Dynamic scene analysis: Recognition action and events

Representing temporal and spatial structure

- Relation to yesterday's lectures:

  – Jim Rehg: Further analysis of the problem of action/event detection in videos

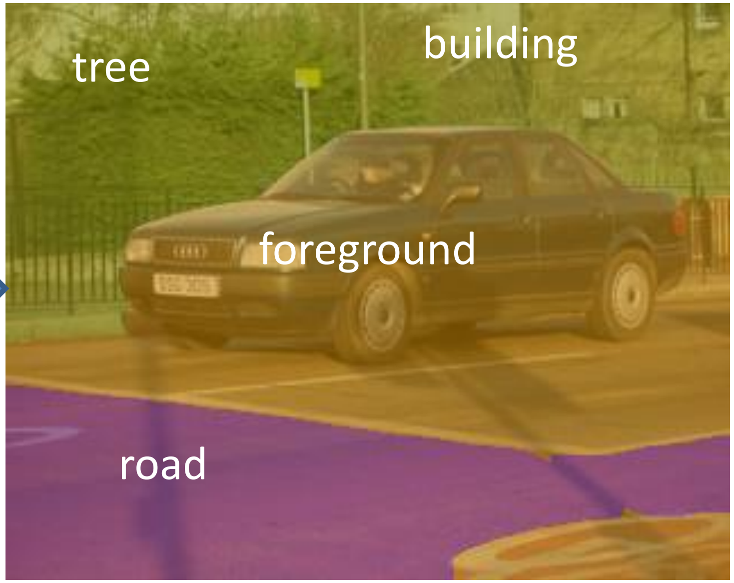
  – Jinxiang Chai: Models for classification/detection in input videos vs. synthesis and human motion model acquisition

First, a little bit of philosophy

First, a little bit of philosophy

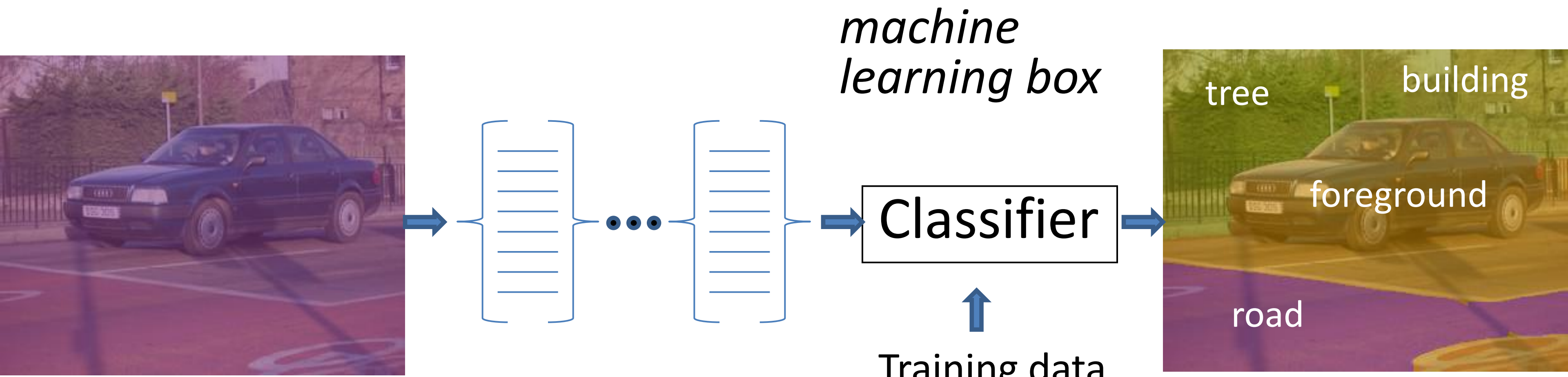
Let's look at an image labeling problem first

Input image



First, a little bit of philosophy

Let’s look at an image labeling problem first



Input image

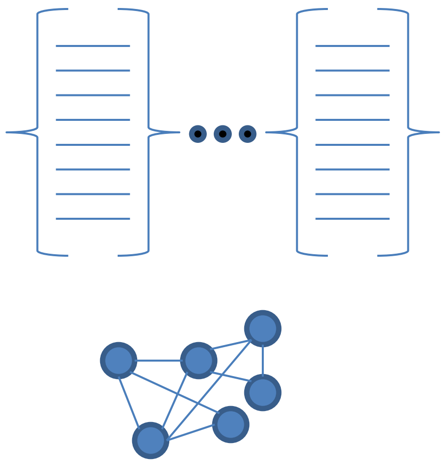
Set of feature of vectors

First, a little bit of philosophy

Let's look at an image labeling problem first

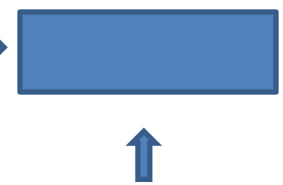
Machine learning box

Input image



Set of feature of vectors

+ additional structure (e.g., geometry, relations)



Training data +

Geometry, relational information, physics, domain knowledge

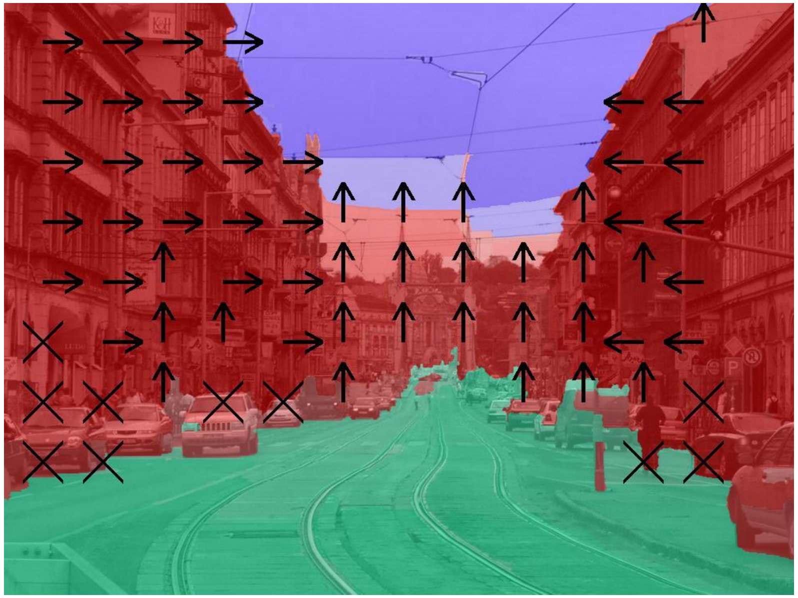
First, a little bit of philosophy

Let’s look at an image labeling problem first

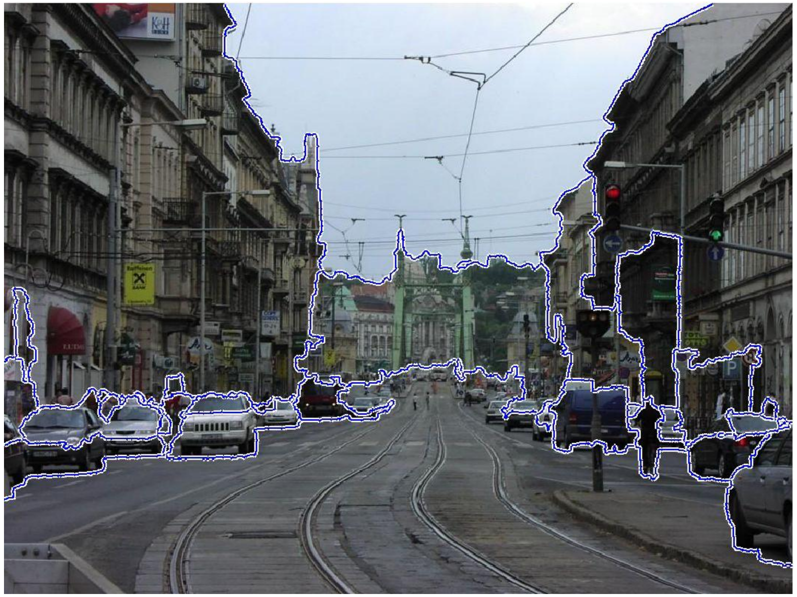
[Gould ICCV’09, Munoz ECCV’10]

Distributions of local features
Region features
Conditional Random Field or hierarchical set of classifiers
Direct interpretation from classification of image features

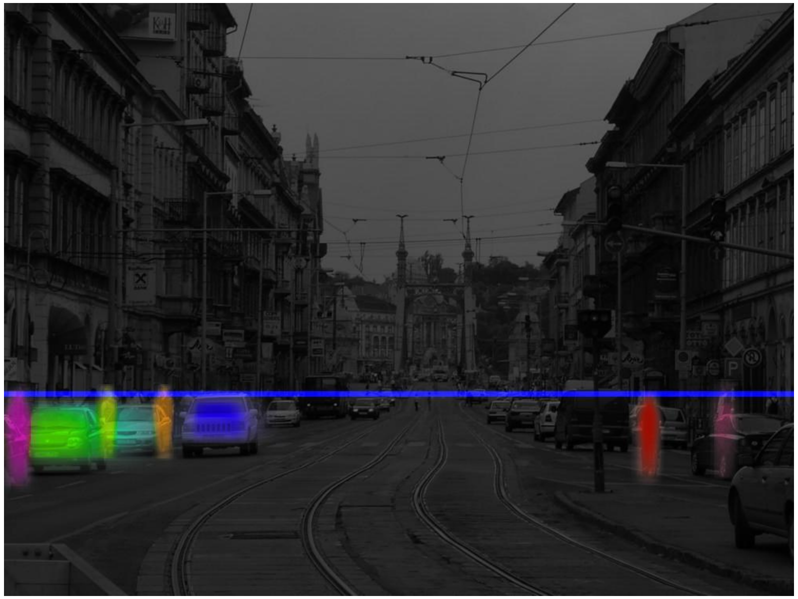
Input



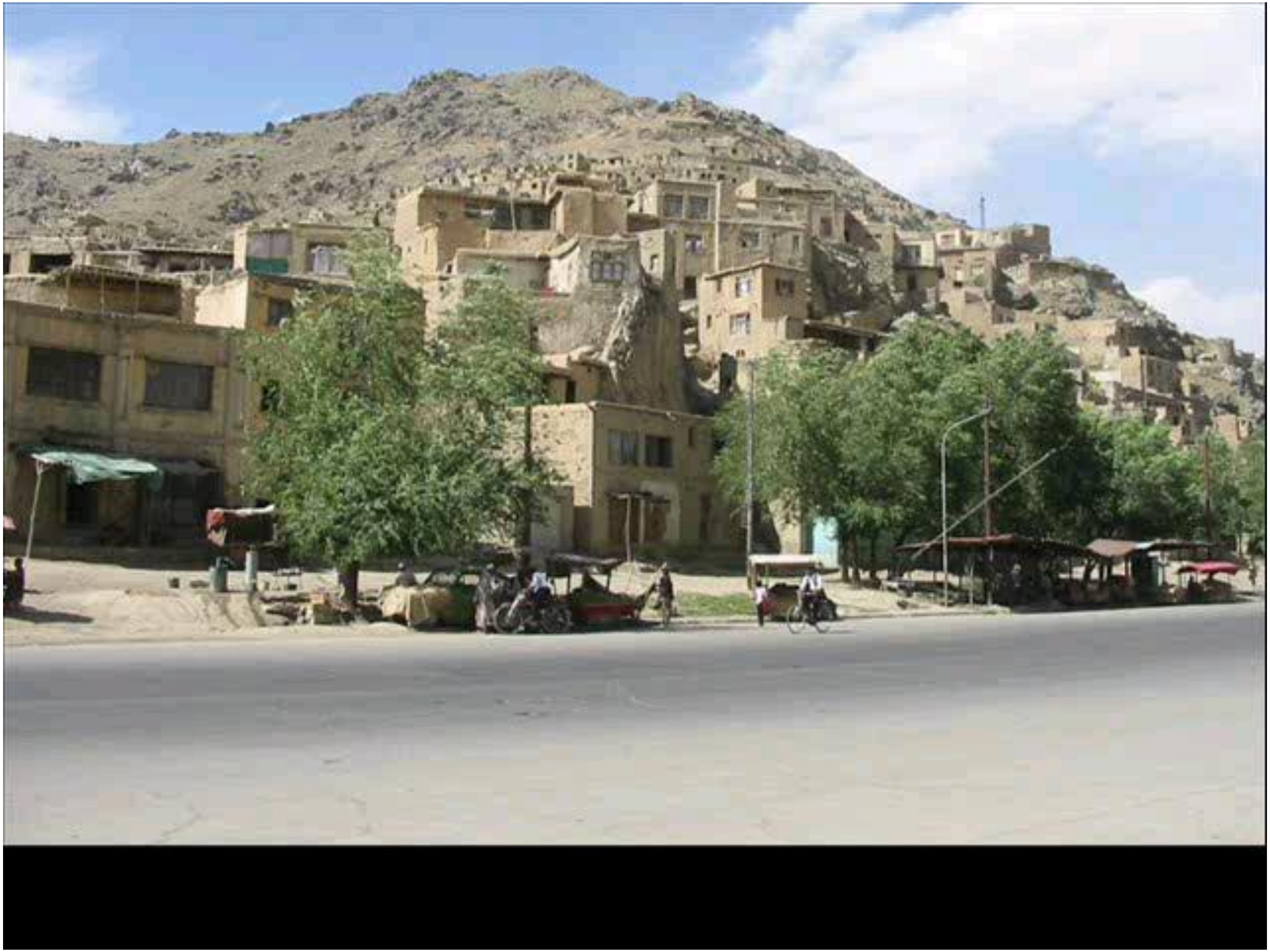
Surfaces



Occlusion Boundaries

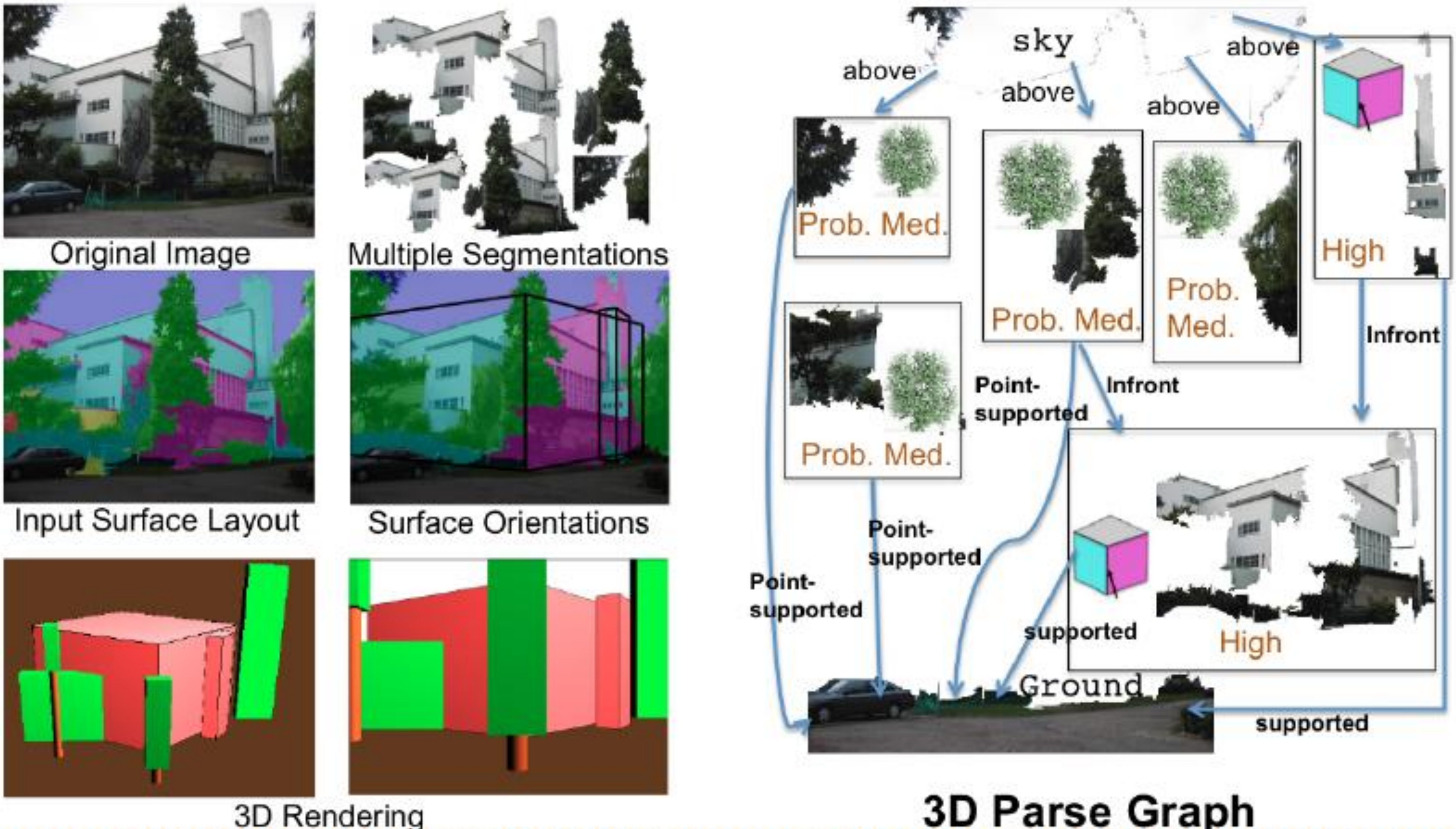


Viewpoint and Objects



[Hoiem, Efros, Hebert, CVPR08, IJCV10]

http://www.cs.cmu.edu/~abhinavg/blocksworld/



[Gupta, Efros, Hebert, ECCV'10]

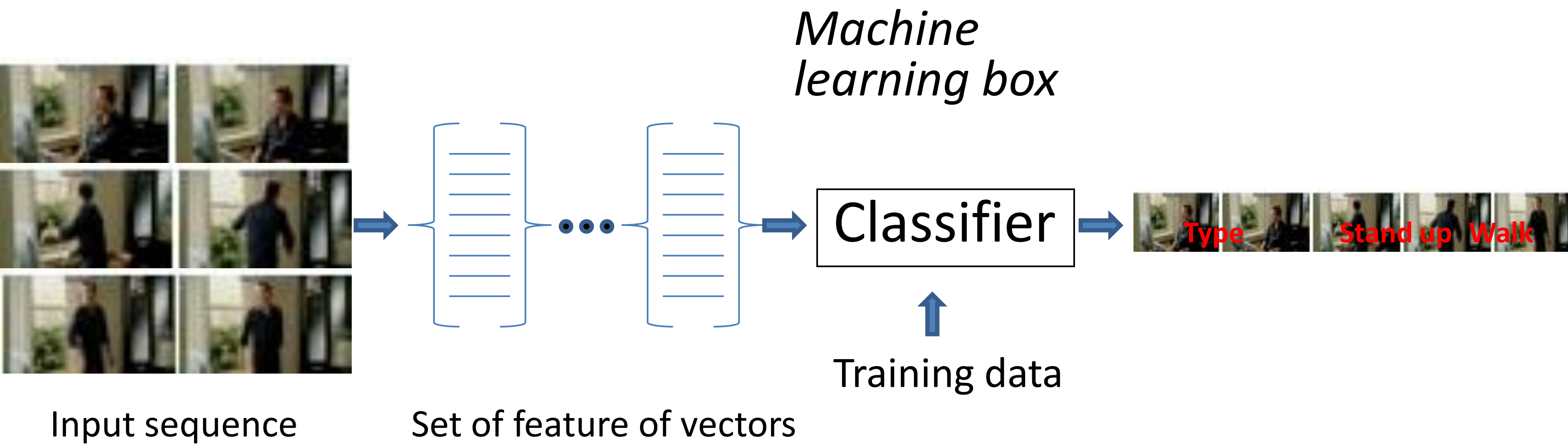
Reasoning about inferred 3D geometry (surfaces, occlusions between objects, physical constraints)

First, a little bit of philosophy

Let’s look at event/action detection for video

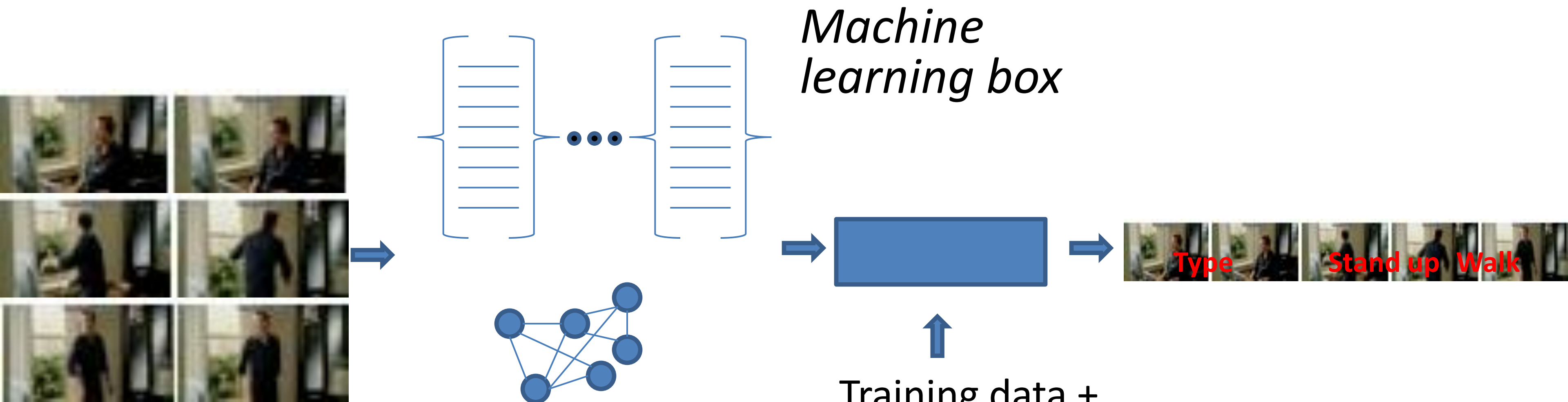
First, a little bit of philosophy

Let's look at event/action detection for video



First, a little bit of philosophy

Let's look at event/action detection for video



Input sequence

Set of feature of vectors

+ additional structure (e.g., geometry, relations)

Training data +

Geometry, relational information, physics, domain knowledge

What representations? What kind of reasoning?

Not much done so far.....

Examples

Hollywood

KTH

Rochester

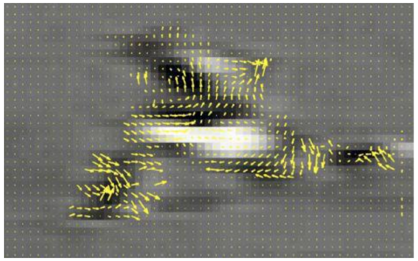
UCF YouTube

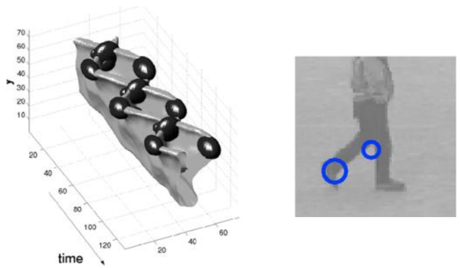
Classification vs. detection

- Classification:
  – Is there a “drinking” event in the input video?
- Detection:
  – Where is (in space and time) the drinking event in the input video?

- What we will see: Profound implications on *bias in training data*

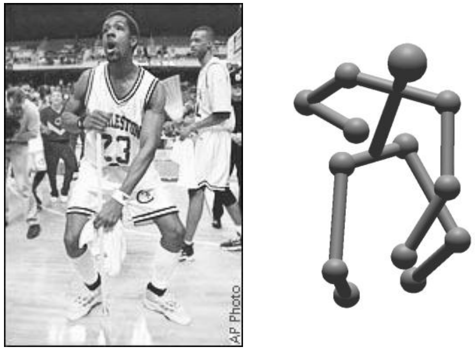

Shape-based


Flow-based

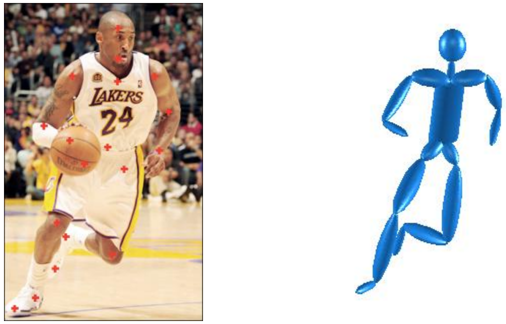


Space-time interest points


Volume-based



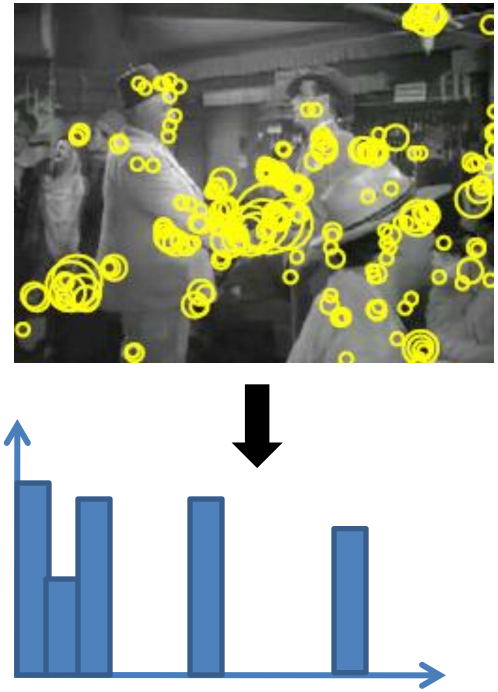
Skeletal models



- Theme today:
  - What are the key trade-offs between the representations?
  - How to balance generalization power, complexity, and spatial and temporal representational power?

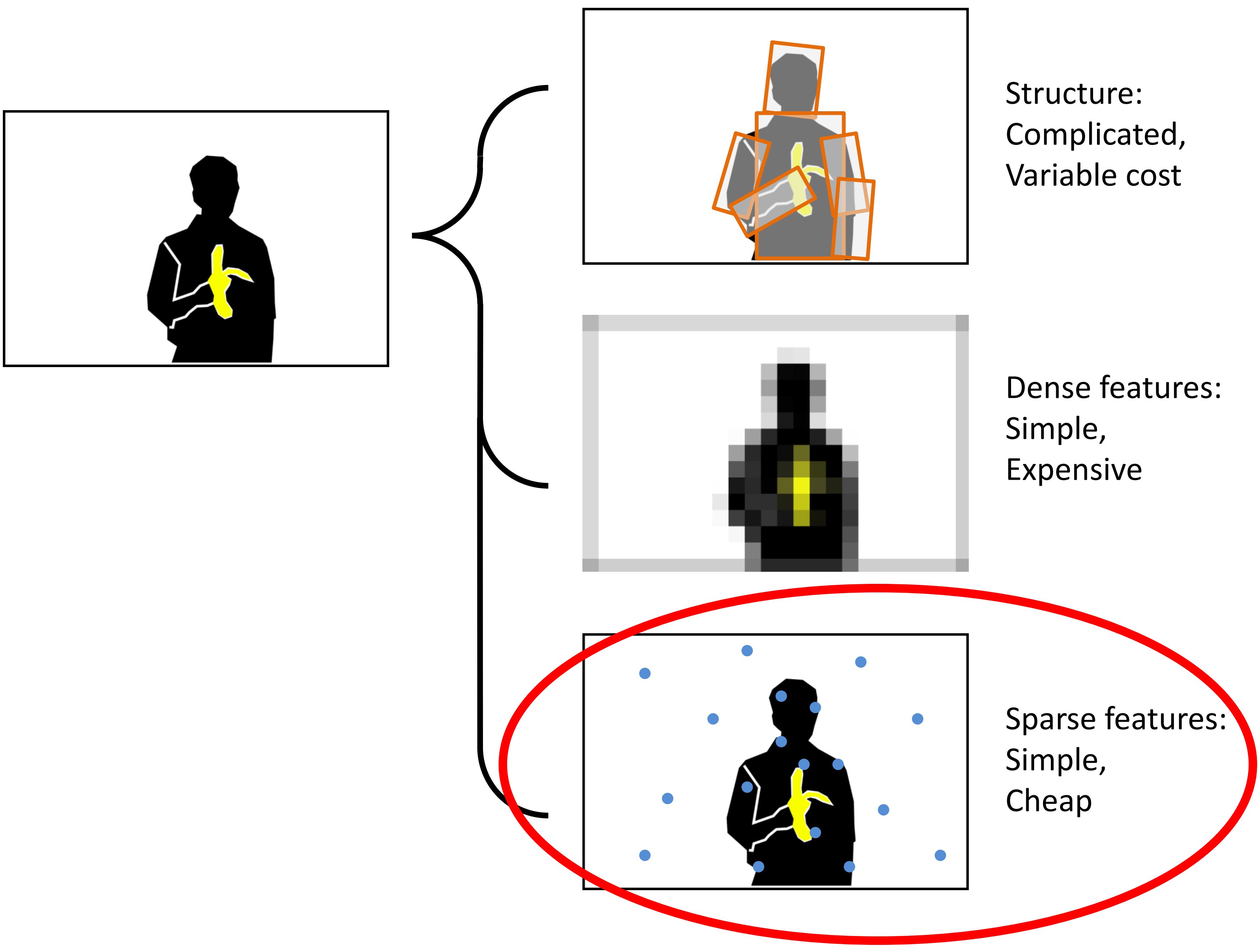
Outline

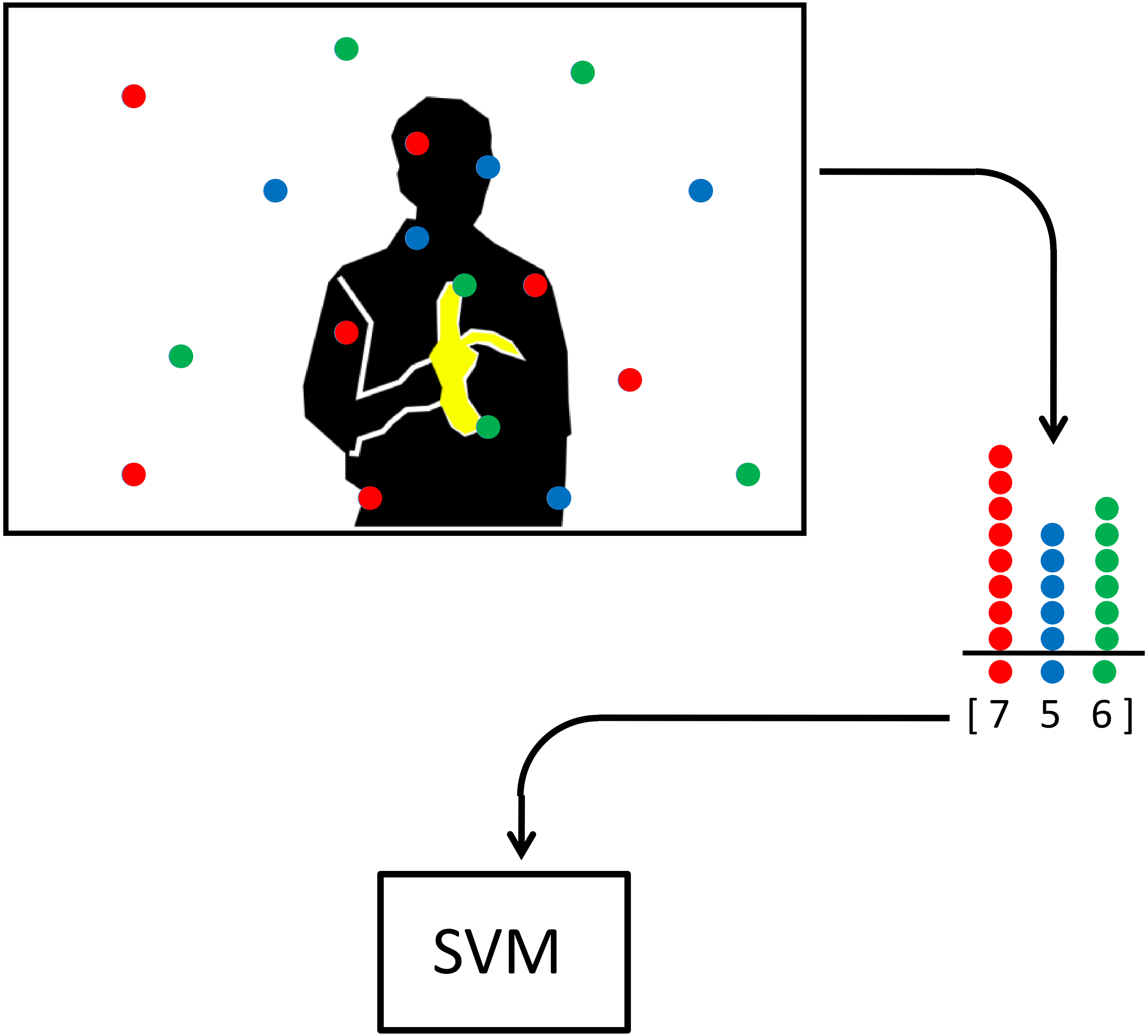
- Quick overview of two standards approaches
  - Statistical BoF approaches
  - Volumetric approaches
- Incorporating temporal information more explicitly
  - Example: Trajectory fragments
- Incorporating spatial information more explicitly
  - Example: Encoding pairwise relations
- Designing stronger structural models
  - Example: “Micro-actions” recognition through implicit 3D reconstruction
- Issues with video training datasets
  - Example: Selecting temporal boundaries
  - Analysis of bias in standard datasets
- Discussion and introduction to proposed challenge problems for afternoon presentations



