Temporal Models

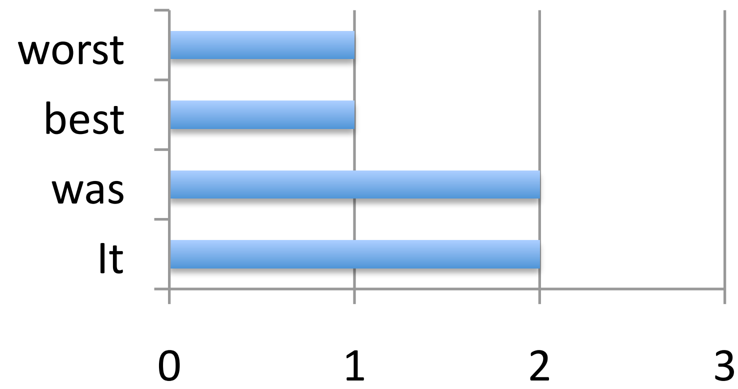
Greg Mori

CMPT 888

SFU Vision and Media Lab

“Bag-of-Words” Models

- Text document models
  - “It was the best of times, it was the worst of times.”



- Bag of Words + Topic Models in Computer Vision
  - Scenes: Fei-Fei & Perona CVPR’05
  - Objects: Sivic et al. ICCV’05, Fergus et al. ICCV’05, Russell et al. CVPR’06
  - Actions: Niebles et al. BMVC’06
  - Human Poses: Bissaco et al. NIPS’06

SFU Vision and Media Lab

Role of Temporal Information

- No temporal info
  - Classify each video frame independently
  - e.g., Efros et al. 03, Shechtman & Irani 05, Fathi & Mori 08

SFU Vision and Media Lab

Role of Temporal Information

? — ? — ? — ? — ? — ? — ? — ? — ? — ?

- Strong temporal info
  - Use hidden Markov Model or grammar on top of video frames
  - e.g. Bobick & Ivanov CVPR98, Yamato et al. CVPR92

SFU Vision and Media Lab

Role of Temporal Information

? ? ? ? ? ? ? ? ? ?

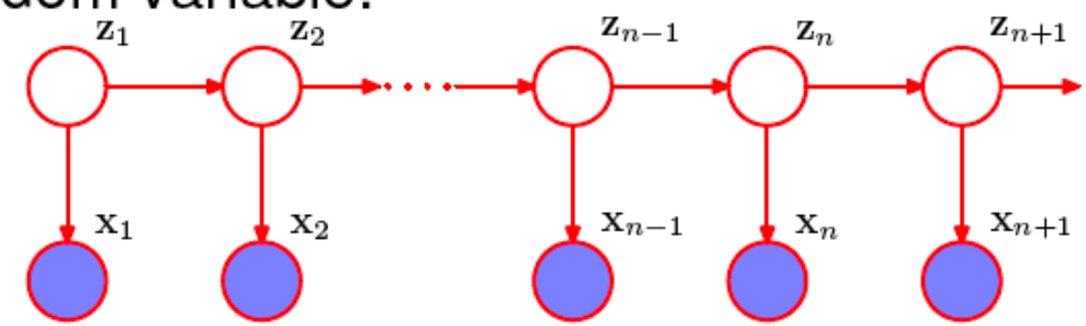
- Y. Wang et al. HUMO/PAMI is somewhere in between
  - Use bag of frames representation
  - Capture some temporal structure (co-occurrences of actions)
  - Simpler than full temporal models

HIDDEN MARKOV MODELS FOR ACTION RECOGNITION

SFU Vision and Media Lab

HMMs

- Sensor Markov assumption: $p(\boldsymbol{x}_t | \boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t-1}) = p(\boldsymbol{x}_t | \boldsymbol{z}_t)$
- Stationary process: transition model $p(\boldsymbol{z}_t | \boldsymbol{z}_{t-1})$ and sensor model $p(\boldsymbol{x}_t | \boldsymbol{z}_t)$ fixed for all t (separate $p(z_1)$)
- HMM special type of Bayesian network, z_t is a single discrete random variable:



- Joint distribution:

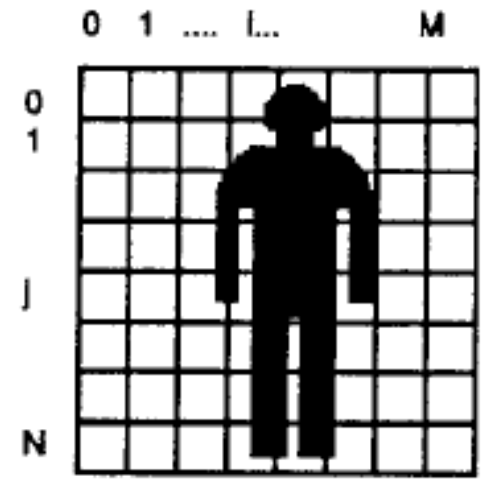
$$p(z_{1:t}, x_{1:t}) = p(z_1) \prod_{i=2:t} p(z_i | z_{i-1}) \prod_{i=1:t} p(x_i | z_i)$$

Using HMMs for Action Recognition (Yamato et al.)

- Each frame is mapped to a discrete symbol
  - Visual word
- For each action category, learn HMM model parameters
  - Transition matrix, emission matrix, prior
- Recognition: compute likelihood of observed symbols under each HMM
  - Choose action category that produces highest likelihood

Features (Yamato et al.)

