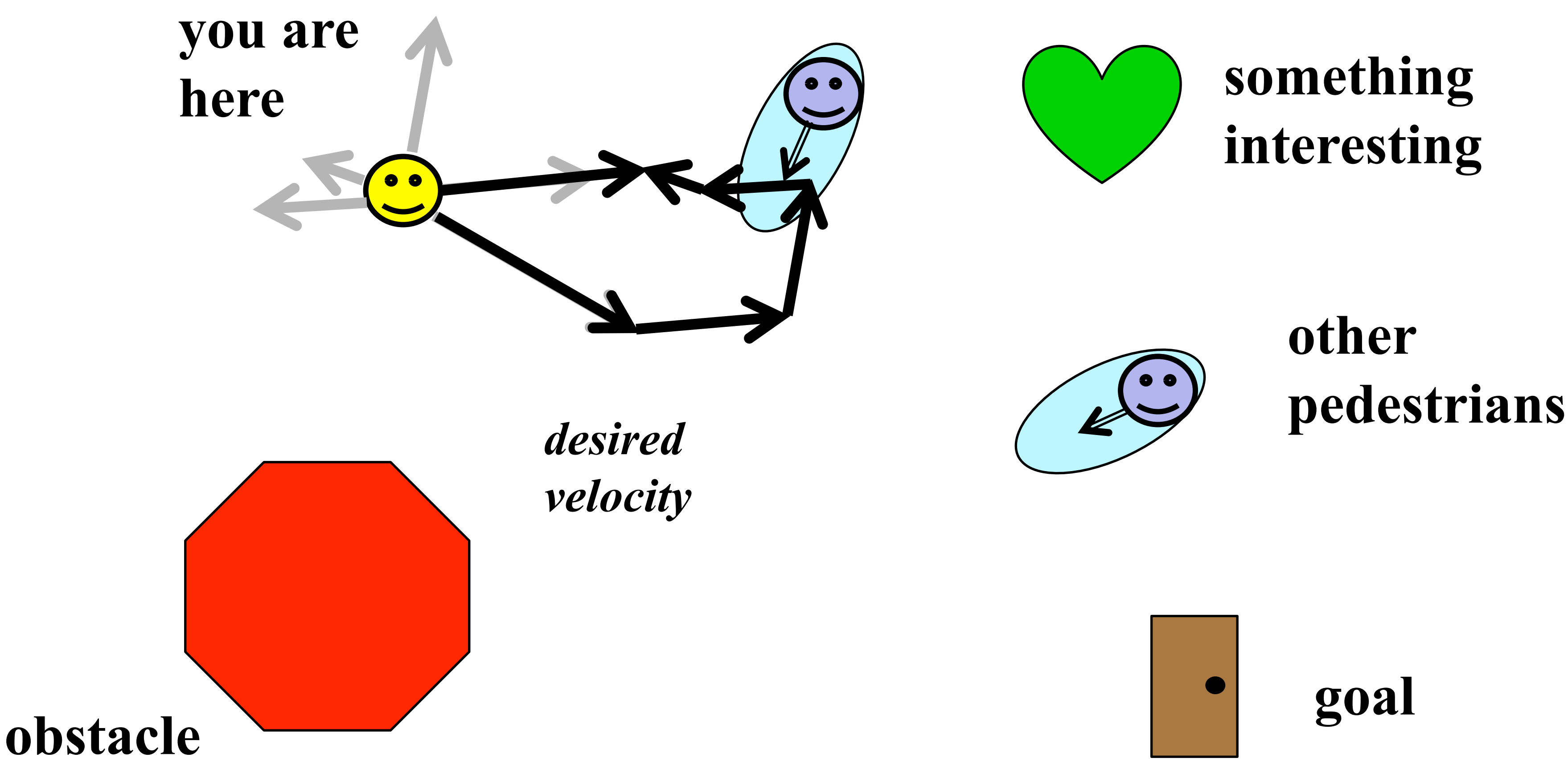


Social Force Model



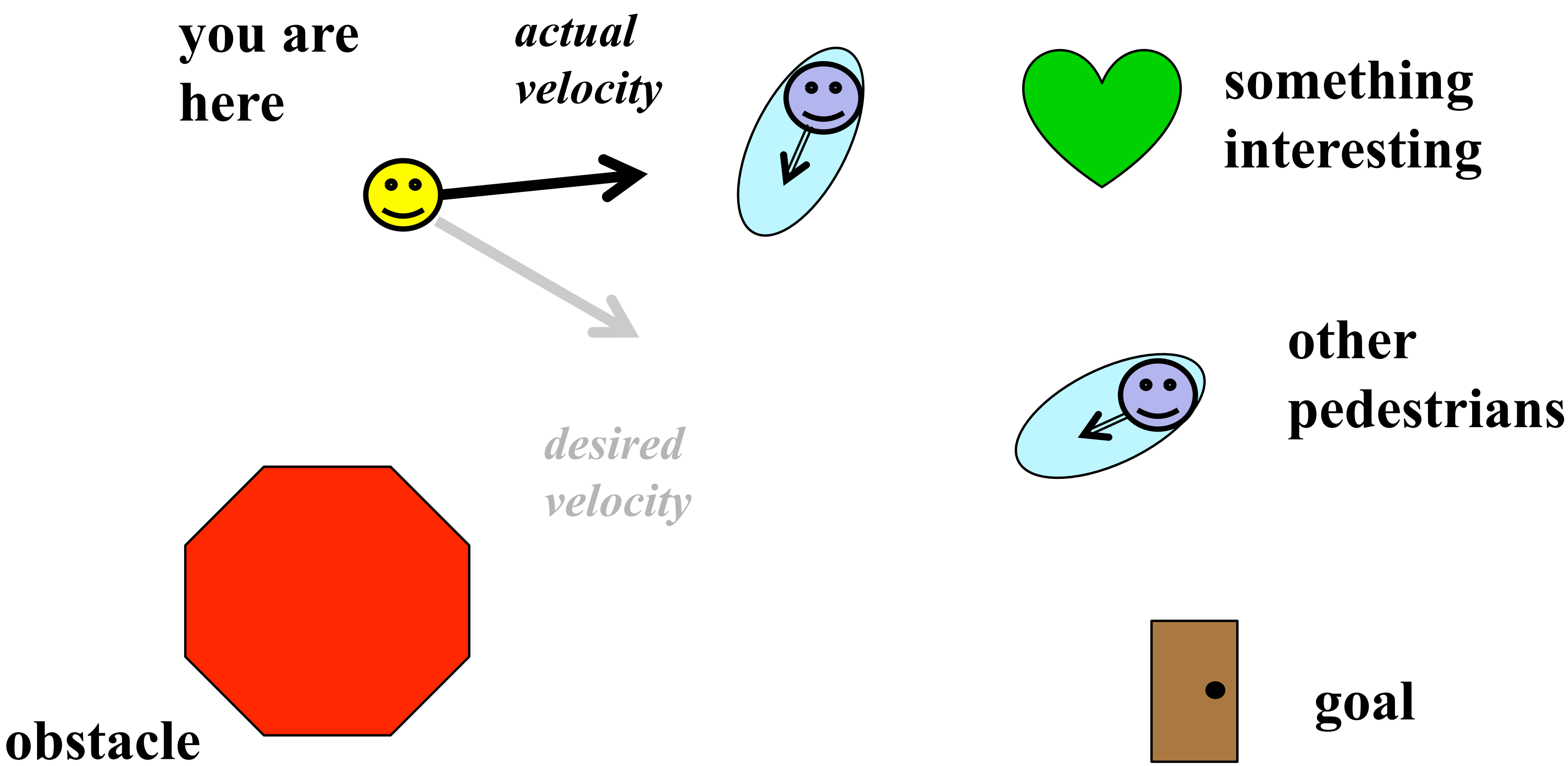
Robert Collins
Penn State

Social Force Model



Robert Collins
Penn State

Social Force Model



Case Study

