Secure Our Society – Computer Vision Techniques for Video Surveillance

Huiyu Zhou
E-mail: h.zhou@qub.ac.uk

24th February, 2016

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Video surveillance…

- Introduction
- Human detection and tracking
- Human profiling
- Activity recognition
- Trajectory clustering
- Summary

Scope of this tutorial

- In this tutorial we talk about the techniques directly used for video surveillance.
- We will go through general concepts, representative methodologies and key stages of the relevant techniques.
- We assume that the audience holds fundamental knowledge in computer vision, computer graphics and image understanding – what happens if not?

To kick-off

- What is “video surveillance”?
- Why is it so important?
- What is the need and technical challenge of this topic?

Definition

- Video surveillance – Wikipedia:

It is a process where video cameras are deployed in order to monitor the behaviour, activities or other change information of people for the purpose of influencing, directing or protecting.

Image courtesy of Ifacility Co.

Categories: generic

- Active: monitoring an area for assisting security officers.
- Passive: an employee monitors a few screens while working on other tasks.
- Recording: collecting information for investigation and evidence purposes.

Citation from V. Gouaillier and A.-E. Fleurant

Categories: specific

- Detection of changes.
- Segmentation of moving objects.
- Tracking of objects.
- Classification and identification of objects.
- Classification of activities and behaviours.

Importance

- Acts as a deterrent to crime.
- Helps apprehend a suspect when a crime occurs.
- Improves the productivity of employees.

http://blog.protectussecurity.com/2012/01/12/video-surveillance-blog-the-benefits-of-video-surveillance/

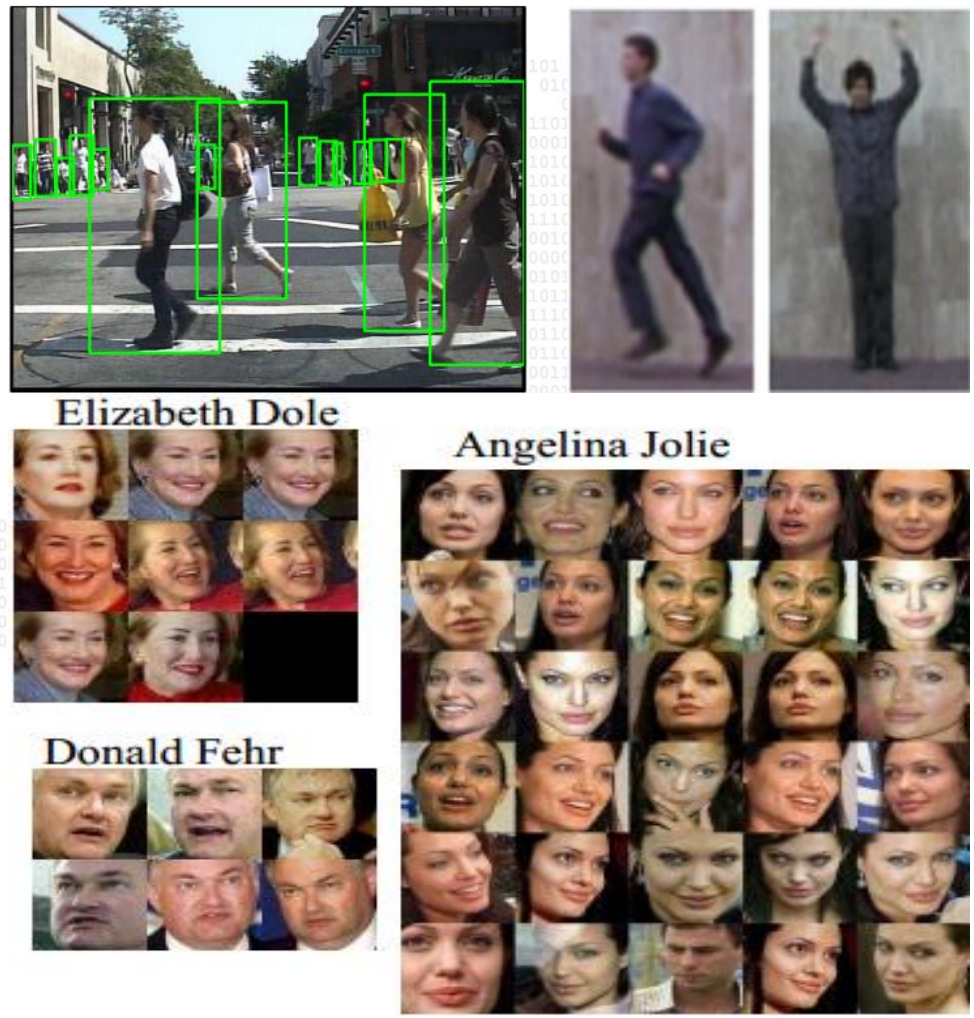
Needs

- Minimising system configuration.
- Good system performance.
- No camera calibration.
- Generic as much as possible.
- Privacy protection.