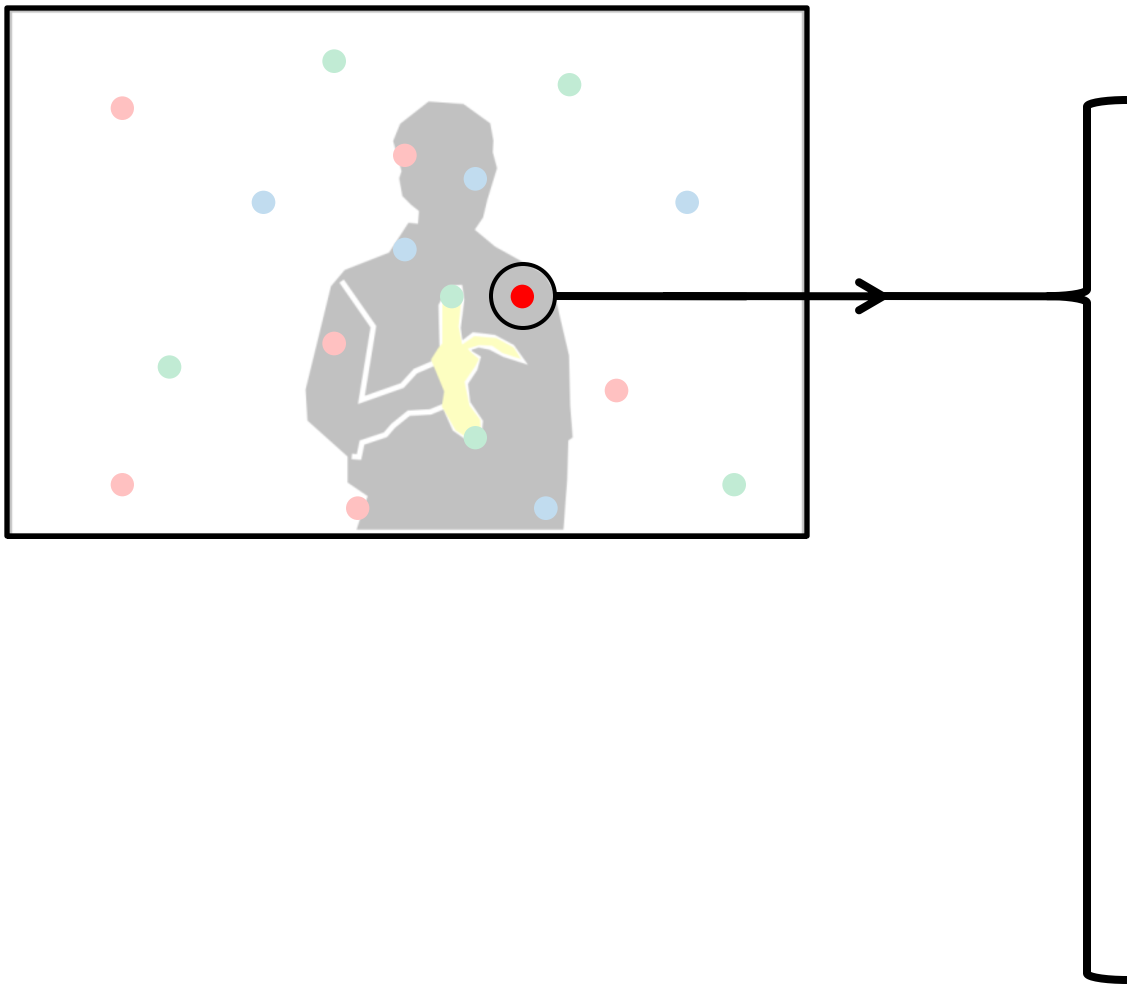
Bags of features
Histograms
……

One extreme:
- Orderless representations
- Efficient, direct extension of BoW approaches for images
- Loses spatial and temporal structure


Structure:
Complicated,
Variable cost
Dense features:
Simple,
Expensive
Sparse features:
Simple,
Cheap



[7 5 6]
SVM

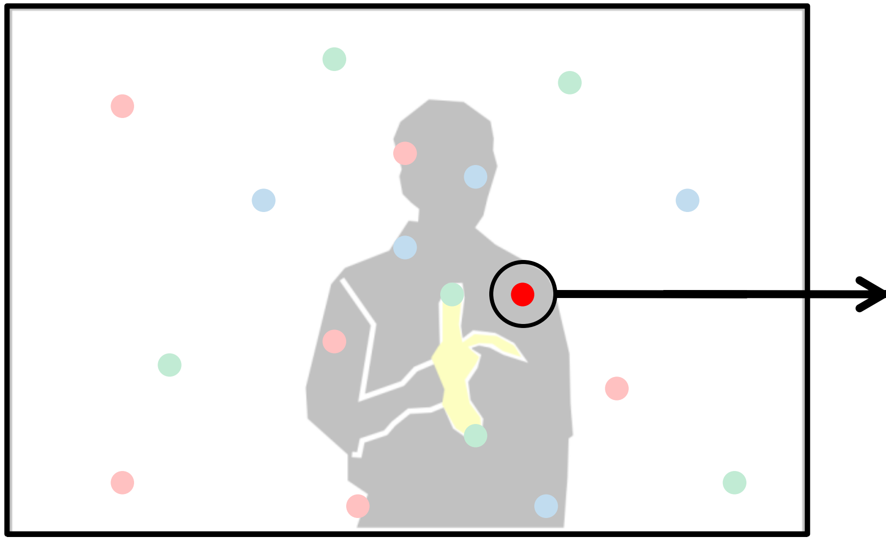


Position (x,y,t)

Quantize to *S* values

Label (color)

L discrete values



SIFT

or

MOSIFT

or

STIP

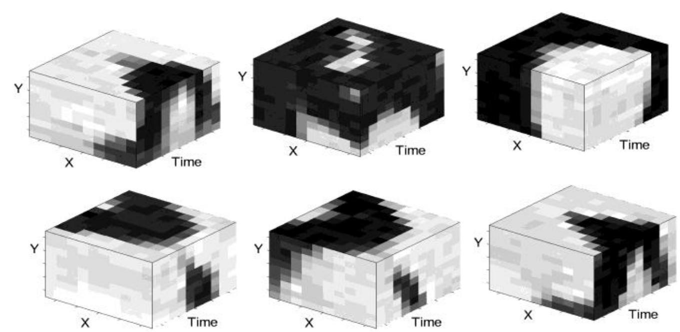
etc.

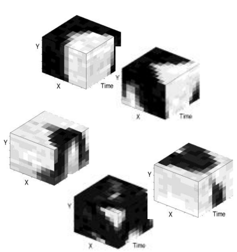
Example: Laptev

- I. Laptev. On space-time interest points. *IJCV*, 64 (2/3):107–123, 2005.
- P. Doll´ar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.
- C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, 2004.
- J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

Bag of space-time features + multi-channel SVM

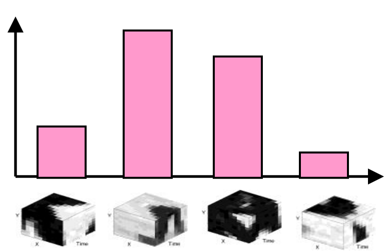
Collection of space-time patches





HOG & HOF patch descriptors

Histogram of visual words



Multi-channel SVM Classifier

Slide adapted from Laptev, CVPR08

Examples

	Clean	Automatic	Chance
AnswerPhone	32.1%	16.4%	10.6%
GetOutCar	41.5%	16.4%	6.0%
HandShake	32.3%	9.9%	8.8%
HugPerson	40.6%	26.8%	10.1%
Kiss	53.3%	45.1%	23.5%
SitDown	38.6%	24.8%	13.8%
SitUp	18.2%	10.4%	4.6%
StandUp	50.5%	33.6%	22.6%

Laptev, CVPR08

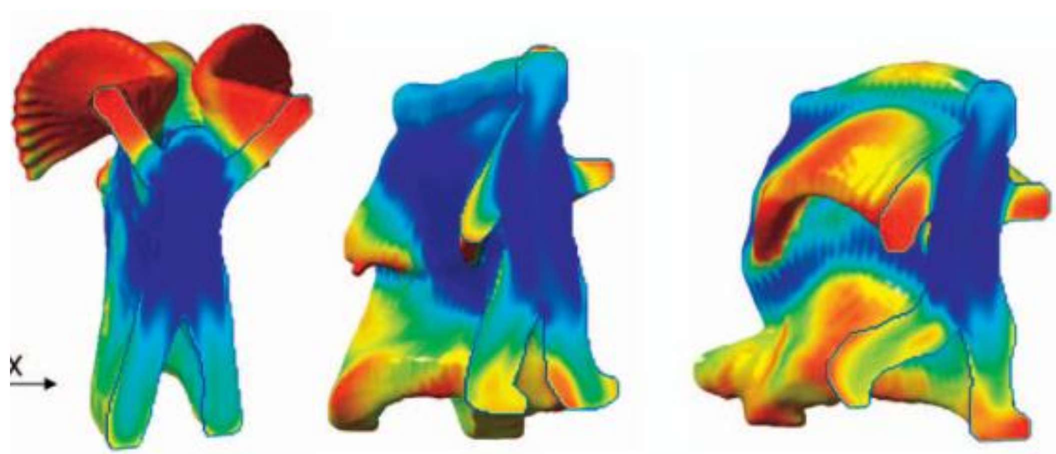
- I. Laptev. On space-time interest points. *IJCV*, 64 (2/3):107–123, 2005.
- P. Doll´ar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.
- C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, 2004.
- J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

Lessons?

- Plus:
  - Can generalize well, e.g., can learn from large sets of examples
  - Fast, can reuse most data across classes
  - Well suited for classification tasks
- Minus:
  - Does not incorporate strong representation of spatial and temporal structure
  - Cannot operate with very few examples
  - Not well suited for detection tasks

At other extreme: Volumetric representations

- Template-based representation
- Preserves strong spatial/temporal structure
- Difficult to generalize to variations in viewpoint



Another extreme:

- Template-based representation
- Preserves strong spatial/temporal structure
- Difficult to generalize to variations in viewpoint

A *few* examples:

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(3).

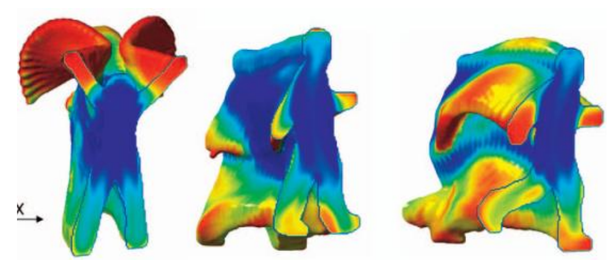
Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space-time shapes. In *Proc. ICCV*.

Ke, Y., Sukthankar, R., Hebert, M. (2010). Volumetric Features for Video Event Detection. International Journal of Computer Vision.

Shechtman, E., & Irani, M. (2007). Space-time behavior based correlation; How to tell if two underlying motion fields are similar without computing them? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(11).

Weinland, D., Ronfard, R., & Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, *104*(2).

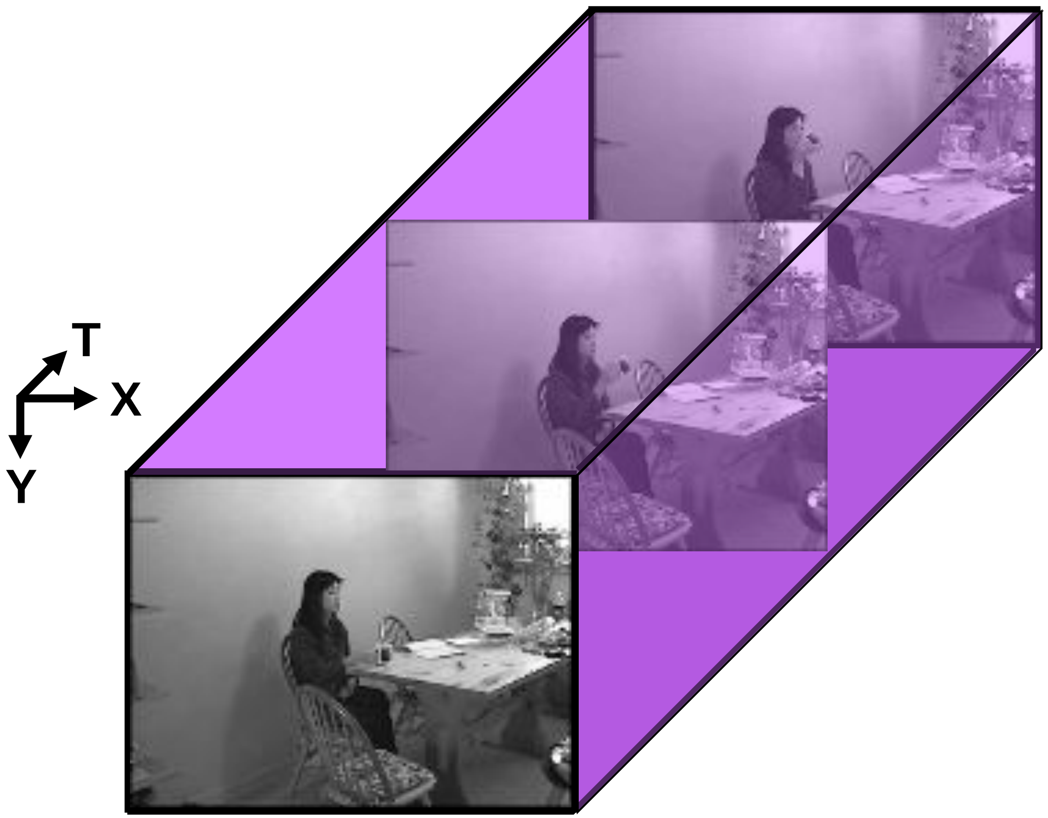
Yilmaz, A., & Shah, M. (2005). Actions as objects: A novel action representation. In *Proc. CVPR*.



Space-time volume
Flow/shape comparison
…….

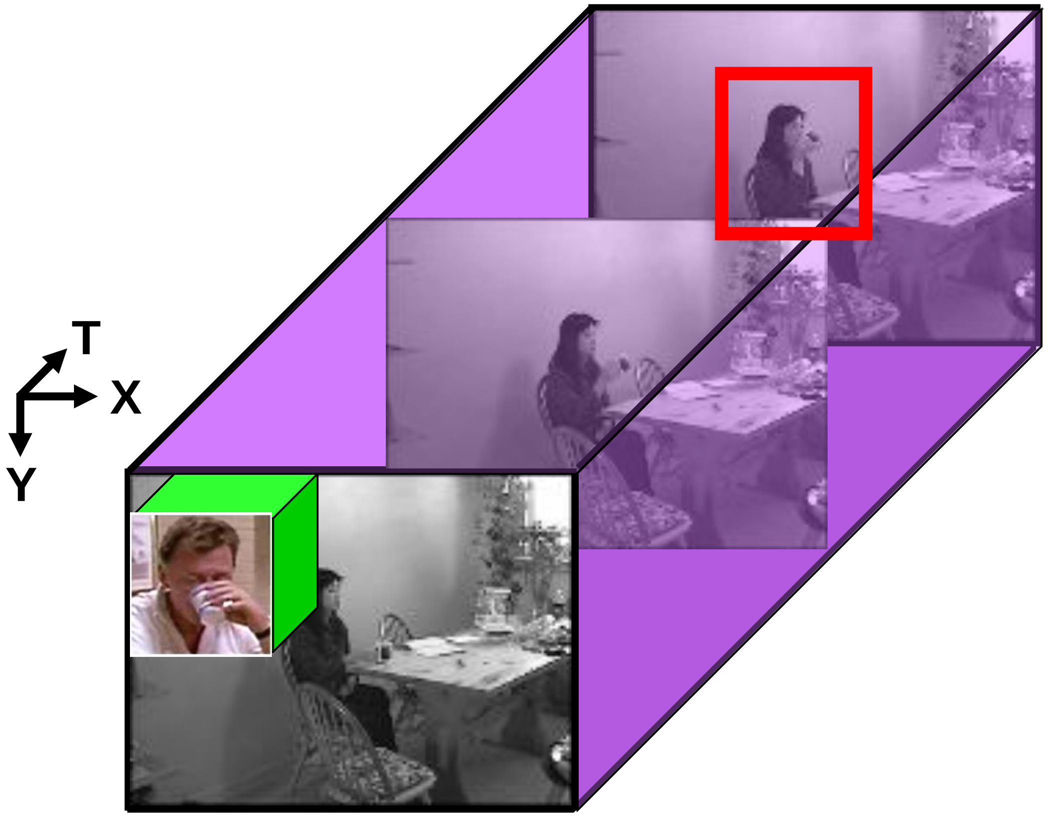
Using space-time volumes: General idea

Model



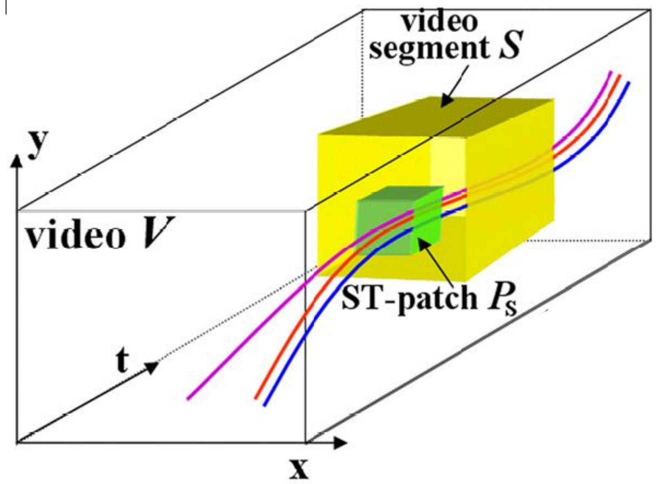
Grab-Cup Event

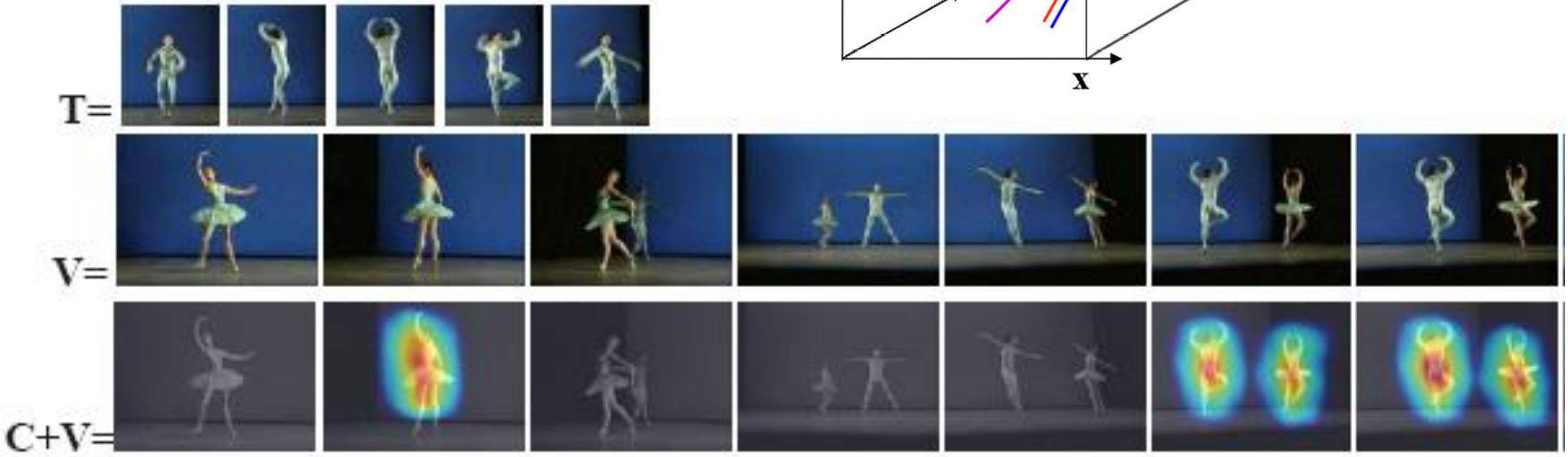
Using space-time volumes: General idea



Grab-Cup Event

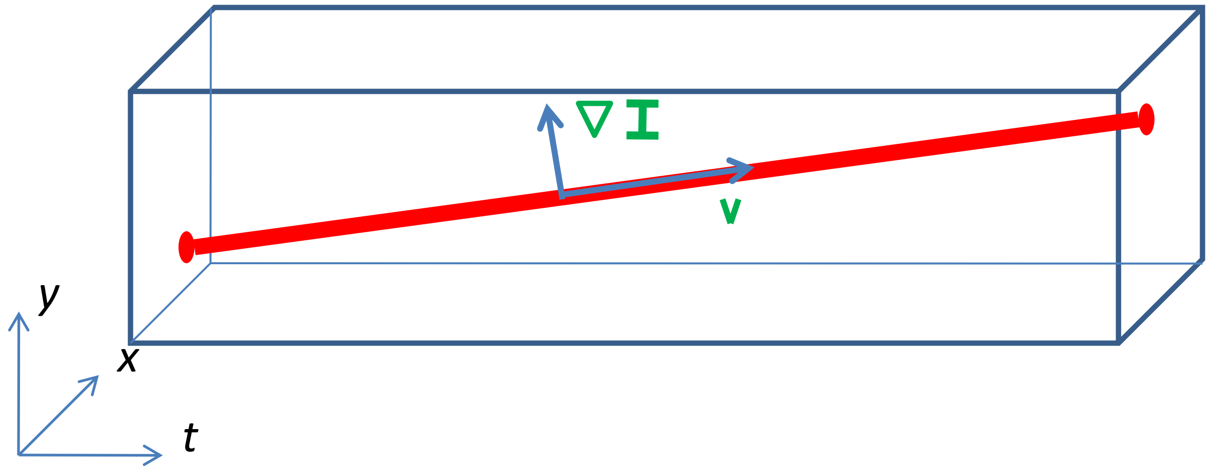
Example





- Compare distribution of motion vectors between 2 blocks (reference action model vs. observed video)
- Trick: Estimate consistency between distributions of motion without estimating motion explicitly

Shechtman, E., & Irani, M. (2007). Space-time behavior based correlation; How to tell if two underlying motion fields are similar without computing them? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(11).



$$\nabla I . v = 0$$

$$Gv = 0$$

$$\underbrace{\begin{bmatrix} P_{x_1} & P_{y_1} & P_{t_1} \\ P_{x_2} & P_{y_2} & P_{t_2} \\ & \cdots & \\ & \cdots & \\ P_{x_n} & P_{y_n} & P_{t_n} \end{bmatrix}_{n\times 3}}_{\mathbf{G}} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1}$$

$$\mathbf{G}^{\mathbf{T}}\mathbf{G} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times 1}$$

[slide adapted from Pyry Matikainen]

$$\mathbf{M} = \begin{bmatrix} \Sigma P_x^2 & \Sigma P_x P_y & \Sigma P_x P_t \\ \Sigma P_y P_x & \Sigma P_y^2 & \Sigma P_y P_t \\ \Sigma P_t P_x & \Sigma P_t P_y & \Sigma P_t^2 \end{bmatrix}$$

$$\mathbf{M}^{\diamond} = \begin{bmatrix} \Sigma P_x^2 & \Sigma P_x P_y \\ \Sigma P_y P_x & \Sigma P_y^2 \end{bmatrix}$$

- Space-Time Harris Matrix

- Upper-left Minor

[slide adapted from Pyry Matikainen]

$$\mathbf{M} = \mathbf{G}^{\mathrm{T}}\mathbf{G} = \begin{bmatrix} \Sigma P_x^2 & \Sigma P_x P_y & \Sigma P_x P_t \\ \Sigma P_y P_x & \Sigma P_y^2 & \Sigma P_y P_t \\ \Sigma P_t P_x & \Sigma P_t P_y & \Sigma P_t^2 \end{bmatrix}$$

Space-Time Harris Matrix

$$\mathbf{M}^{\diamondsuit} = \begin{bmatrix} \Sigma P_x^2 & \Sigma P_x P_y \\ \Sigma P_y P_x & \Sigma P_y^2 \end{bmatrix}$$

Upper-left Minor

If the motion is consistent within the space-time block:
The temporal axis does not affect the rank of $\mathbf{M}^{\diamondsuit}$

$$rank(\mathbf{M}) \approx rank(\mathbf{M}^{\diamondsuit})$$

[slide adapted from Pyry Matikainen]

$\mathbf{G_1}$

$\mathbf{G_2}$

[slide adapted from Pyry Matikainen]

$$\mathbf{G_{12}} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} \mathbf{G_1} \\ \mathbf{G_2} \end{bmatrix}_{2n\times 3} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{2n\times 1}$$

$$\mathbf{M_{12}} \begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times 1}$$

$$\mathbf{M_{12}} = \mathbf{M_1} + \mathbf{M_2} = \mathbf{G_1^T G_1} + \mathbf{G_2^T G_2}$$

[slide adapted from Pyry Matikainen]

$$\Delta r = rank(\mathbf{M}) - rank(\mathbf{M}^{\diamondsuit}) = \begin{cases} 0 & \textit{single motion} \\ 1 & \textit{multiple motions} \end{cases}$$

We use a continuous extension of this measure

$$\Delta\tilde{r} = \frac{\lambda_2 \cdot \lambda_3}{\lambda_1^{\diamondsuit} \cdot \lambda_2^{\diamondsuit}} = \frac{\det(M)}{\det(M^{\diamondsuit}) \cdot \lambda_1} \approx \frac{\det(M)}{\det(M^{\diamondsuit}) \cdot \|M\|_F}$$

$$m_{12} = \frac{\Delta r_{12}}{\min(\Delta r_1, \Delta r_2) + \varepsilon}$$

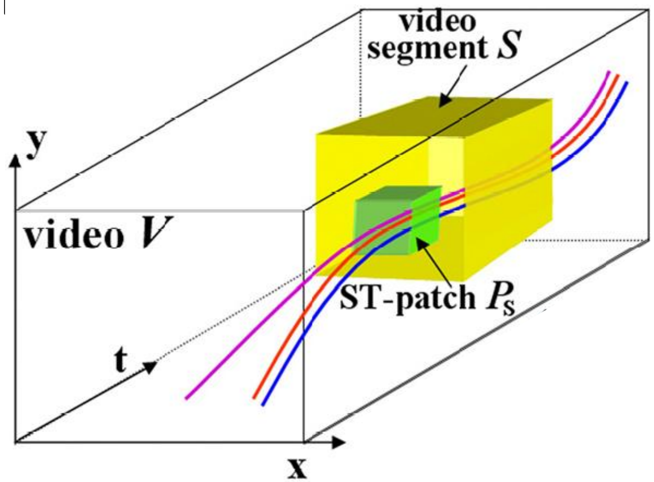
Inconsistency

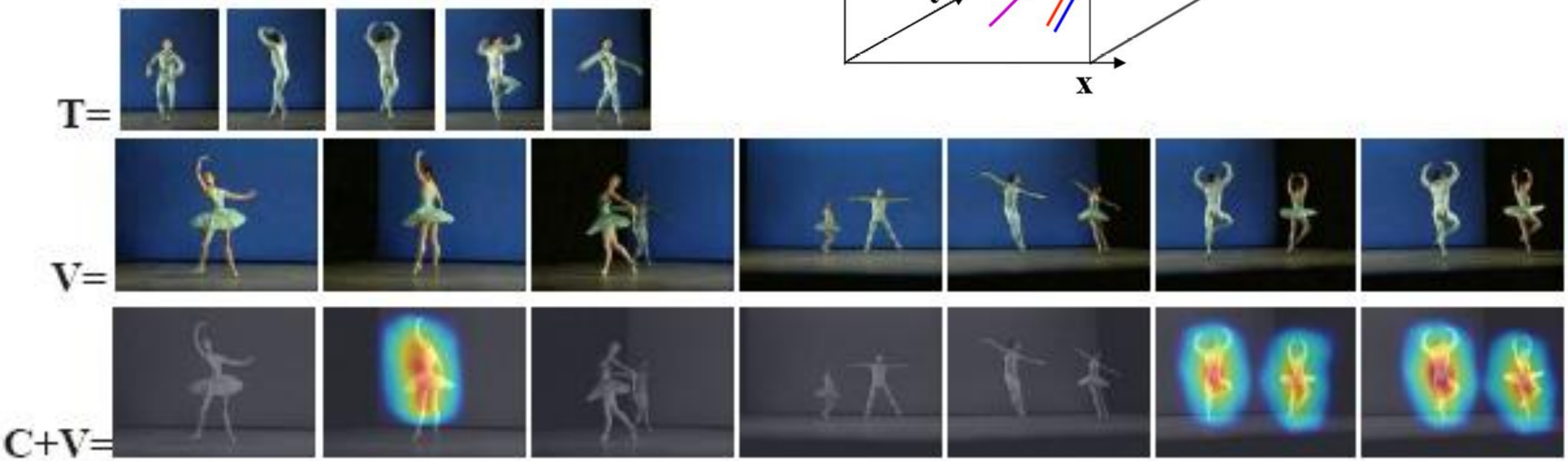
$$c_{12} = 1/m_{12}$$

Consistency

[slide adapted from Pyry Matikainen]

Example

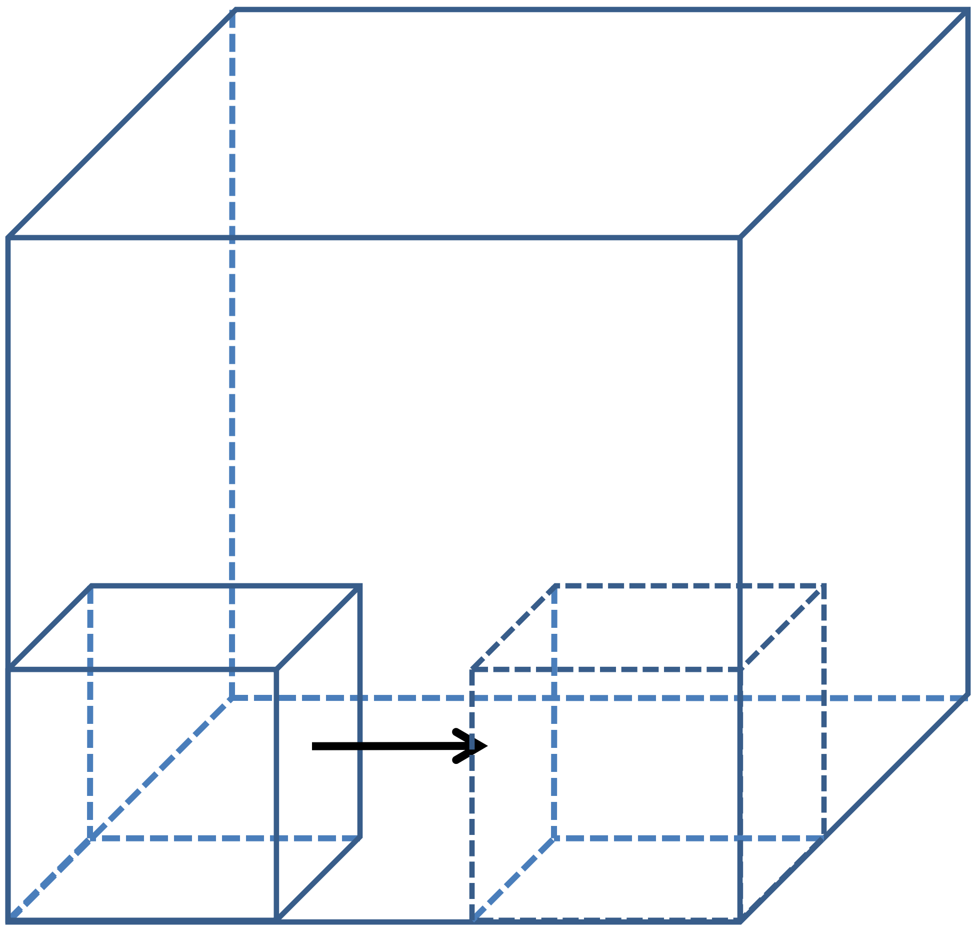




- Compare distribution of motion vectors between 2 blocks (reference action model vs. observed video)
- Trick: Estimate consistency between distributions of motion without estimating motion explicitly

Shechtman, E., & Irani, M. (2007). Space-time behavior based correlation; How to tell if two underlying motion fields are similar without computing them? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(11).