Pellegrini, Ess, Schindler, van Gool *You'll Never Walk Alone*: Modeling Social Behavior for Multi-target Tracking ICCV 2009

Consider two moving pedestrians.

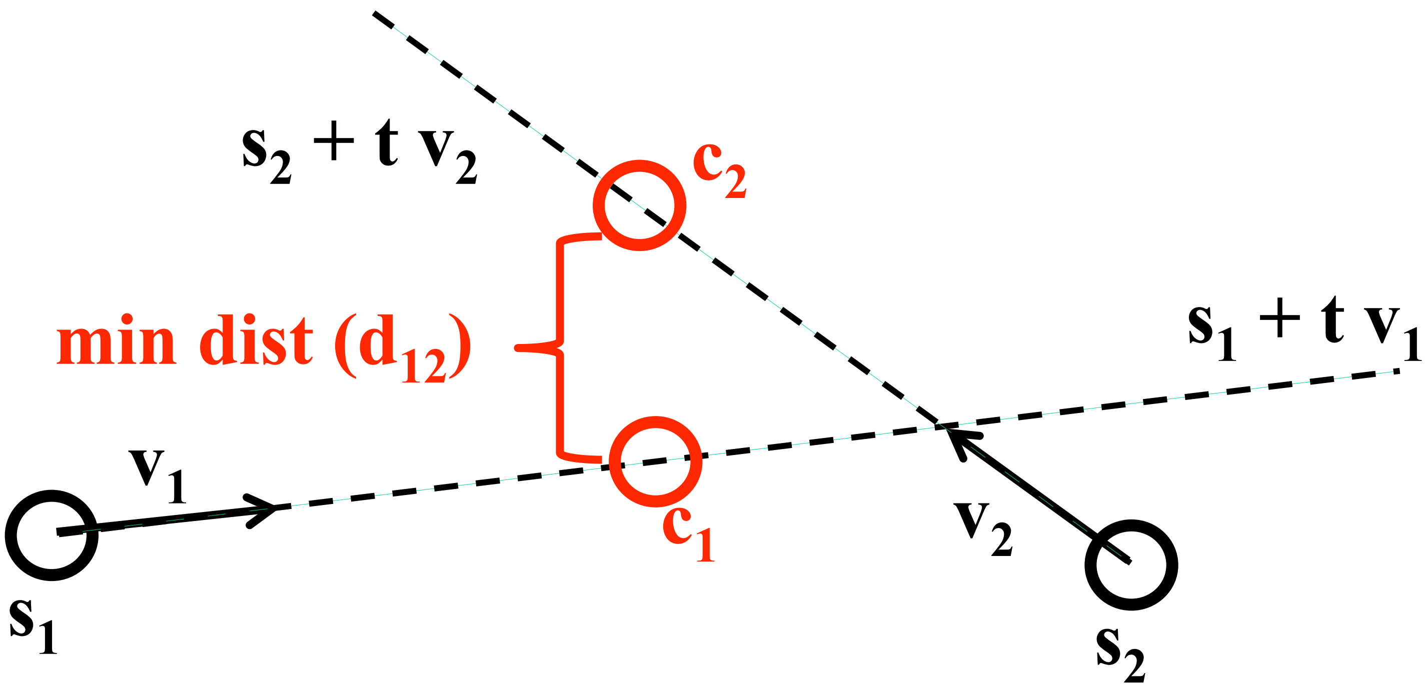
What is their point of closest approach?