- Mesh features
  - Looks like HOG on foreground mask!
- Some form of vector quantization (VQ) is used
  - Manual/random(?) selection of prototypes
  - K-means?
- 1992 vs. 2010? ☺

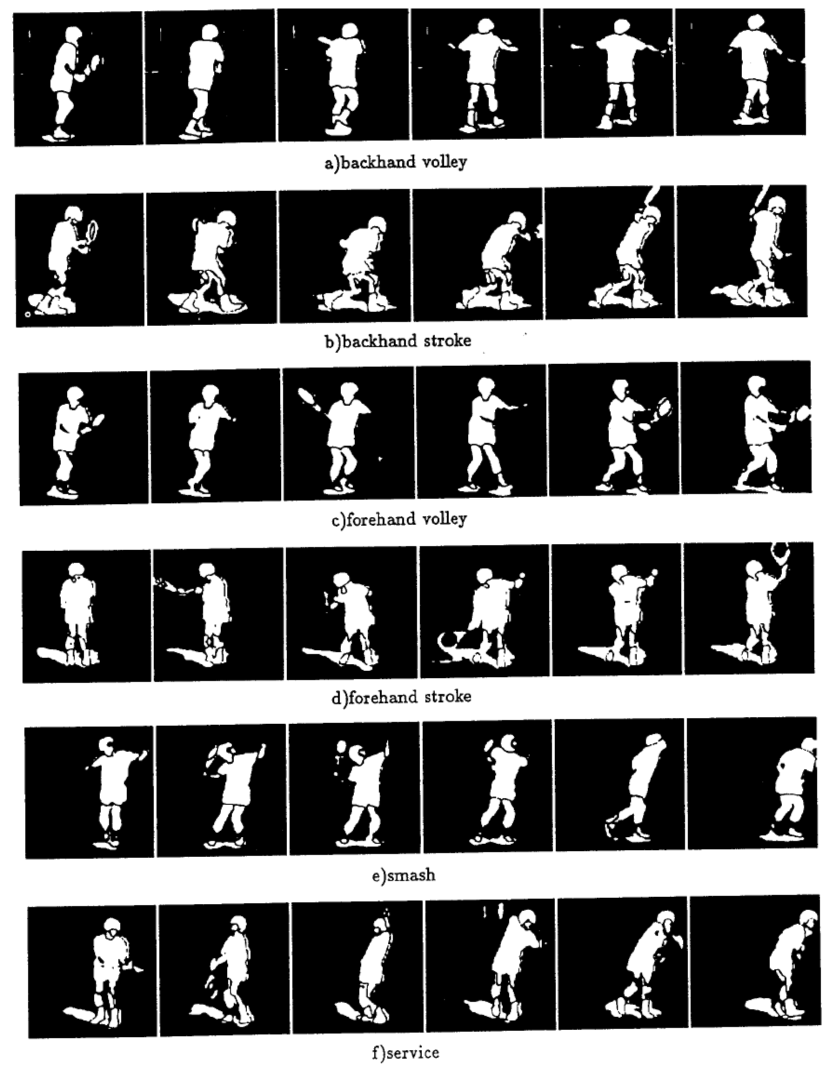


$f=(a_{00}, a_{01}, \ldots, a_{ij}, \ldots a_{MN})$

a_{ij}=number of black mesh(ij)/$M_m N_m$

Experiments

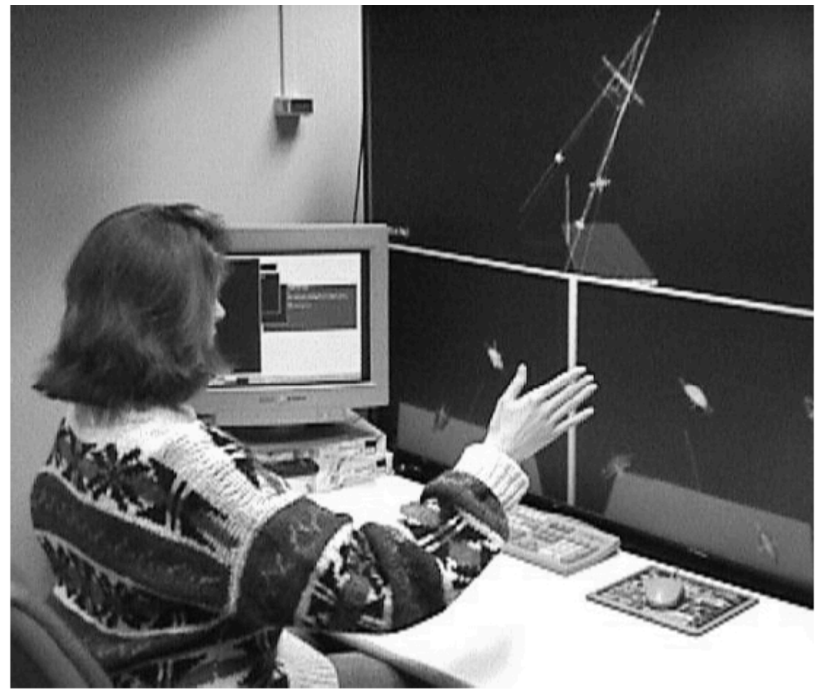
- Limited data/compute available
- 6 actions, more variability than KTH/Weizmann
- Good results
  - Likely would be quite accurate with more training data / parameter CV



SFU Vision and Media Lab

Other work

- Bobick & Ivanov CVPR 98
  - Recognize hand gestures
  - Grammar to describe a gesture