Potentially expensive?



[slide adapted from Pyry Matikainen]

- Typical situation:

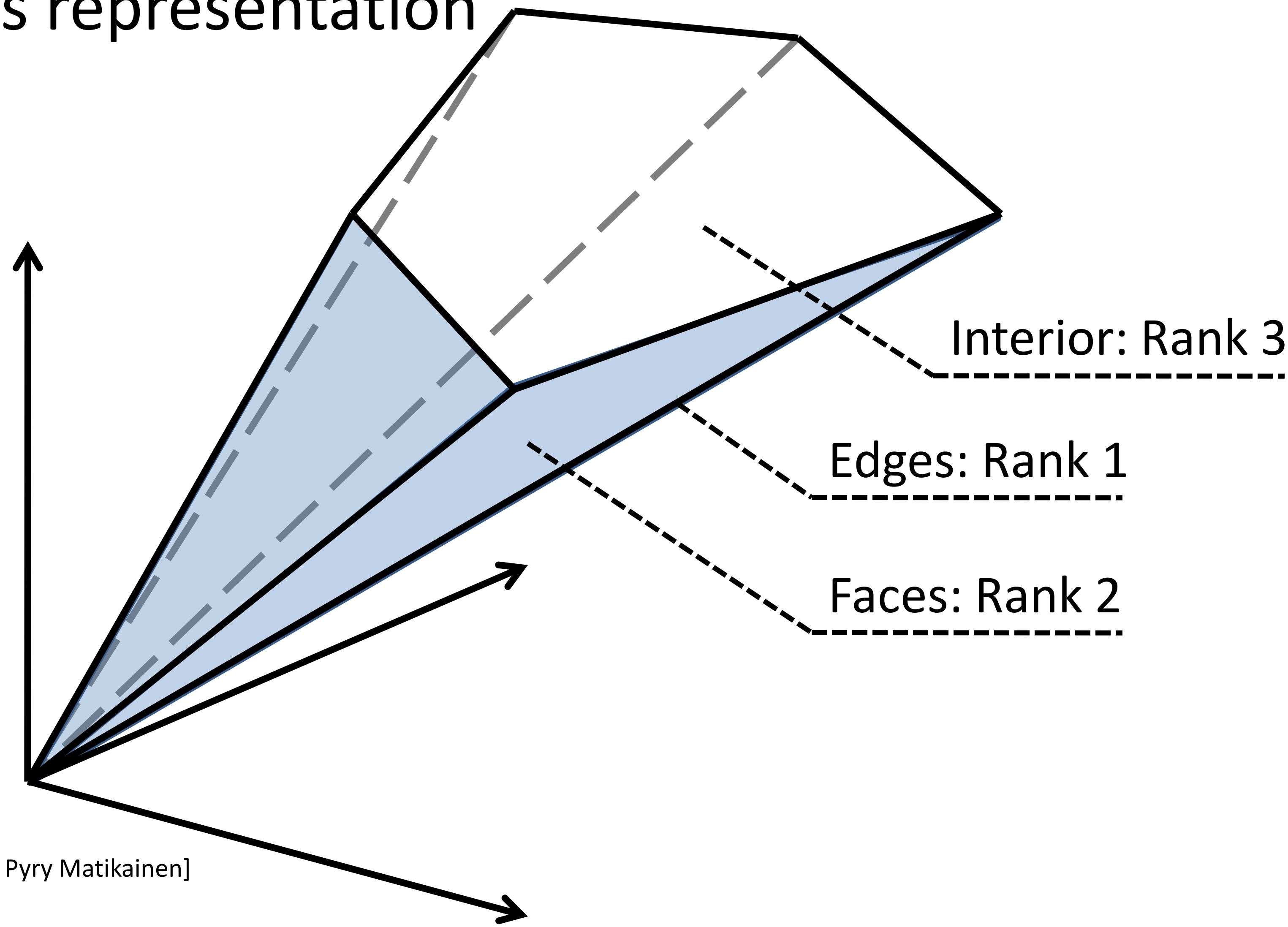
180x144x200 video (6s)

60x30x30 template (1s)

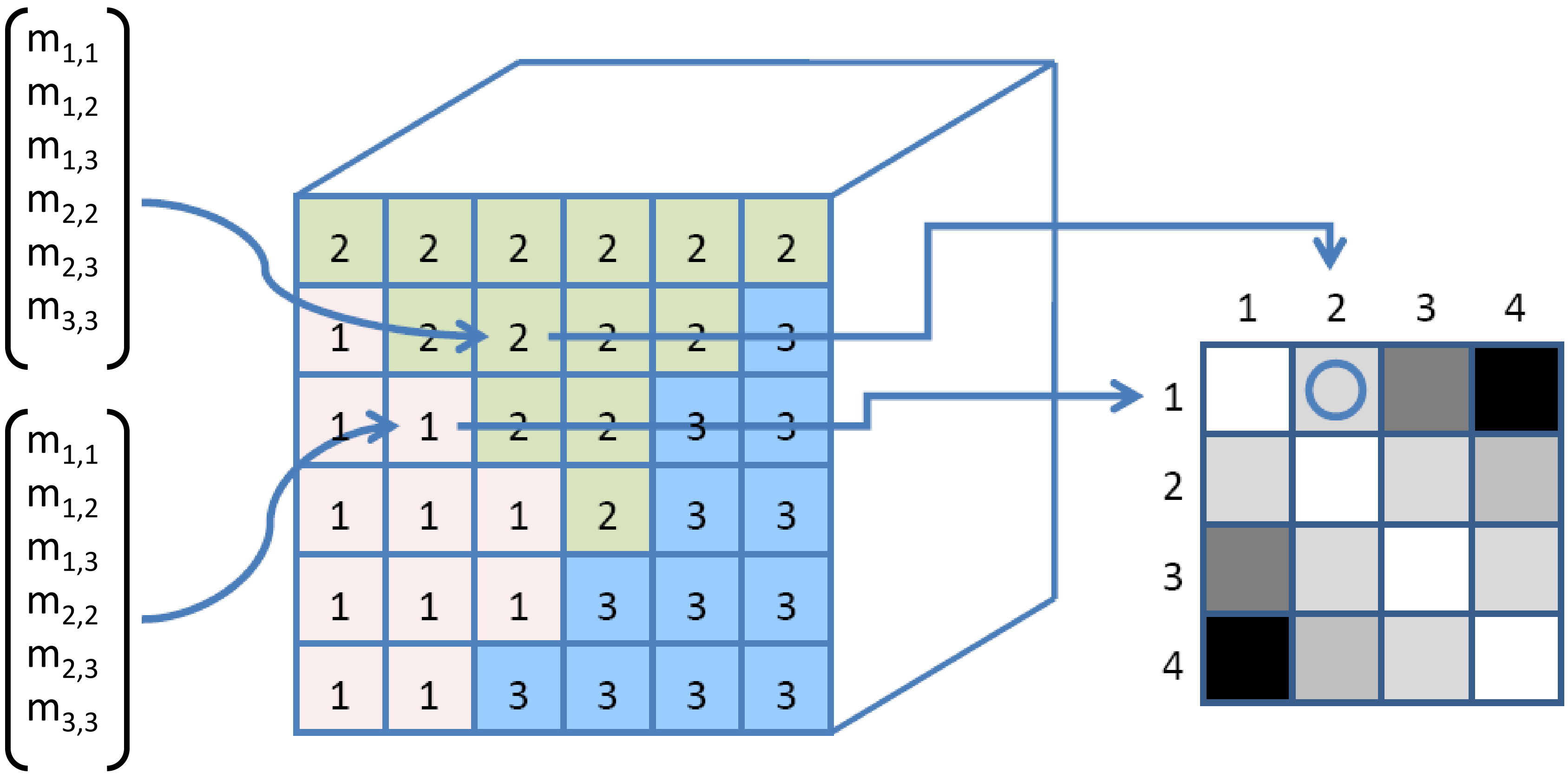
=

279,936,000,000 consistency comparisons

- The matrices involved are semi definite positive
- Bounded domain with proper normalization
- Idea: Use quantized representation instead of continuous representation



[slide adapted from Pyry Matikainen]

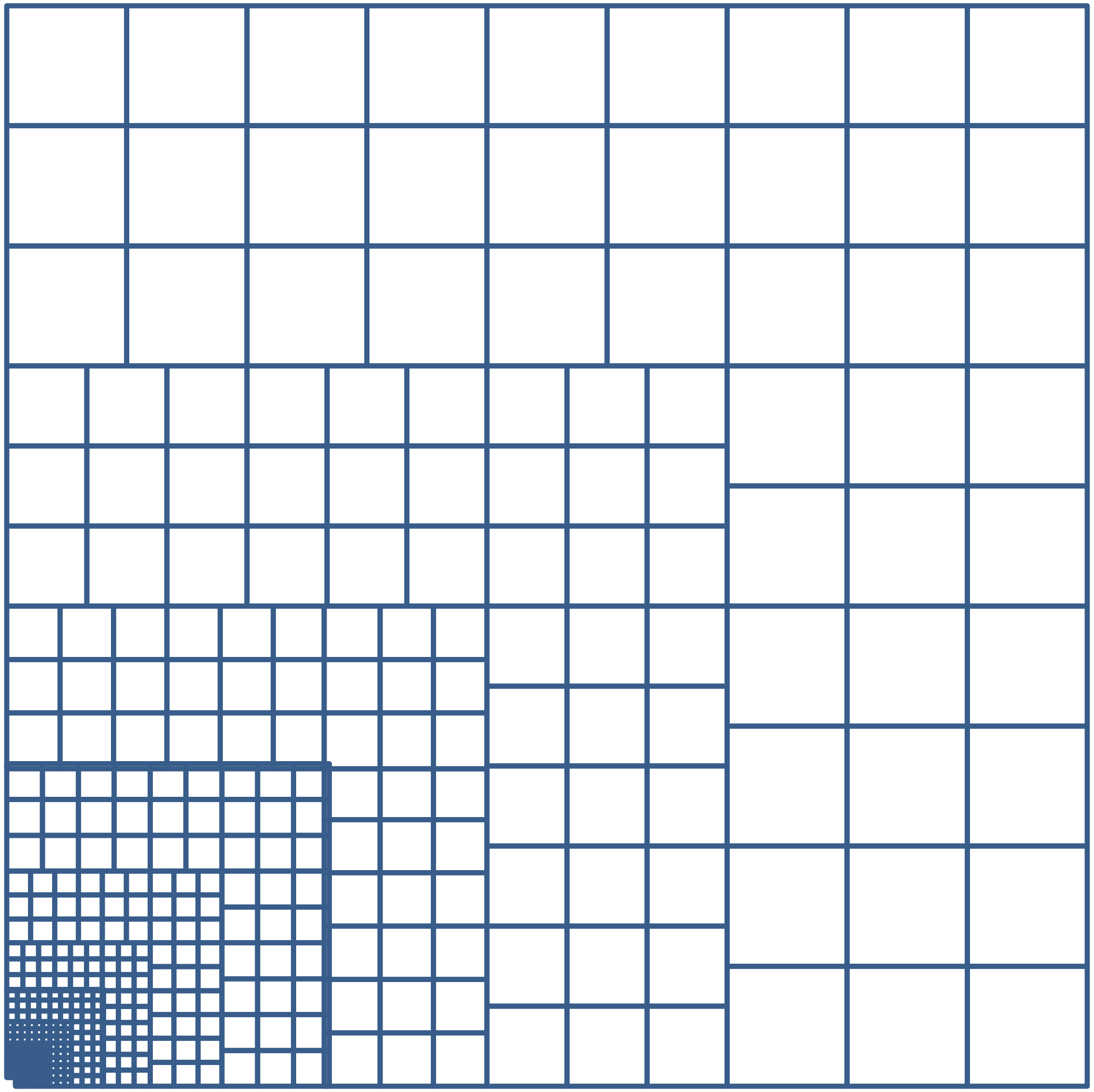


ST-Harris matrices

Quantization and label assignment

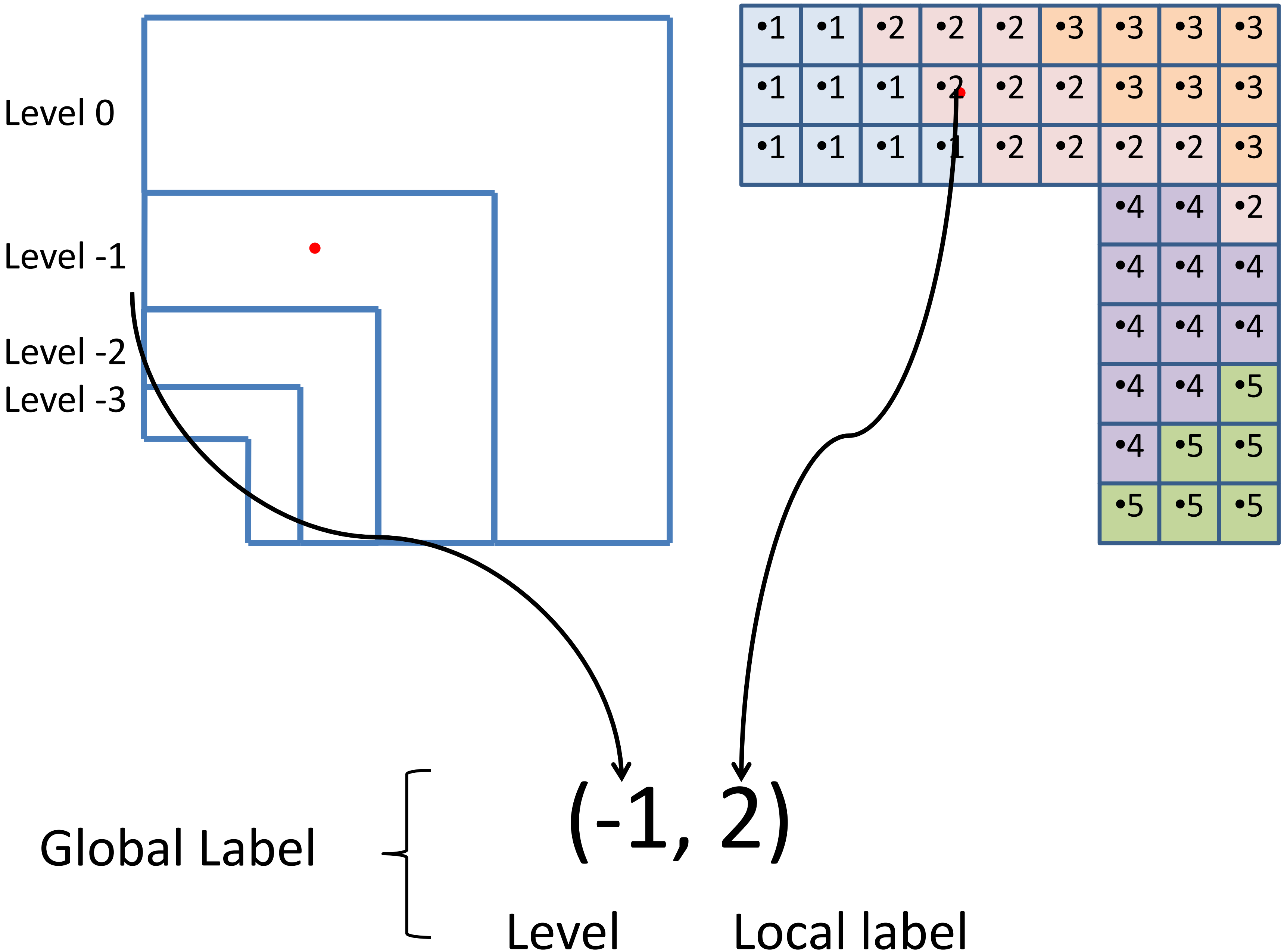
Efficient motion consistency computation through table lookup

[slide adapted from Pyry Matikainen]

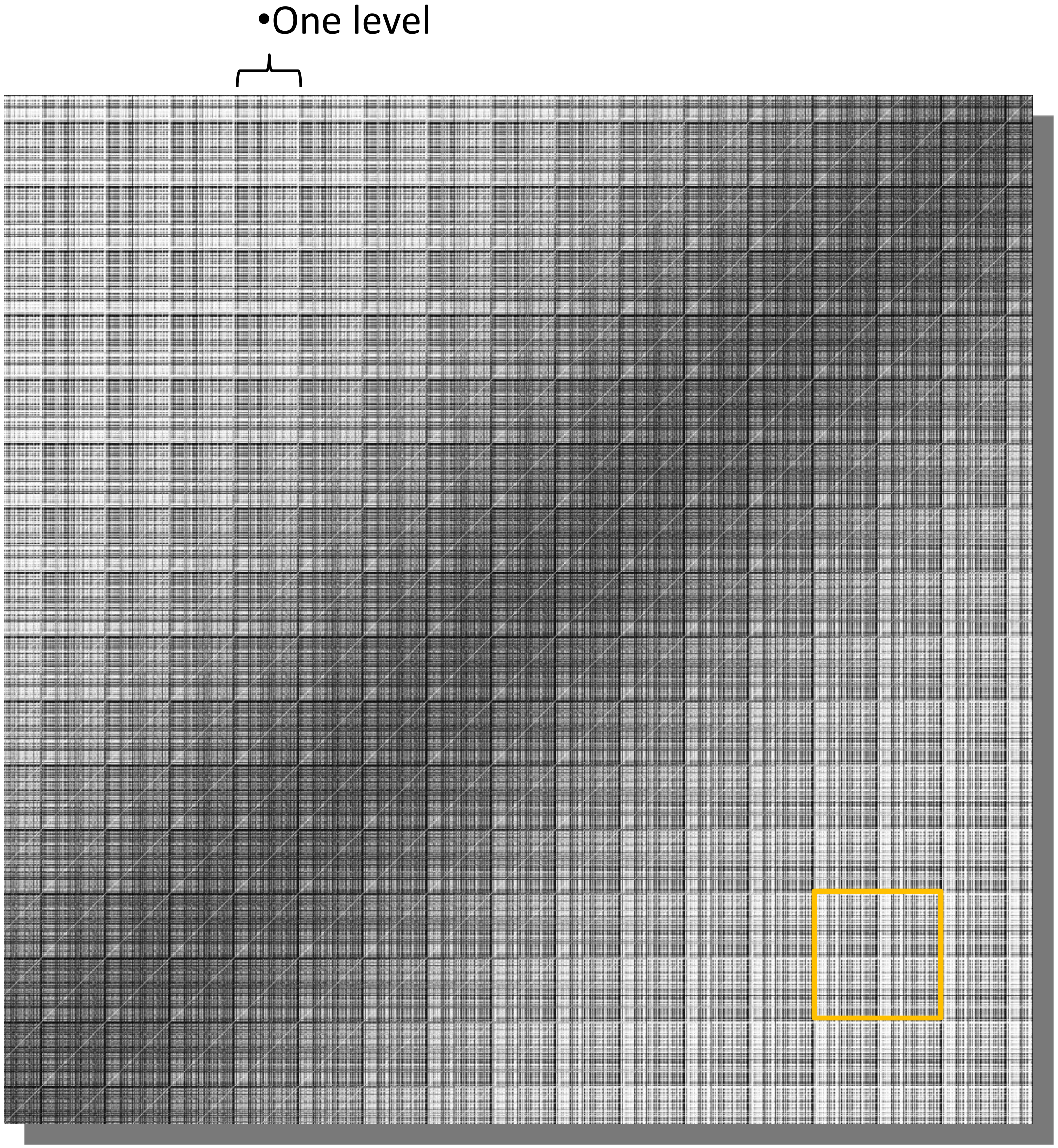


Hierarchical table

[slide adapted from Pyry Matikainen]



[slide adapted from Pyry Matikainen]



Actual consistency table:

1600x1600 =

(100 centers)

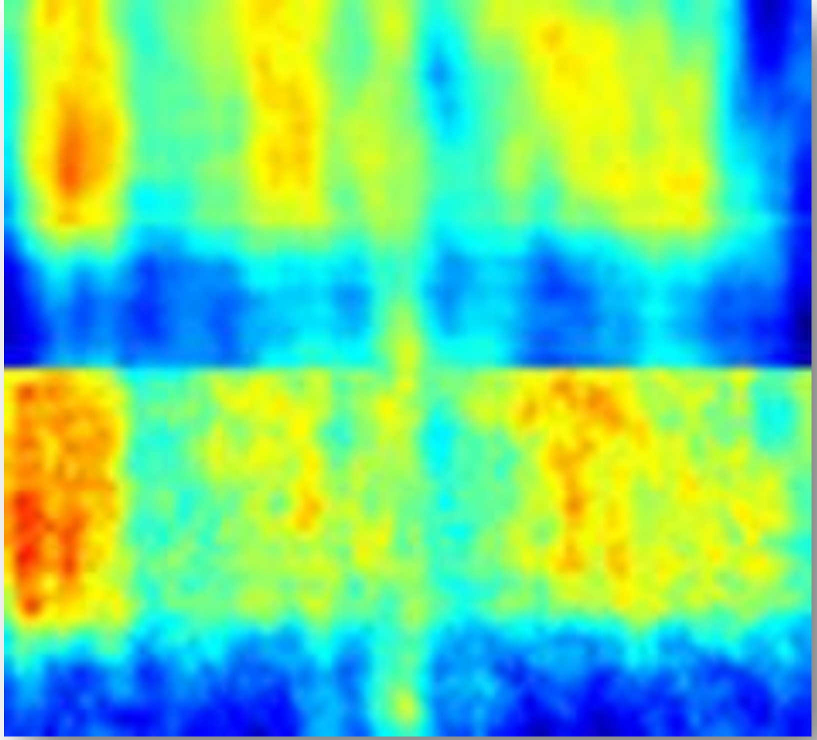
x

(16 levels)

[slide adapted from Pyry Matikainen]

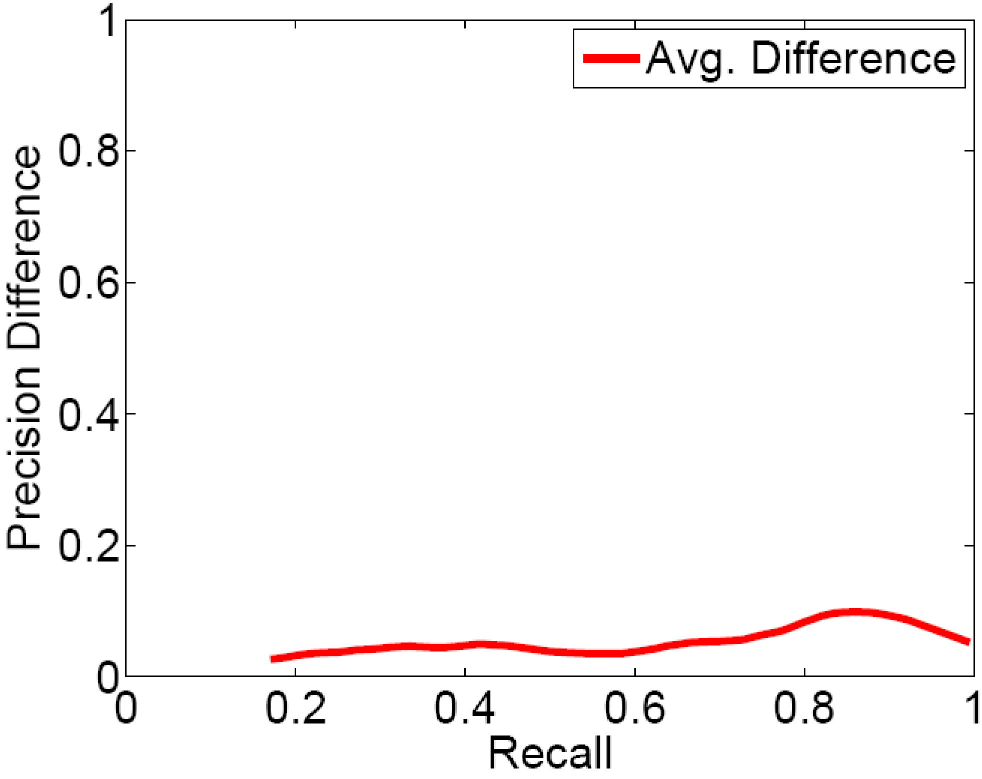
No quantization

With quantization



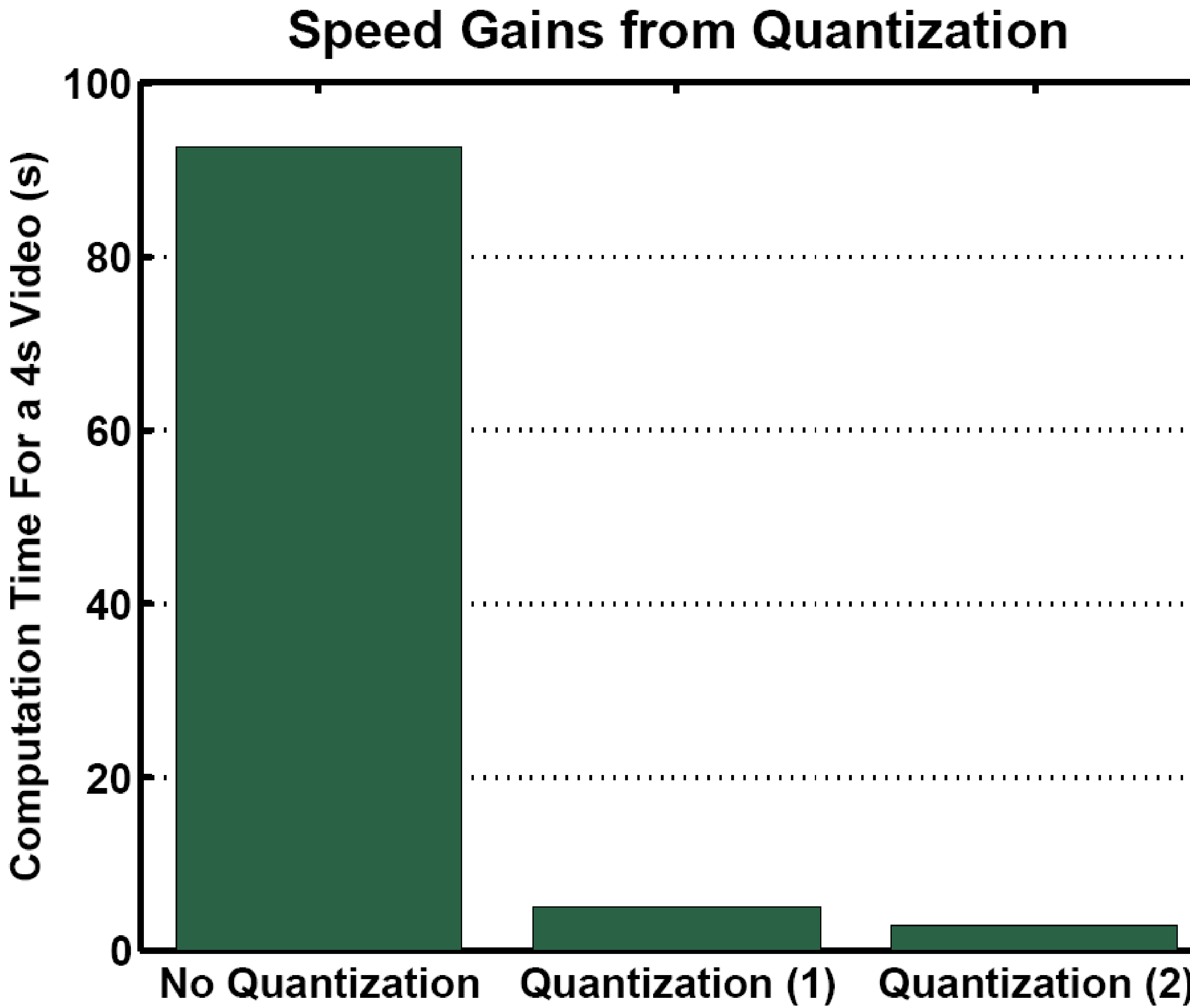
[Matikainen, Hebert, Sukthankar Ke, Fast Motion Consistency Through Matrix Quantization, 2009]

Does quantization affect performance?