(assuming they move with constant velocity)

Robert Collins
Penn State

Case Study

Pellegrini, Ess, Schindler, van Gool *You'll Never Walk Alone*: Modeling Social Behavior for Multi-target Tracking ICCV 2009



$$p_1(t) = s_1 + t\ v_1 \qquad p_2(t) = s_2 + t\ v_2$$

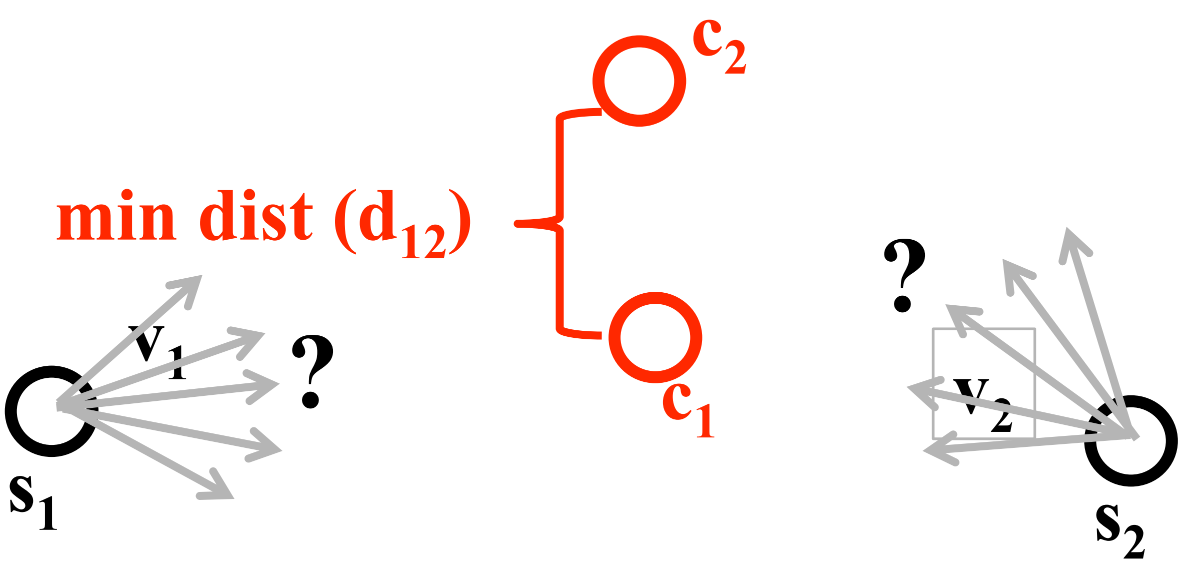
$$t^* = \mathrm{argmin}_{(t>0)} \| p_2(t) - p_1(t) \|$$

$$c_1 = s_1 + t^*\ v_1 \qquad c_2 = s_2 + t^*\ v_2$$

Robert Collins
Penn State

Case Study

Pellegrini, Ess, Schindler, van Gool *You'll Never Walk Alone*: Modeling Social Behavior for Multi-target Tracking ICCV 2009



intuition: we want to adjust v1 and v2 to keep a "comfortable" distance d_{12} between them, while maintaining roughly the original desired directions and speeds.

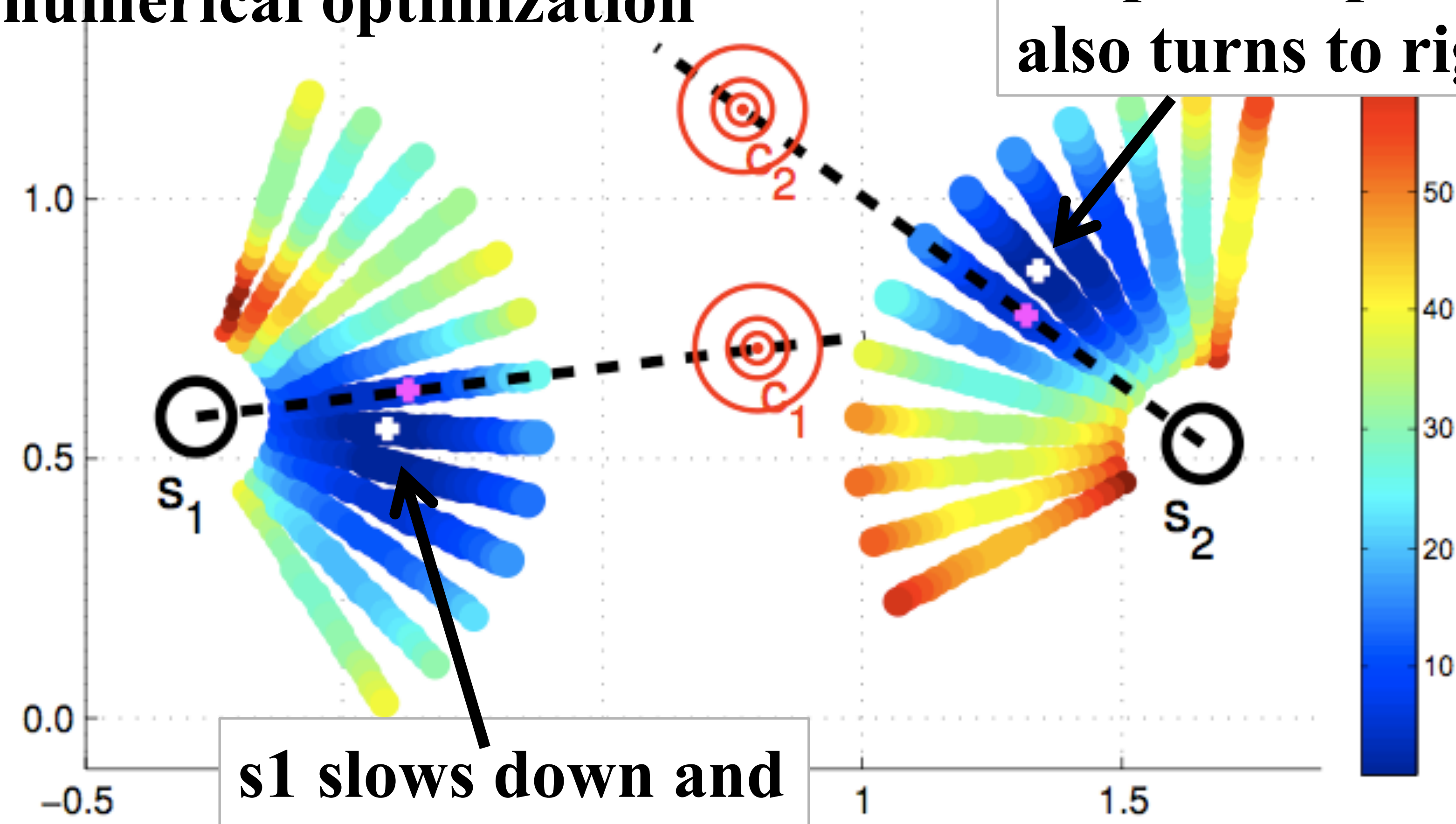
Robert Collins
Penn State

Case Study

Pellegrini, Ess, Schindler, van Gool *You'll Never Walk Alone*: Modeling Social Behavior for Multi-target Tracking ICCV 2009