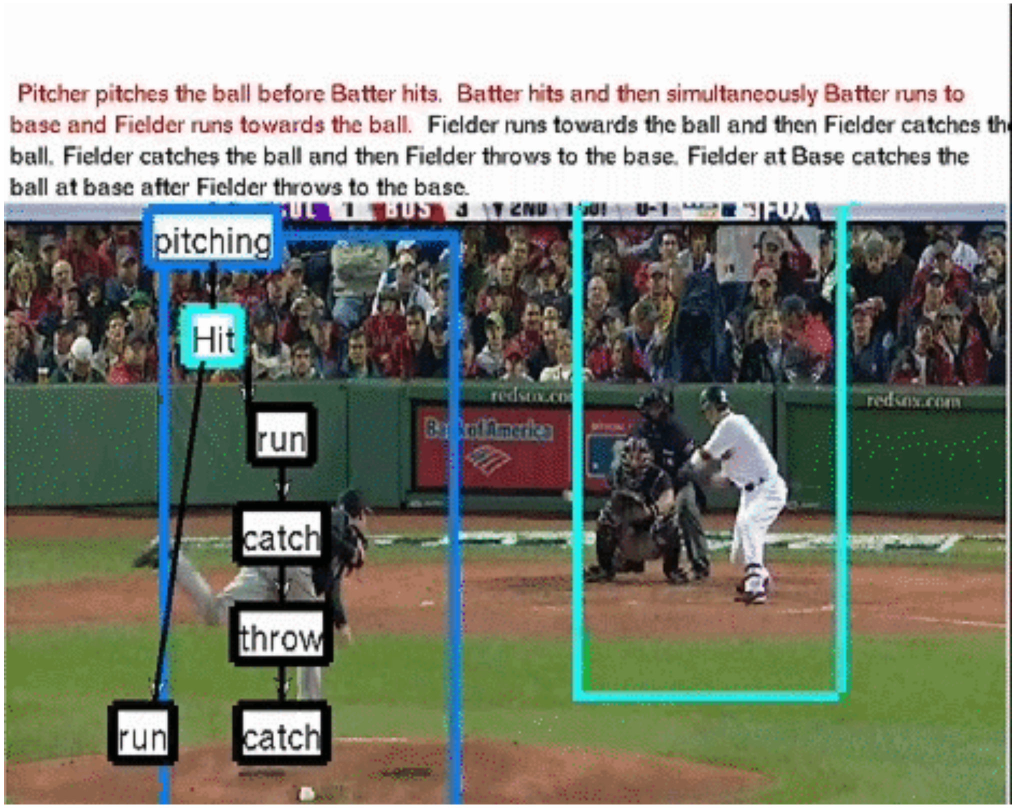


G_{square} :

SQUARE	$\rightarrow$		RH	[0.5]
		|	LH	[0.5]
RH	$\rightarrow$		TOP UD BOT DU	[1.0]
LH	$\rightarrow$		BOT DU TOP UD	[1.0]
TOP	$\rightarrow$		LR	[0.5]
		|	RL	[0.5]
BOT	$\rightarrow$		RL	[0.5]
		|	LR	[0.5]
LR	$\rightarrow$		left-right	[1.0]
UD	$\rightarrow$		up-down	[1.0]
RL	$\rightarrow$		right-left	[1.0]
DU	$\rightarrow$		down-up	[1.0]