Evaluated over all actions in KTH

[slide adapted from Pyry Matikainen]



[Matikainen, Hebert, Sukthankar, Ke, Fast Motion Consistency Through Matrix Quantization, 2009]

Lessons learned

- Added temporal dimension increases complexity
- Clever quantization scheme can be crucial for efficient computation
- Better quantization is often more relevant than blind clustering

- More later on using quantization schemes for efficient representations of spatial and temporal relations

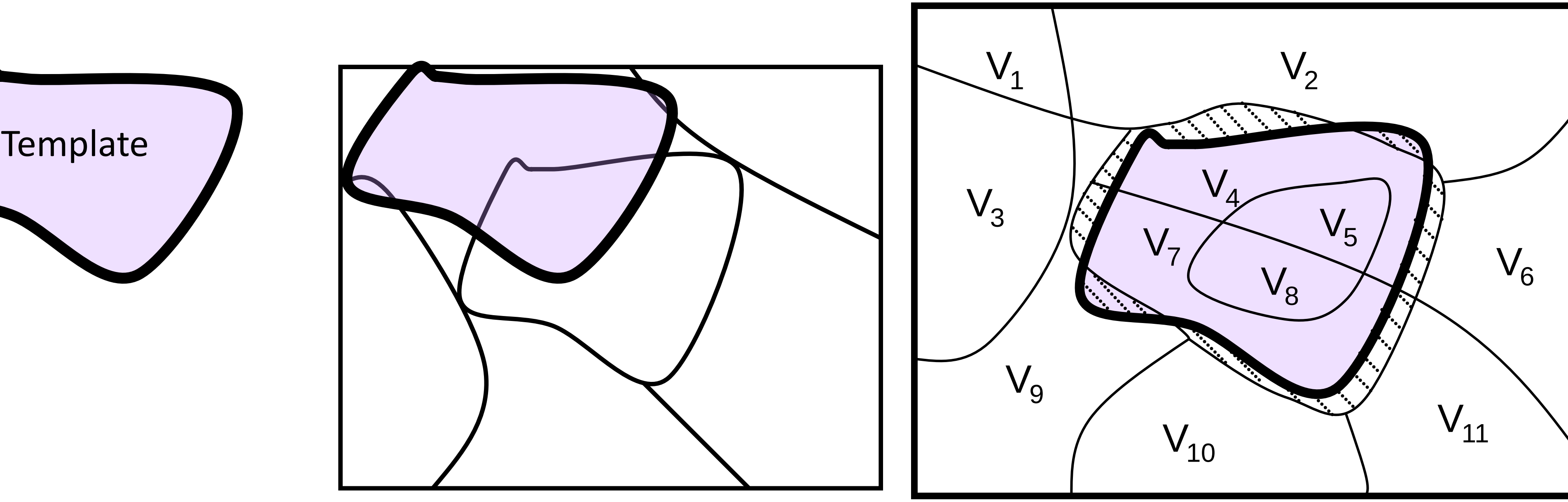
Example

- Compare volume with over-segmentation of video
  - Alignment of regions +
  - Consistency of motion distributions

Ke *et al.* IJCV 2010

Example: Naïve volumetric approach

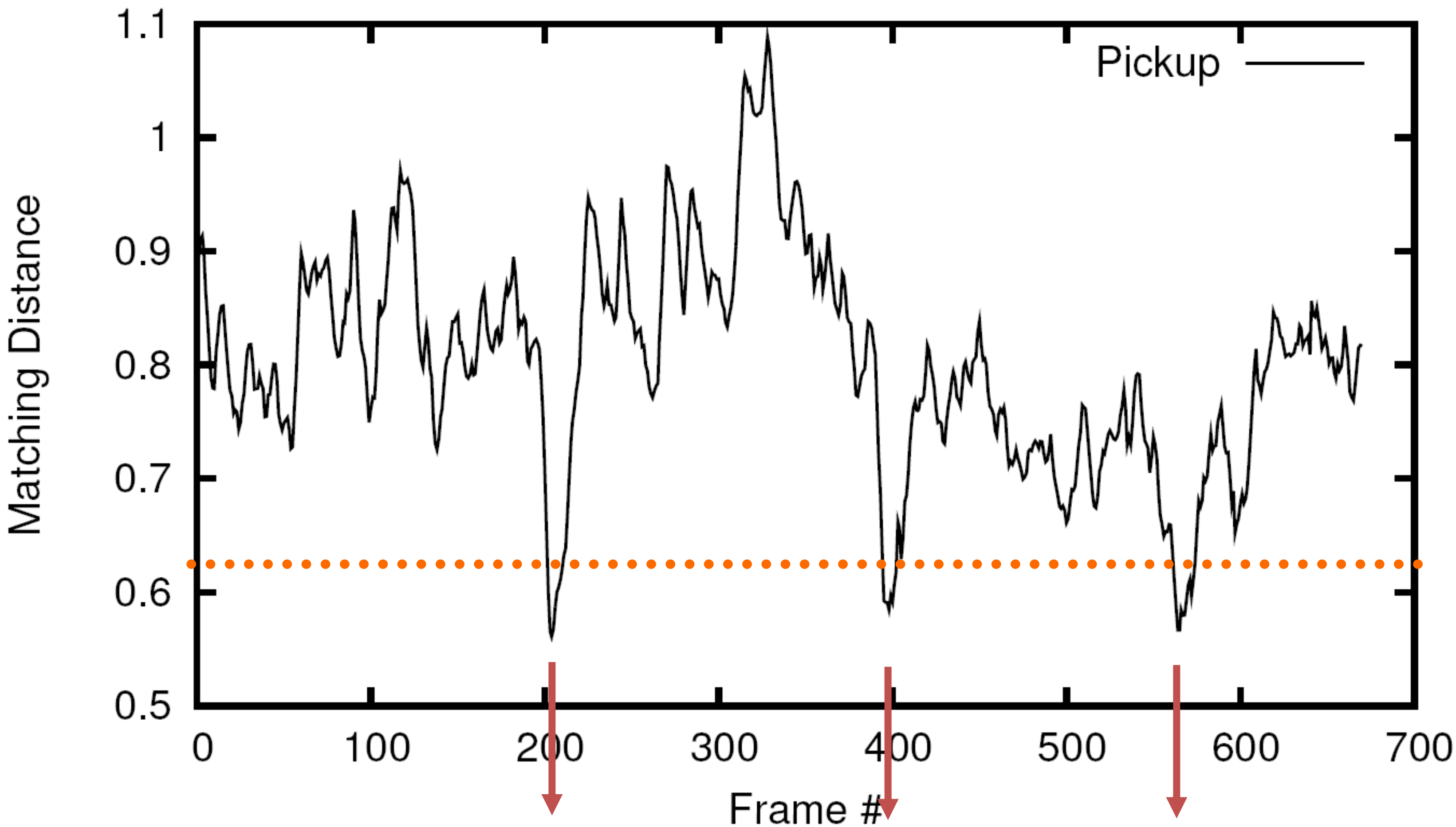
$V = V_1 \cup V_2 \cup \ldots \cup V_n$



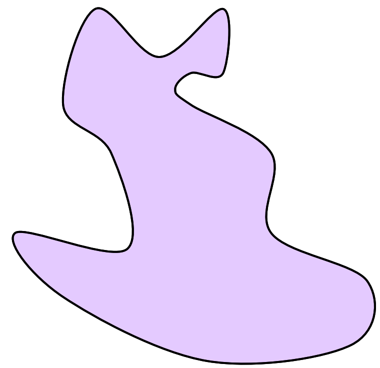
$$d(T,V_i) = \begin{cases} |T \cap V_i| & \text{if } |T \cap V_i| < |V_i|/2 \\ |V_i - T \cap V_i| & \text{otherwise} \end{cases}$$

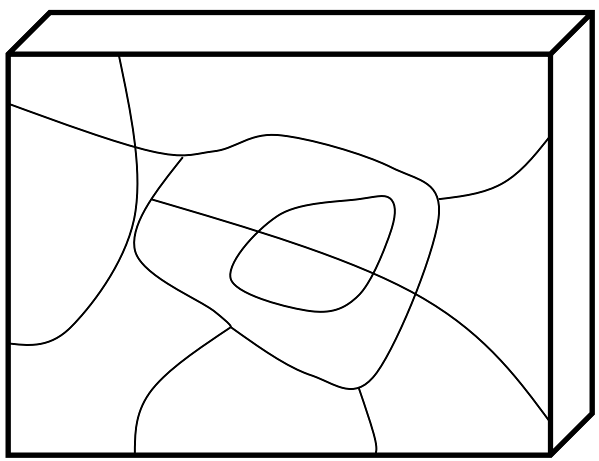
$$d(T,V) = \sum_i d(T,V_i)$$

Naïve volumetric approach

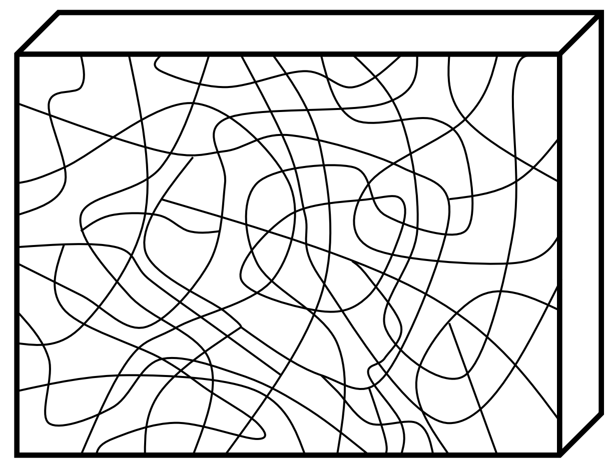


A little better: Normalization for variation in granularity of space-time segmentation





•Normal Over-segmentation

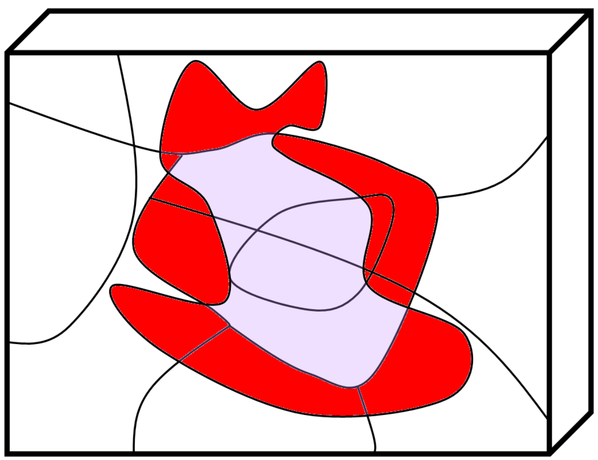


•Extreme Over-segmentation

A little better: Normalization for variation in granularity of space-time segmentation

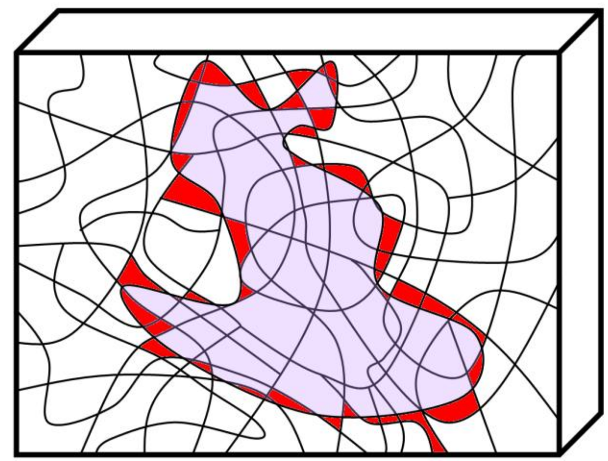
$$d_{shape} = \frac{d(T,V)}{E_{\mathcal{T}}[d(\cdot,V)]}$$

•E [distance] is large



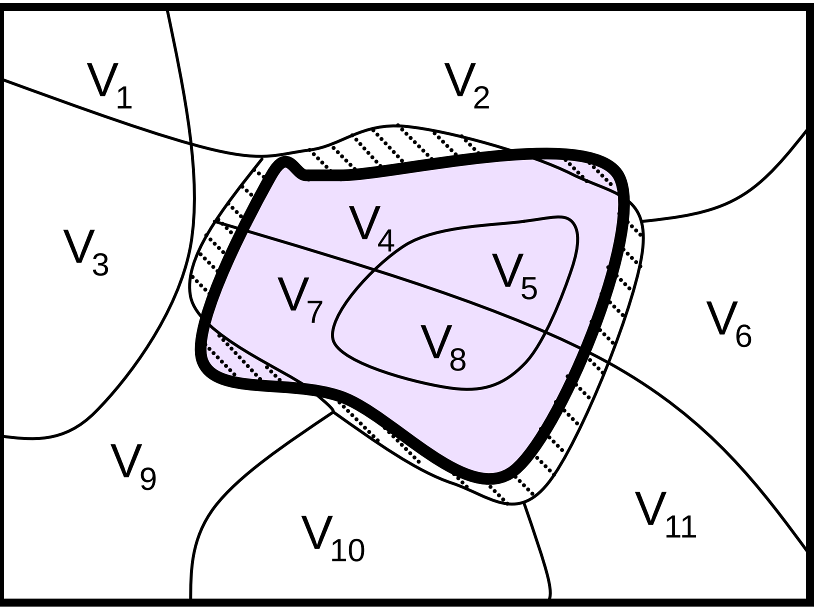
•Normal Over-segmentation

•E [distance] is small



•Extreme Over-segmentation

Much better: Incorporate motion consistency



$$d_{shape} = \frac{d(T,V)}{E_{\mathcal{T}}[d(\cdot,V)]}$$

$$d_{flow}(T,V) = \sum_{P_j \subset T} d_{ST}(P_j)$$

$d_{ST}(P)$ = motion inconsistence within small block P between template and input space-time volume

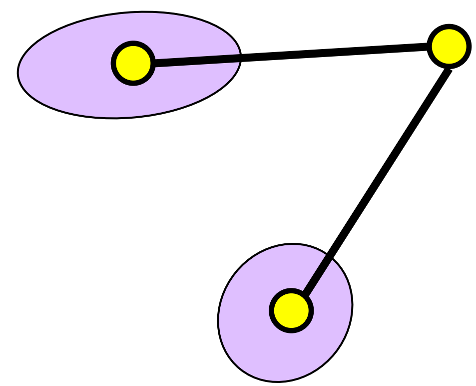
$$\boldsymbol{d}(\boldsymbol{T},\boldsymbol{V}) = \alpha \boldsymbol{d}_{shape}(\boldsymbol{T},\boldsymbol{V}) + (1-\alpha)\boldsymbol{d}_{flow}(\boldsymbol{T},\boldsymbol{V})$$

Issues with generalization and one possible fix (but not very satisfactory)

$$L^* = \underset{L}{\operatorname{argmin}} \left(\sum_{i=1}^{n} a_i(l_i) \quad + \sum_{(v_i,v_j)\in E} d_{ij}(l_i, l_j) \right)$$

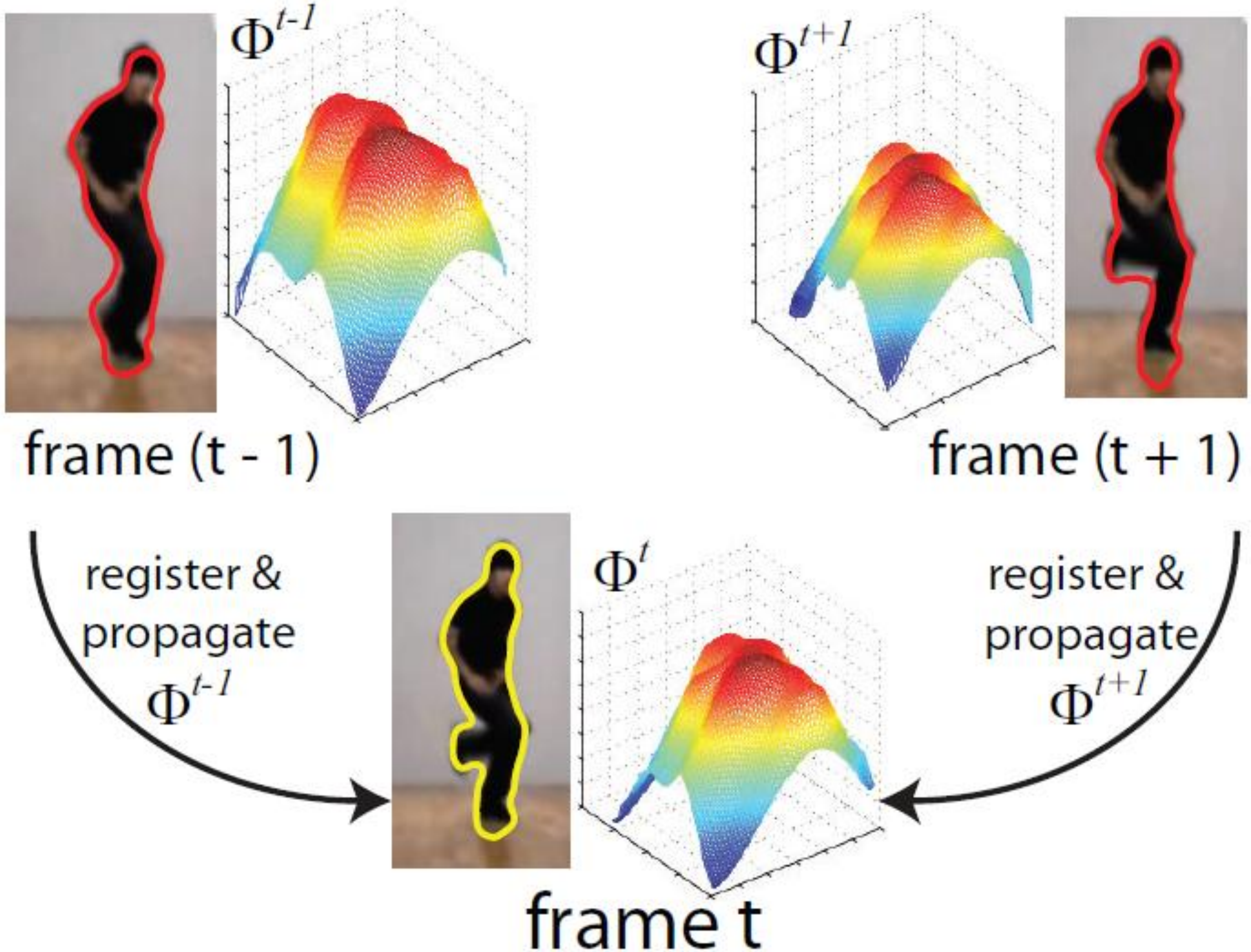
Shape + Flow Correlation

Gaussian Distribution



- Not robust to variations (different actors, viewpoint, speed…)
- Attempted fix: Parts-based representation + representation of deformations

About extracting space-time volumes: One example



Juan Carlos Nieble, Bohyung Han, Li Fei-Fei . Efficient Extraction of Human Motion Volumes by Tracking. CVPR 2010.

About extracting space-time volumes: One example

Juan Carlos Nieble, Bohyung Han, Li Fei-Fei . Efficient Extraction of Human Motion Volumes by Tracking. CVPR 2010.

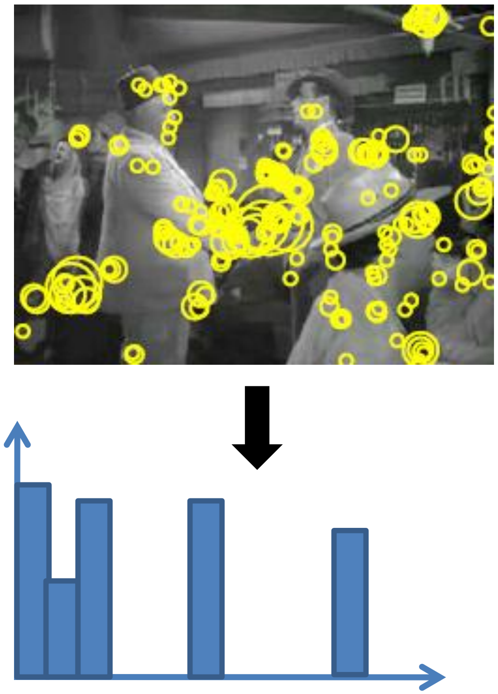
Lessons?

- Plus:
  - Can operate with very few examples
  - Incorporate explicit representation of spatial and temporal structure
  - Does not require explicit tracking, motion estimation, or feature points
  - Well suited for detection tasks
- Minus:
  - Cannot generalize well, i.e., build models from many examples
  - Expensive? Cannot reuse data across models
  - Not well suited for classification tasks

What to do?

- Plus:
  - Can generalize well, e.g., can learn from large sets of examples
  - Fast, can reuse most data across classes
  - Well suited for classification tasks
- Minus:
  - Does not incorporate strong representation of spatial and temporal structure
  - Cannot operate with very few examples
  - Not well suited for detection tasks

- Plus:
  - Can operate with very few examples (1!)
  - Incorporate explicit representation of spatial and temporal structure
  - Does not require explicit tracking, motion estimation, or feature points
  - Well suited for detection tasks
- Minus:
  - Cannot generalize well, i.e., build models from many examples
  - Expensive? Cannot reuse data across models
  - Not well suited for classification tasks



2D spatial relations
Temporal consistency
3D spatial relations

Bags of features
Histograms
……

Trajectory fragments

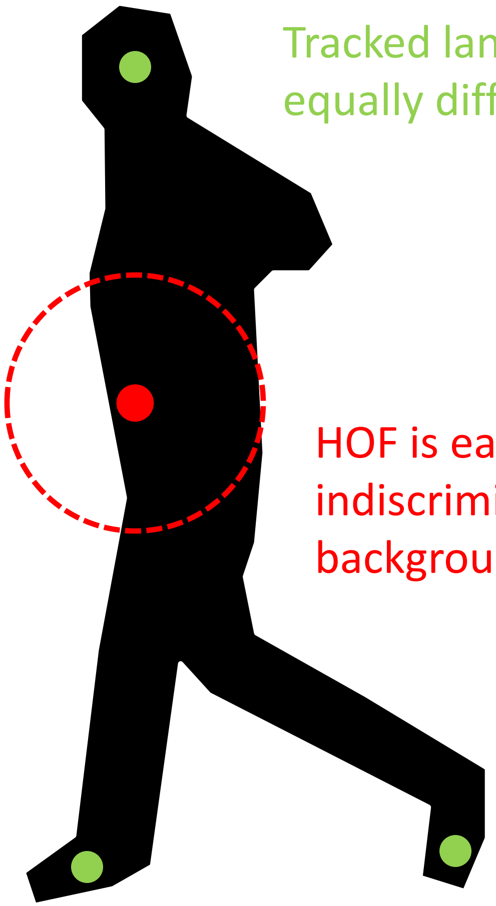
Space-time volume
Flow/shape comparison
…….

Issues and examples

- How to represent distribution of local motion patterns?
- How to represent both temporal and spatial consistency?
- How to train classifiers?
- Examples:
  - Using trajectory fragments
  - Using implicit human motion model

Silhouettes are nicely attached to the action, but difficult to compute

Tracked landmarks are also attached, but equally difficult to compute



HOF is easier to compute, but indiscriminately lumps foreground and background together

Could we find a feature that is both easy to compute and attached?

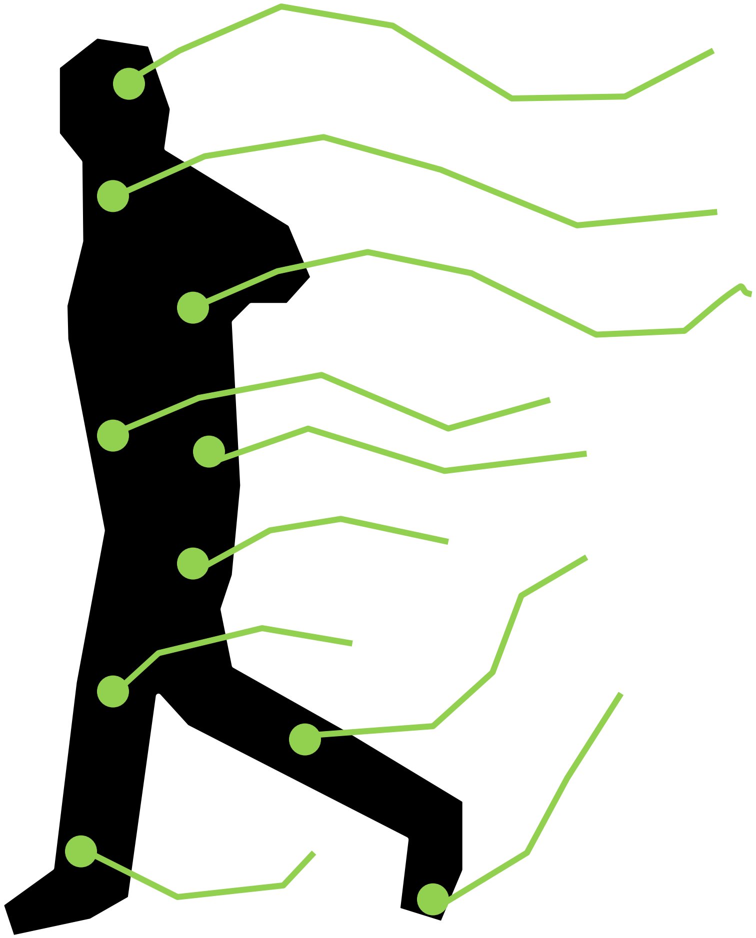
Pyry Matikainen

Quantized trajectory fragments

Avoids difficulty of tracking known landmarks by blindly tracking with KLT

Trajectories are intrinsically attached

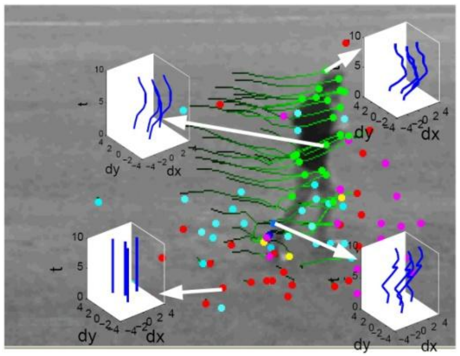
Treats trajectories statistically rather than structurally

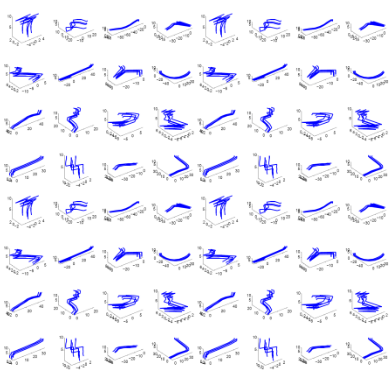


Pyry Matikainen

Overview

Trajectory fragment extraction





Fragment dictionary (codebook)