the new velocities are found by numerical optimization

s2 speeds up and also turns to right

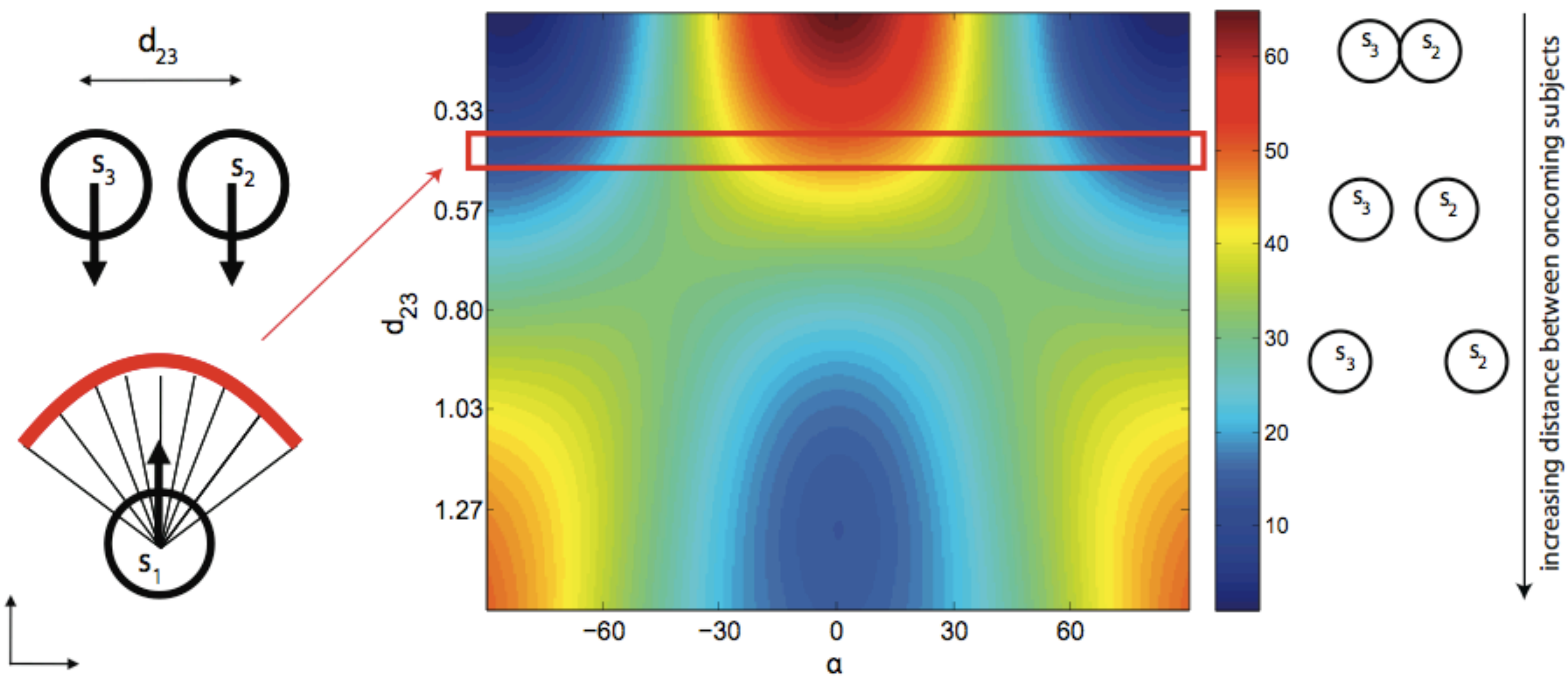


s1 slows down and turns to right

Robert Collins
Penn State

Model Yields Intuitive Behavior

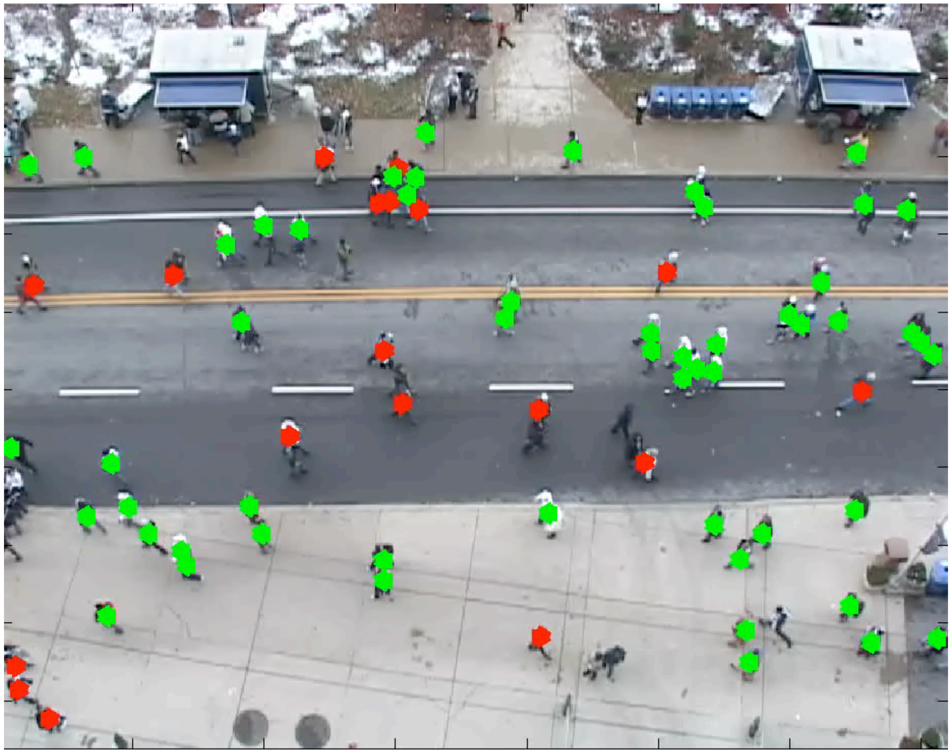
Pellegrini, Ess, Schindler, van Gool *You'll Never Walk Alone*: Modeling Social Behavior for Multi-target Tracking ICCV 2009