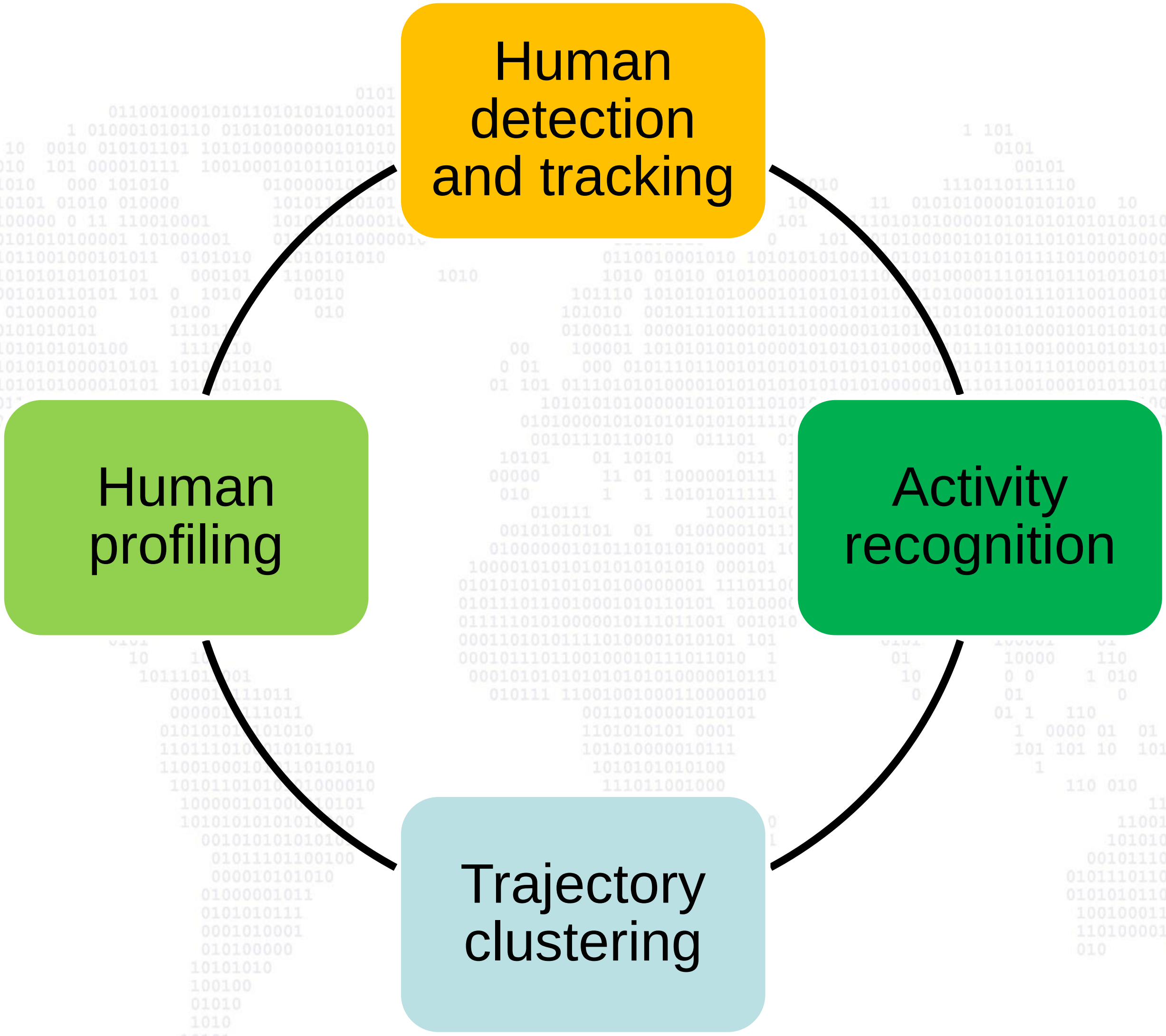
Citation from V. Gouaillier and A.-E. Fleurant

Challenges

- Real-time human detection and tracking.
- Consistent human identification and recognition.
- Reliable behaviour/activity understanding and interpretation.



Key components



- Introduction
- Human detection and tracking
- Human profiling
- Activity recognition
- Trajectory clustering
- Summary

Human detection

Overview

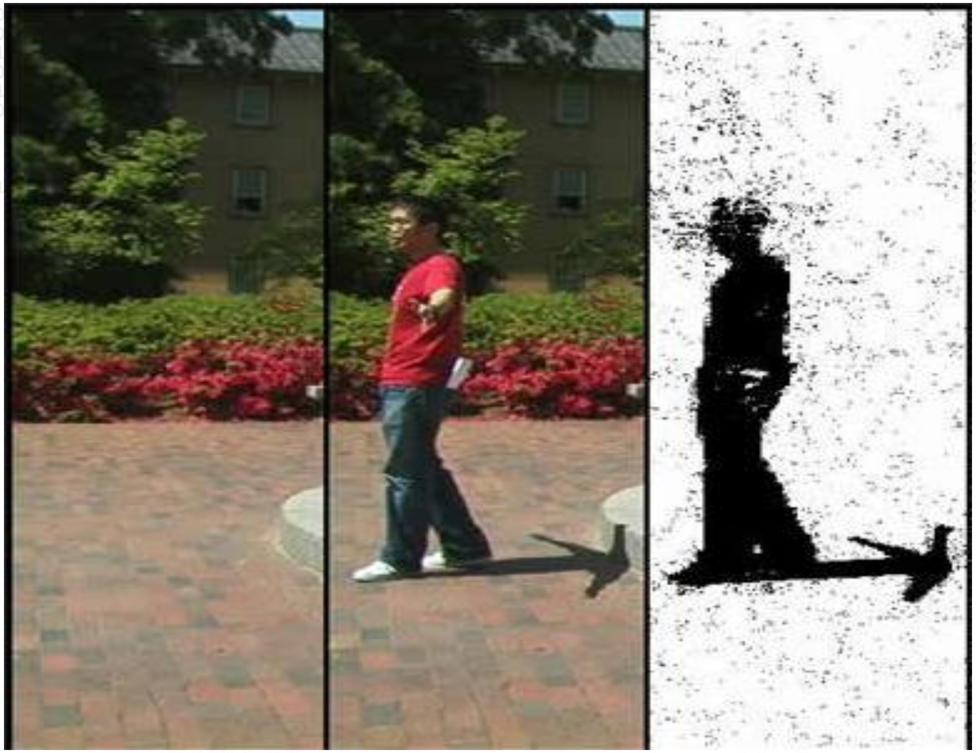
- Overview
- Background subtraction
- Viola-Jones method
- Histograms of Oriented Gradients (HoG)
- Shape context