Nearest neighbor

Fragment Labels

Accumulate

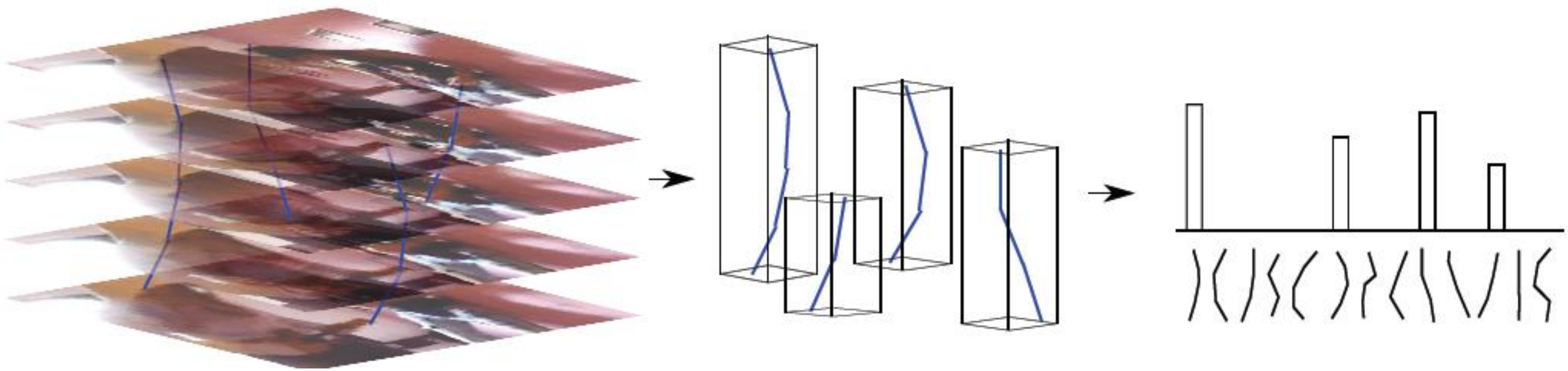
Label Histogram

SVM

Classification

Pyry Matikainen

Overview

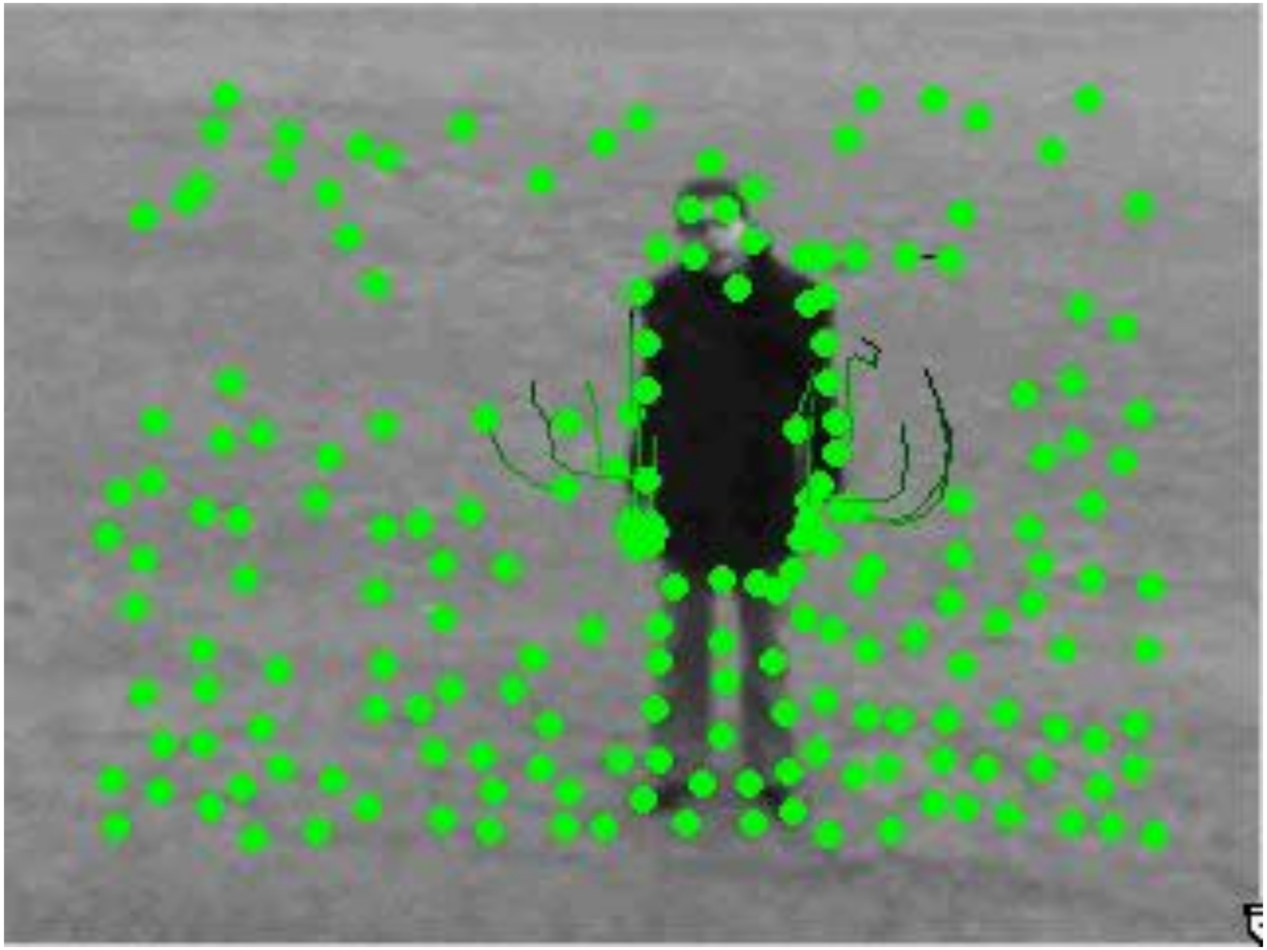


Feature tracking

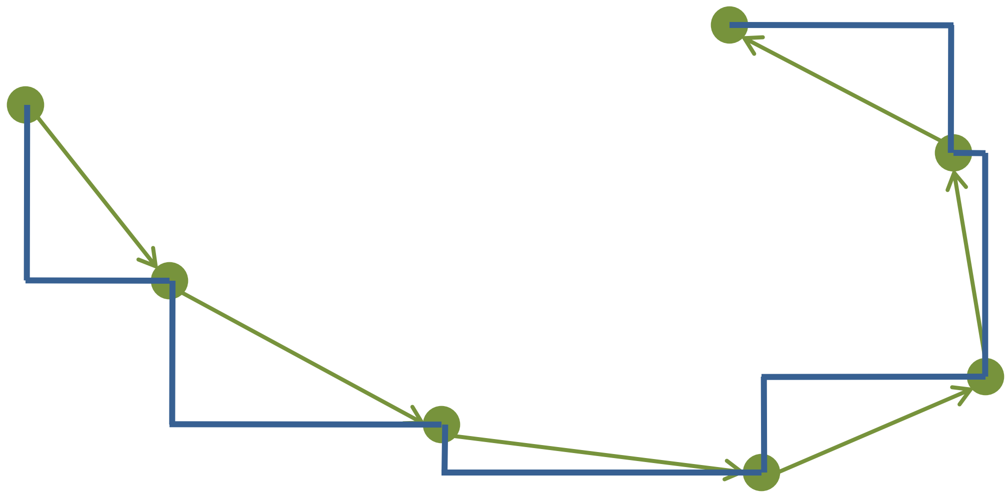
Bags of trajectories

Quantized trajectory clusters

Pyry Matikainen



Pyry Matikainen

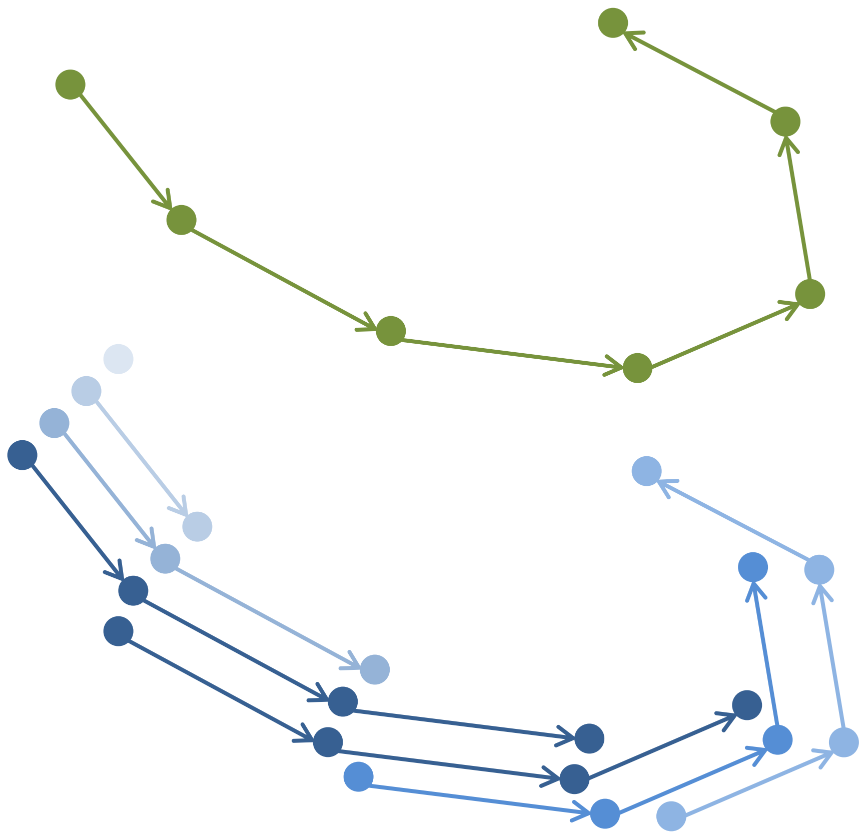


dx	[0,	0,	0,	3,	5,	6,	4,	-1,	-4]
dy	[0,	0,	0,	-4,	-3,	-1,	2,	5,	3]

[0, 0,	0, 0,	0, 0,	3, -4,	5, -3,	6, -1,	4, 2,	-1, 5,	-4, 3]
[V_{-9}	V_{-8}	V_{-7}		…			V_{-1}	V_0]

Trajectory fragment: derivatives packed into a vector

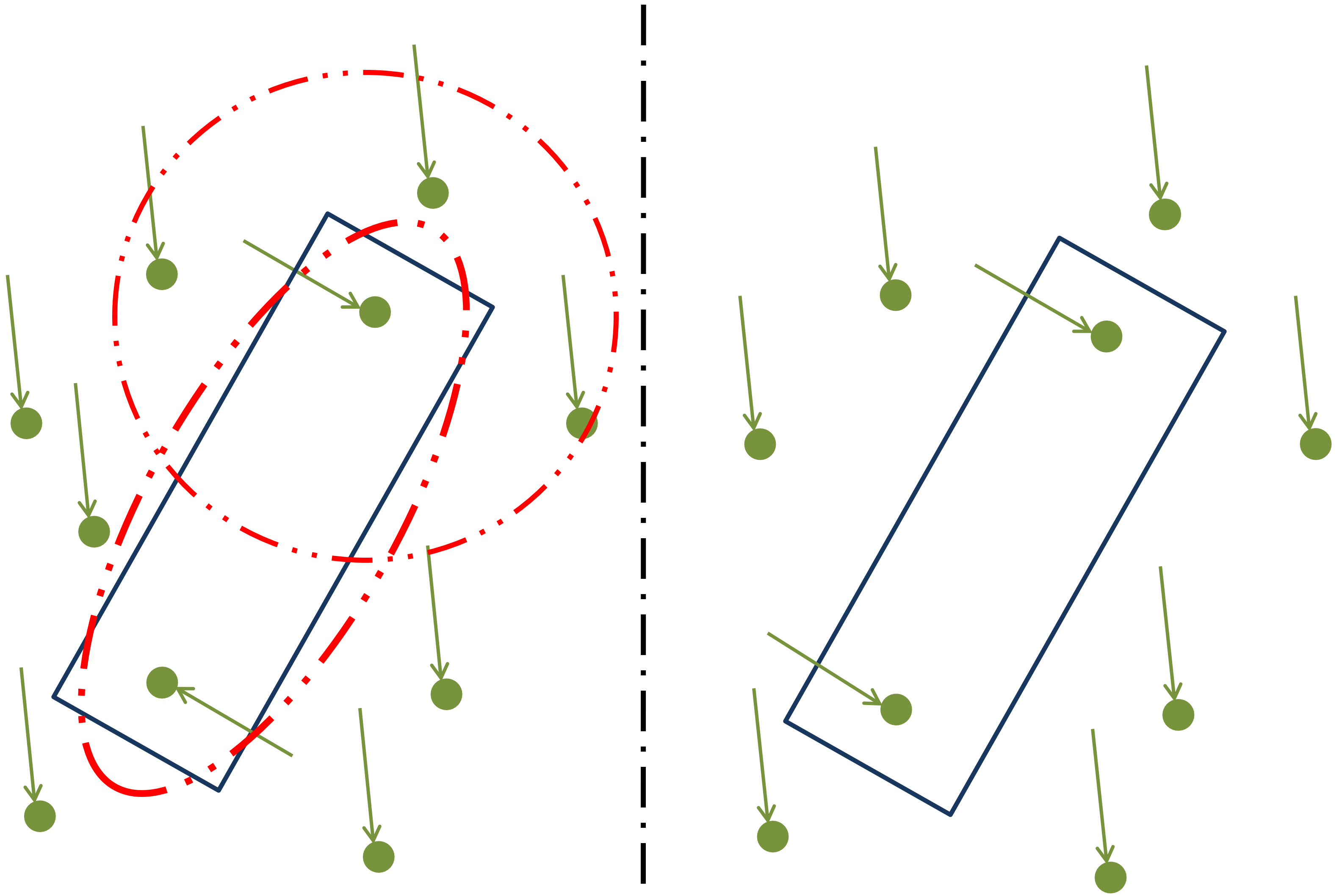
Pyry Matikainen



Each trajectory produces many fragments

#fragments ~ #features x #frames

Pyry Matikainen

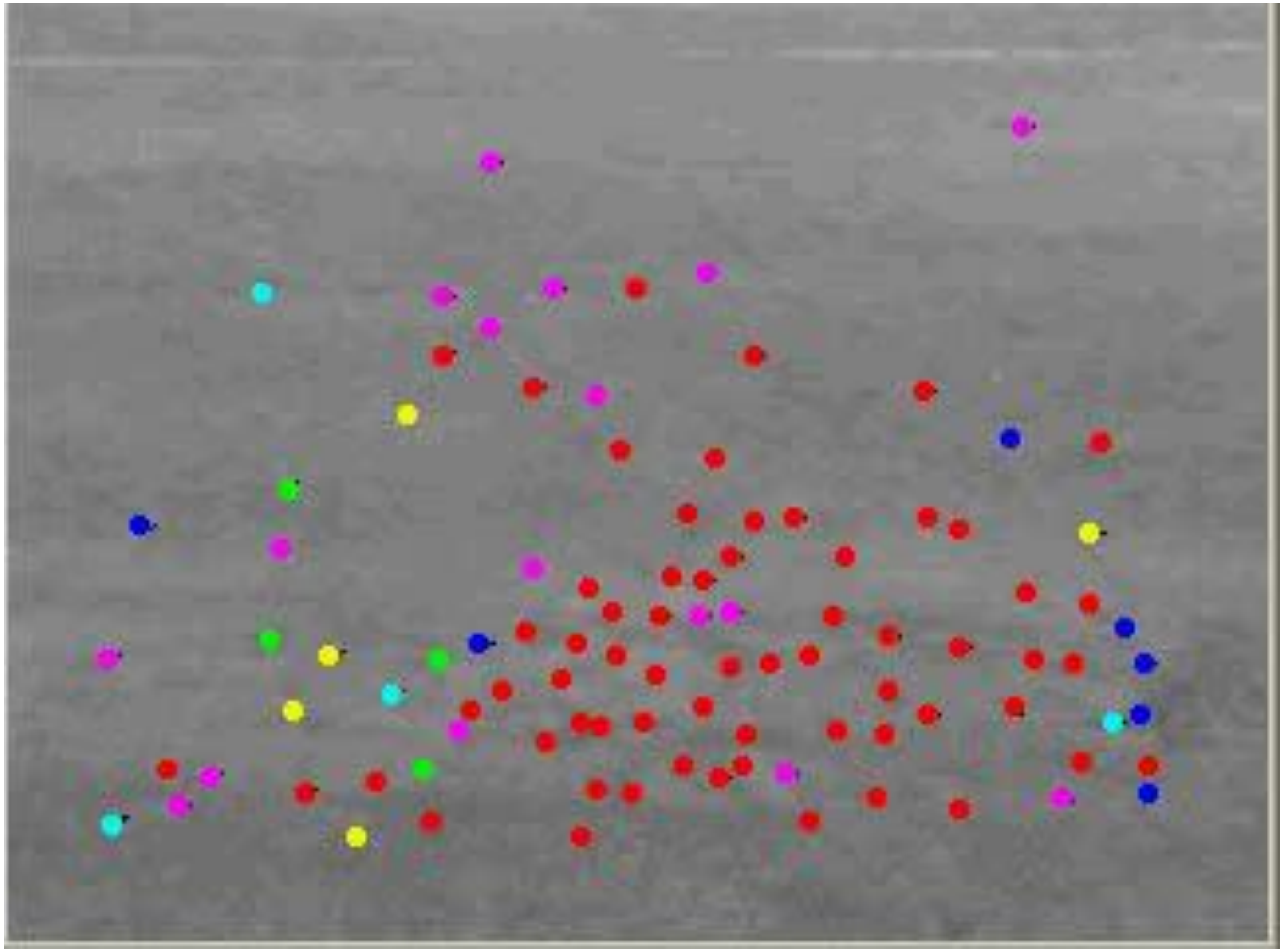


Problem: trajectory fragments have no local context

Solution: augment with transforms of their motion clusters

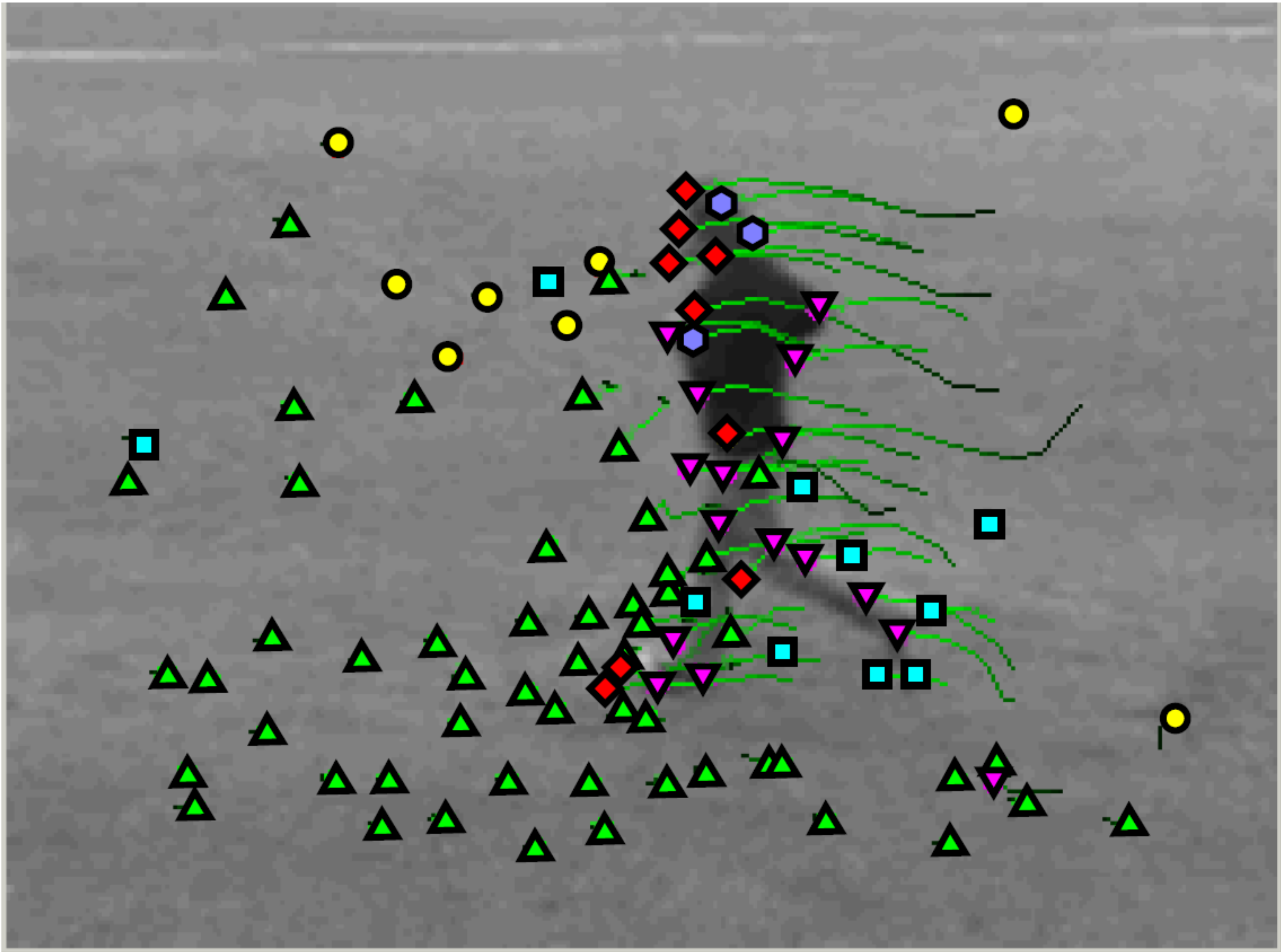
Pyry Matikainen

Example Motion Clustering

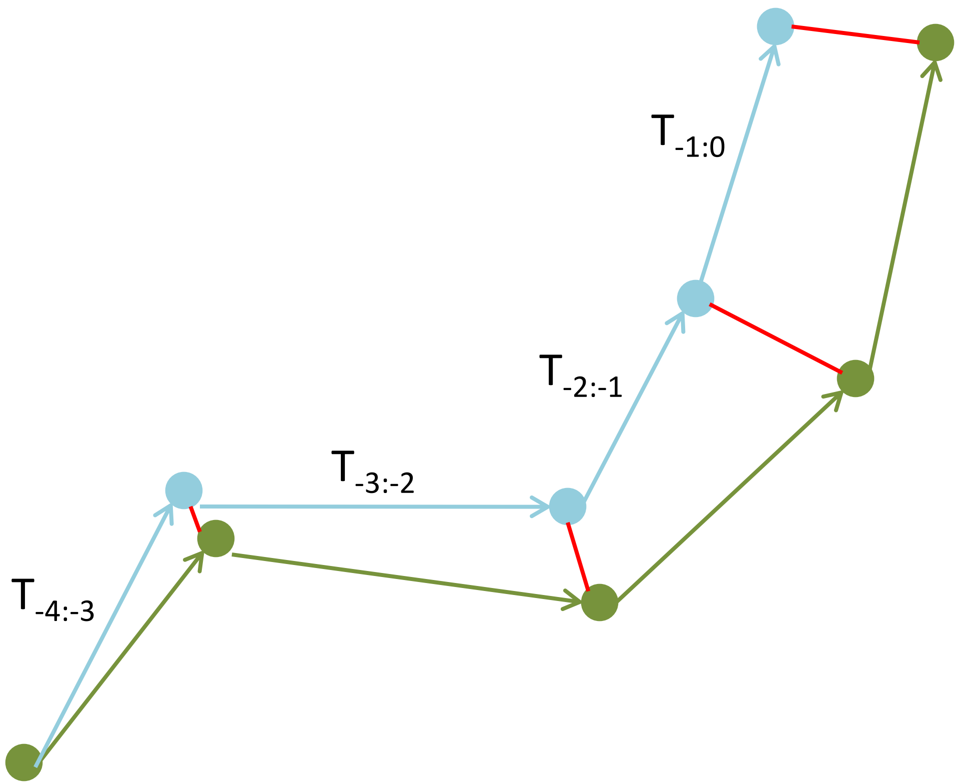


Pyry Matikainen

Example Motion Clustering

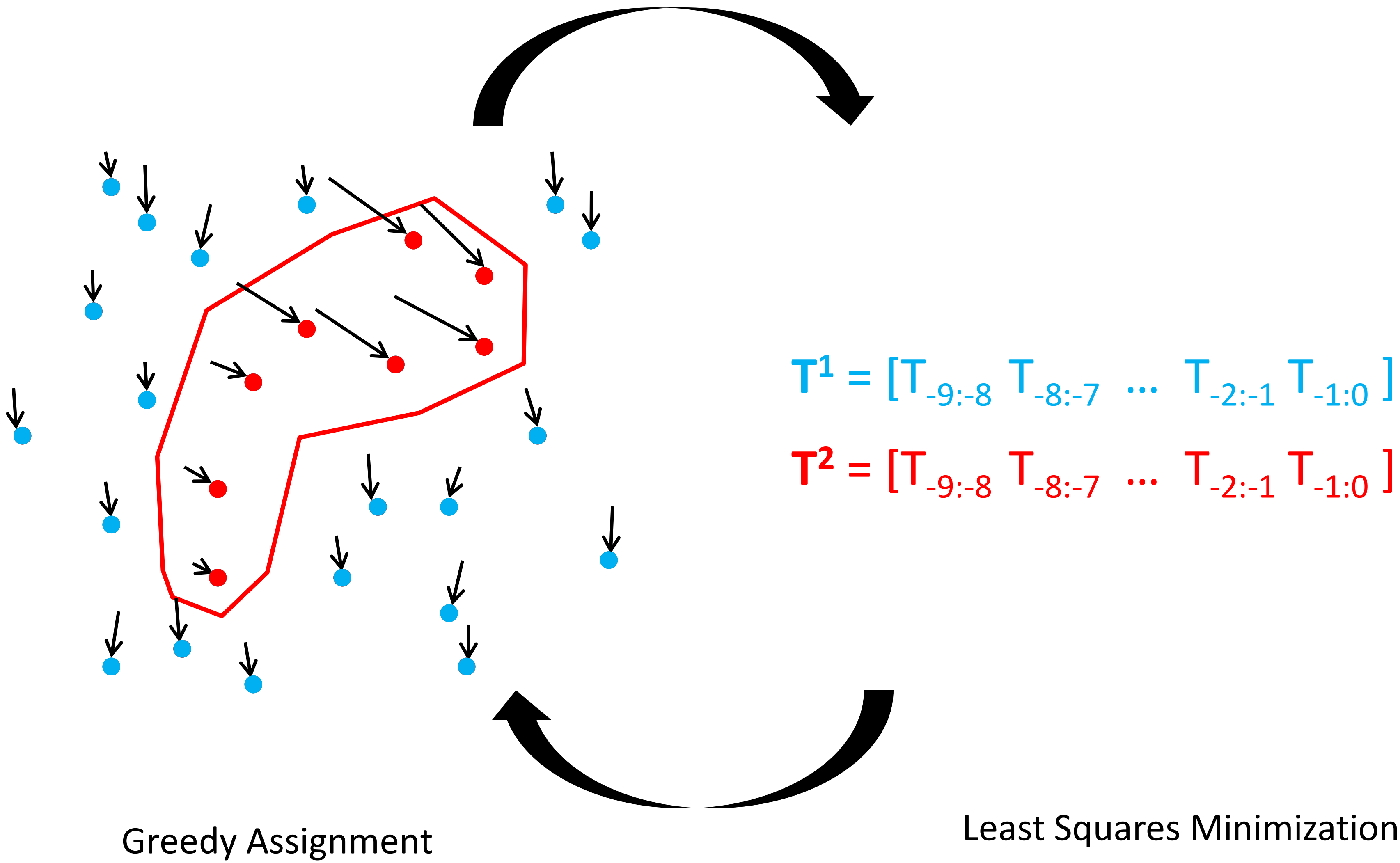


Pyry Matikainen



Error of a fragment given a set of transforms

Pyry Matikainen



Motion cluster estimation

Pyry Matikainen

$$[\ V_{-9}\quad V_{-8}\quad V_{-7}\quad \underbrace{\dots\quad V_{-i}\quad \dots}_{[\,dx\ \ dy\,]}\quad V_{-1}\quad V_0\ ]$$

Trajectory fragment

$$[\ V_{-9}\quad V_{-8}\quad V_{-7}\quad \underbrace{\dots\quad V_{-i}\quad \dots}_{[\,dx\ \ dy\ \ \underbrace{a_{1,1}\ \ a_{2,1}\ \ a_{1,2}\ \ a_{2,2}}\,]}\quad V_{-1}\quad V_0\ ]$$

Affine-Augmented
Trajectory fragment

$$A_{-i} = T_{-(i+1):-i} = \begin{bmatrix} a_{1,1} & a_{1,2} & t_x \\ a_{2,1} & a_{2,2} & t_y \\ 0 & 0 & 1 \end{bmatrix}$$

Motion cluster transforms

AA (affine augmented) fragment

Trajectory fragment and affine transform packed into vector

Pyry Matikainen

$[\ V_{-9}\quad V_{-8}\quad V_{-7}\quad \ldots\quad V_{-i}\quad \ldots\quad V_{-1}\quad V_{0}\ ]$

$[\ V_{-9}\quad V_{-8}\quad V_{-7}\quad \ldots\quad V_{-i}\quad \ldots\quad V_{-1}\quad V_{0}\ ]$

$[\ V_{-9}\quad V_{-8}\quad V_{-7}\quad \ldots\quad V_{-i}\quad \ldots\quad V_{-1}\quad V_{0}\ ]$

$[\ V_{-9}\quad V_{-8}\quad V_{-7}\quad \ldots\quad V_{-i}\quad \ldots\quad V_{-1}\quad V_{0}\ ]$

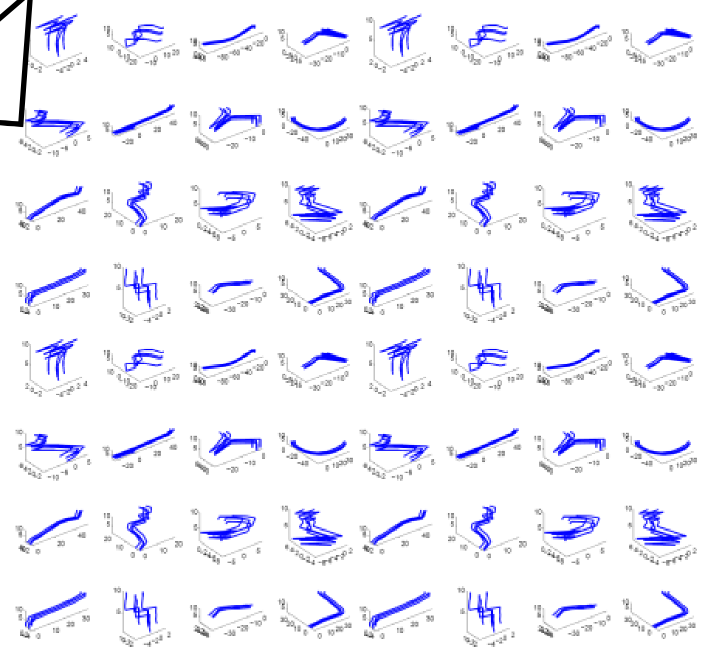
$[\ V_{-9}\quad V_{-8}\quad V_{-7}\quad \ldots\quad V_{-i}\quad \ldots\quad V_{-1}\quad V_{0}\ ]$

$[\ V_{-9}\quad V_{-8}\quad V_{-7}\quad \ldots\quad V_{-i}\quad \ldots\quad V_{-1}\quad V_{0}\ ]$