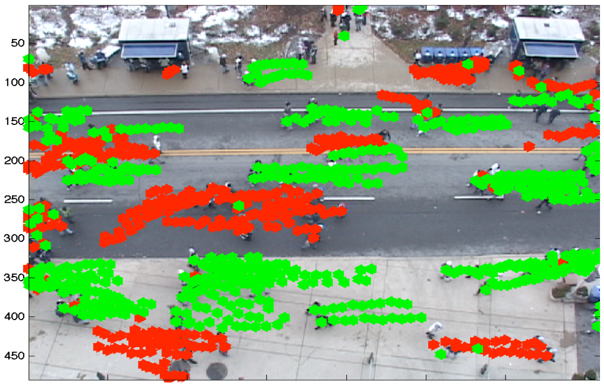
Depending on distance between s2 and s3, pedestrian s1 will either try to pass between them, or around them.

Pedestrian Fingering

Helbing's social force model also predicts "fingering" in areas of bidirectional motion. People tend to follow others to minimize collisions (maximize throughput).

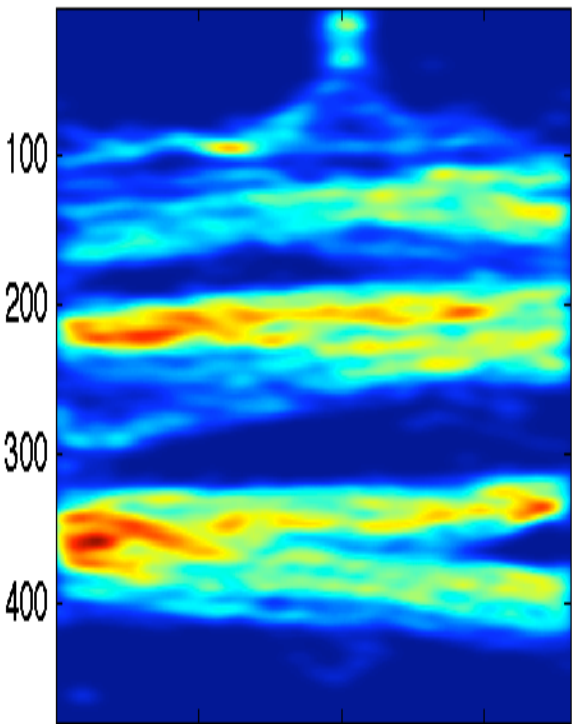


Green: leftward moving. Red: rightward moving,

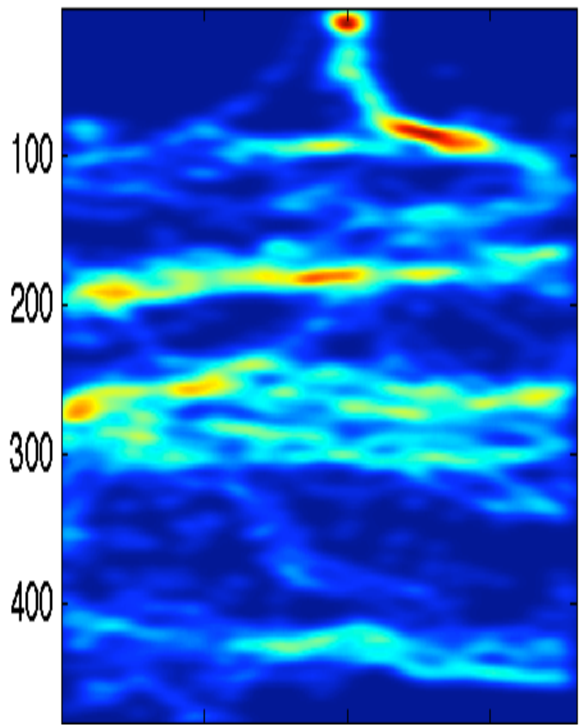


Fingering Effect

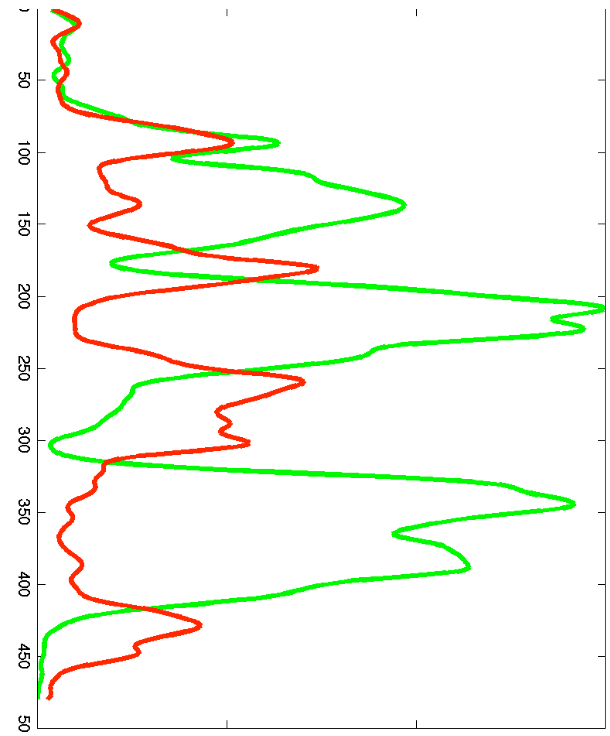
collective behavior emerges
from independent decisions