Overview

- Feature representation:
  - Haar wavelets (Viola et al, 2003; Pyun, et al, 2014).
  - Edges (Gavrila and Philomin, 1999; Shen, et al, 2015).
  - Gradient orientations (Dalal and Triggs, 2005; Tzimiropoulos, et al, 2012).
  - Gradient and second derivatives (Ronfard et al, 2002).
  - Regions (Mori at al, 2004).
- Feature classification:
  - Template matching (Gavrila and Philomin, 1999; Dekel, et al, 2015).
  - Support Vector Machine (Ronfard et al, 2002; Zhou, et al, 2011).
  - Adaboost (Viola et al, 2003; Cai, et al, 2015).
  - Grouping (Mori et al, 2004).
  - Bayesian, Neural Network/Deep Learning, MCMC, etc..

Background subtraction

- Naïve approach: foreground objects ARE the difference between the current frame and a clean reference image.



Background subtraction

- Challenges:
  - Illumination changes, e.g. shadows.
  - Motion changes, e.g. background objects change.
  - Changed background geometry, e.g. moving cars.

Background subtraction

- Improved versions of the naïve version
  - **Average** or **median** of previous *n* frames (Lo and Velastin, 2000; Cucchiara at al, 2003)
    - Pros: fast.
    - Cons: memory consuming.
  - **Running average**

$$B_{i+1} = \alpha * F_i + (1 - \alpha) * B_i$$

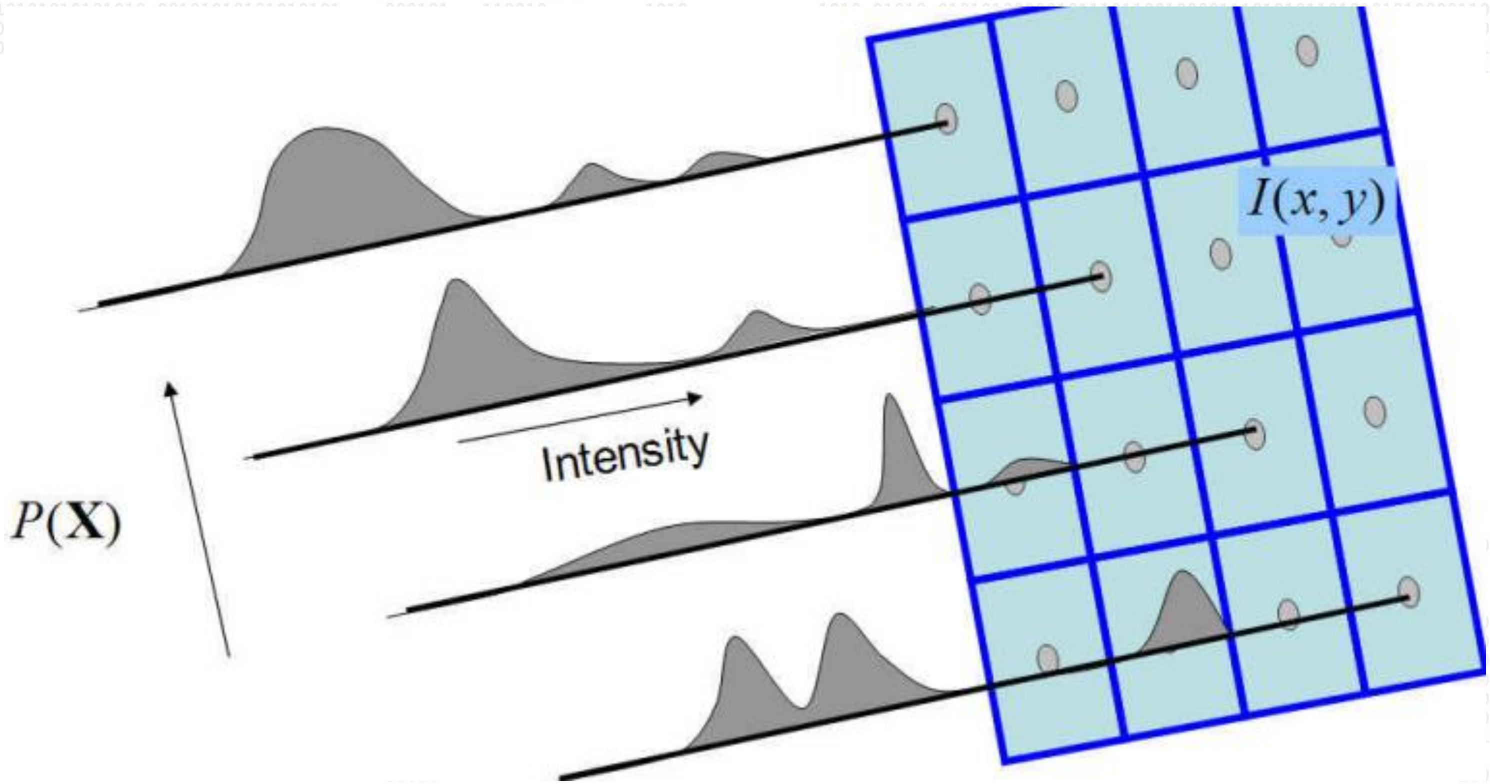
Where α is the learning rate, *F* is the foreground and *B* the background.