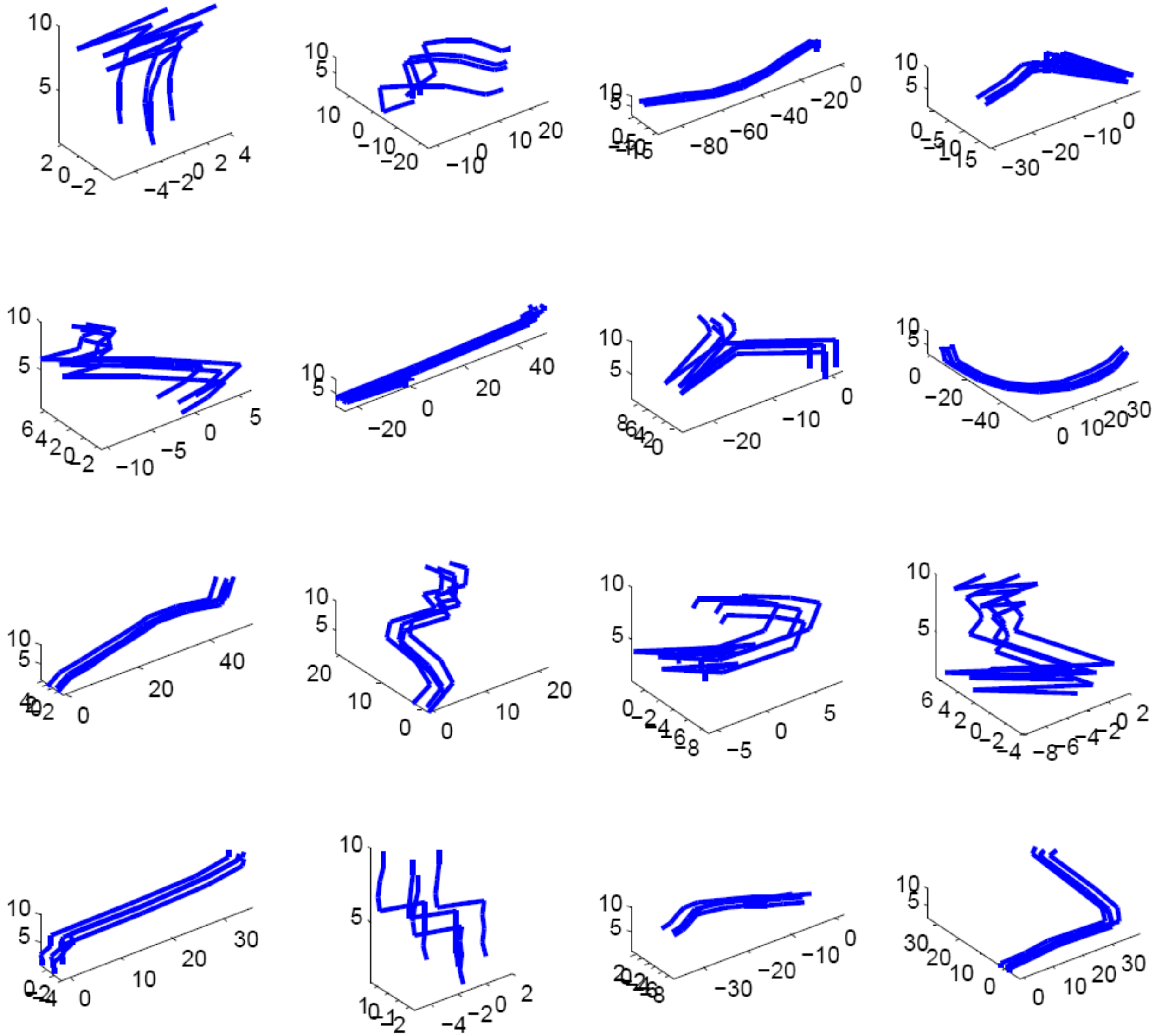
Training set fragments

k-means clustering

Library

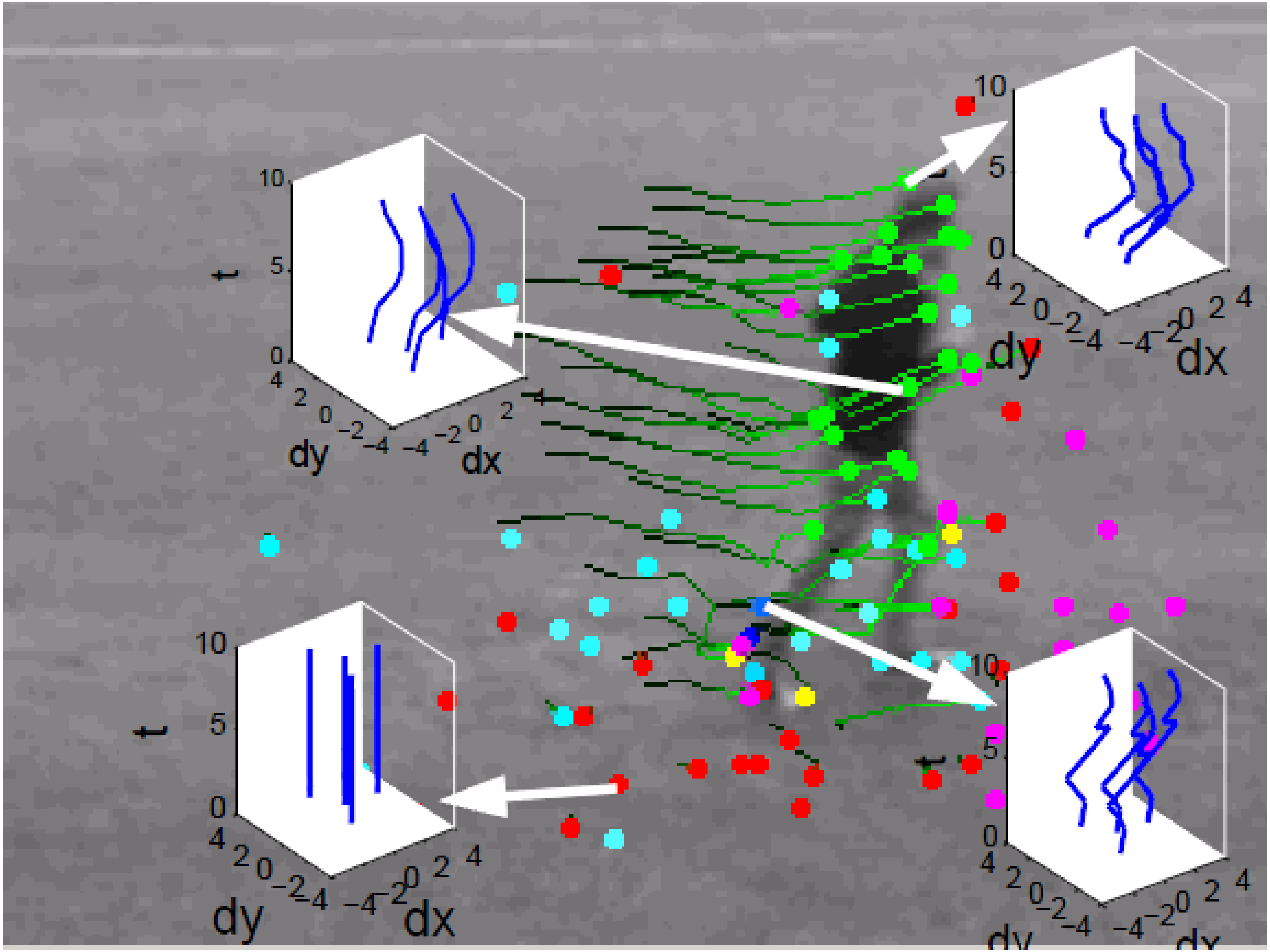


Pyry Matikainen



Example dictionary

Pyry Matikainen



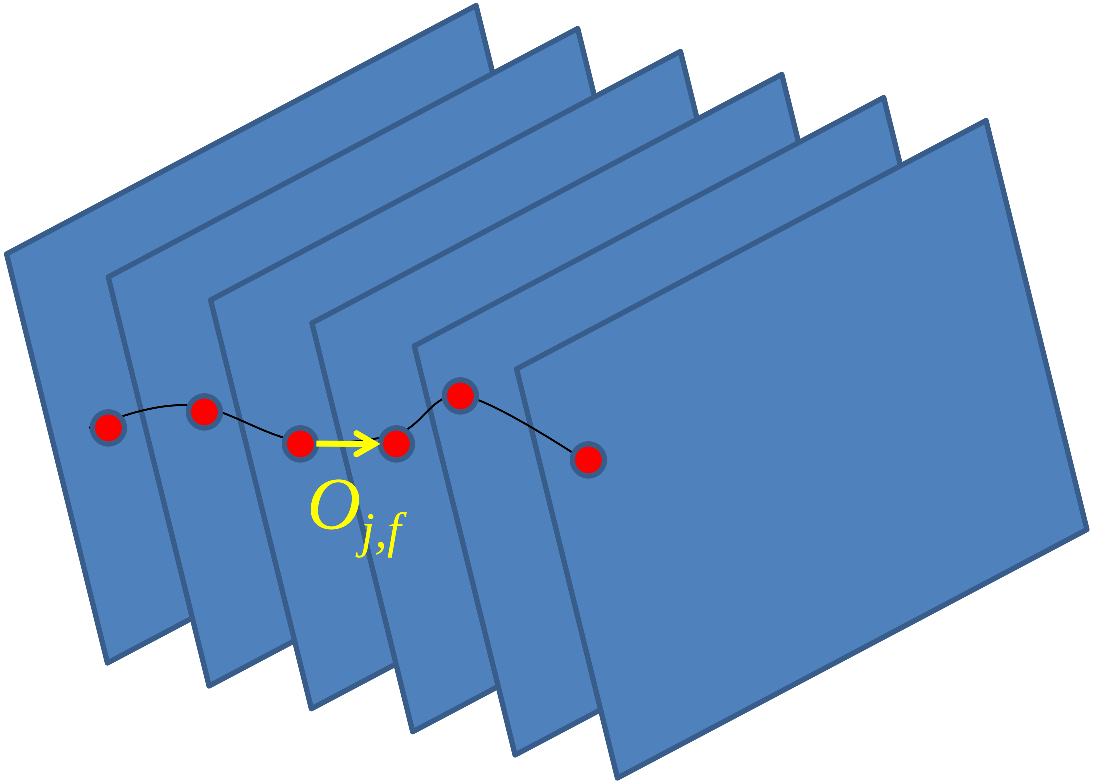
Pyry Matikainen

Hollywood Actions Dataset [Laptev et al. 2008]

- 1000 AA quantized fragments, 100 features, 6 motion clusters
- Linear SVM classification on histograms
- Histograms accumulated over entire clips

Action	Quantized fragments	Quantized fragments (lax SVM)	HoF
Total	**31.1**	**27.2**	**27.1**
SitDown	4.5	13.6	20.7
StandUp	69.0	42.9	40.0
Kiss	71.4	42.9	36.5
AnswerPhone	0.0	35.5	24.6
HugPerson	0.0	23.5	17.4
HandShake	5.3	5.3	12.1
SitUp	11.1	11.1	5.7
GetOutCar	7.7	7.7	14.9

Another example



Feature = sequence O of velocities over fixed length (e.g., 500)

$O_{j,f}$ = velocity at frame j of feature f

[Messing, Pal, and Kautz, ICCV 2009]

Probabilistic model

- Each feature generated by a set of mixture components
- M = set of N_m feature component
- $M_{i,f}$ = feature f is generated by mixture component i

$$P(O_f \mid A) = \sum_{i=1}^{i=N_m} \underbrace{P(M_{i,f} \mid A)}_{\text{Mixture weights for action } A} \underbrace{P(O_f \mid M_{i,f})}_{\text{Model for mixture component } i}$$

Mixture weights for action A

Model for mixture component i

[Messing, Pal, and Kautz, ICCV 2009]

Example mixture components

Probabilistic model

- Markov model for the velocity features:

$$P(O_f \mid M_{i,f}) = P(O_{0,f} \mid M_{i,f}) \prod_{t=1}^{t=T} P(O_{t,f} \mid O_{t-1,f}, M_{i,f})$$

Initial velocity model

Prediction model from time t-1 to time t

$$P(O_f \mid A) = \sum_{i=1}^{i=N_m} P(M_{i,f} \mid A) P(O_f \mid M_{i,f})$$

$$P(O_f \mid A) = \sum_{i=1}^{i=N_m} P(M_{i,f} \mid A) P(O_f \mid M_{i,f}) P(O_{0,f} \mid M_{i,f}) \prod_{t=1}^{t=T} P(O_{t,f} \mid O_{t-1,f}, M_{i,f})$$

[Messing, Pal, and Kautz, ICCV 2009]

Probabilistic model

- Assuming Naïve Bayes independence of the trajectories

$$P(O \mid A) = \prod_{f=1}^{f=N_f} P(O_f \mid A)$$

$$P(O_f \mid A) = \sum_{i=1}^{i=N_m} P(M_{i,f} \mid A) P(O_f \mid M_{i,f}) P(O_{0,f} \mid M_{i,f}) \prod_{t=1}^{t=T} P(O_{t,f} \mid O_{t-1,f}, M_{i,f})$$

$$P(O \mid A) = \prod_{f=1}^{N_f} \sum_{i=1}^{i=N_m} P(M_{i,f} \mid A) P(O_f \mid M_{i,f}) P(O_{0,f} \mid M_{i,f}) \prod_{t=1}^{t=T} P(O_{t,f} \mid O_{t-1,f}, M_{i,f})$$

[Messing, Pal, and Kautz, ICCV 2009]

Probabilistic model

- Training:
  - Learn mixture weights and mixture models from labeled data
- Testing:
  - Find A such that $\arg\max_A P(O \mid A)$
  - Note: Classification only

$$P(O_f \mid A) = \sum_{i=1}^{i=N_m} P(M_{i,f} \mid A) P(O_f \mid M_{i,f})$$

Mixture weights for action A

Model for mixture component i

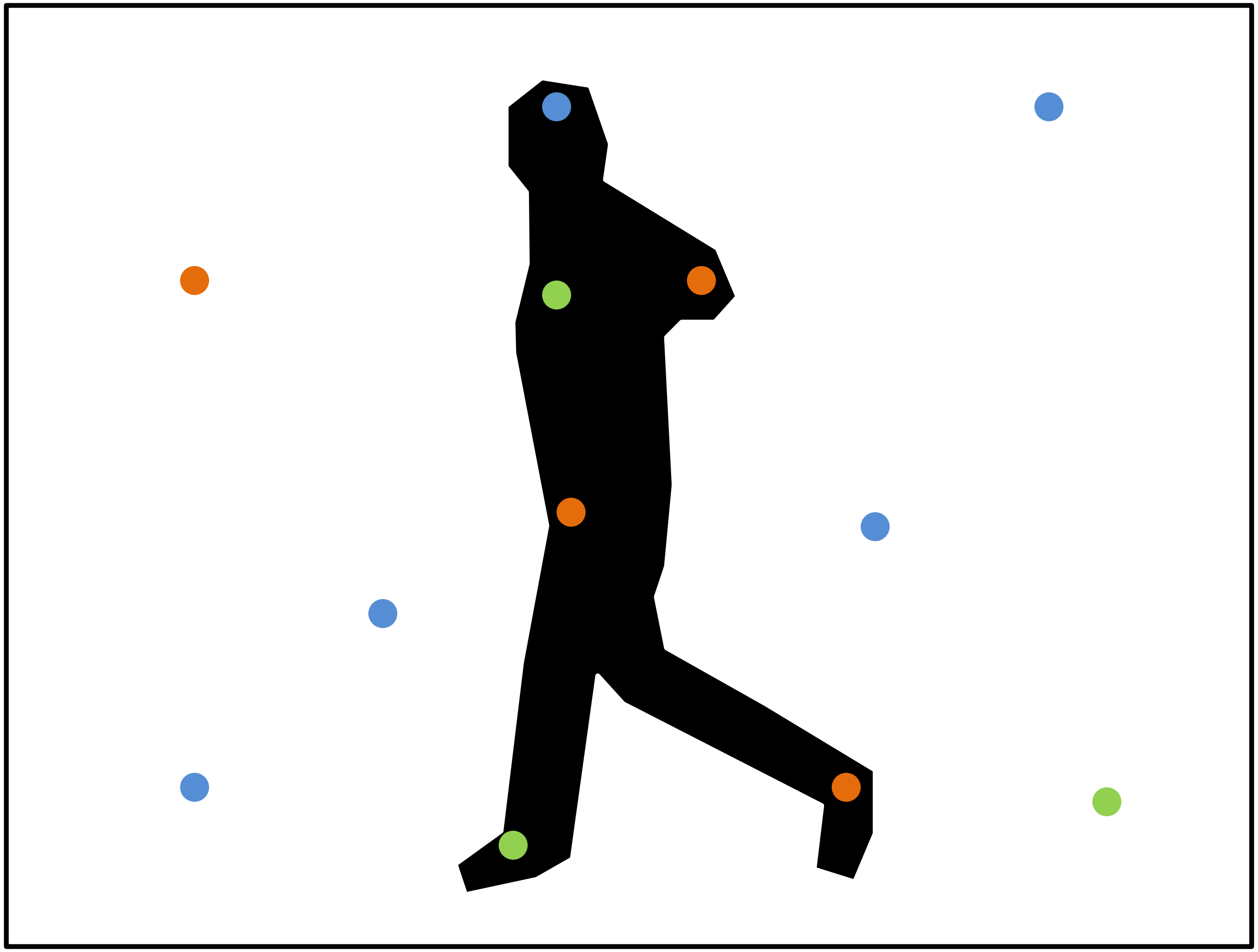
[Messing, Pal, and Kautz, ICCV 2009]

Can we add more structure?

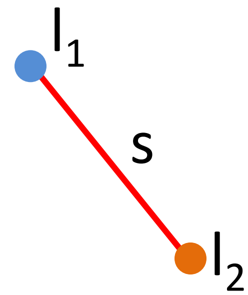
- Using explicit trajectory fragments allowed to make explicit some temporal information
- Can we now add some spatial consistency information?

• P. Matikainen, M. Hebert, and R. Sukthankar. Representing Pairwise Spatial and Temporal Relations for Action Recognition. In ECCV, 2010.
• R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In ICCV, 2009.
• A. Gilbert, J. Illingworth and R. Bowden. Fast Realistic Multi-Action Recognition using Mined Dense Spatio-temporal Features. In ICCV 2009.
• S. Maji and J. Malik. Object detection using a max-margin Hough transform. In CVPR, 2009.

Detailed explanation in next set of slides: Courtesy Pyry Matikainen



Pyry Matikainen



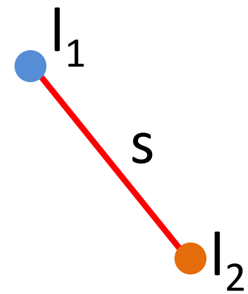
Naïve method: quantize pairs
quantize relationships

Produce new labels which are $(\underbrace{l, l}_{\text{Original labels}}, \underbrace{s}_{\text{Quantized spatial relationships}})$ triples

Original labels

Quantized spatial relationships

Pyry Matikainen



Naïve method: quantize pairs
quantize relationships

$O(L^2S)$ pair labels

# of possible quantized spatial relationships

# of possible feature labels

Pyry Matikainen

$O(L^2S)$ pair labels

1000 labels, 10 relationships

10,000,000 pair labels

Pyry Matikainen

Some approaches to dealing with $O(L^2S)$ pair labels

Restrict the number of relationships

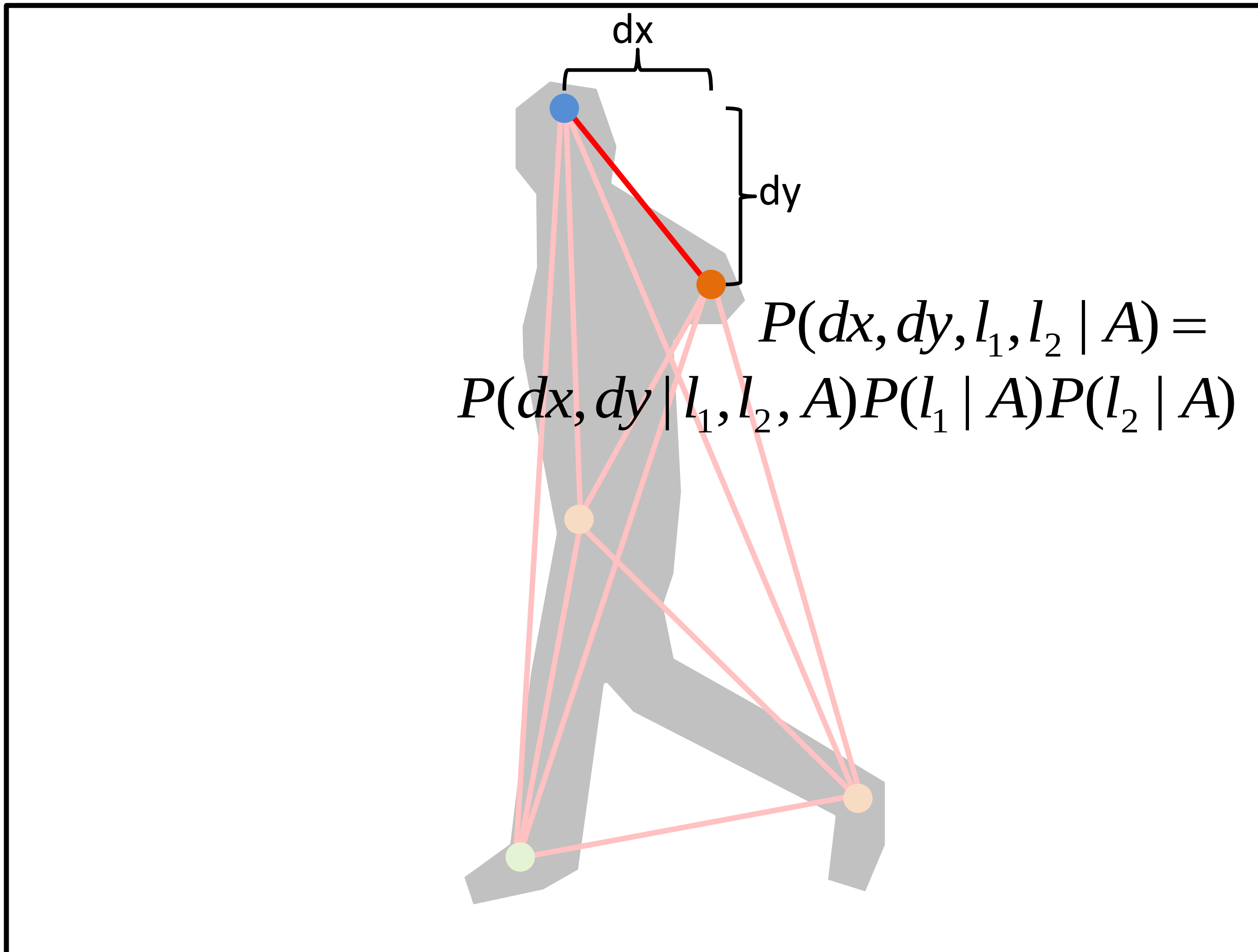
Proximity only, ahead / behind, etc.

(Ryoo and Aggarwal 2009, Savarese et al . 2008)

Restrict the number of labels

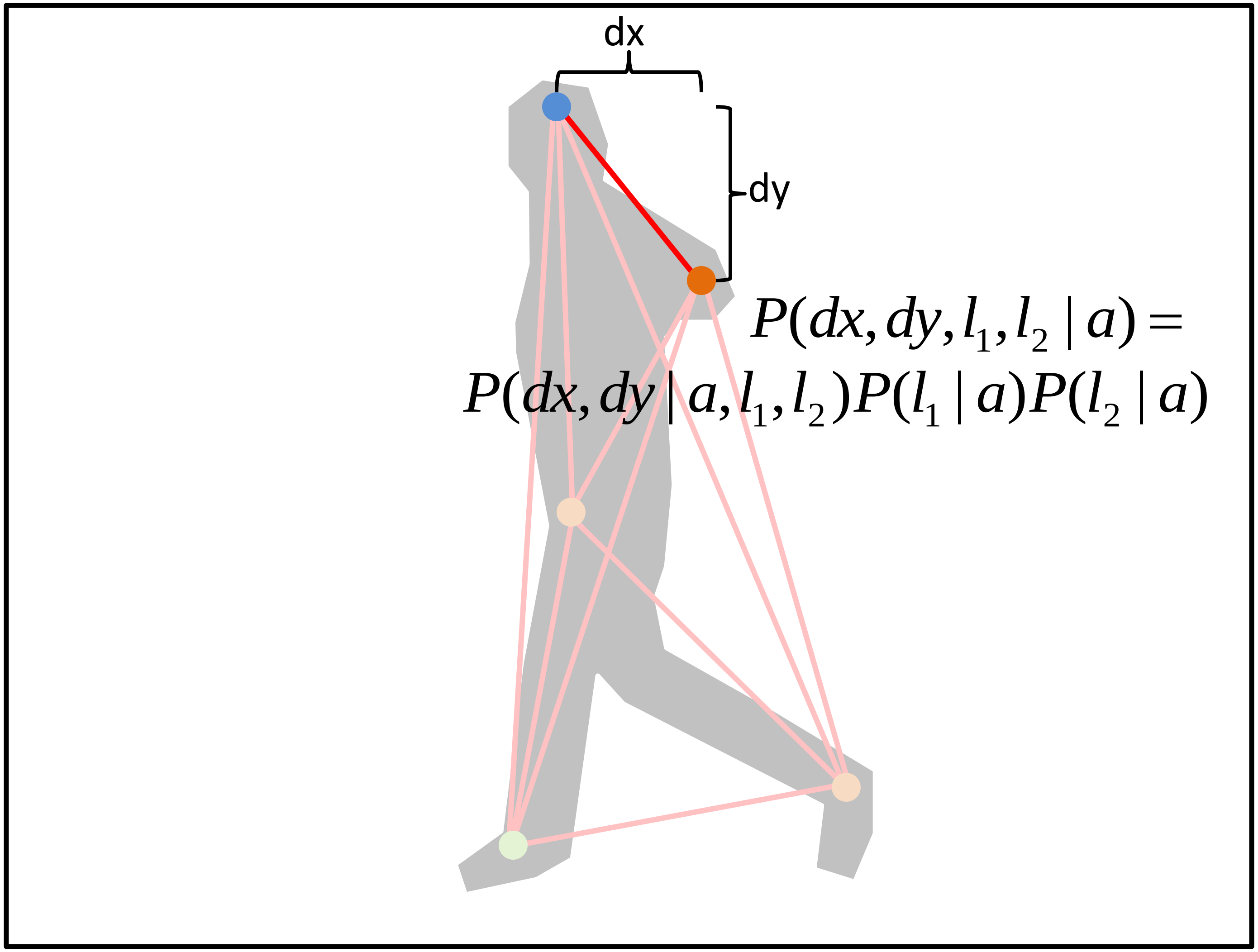
Aggressive quantization, only one label

(Gilbert et al. 2008)



$$P(dx, dy, l_1, l_2 \mid A) = P(dx, dy \mid l_1, l_2, A) P(l_1 \mid A) P(l_2 \mid A)$$

From Matikainen, Hebert, Sukthankar, ECCV 2010



$$P(F|a) = \prod_{f_i \in F} P(l_i|a) \prod_{f_j \in F} P(l_j|l_i, a) P(dx, dy|a, l_i, l_j),$$

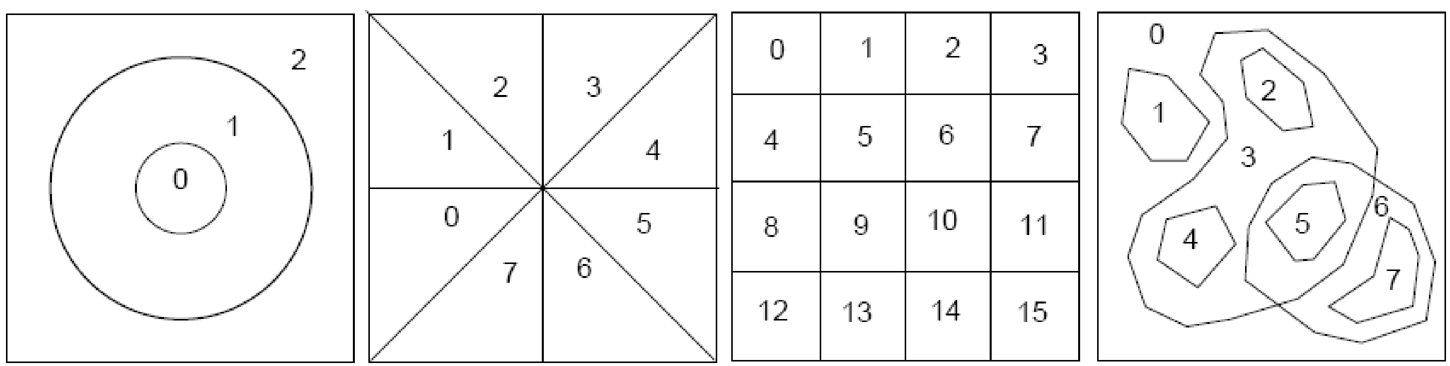
From Matikainen, Hebert, Sukthankar, ECCV 2010

$$P(F|a) = \prod_{f_i \in F} P(l_i|a) \prod_{f_j \in F} P(l_j|l_i, a) P(dx, dy|a, l_i, l_j),$$

$$\log(P(F|a)) = \sum_{f_i \in F} \sum_{f_j \in F} \underbrace{\log(P(dx, dy|a, l_i, l_j))}_{\text{These are the edge weights}} + C$$

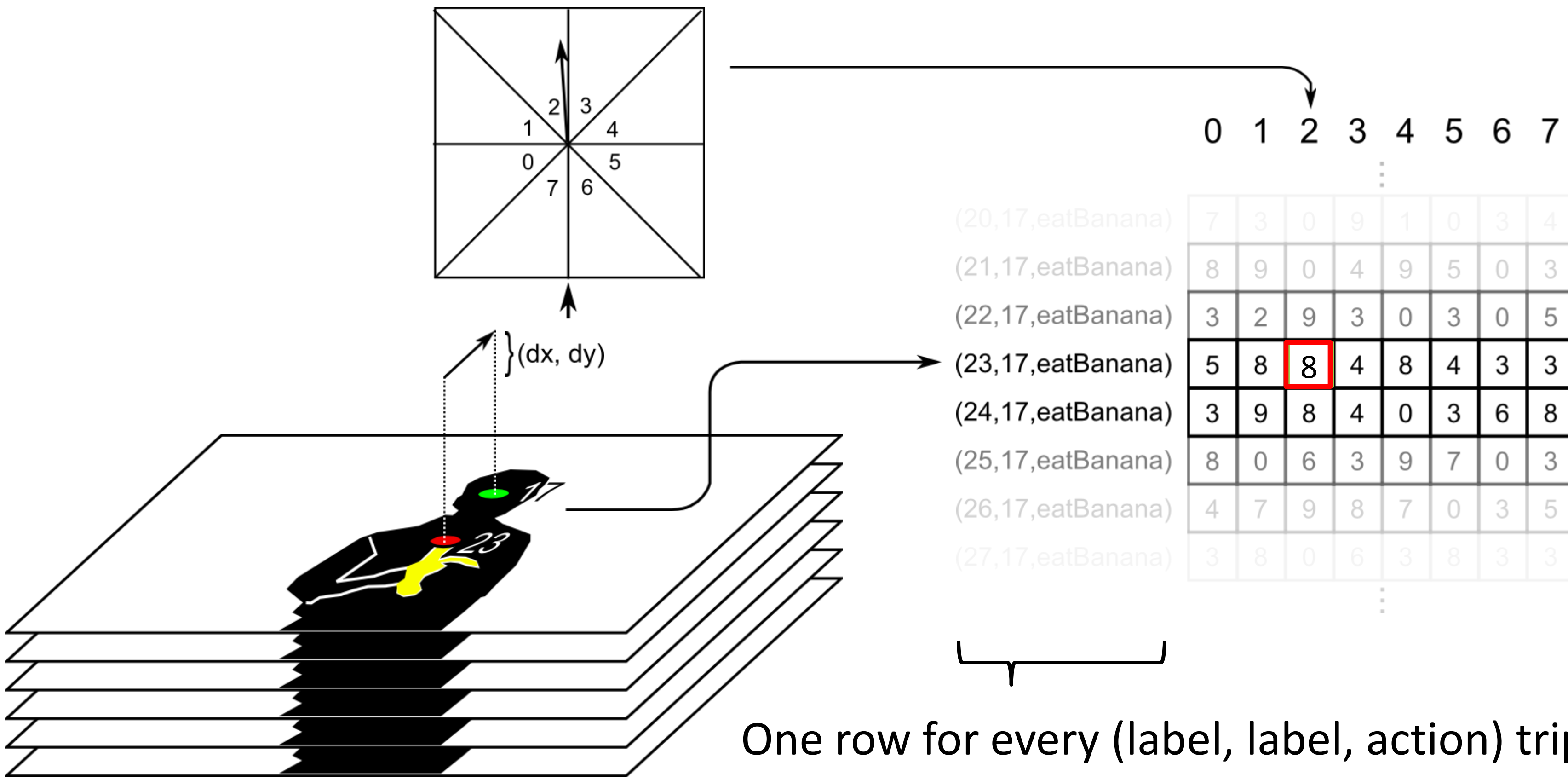
From Matikainen, Hebert, Sukthankar, ECCV 2010

$$\log(P(dx, dy|a, l_i, l_j)) = T_{a,l_i,l_j}[M(dx, dy)]$$



From Matikainen, Hebert, Sukthankar, ECCV 2010

[Matikainen, Hebert, Sukthankar. Representing Pairwise Spatial and Temporal Relations for Action Recognition, ECCV 2010]



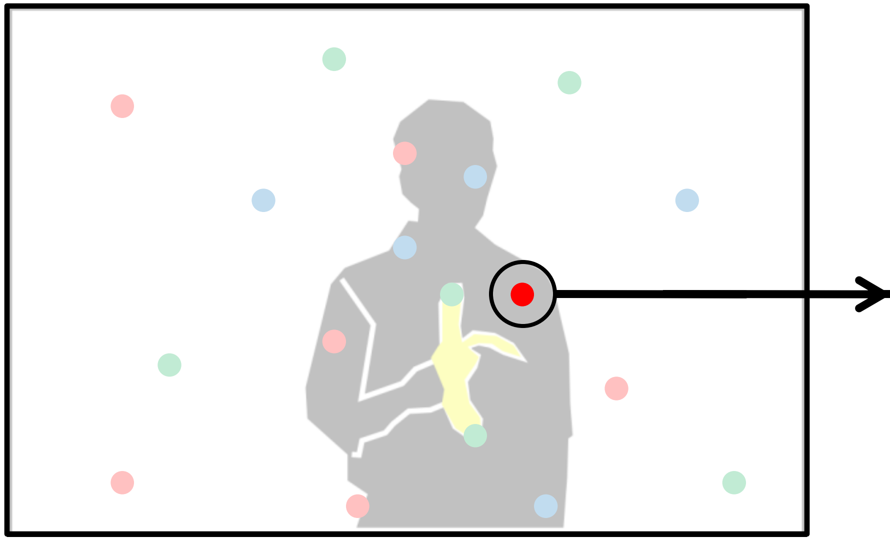
One row for every (label, label, action) triple

Classify on:

$$B_{a,l} = \sum_{f_i \in l} \sum_{f_j} \log(P(dx, dy \mid a, l, l_j))$$

S. Maji and J. Malik. Object detection using a max-margin Hough transform. In CVPR, 2009.

Pyry Matikainen



SIFT

or

MOSIFT

or

STIP

or

Trajectons

etc.