Other work

- Gupta et al. CVPR 2009, others
- “Storyline” model explaining video

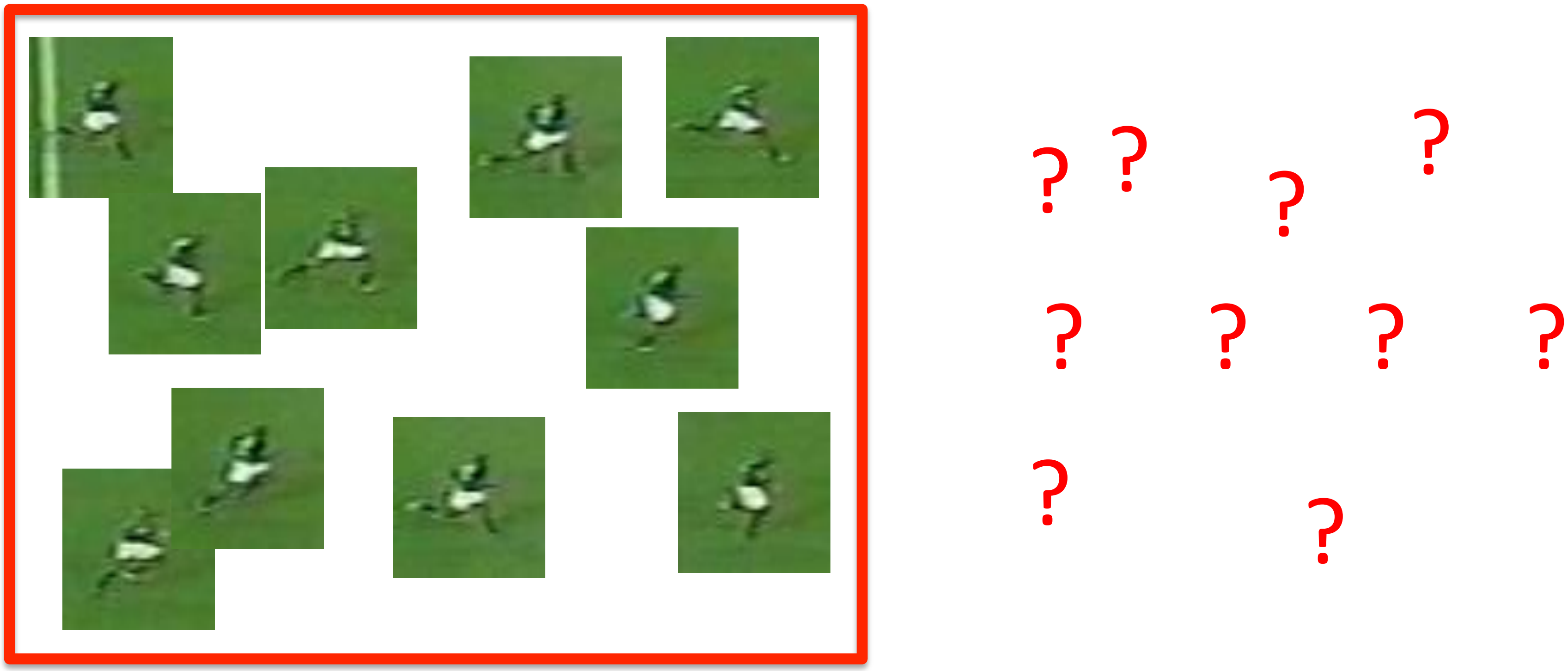


SFU Vision and Media Lab

BAG OF FRAMES MODEL

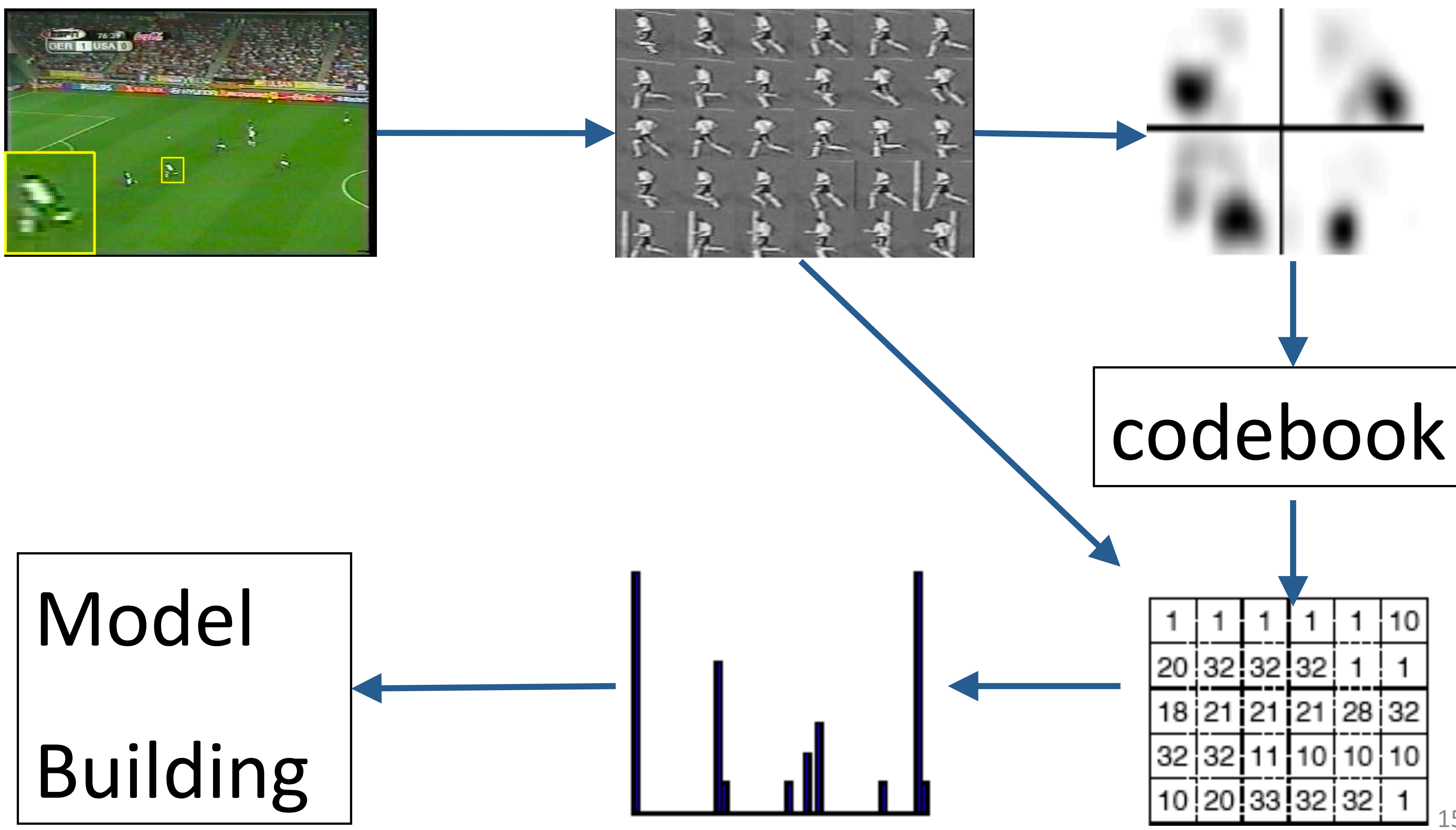
SFU Vision and Media Lab