Background subtraction

- Major problems of the naïve methods
  - – No strategy available to choose the threshold.
  - – Cannot cope with multiple background distributions.

Background subtraction

- Mixture of Gaussians (Stauffer and Grimson, 1999):
  - Each pixel value in an image can be modelled by a mixture of Gaussian distributions.



Mixture of Gaussian

- The values of a particular pixel is modeled as a mixture of adaptive Gaussians.
  - Why mixture? Multiple surfaces appear in a pixel.
  - Why adaptive? Lighting conditions change.
- At each iteration Gaussians are evaluated using a simple heuristic to determine which ones are mostly likely to correspond to the background.
- Pixels that do not match with the “background Gaussians” are classified as foreground.
- Foreground pixels are grouped using 2D connected component analysis.

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Demo of MoG

Courtesy of the algorithmic developer

Variants of MoG

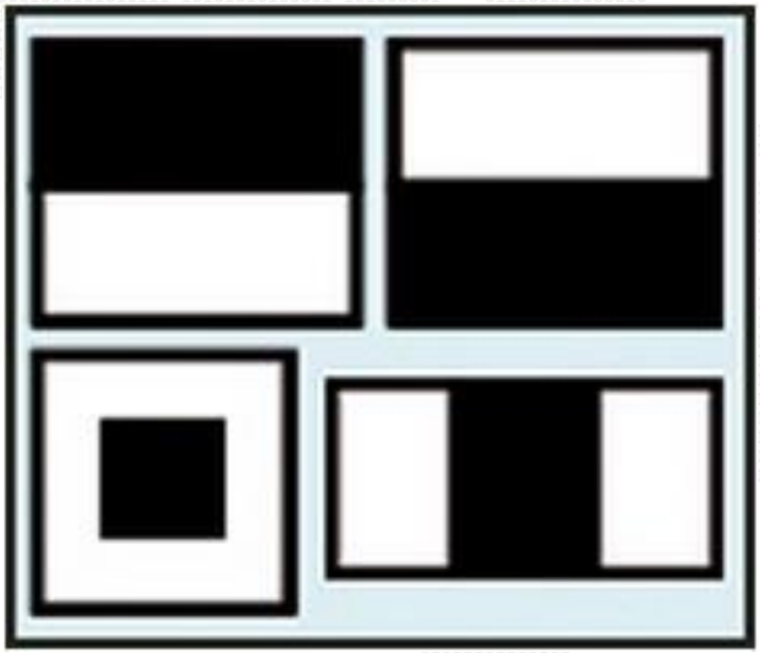
- Regularised region-based MoG (Varadarajan et al, 2014 and 2015).

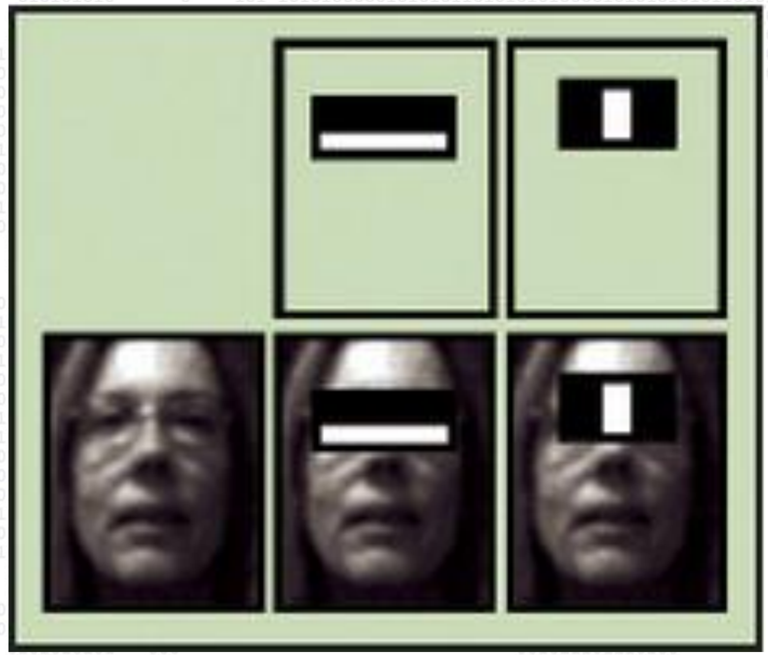
Other approaches

- Kernel density estimators (Elgammal et al, 2000; Narayana et al, 2013).
- Mean-shift (Han et al, 2004; Cho and Kang, 2011).
- Eigenbackgrounds (Oliver et al, 2000; Hu, et al, 2011).