Evaluation Features

STIP – Appearance, k-means quantization
Trajectons – Motion, fixed quantization

Evaluation Datasets

UCF-YT – YouTube videos, low-res, complex
Rochester– Kitchen videos, high-res, simple

(150 videos, 10 classes)

(1600 videos, 11 classes)

Method	UCF-YT	Rochester
STIP-HoG (single) (Laptev *et al*. [1])	55.0%	56.7%
STIP-HoG (NB-pairwise alone)	16.4%	20.7%
STIP-HoG (D-pairwise alone)	46.6%	46.0%
STIP-HoG (single + D-pairwise)	59.0%	64.0%
STIP-HoG-Norm (single) (Laptev *et al*. [1])	42.6%	40.6%
SCM-Traj (single)	42.3%	37.3%
SCM-Traj (NB-pairwise alone)	14.3%	70.0%
SCM-Traj (D-pairwise alone)	40.0%	48.0%
SCM-Traj (single + D-pairwise)	47.1%	50.0%

NB = Naïve Bayes

D = discriminative approach

From Matikainen, Hebert, Sukthankar, ECCV 2010

(150 videos, 10 classes)

(1600 videos, 11 classes)

Method	UCF-YT	Rochester
STIP-HoG (single) (Laptev *et al.* [1])	55.0%	56.7%
STIP-HoG (NB-pairwise alone)	16.4%	20.7%
STIP-HoG (D-pairwise alone)	46.6%	46.0%
STIP-HoG (single + D-pairwise)	59.0%	64.0%
STIP-HoG-Norm (single) (Laptev *et al.* [1])	42.6%	40.6%
SCM-Traj (single)	42.3%	37.3%
SCM-Traj (NB-pairwise alone)	14.3%	70.0%
SCM-Traj (D-pairwise alone)	40.0%	48.0%
SCM-Traj (single + D-pairwise)	47.1%	50.0%

(150 videos, 10 classes)

(1600 videos, 11 classes)

Method	UCF-YT	Rochester
STIP-HoG (single) (Laptev *et al.* [1])	55.0%	56.7%
STIP-HoG (NB-pairwise alone)	16.4%	20.7%
STIP-HoG (D-pairwise alone)	46.6%	46.0%
STIP-HoG (single + D-pairwise)	59.0%	64.0%
STIP-HoG-Norm (single) (Laptev *et al.* [1])	42.6%	40.6%
SCM-Traj (single)	42.3%	37.3%
SCM-Traj (NB-pairwise alone)	14.3%	70.0%
SCM-Traj (D-pairwise alone)	40.0%	48.0%
SCM-Traj (single + D-pairwise)	47.1%	50.0%

(150 videos, 10 classes)

(1600 videos, 11 classes)

Method	UCF-YT	Rochester
STIP-HoG (single) (Laptev *et al.* [1])	55.0%	56.7%
STIP-HoG (NB-pairwise alone)	16.4%	20.7%
STIP-HoG (D-pairwise alone)	46.6%	46.0%
STIP-HoG (single + D-pairwise)	59.0%	64.0%
STIP-HoG-Norm (single) (Laptev *et al.* [1])	42.6%	40.6%
SCM-Traj (single)	42.3%	37.3%
SCM-Traj (NB-pairwise alone)	14.3%	70.0%
SCM-Traj (D-pairwise alone)	40.0%	48.0%
SCM-Traj (single + D-pairwise)	47.1%	50.0%

(150 videos, 10 classes)

(1600 videos, 11 classes)

Method	UCF-YT	Rochester
STIP-HoG (single) (Laptev *et al.* [1])	55.0%	56.7%
STIP-HoG (NB-pairwise alone)	16.4%	20.7%
STIP-HoG (D-pairwise alone)	46.6%	46.0%
STIP-HoG (single + D-pairwise)	59.0%	64.0%
STIP-HoG-Norm (single) (Laptev *et al.* [1])	42.6%	40.6%
SCM-Traj (single)	42.3%	37.3%
SCM-Traj (NB-pairwise alone)	14.3%	70.0%
SCM-Traj (D-pairwise alone)	40.0%	48.0%
SCM-Traj (single + D-pairwise)	47.1%	50.0%

(150 videos, 10 classes)

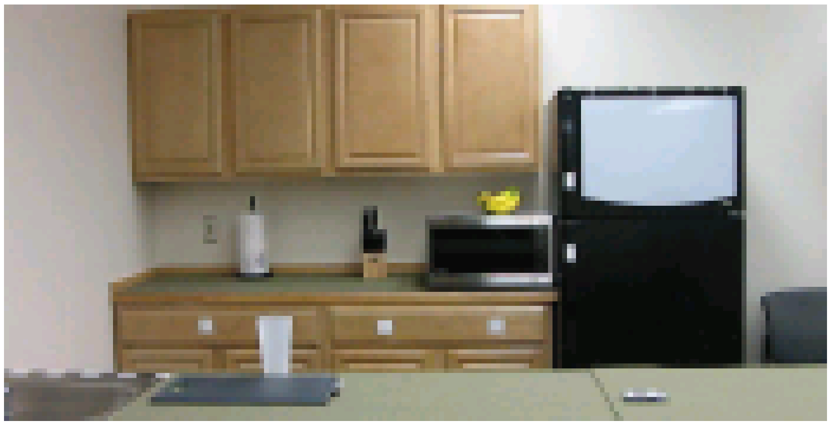
(1600 videos, 11 classes)

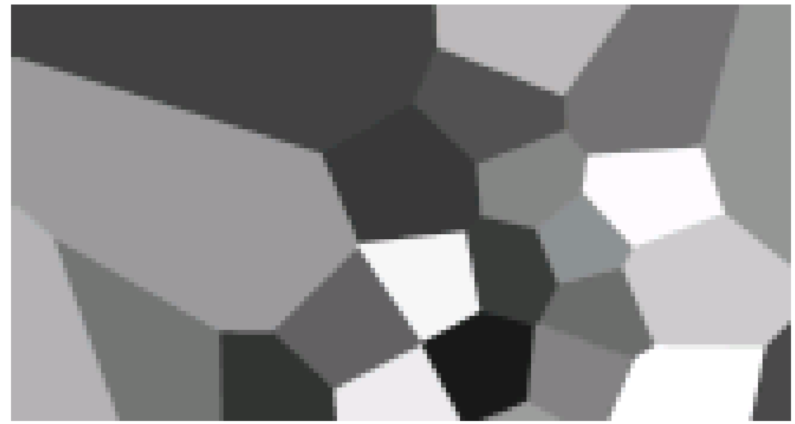
Method	UCF-YT	Rochester
STIP-HoG (single) (Laptev *et al.* [1])	55.0%	56.7%
STIP-HoG (NB-pairwise alone)	16.4%	20.7%
STIP-HoG (D-pairwise alone)	46.6%	46.0%
STIP-HoG (single + D-pairwise)	59.0%	64.0%
STIP-HoG-Norm (single) (Laptev *et al.* [1])	42.6%	40.6%
SCM-Traj (single)	42.3%	37.3%
SCM-Traj (NB-pairwise alone)	14.3%	70.0%
SCM-Traj (D-pairwise alone)	40.0%	48.0%
SCM-Traj (single + D-pairwise)	47.1%	50.0%

Yellow = high
P(pair | answerPhone)

From Matikainen, Hebert, Sukthankar, ECCV 2010

Relative vs. absolute information





Cluster of feature locations from training data

- Using (x,y) instead of (dx,dy) emphasizes correlations to fixed environment/camera

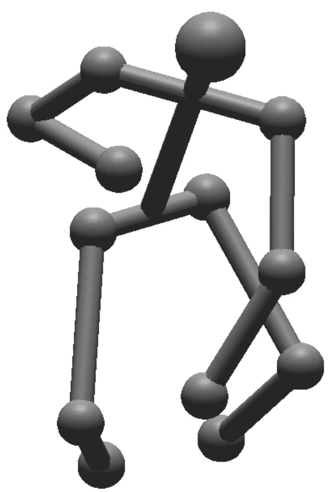
- We can now make more explicit both temporal and spatial relations
- But can we make structure even more explicit without compromising generalization (i.e., without going back to strong templates)?

Using trajectory elements

- Critique: Better representation of temporal and spatial structure, but:
  - Still along the lines of statistical representations
  - Still weak, implicit representation of structure
  - Does not exploit skeletal knowledge
- Solution:
  - Use strong underlying limb-based skeletal model
  - Enormous literature (not reviewed here!) on human body tracking, pose recovery, 3D reconstruction of semi-deformable bodies, etc.

Dilemma

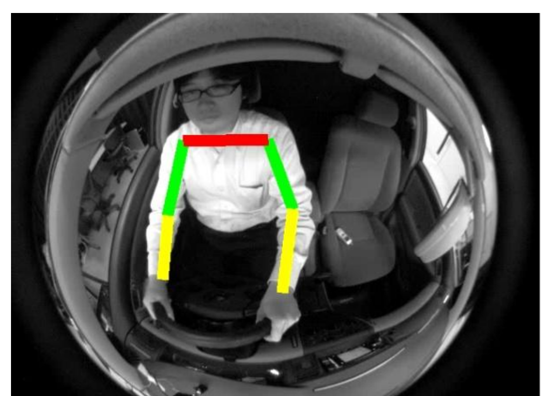
- Enormous literature on human body tracking, pose recovery, 3D reconstruction of semi-deformable bodies, etc.
- But:
  - If we knew the 3D structure (which point corresponds to which limb) we could (maybe) compare to an action model
  - But we don't know the associations or the 3D structure
- Solution:
  - Estimate consistency between a single camera model and an action model with *implicit* 3D reconstruction



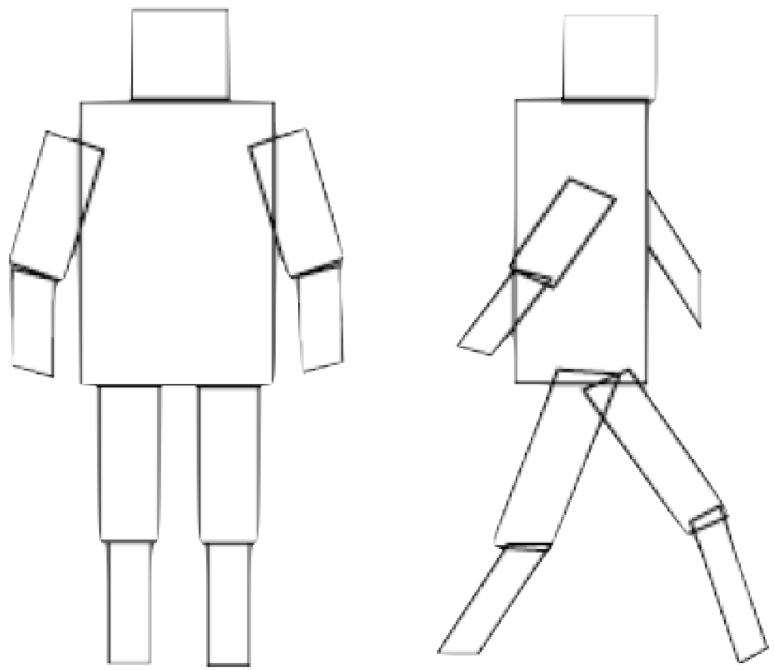
Examples (*very* small sample)

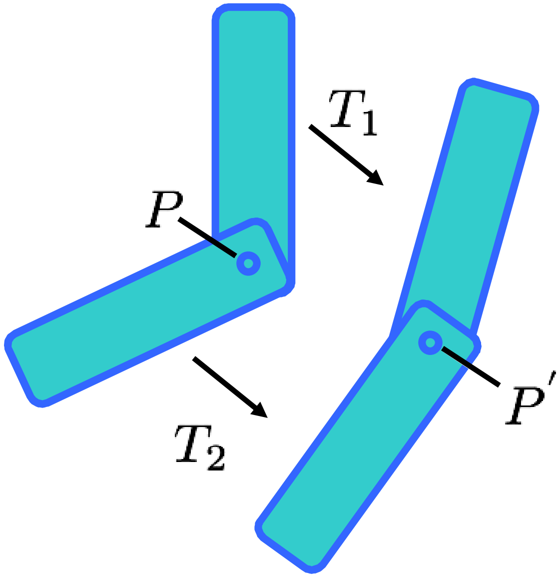
- A. Datta. Closed-Form Analysis of Human Motion in Monocular Videos. Ph.D. Dissertation. CMU/RI. 2010
- V. Parameswaran and R. Chellappa. View independent human body pose estimation from a single perspective image. *CVPR*, 2004.
- X. K. Wei and J. Chai. Modeling 3d human poses from uncalibrated monocular images. *ICCV*, 2009.
- R. Rosales and S. Sclaroff. Inferring body pose without tracking body parts. *CVPR*, 2000.
- A. Agarwal and B.Triggs. 3d human pose from silhouettes by relevance vector regression. *CVPR*, 2004.
- Ahmed Elgammal and Chan-Su Lee. Inferring 3d body pose from silhouettes using activity manifold learning. *CVPR*, 2004.

Next few slides from Ankur Datta and Yaser Sheikh

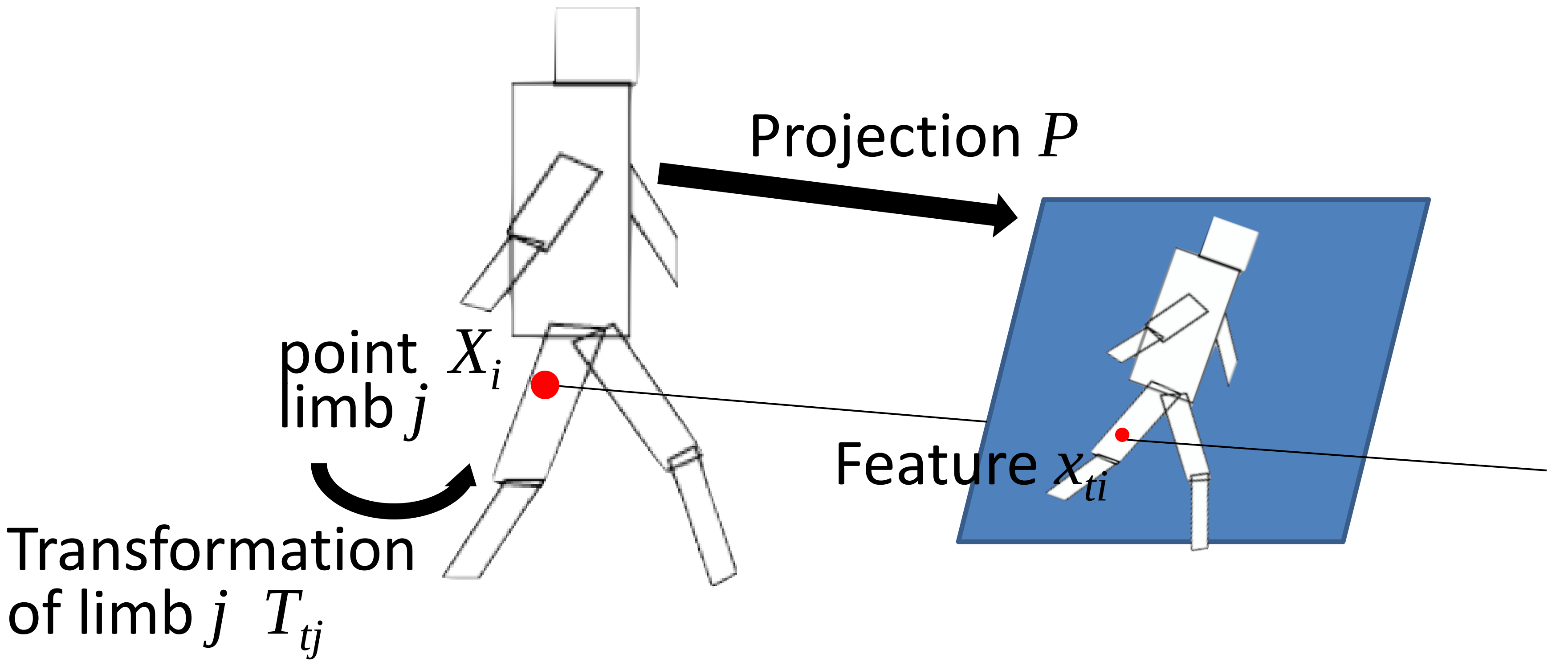


Simple model





Micro-action model



$$x_{ti} = P T_{tj}^{a} X_i$$

Example: Micro-action recognition

Problem Formulation

$$\min_{\mathbf{P},\mathbf{X}_k,z_{ij}} \; \sum_{i=1}^{N} \sum_{j=1}^{M} z_{ij} \sum_{t\in\Omega_i} \|\mathbf{x}_{ti} - \mathbf{P}\mathbf{T}_{tj}^{a}\mathbf{X}_i\|^2$$

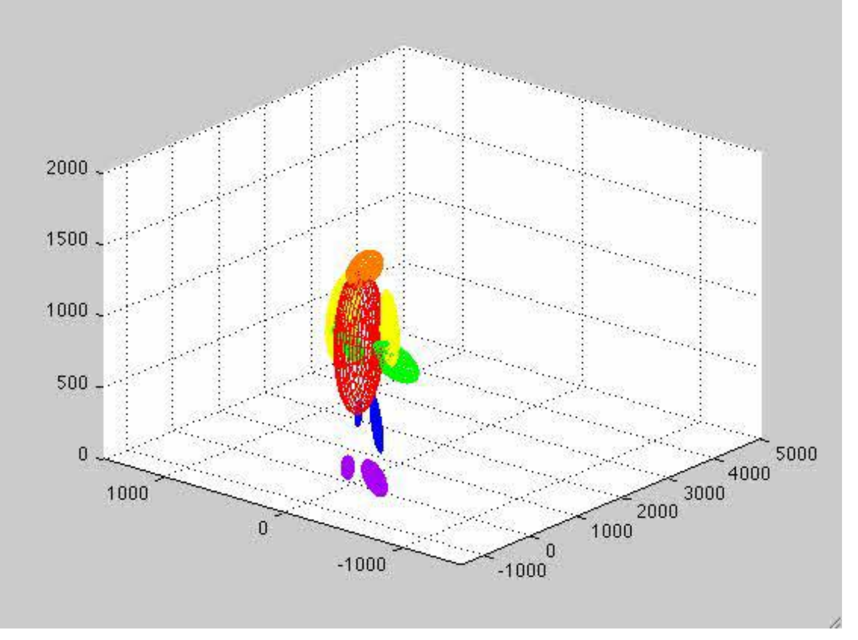
$$\text{subject to} \quad z_{ij} \in \{0,1\} \quad \text{and} \quad \sum_{j=1}^{M} z_{ij} = 1$$

x: 2D trajectory
X: 3D point (homogeneous coordinates)
T: rigid transformation of a limb
P: affine camera
M: number of limbs
N: number of trajectories
z: binary association variables of trajectory to limb

Slide adapted from Ankur Datta

Chandraker, 2008

Micro-Action: Gaussian Man



3D Models constructed from motion-capture data

Assumption:

- Points on the same limb move rigidly and are distributed according to a Gaussian.

Approach

- If we knew
  - Which feature correspond to which limb
  - The 3D position X of the features
- Then we could estimate the camera P
- There exist a P only if the motion is consistent with the transformation model of action a

$$\min_{\mathbf{P},\mathbf{X}_k,z_{ij}} \; \sum_{i=1}^{N} \sum_{j=1}^{M} z_{ij} \sum_{t\in\Omega_i} \|\mathbf{x}_{ti} - \mathbf{P}\mathbf{T}_{tj}^{a}\mathbf{X}_i\|^2$$

$$\text{subject to} \quad z_{ij} \in \{0,1\} \quad \text{and} \quad \sum_{j=1}^{M} z_{ij} = 1$$

Micro-Action: Camera and Action

<u>Leveraging Gaussian Man</u>

Treat **X** as a nuisance parameter

$$\min_{\mathbf{p}} \sum_{k=1}^{K} \sum_{t \in \Omega_k} \int p(\mathbf{X}_k | j(k)) \; \|\mathbf{x}_{tk} - \mathbf{P}\mathbf{T}^{a}_{tj}\mathbf{X}_k\|^2 \; d\mathbf{X}_k,$$

where,

$$p(\mathbf{X}_k | j(k)) = \mathcal{N}(\mathbf{X}_k; \mu_{j(k)}, \mathbf{\Sigma}_{j(k)})$$

K: number of sample trajectories

$j(k)$: limb association for the k^{th} trajectory

Slide adapted from Ankur Datta

Micro-Action: Camera and Action

<u>Leveraging Gaussian Man</u>

$$\min \sum_{k=1}^{K} \sum_{t \in \Omega_k} \mathrm{tr}\left\{\mathbf{P}\mathbf{T}^a_{tj(k)}\mathbf{A}_{j(k)}(\mathbf{T}^a_{tj(k)})^T\mathbf{P} - 2\mathbf{P}\mathbf{T}^a_{tj(k)}\mu_{j(k)}\mathbf{x}_{tk}\right\}$$

where,

$$\mathbf{A}_{j(k)} = \begin{bmatrix} \mathbf{\Sigma}_{j(k)} + \mu_{j(k)}\mu^T_{j(k)} & \mu_{j(k)} \\ \mu^T_{j(k)} & 1 \end{bmatrix}$$

Linear system that can be solved efficiently.

Slide adapted from Ankur Datta

Action Recognition Algorithm

RANSAC-based Optimization

For all actions:

For all samples:

Sample *K* trajectories and their limb associations

Solve for camera parameters

Compute consensus error

end For

end For

OUTPUT:

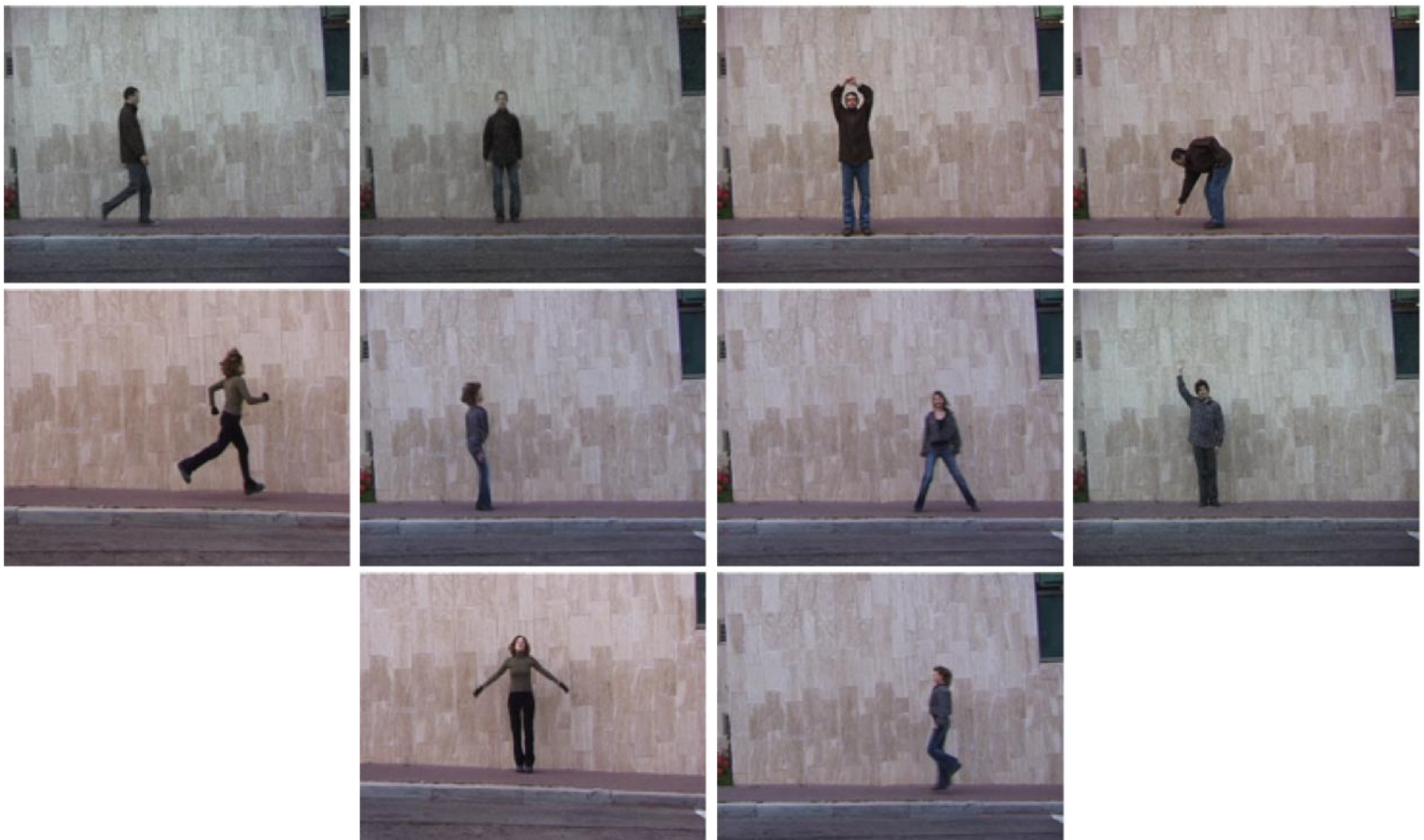
Action and Camera with the least consensus error

Slide adapted from Ankur Datta

Weizmann Dataset

10 Actions, 9 actors per action

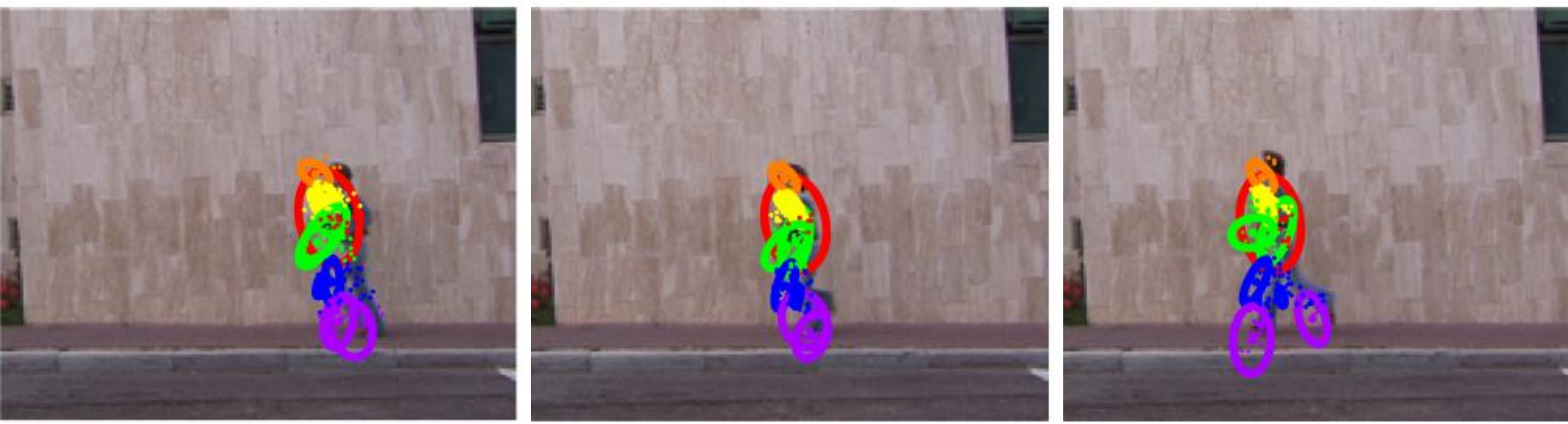
Note: Really boring but easy to verify output