<-- Leftward



Rightward -->

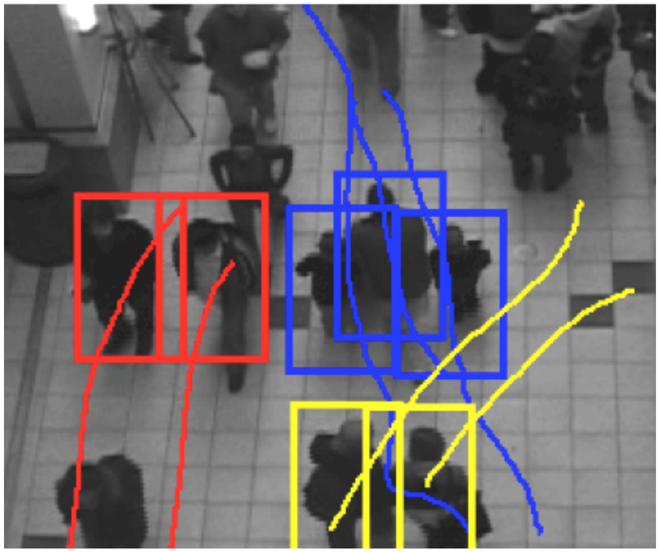


Density by image row

Green (leftward); Red (rightward)

Collective Locomotion

- **Find small groups traveling together**
  - Sociological hypothesis: validating that the majority of people in the crowd cluster in small groups
  - Public safety: improving situation awareness and emergency response during public disturbances



McPhail and Wohlstein, 1982

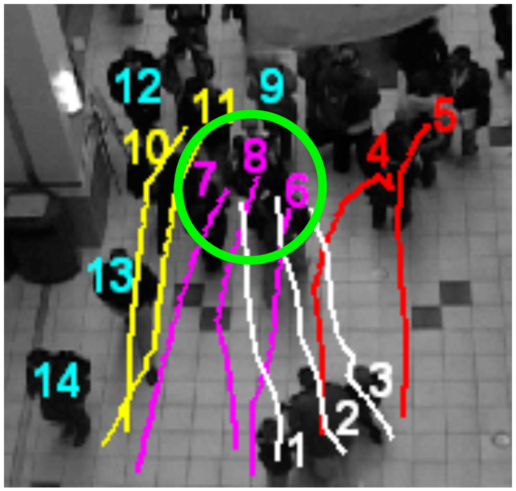
- Group membership is determined via a cascaded set of three tests:
  1. Any two people who are within 7 feet of each other and not separated by another individual are considered to be contiguous
  2. Any two contiguous people whose speeds are the same to within .5 feet per second are judged to have the same speed
  3. Any two contiguous people traveling at the same speed whose directions of motion are the same to within 3 degrees are judged to have the same direction
- Another procedure tests whether a new individual should be added to an existing group to form a larger group

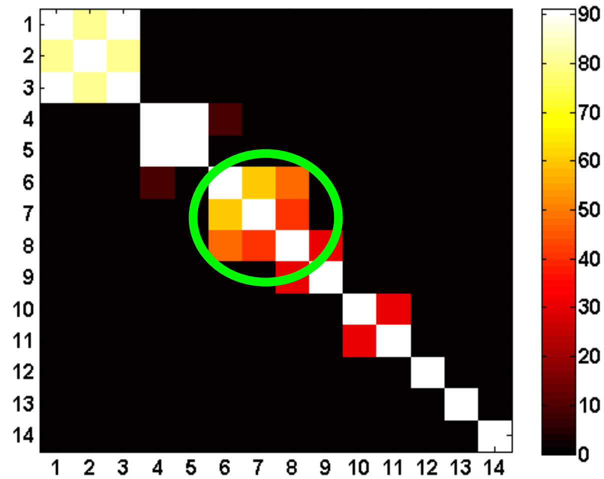
- Limitations
  - Hundreds of person-hours needed to hand code just minutes of film
  - Difficult for dense crowds/long sequences

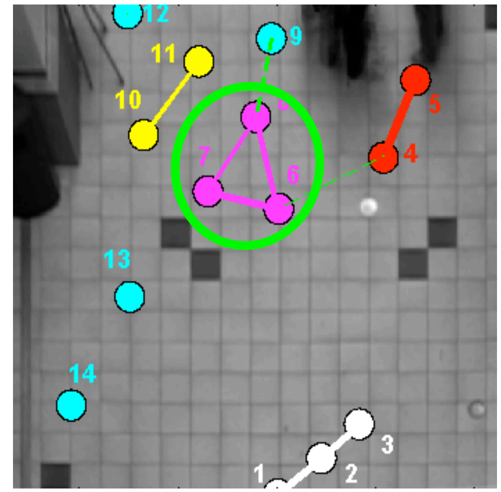
C. McPhail and R.T. Wohlstein. Using film to analyze pedestrian behavior. *Sociological Methods and Research*, 10:347–375, 1982.