Viola-Jones method

1. Rectangular features, called Haar features.

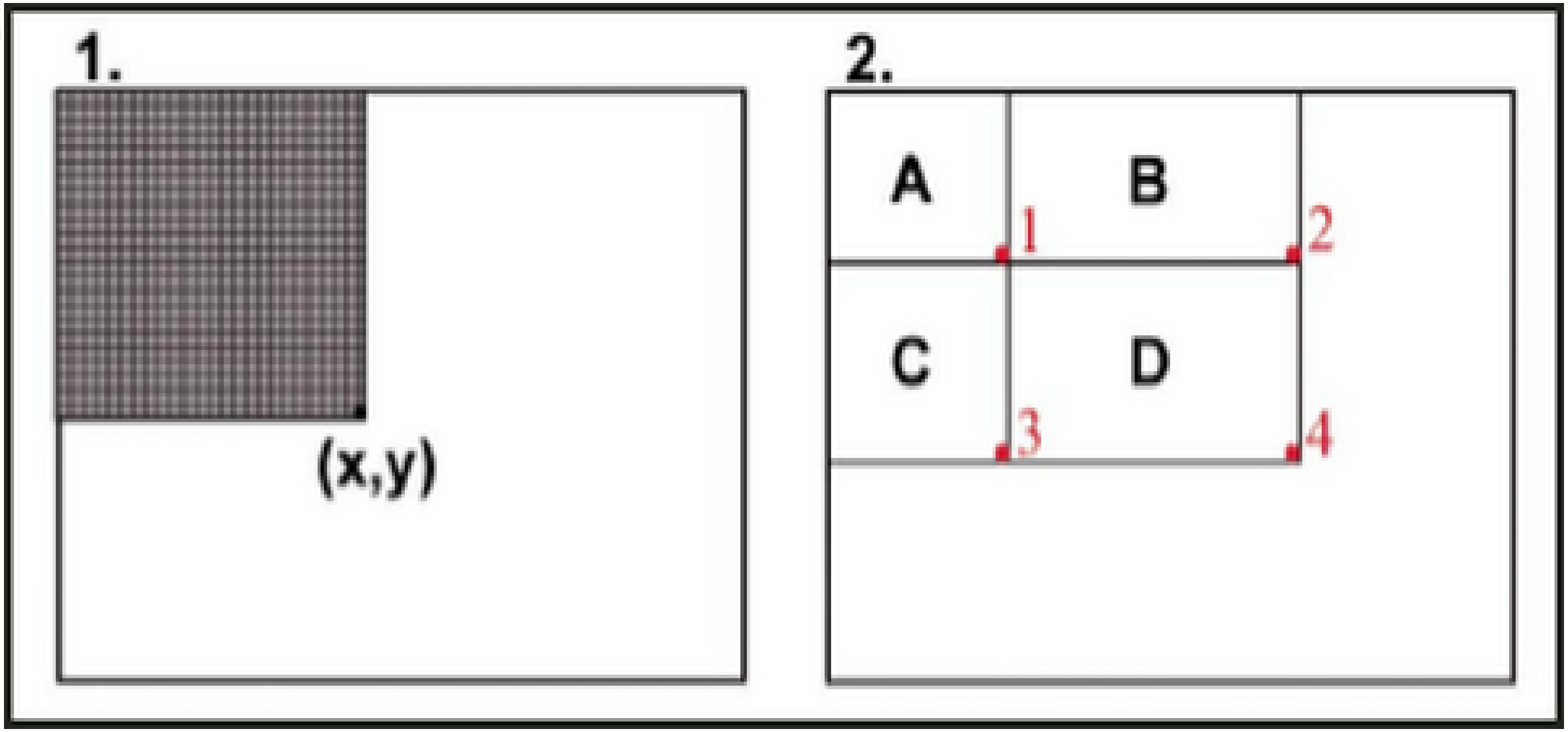




http://www.cognotics.com/opencv/servo_2007_series/part_2/sidebar.html

Viola-Jones method

2. An integral image for rapid feature detection:
   – Integral value of each pixel is the sum of all the pixels above it and to its left.



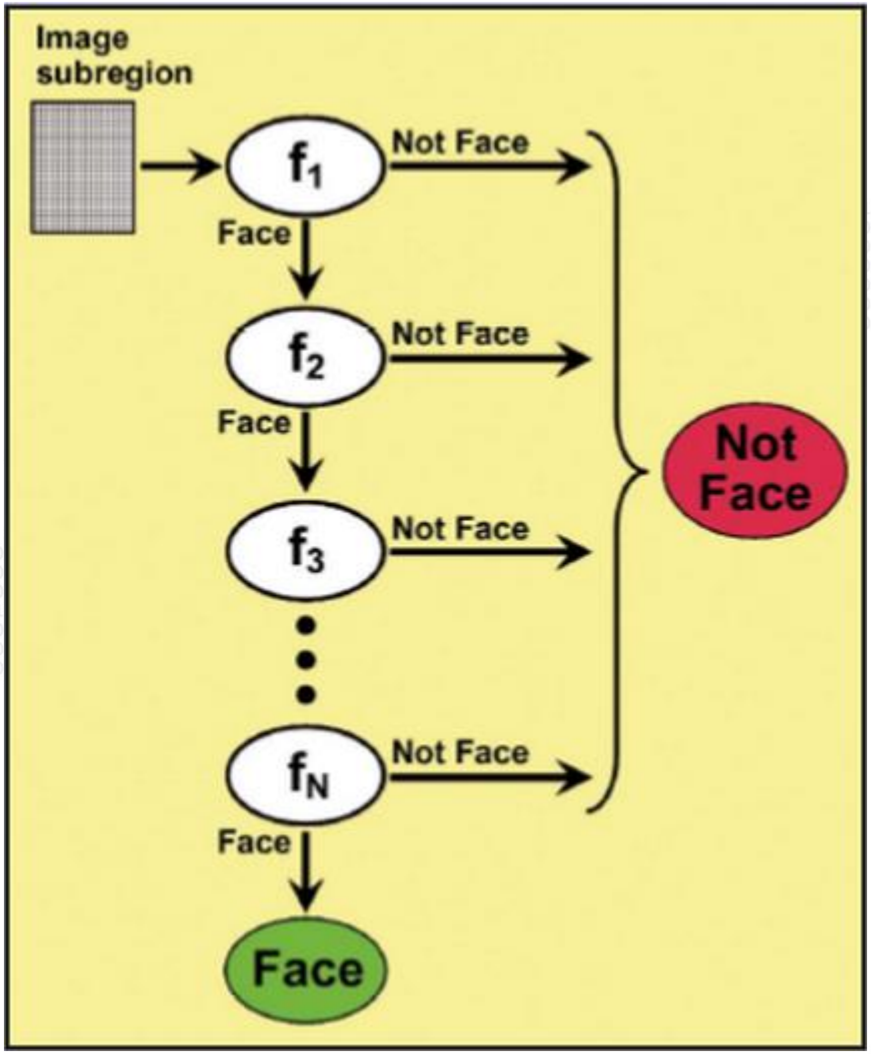
http://www.cognotics.com/opencv/servo_2007_series/part_2/sidebar.html

Viola-Jones method

3. Adaboost method:
   - Selecting a set of weak classifiers to combine and assigning a weight to each.
   - The weighted combination is the stronger classifier.