Role of Temporal Information

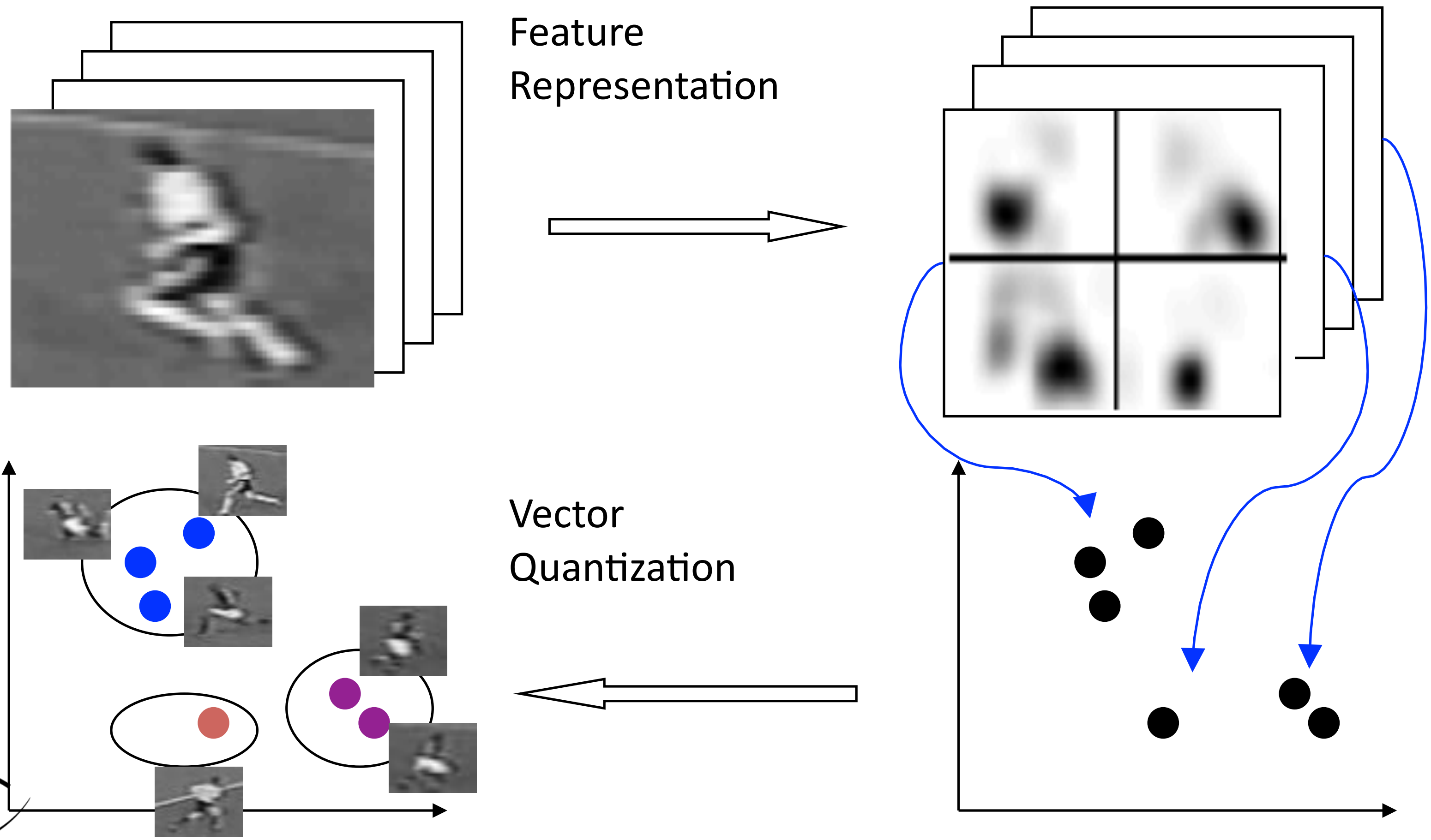


- Y. Wang et al. HUMO/PAMI is somewhere in between
  - Use bag of frames representation
  - Capture some temporal structure (co-occurrences of actions)
  - Simpler than full temporal models