Automated Group Testing by Agglomerative Clustering

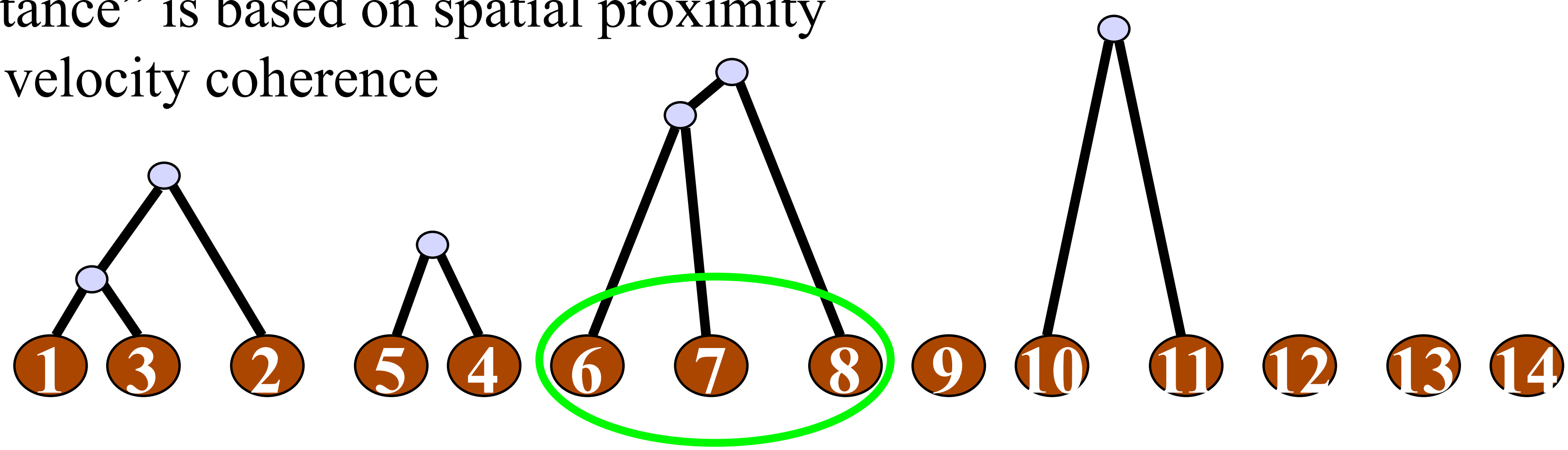
W.Ge, R.Collins and B.Ruback, "Vision-based Analysis of Small Groups in Pedestrian Crowds," *IEEE Trans Pattern Analysis and Machine Intelligence,* Vol 34(5), 2012, pp.1003-1016.





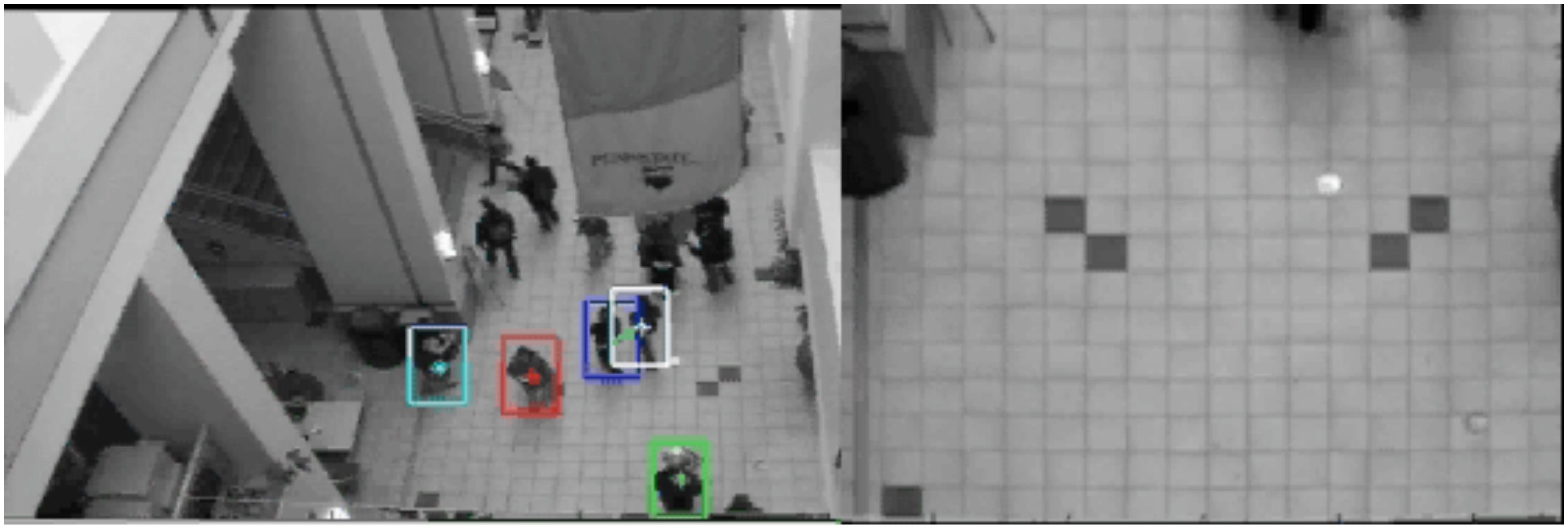


"distance" is based on spatial proximity and velocity coherence



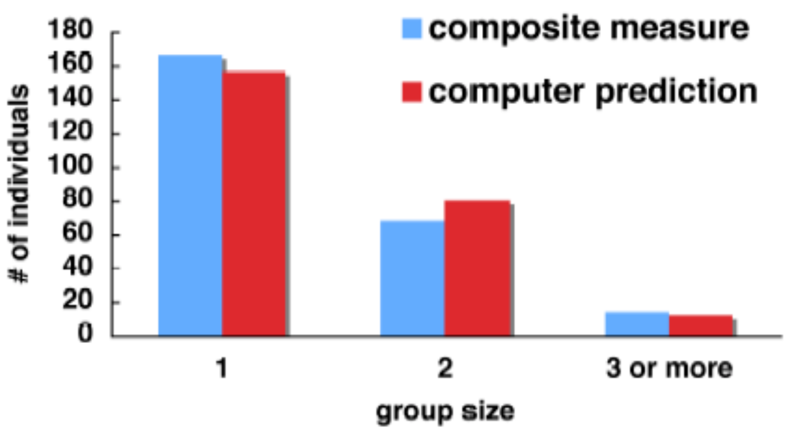
Robert Collins
Penn State

Sample Results



note: computer only sees this view!

Evaluation reveals substantial agreement between computer-generated groupings and those found by human coders (ground truth)



	match rate	$\chi^2(4, 248)$	Cohen's κ
trichotomous	85%	219.98	.69
dichotomous	89%	138.26	.75

$p < .001$

More Grouping Results

Likely Group Shapes

Are some group configurations more likely than others? Of course!