Viola-Jones method

4. A cascaded classifier to combine features.



HoG (Dalal and Triggs, 2005)

- Motivation of the development:
  - Human shape is characterised by the distribution of local intensity gradient or edge directions.

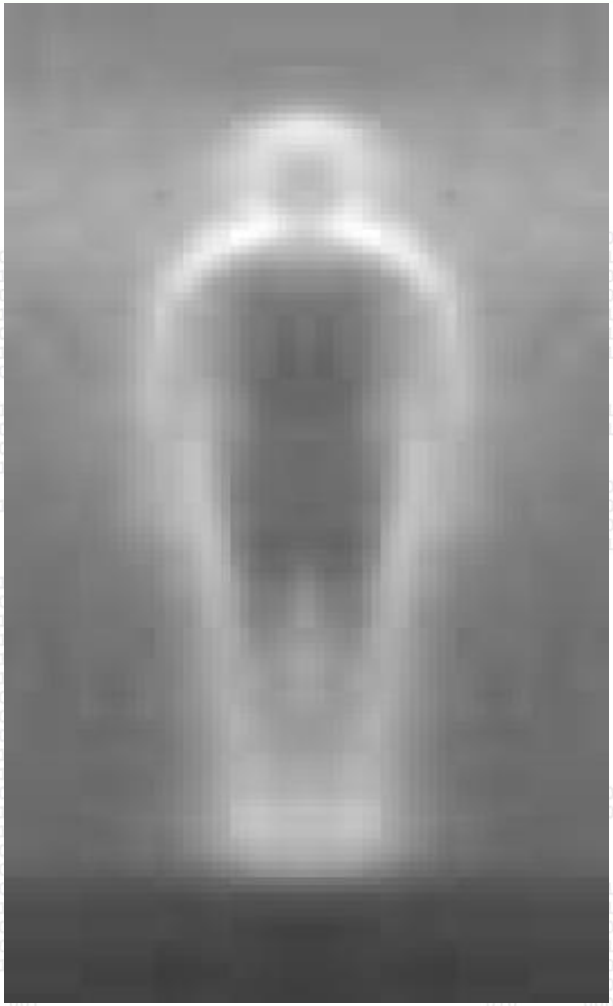
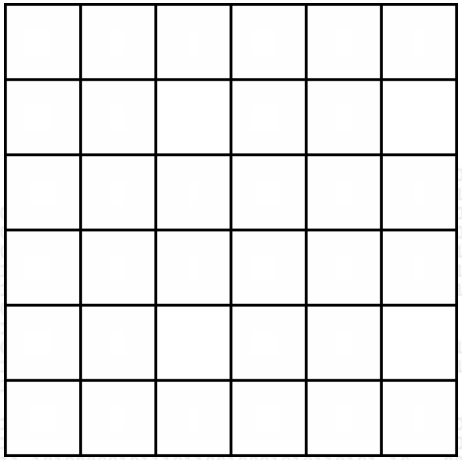


Image courtesy of Tsai

HoG

- Divide the image into small cells.
- Cells can be rectangle or radial.
- Accumulating a weighted local 1-D histogram of gradient directions over the pixels of the cell.



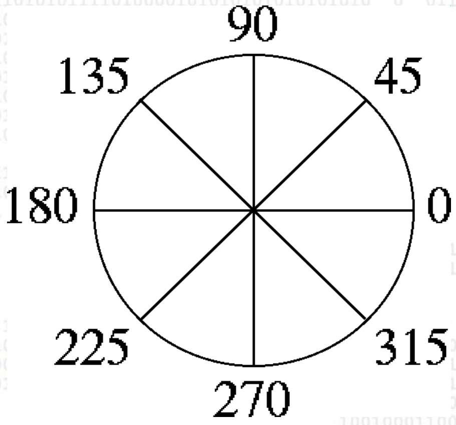


Image courtesy of Tsai

HoG

- Contrast-normalise local responses for illumination invariance.
- Accumulating a local histogram over a larger region to normalise all the cells.