Bag-of-Words Sequence Model

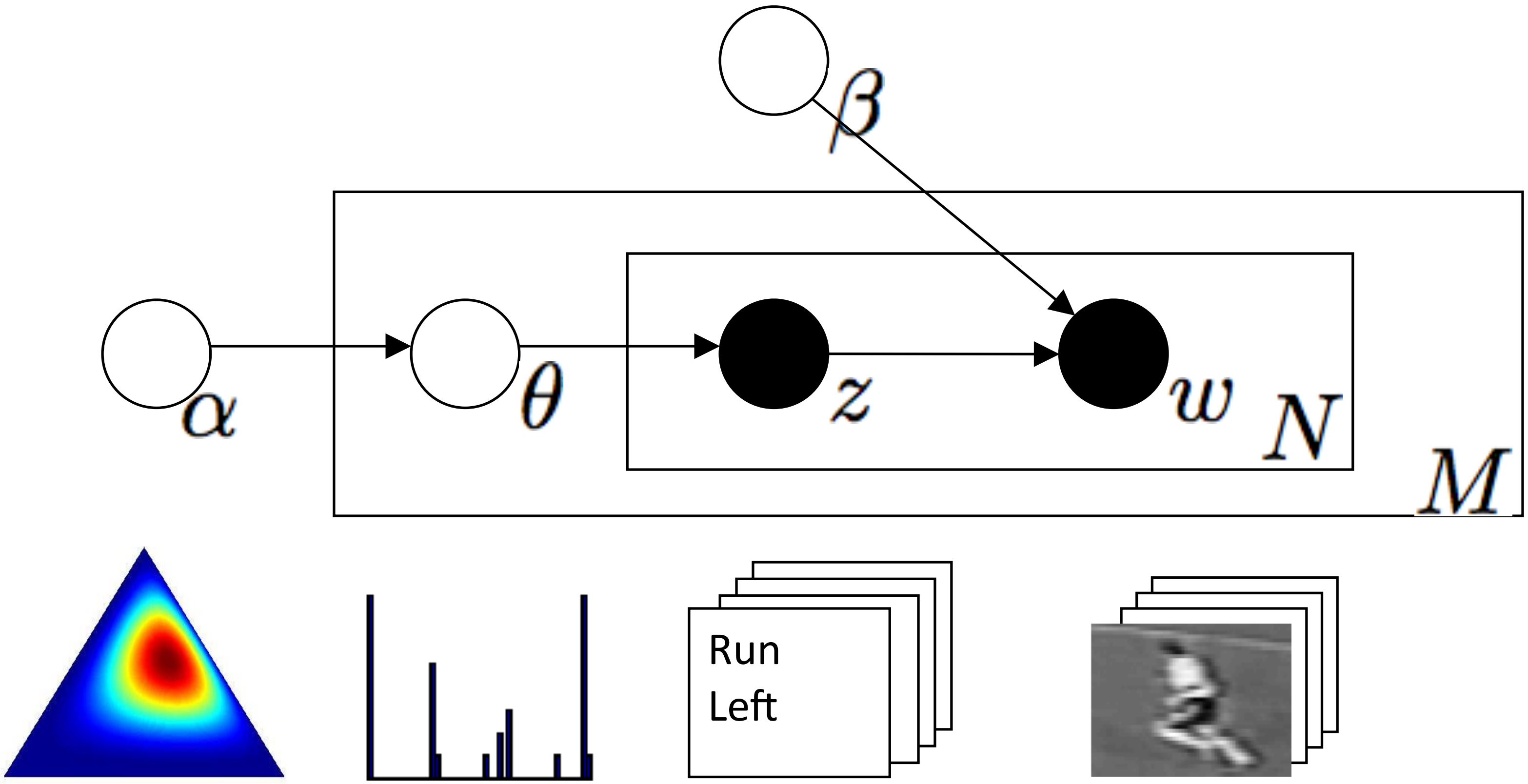


Codebook Formation



SFU Vision and Media Lab

Semi-Latent Dirichlet Allocation

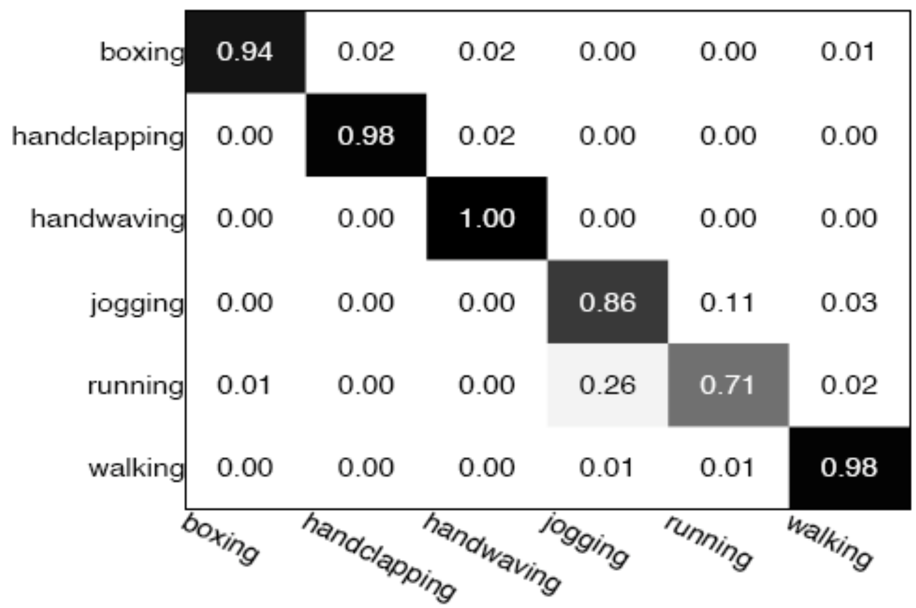


Learning is easier due to decoupling of model parameters
cf. Blei et al. JMLR 2003

Experiments: KTH dataset

- Benchmark dataset
  - 6 actions
  - 25 subjects
  - 4 scenarios

Method	Accuracy
Ours (sLDA)	91.2%
Liu & Shah CVPR08	94.2%
Jhuang and Poggio ICCV07	91.7%
Niebles & Fei-Fei BMVC06	81.5%
Schuldt & Laptev ICPR04	71.7%



Experiments: Soccer Dataset

- Real actions, moving camera, poor video
- 8 classes of actions
- 4500 frames of labeled data

Action	Our method (sLDA)	Efros et al. (k-NN)
Run left 45	0.64	0.67
Run left	0.77	0.58
Walk left	1.00	0.68
Walk in/out	0.86	0.79
Run in/out	0.81	0.59
Walk right	0.86	0.68
Run right	0.71	0.58
Run right 45	0.66	0.66

SFU Vision and Media Lab

Experiments: Irregularity detection

- sLDA is full probabilistic model
- Can detect most unusual sequences via likelihood
  - Sequences with lowest likelihood under model shown