Robert Collins
Penn State

Analysis of Group Shape

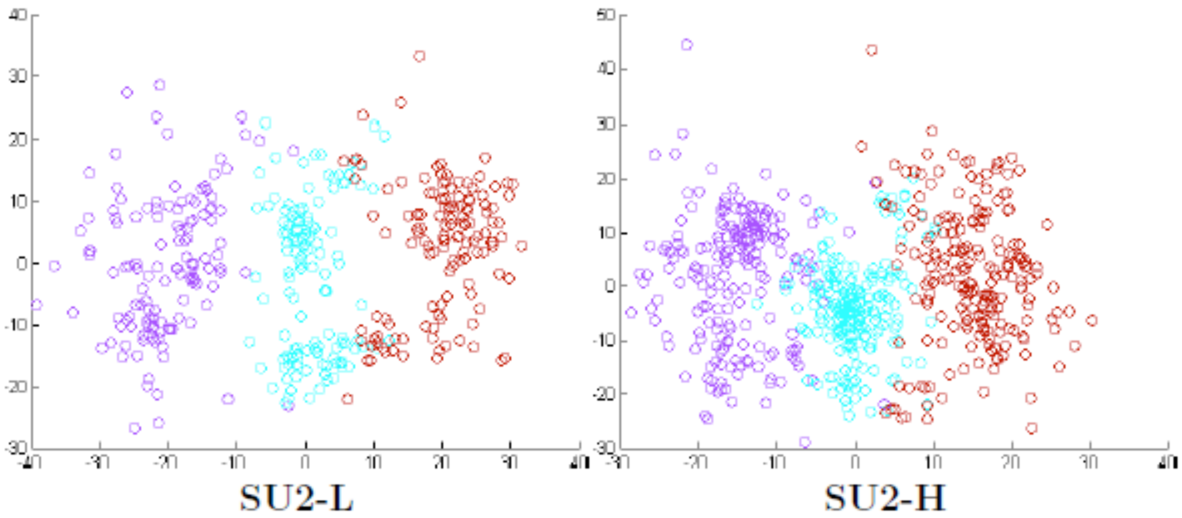
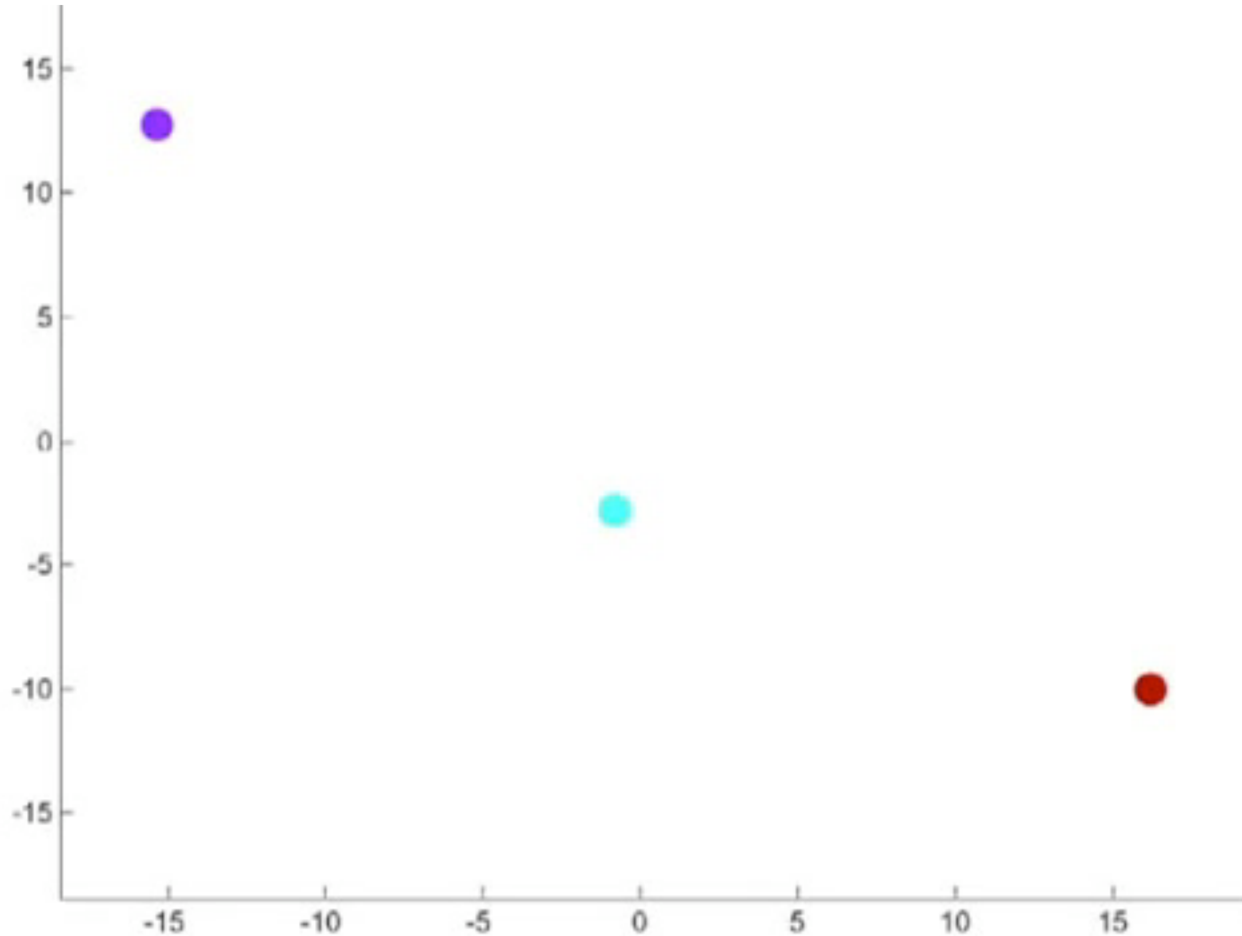


Figure 5.14. The configurations of groups of size three are aligned with respect to their group centers and moving directions. The three members are plotted with three different colors after a data association procedure that matches points across different configurations. Edges indicating the group configuration are omitted for clarity.

Analysis of Group Shape

analyzing groups of three people



Procrustes Analysis, first four modes of variation

Research Questions

Is multitarget tracking of human crowds any different than tracking crowds of animals? bats? cells?