Slide adapted from Ankur Datta

Micro-Action Alignment Results

Micro-Action Alignment Results

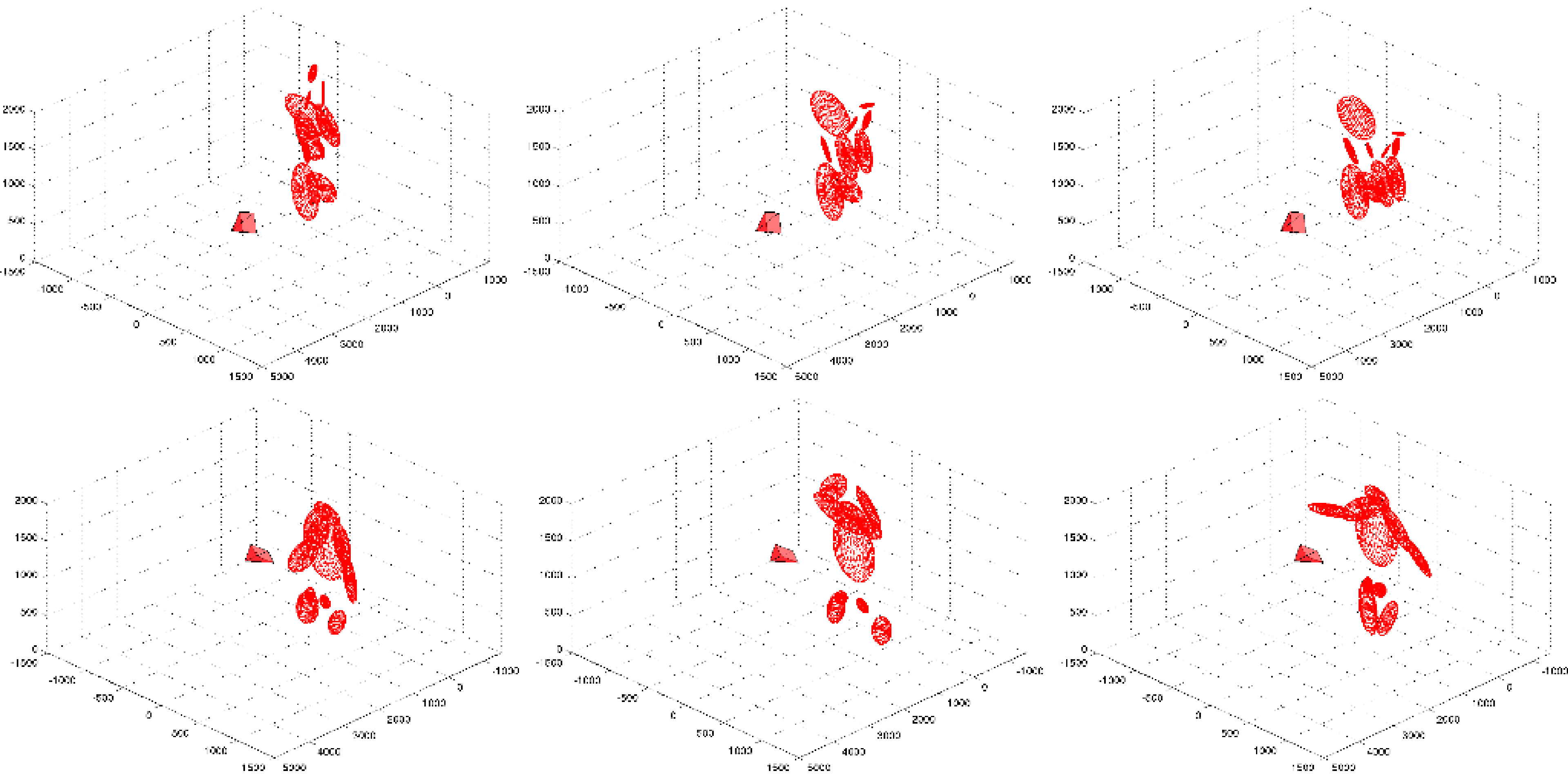


Slide adapted from Ankur Datta

Micro-Action Alignment Results: Challenges

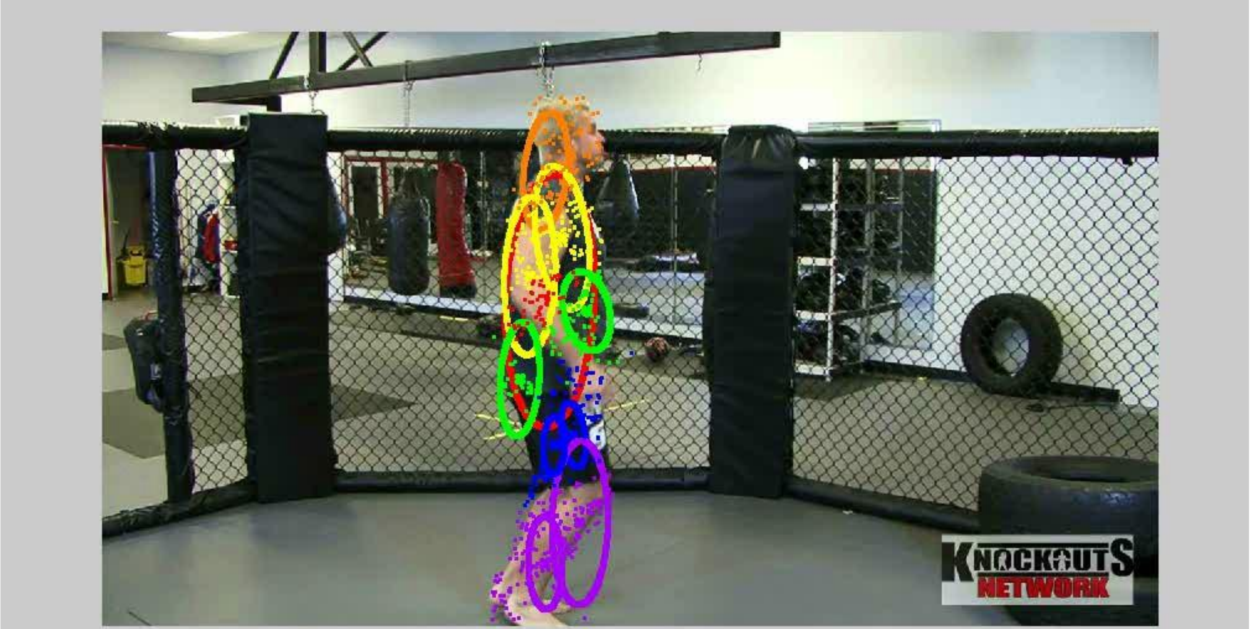
Stylized differences in exhibition of micro-action

Camera Recovery



Slide adapted from Ankur Datta

High-Resolution Data



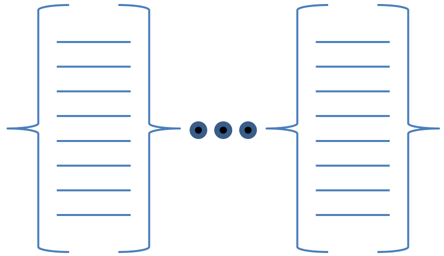
Discussion

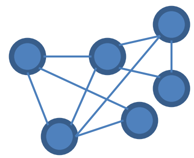
- Simple and fast
- Robustness to background features?
- Depends on strong models (e.g., from motion capture)
- Rigid definition of action; generalization to broad classes questionable?

Outline

- Quick overview of two standards approaches
  - Statistical BoF approaches
  - Volumetric approaches
- Incorporating temporal information more explicitly
  - Example: Trajectory fragments
- Incorporating spatial information more explicitly
  - Example: Encoding pairwise relations
- Designing stronger structural models
  - Example: “Micro-actions” recognition through implicit 3D reconstruction
- Issues with video training datasets
  - Example: Selecting temporal boundaries
  - Analysis of bias in standard datasets
- Discussion and introduction to proposed challenge problems for afternoon presentations

Input sequence





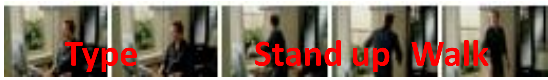
Set of feature of vectors

+ additional structure (e.g., geometry, relations)

machine learning box

Training data +

Geometry, relational information, physics, domain knowledge



We never talk about training data for all this. Any issues there?

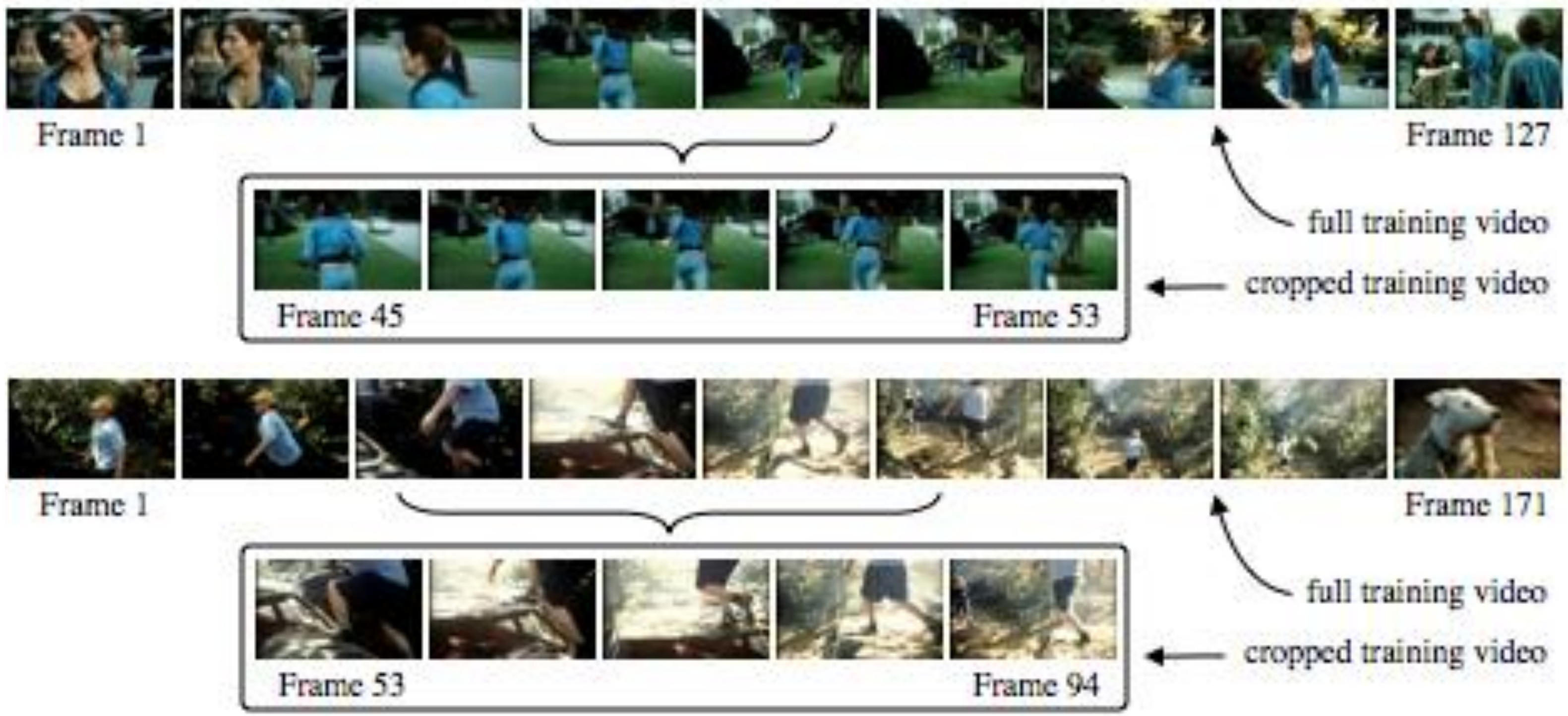
Automatic refinement from imperfect training samples

- Problem: Temporal boundaries of actions are ill-defined
- System relies on templates (video clips) selected by user
- Video section selected by user is not optimized for good detection performance
- Same issue with automatic selection of training samples based on caption or text annotations
- Solution: Is it possible to adjust the temporal boundaries in order to maximize classification performance

Example: The action is very short compared to the selected clip

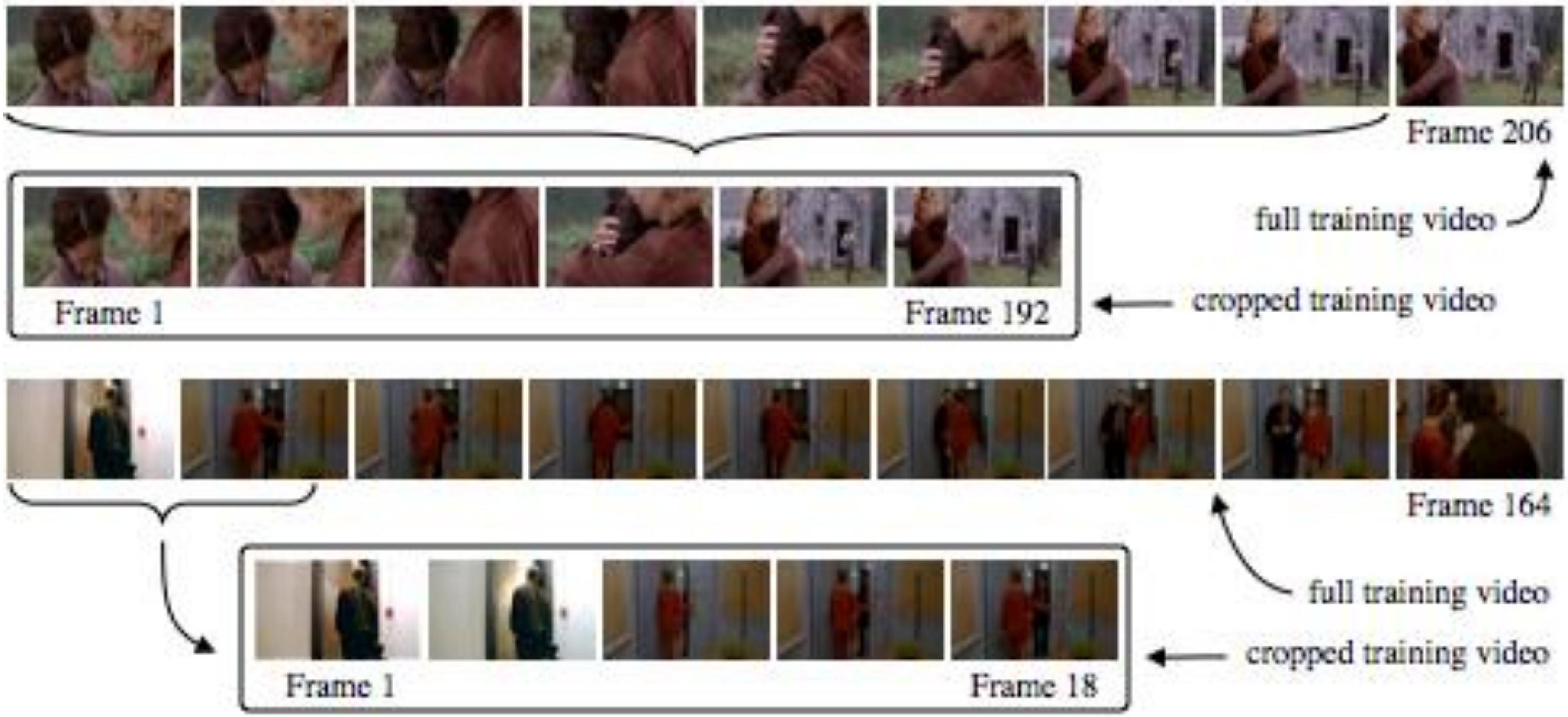
- Many actions occur quickly, taking only a few frames to complete
- The “Stand up” action above takes less than a second
- Issue: automatically cropped the instant the action occurs from the other frames of the video

Example: The selected clip include multiple actions



- Two videos of the action “running” for which cropping was shown to improve the system performance
- Note the discriminative and unambiguous portions of each video which were selected

Example: The selected clip is much too long

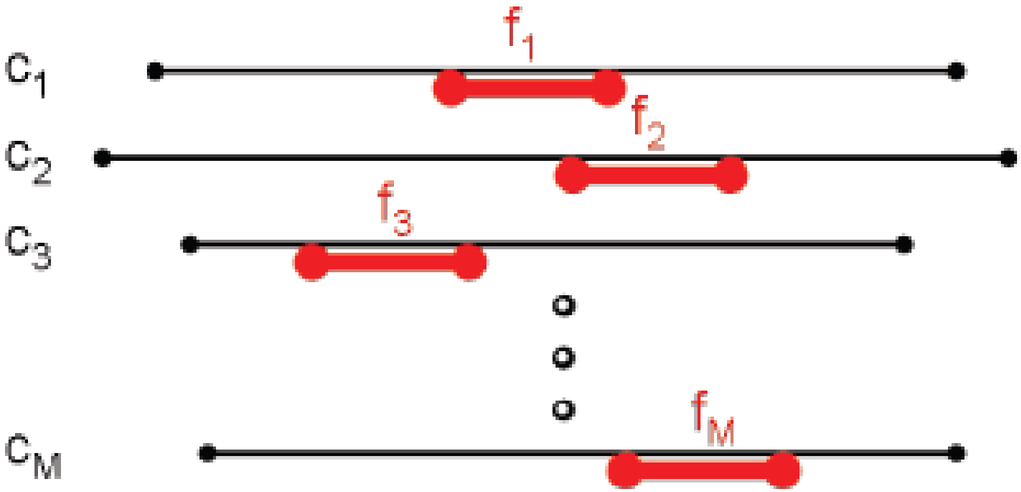


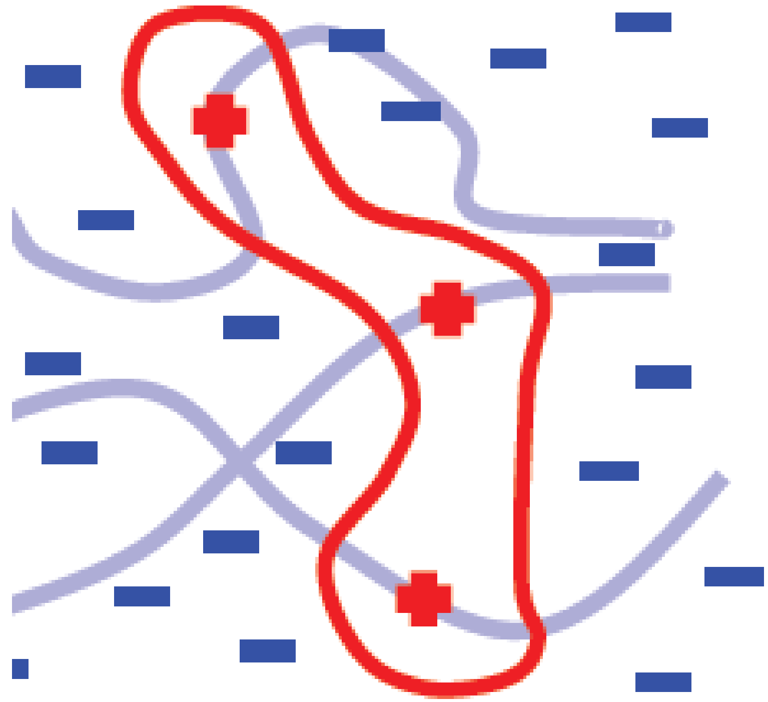
- "Hugging" video for which almost the entire video was selected, indicating the initial cropping was adequate
- "Opening door" video for which only the first few frames were necessary to sufficiently model the action

Sample approaches

- *Multiple instance learning*: Buehler, P., Zisserman, A., Everingham, M.: Learning sign language by watching TV (using weakly aligned subtitles). In: CVPR. (2009)
- *No constraints on temporal connectivity*: Nowozin, S., Bakir, G., Tsuda, K.: Discriminative subsequence mining for action classification. In: CVPR. (2007)
- *Specific to STIP:* Yuan, J., Liu, Z., Wu, Y.: Discriminative subvolume search for efficient action detection. In: CVPR. (2009)
- *Optimizes croppings with respect to human performance*: Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J.: Automatic annotation of human actions in video. In: ICCV. (2009)
- *Attempts to find most discriminative croppings (Examples in this presentation)*: Satkin, S. and Hebert, M.: Modeling the Temporal Extent of Actions. In: ECCV. (2010)

Example





- Temporally localize a video segment in each clip containing the action
- Treated as semi-supervised clustering

Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J.: Automatic annotation of human actions in video. In: ICCV. (2009)

Example

Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J.: Automatic annotation of human actions in video. In: ICCV. (2009)

Possible overall approach

$$\underset{\{\forall_i:(f_i^0,f_i^1)\}}{\operatorname{argmax}} \sum_{i=1}^{n} \text{classify}\left(\text{train}(\mathcal{F}_{(1\ldots n)\neq i}, f_i^0, f_i^1),\ \mathcal{F}_i\right) = \mathcal{C}_i$$

- Find the set of croppings (f^0_i, f^1_i) that maximizes leave-one-out cross-validation performance

[Satkin & Hebert, Modeling the Temporal Extent of Actions, ECCV2010]

Possible overall approach

$$\underset{\{\forall_i:(f_i^0,f_i^1)\}}{\operatorname{argmax}} \sum_{i=1}^{n} \operatorname{classify}\left(\operatorname{train}(\mathcal{F}_{(1\ldots n)\neq i}, f_i^0, f_i^1),\ \mathcal{F}_i\right) = \mathcal{C}_i$$

Intractable because of the exponential number of possible croppings.

Observation:

- Portions of videos which are most confidently and correctly classified by a trained action recognition system are highly correlated with actions of the same class and differ from actions of other classes.
- These portions of the videos are discriminative and are a good choice for training our classifier.

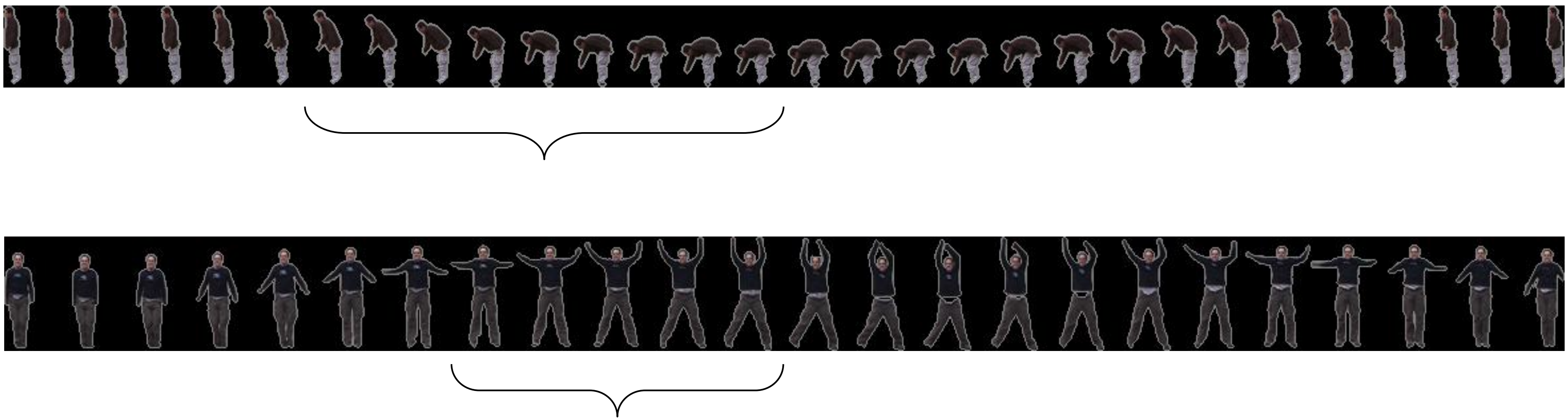
Possible overall approach

$$\operatorname*{argmax}_{\{\forall_i:(f_i^0, f_i^1)\}} \sum_{i=1}^{n} \text{classify}\left(\text{train}(\mathcal{F}_{(1\ldots n)\neq i}, f_i^0, f_i^1),\ \mathcal{F}_i\right) = \mathcal{C}_i$$

Intractable because of the exponential number of possible croppings. Approximation:

1. Split the video we aim to crop into its $|f|^2/2$ possible temporal croppings.
2. Train a classifier on the remaining training videos, excluding the one from step 1.
3. Evaluate this classifier on each of the $|f|^2/2$ croppings.
4. Select the individual cropping that was correctly classified with the highest level of confidence.

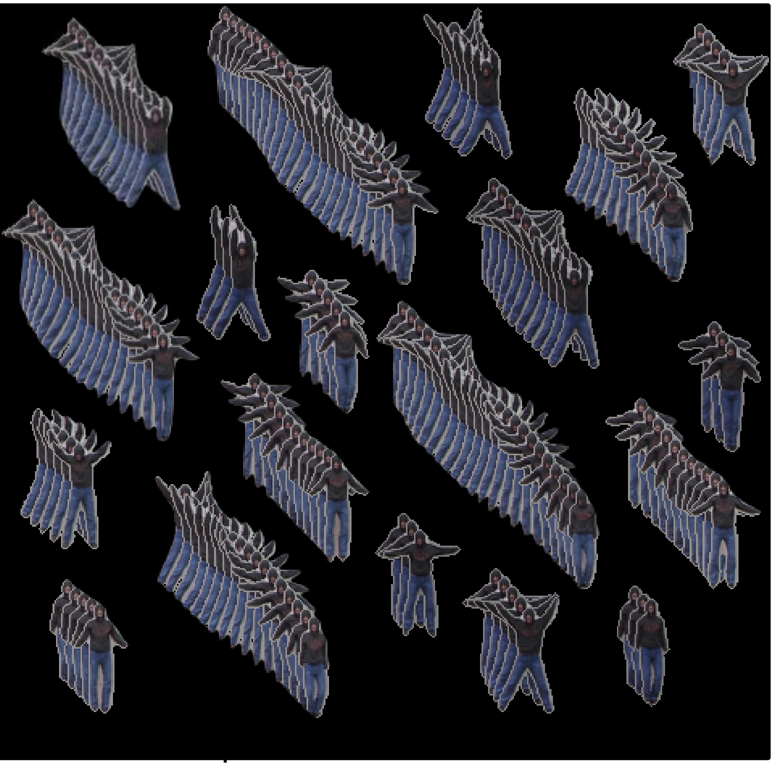
Experiment



Find most discriminative croppings for a simple action dataset

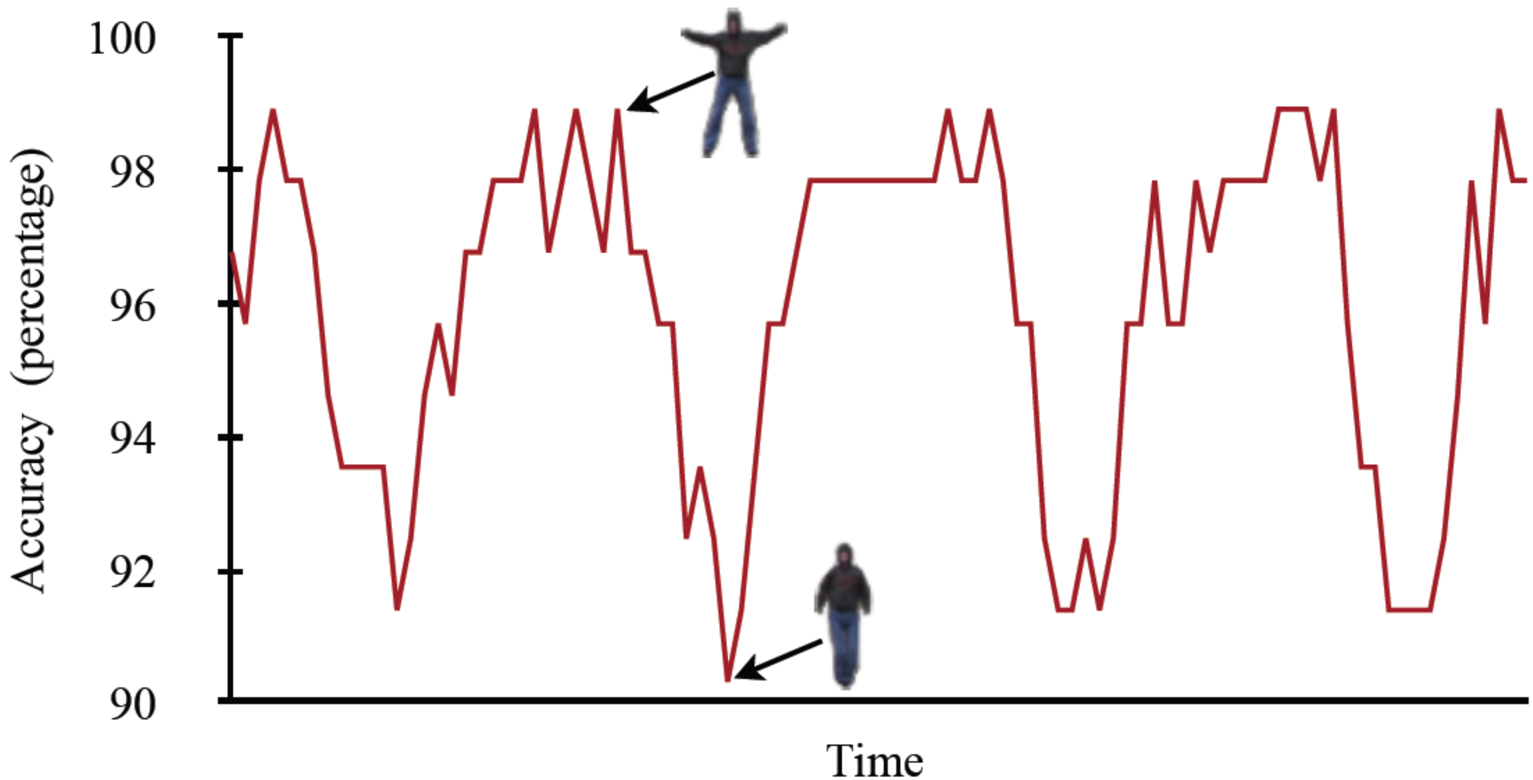
From Satkin & Hebert, ECCV2010]

Experiment



Generate all possible cropped clips from an initial user-selected example → Compare each template against a set of test videos → Select the templates with highest performance (detection rate and localization)

Experiment



- Hypothesis: Using “optimal” cropping of training samples boosts accuracy
- Proof-of-concept: Brute force search through possible croppings by using volumetric matching

From Satkin & Hebert, ECCV2010]

Proof-of-concept

Action	Worst Cropping Accuracy	Best Cropping Accuracy
Bend	90.63	98.00
Jumping Jack	90.94	97.70
Run	93.39	96.47
Walk	93.55	95.70
10-class Average	91.98	95.76

- Hypothesis verified: Substantial performance gain when selecting the best cropping

From Satkin & Hebert, ECCV2010]

More practical approach

$$\underset{\{\forall i:(f_i^0, f_i^1)\}, \mathbf{w}, b, \xi}{\operatorname{argmin}} \left(\frac{1}{2} ||\mathbf{w}||^2 + C \sum_{i=1}^{n} \xi_i \right),$$

$$\text{subject to:} \quad \forall i: \quad y_i \left(\mathbf{w} \cdot \phi \left(\sum_{f=f_i^0}^{f_i^1} H_i(f) \right) + b \right) \geq 1 - \xi_i$$

- Tractable approach:
  - Train on all uncropped videos excluding one
  - Test resulting model on all $|f|^2/2$ possible cropping of that video
  - Select the best cropping for that video

From Satkin & Hebert, ECCV2010]

Examples

From Satkin & Hebert, ECCV2010]

Hollywood




	Baseline Accuracy (using full videos)	Our Accuracy (cropped videos)	Absolute Change cropped - full	% Improvement (cropped - full)/full
Trajectons	37.84	41.85	4.01	10.60
HOG	33.08	33.71	0.63	1.90
HOF	38.47	43.48	5.01	13.02

Rochester




	Baseline Accuracy (using full videos)	Our Accuracy (cropped videos)	Absolute Change cropped - full	% Improvement (cropped - full)/full
Trajectons	46.00	54.00	8.00	17.39
HOG	54.67	60.00	5.33	9.75
HOF	79.33	80.00	0.67	0.84

Lessons?

- Temporal boundaries are not well defined (unlike boundaries of physical object)
- May be possible to define "optimal" temporal boundaries (croppings) based on discriminability
- But: intractable
- What is the "right" approximation to the problem
- Current approximation seems very coarse
- Any other ideas?

Outline

- Quick overview of two standards approaches
  - Statistical BoF approaches
  - Volumetric approaches
- Incorporating temporal information more explicitly
  - Example: Trajectory fragments
- Incorporating spatial information more explicitly
  - Example: Encoding pairwise relations
- Designing stronger structural models
  - Example: “Micro-actions” recognition through implicit 3D reconstruction
- Issues with video training datasets
  - Example: Selecting temporal boundaries
  - Analysis of bias in standard datasets
- Discussion and introduction to proposed challenge problems for afternoon presentations