SFU Vision and Media Lab

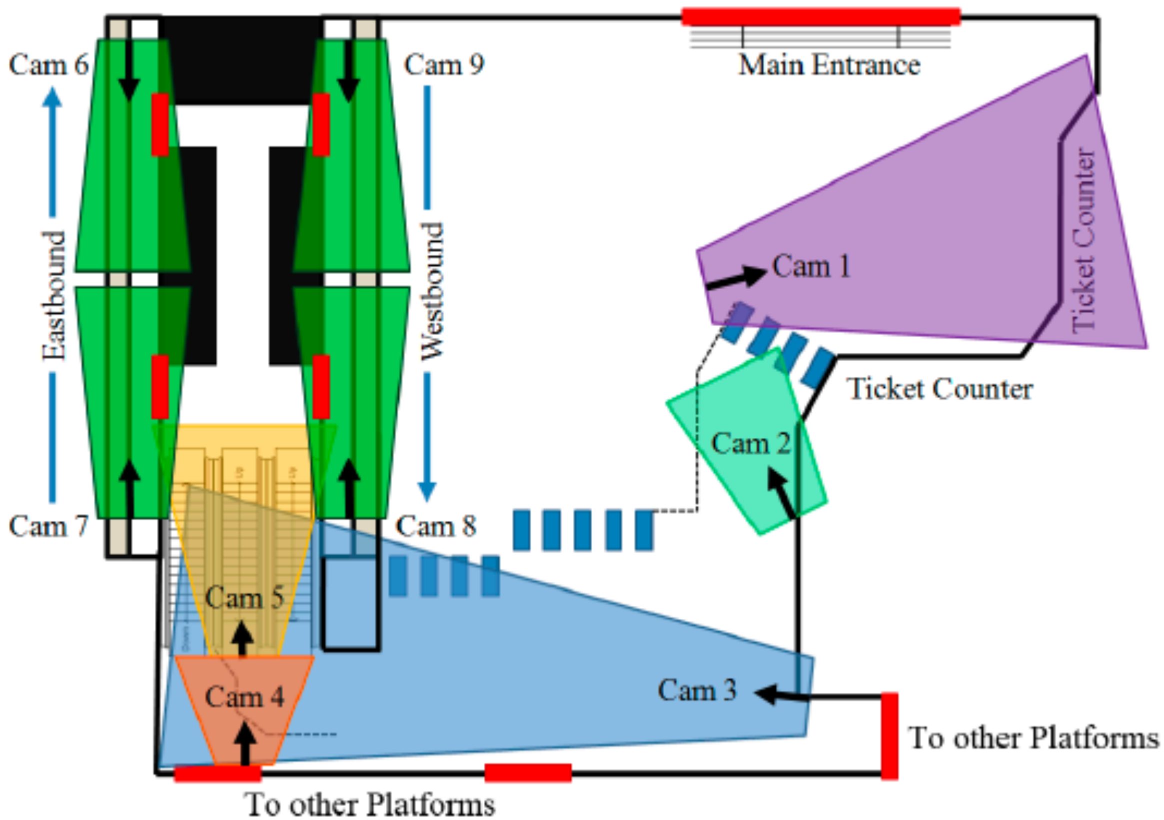
CAMERA NETWORKS

SFU Vision and Media Lab

Multiple Cameras

- In many situations, we have a set of cameras views of a scene
  - Not necessarily overlapping
- Need models for activities that span these different views
- Loy et al. ICCV09 paper is an example of this type of work

Underground Scenario



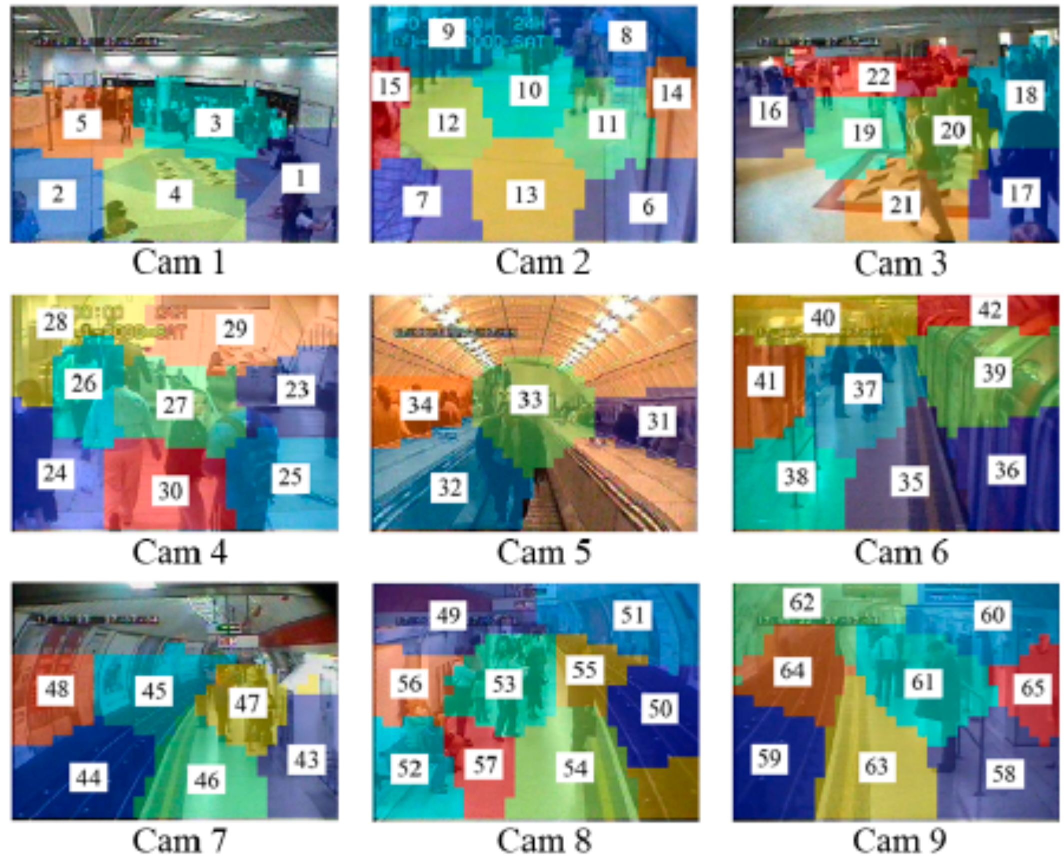
Detect Abnormal Events

- Approach
  - Build model of "normal"
    - Incorporate time delayed relationships over scene
    - Learn structure of these relationships
  - Score clips by likelihood under this model

SFU Vision and Media Lab

Camera View Regions (CVPR09)

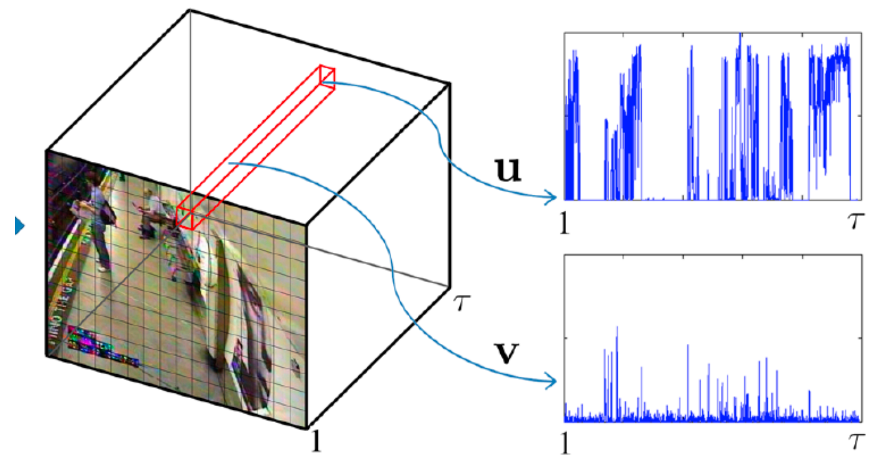
- Decompose each camera view into regions

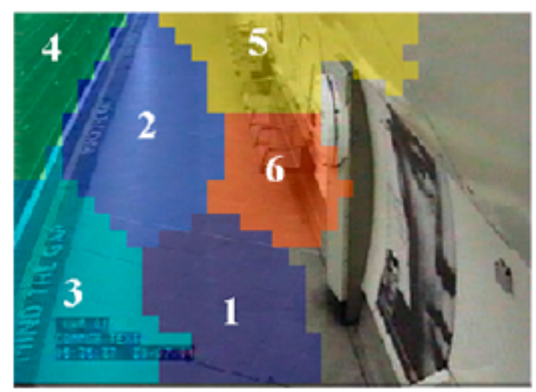


SFU Vision and Media Lab

Features

- Divide camera view into 10x10 pixel blocks
- Count:
  - #foreground pixels in block
  - #moving pixels in block
- Aggregate over time
- Compute correlation between time series for pairs of blocks
- Perform spectral clustering on blocks





SFU Vision and Media Lab

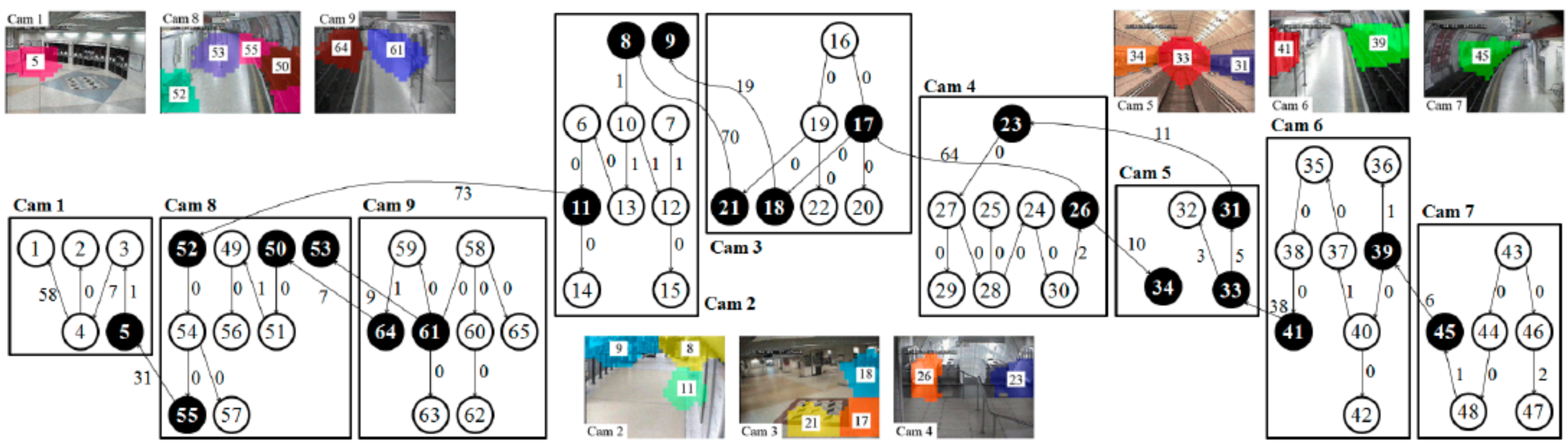
Time Delayed Analysis

- Given all the regions in all camera views, what are the relationships between them?
  - Describe each region with Gaussian mixture model on same foreground/moving features
  - Compute mutual information between pairs of time series
  - Search for best temporal offset between pairs of time series
    - Offset that maximizes mutual information

SFU Vision and Media Lab

Structure Learning

- Given all regions, and MI between each pair, build a Bayesian Network
  - Use Chow-Liu algorithm
  - Finds tree-structured BN that keeps most important edges



Other details

- Bayesian parameter learning
- Aggregate likelihoods over time to smooth out noise (Cumulative Abnormality Score)

SFU Vision and Media Lab