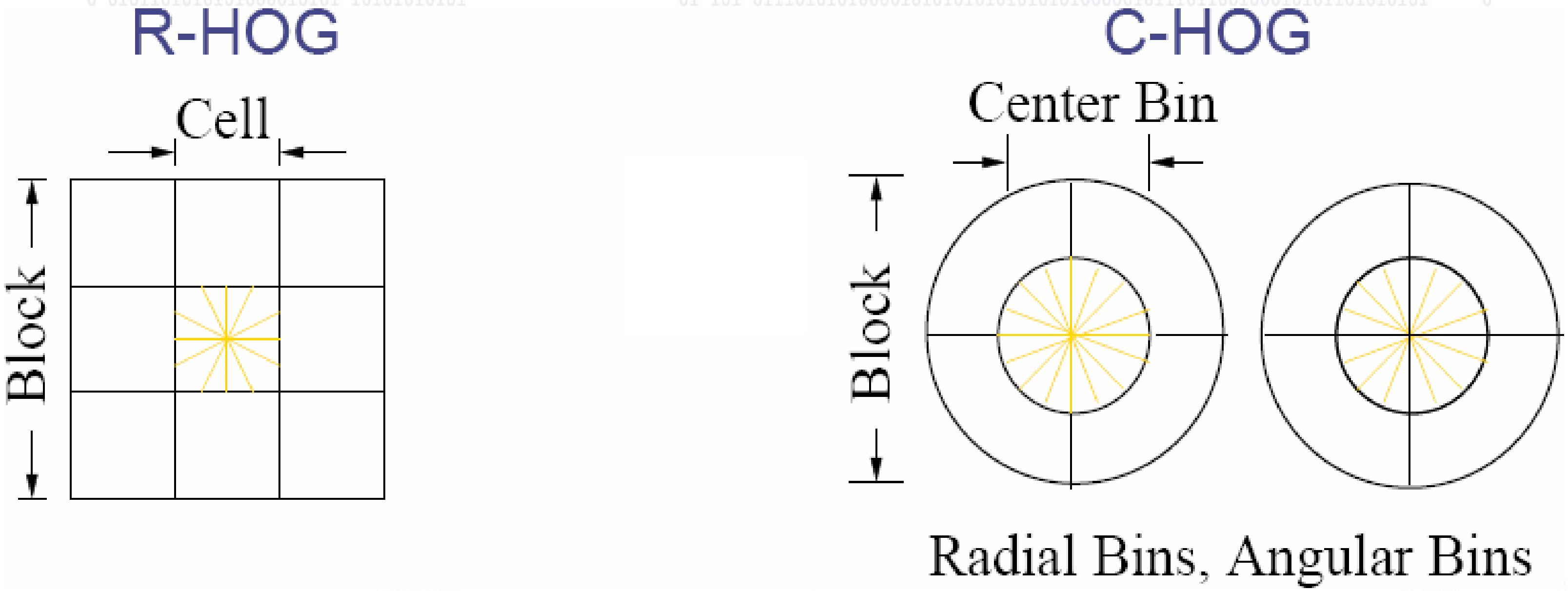


Image courtesy of Tsai

Shape context

1

- N-samples from edges

2

- Euclidean-distance r and angles from one to the remainder

3

- Normalise r and angle on x-axis

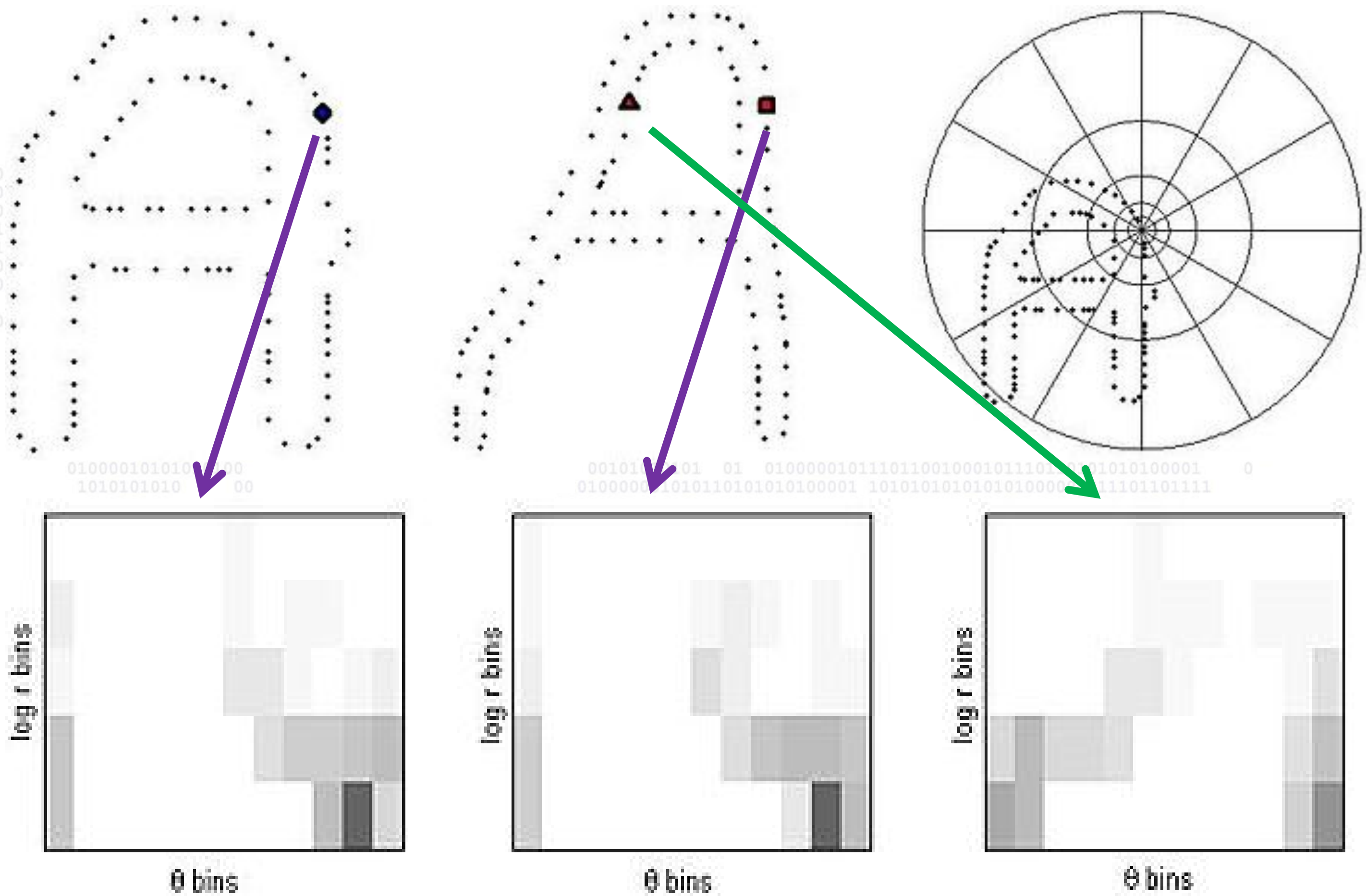
4

- Log of r and discretisation of distance/angle

5

- No. of points in a bin

Matching shapes

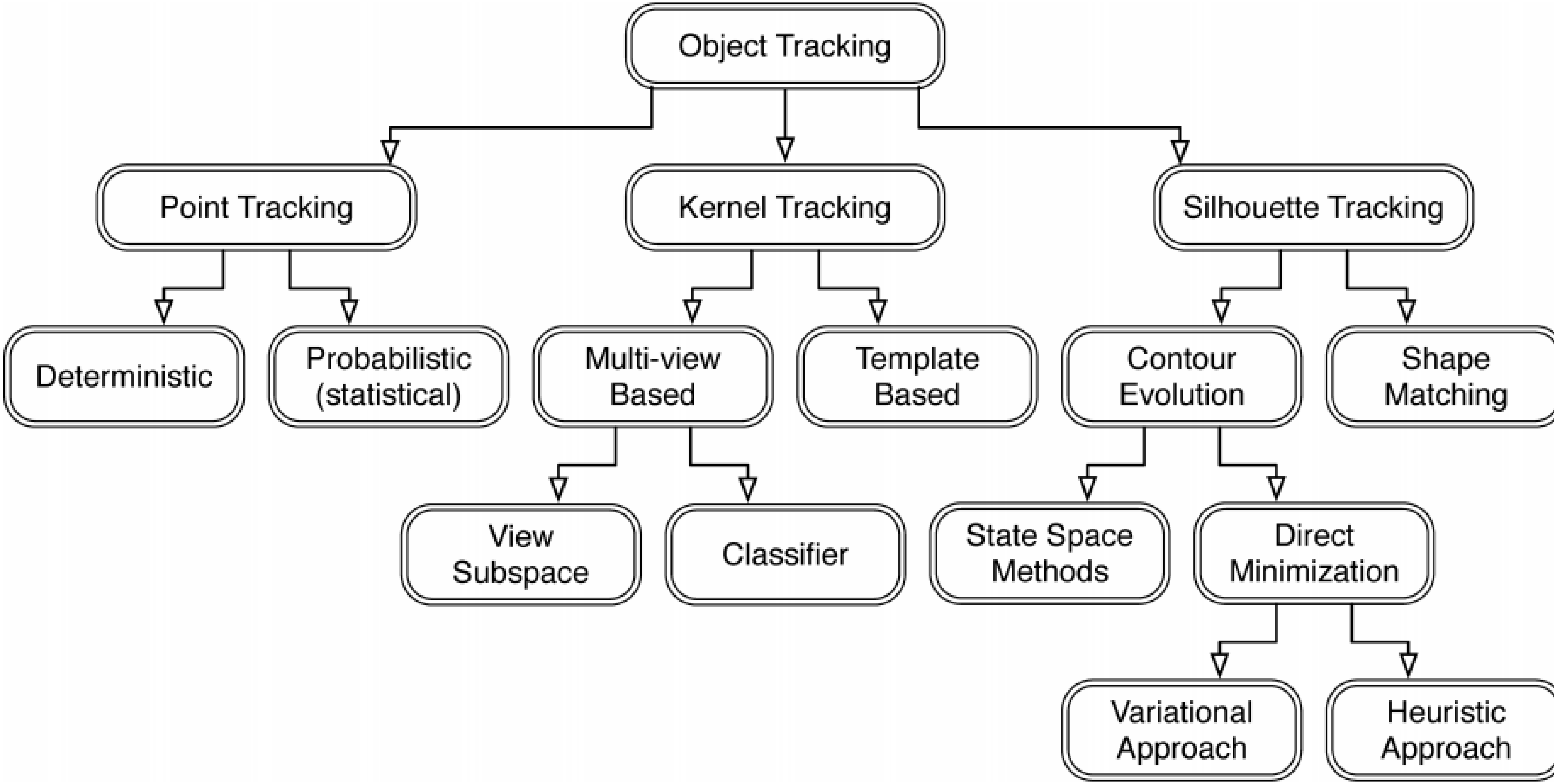


Human tracking

Overview

- Established techniques.
- Exemplar approaches.
- Incremental learning for visual tracking.
- Tracking with online multiple instance learning.
- Combining local features with kernel tracking.
- Audiovisual tracking.

Established techniques



Yilmaz et al, 2006

Exemplar approaches

- Point tracking:
  - Kalman filter (Broida and Chellappa, 1986; Zhou, et al, 2008)
  - JPDAF (Bar-Shalom and Foreman, 1998; Zhou, et al, 2008)
  - PMHT (Streit and Luginbuhl, 1994)
- Kernel tracking:
  - Mean-Shift (Comaniciu et al, 2003; Zhou, et al, 2009)
  - KLT (Shi and Tomasi, 1994; Zhou, et al, 2009)
  - Muti-view: Eigentracking (Black and Jepson, 1998)
- Silhouette tracking:
  - State space model (Isard and Blake, 1998)
  - Hough transfer (Sato and Aggarwal, 2004)
  - Graph cuts (Ma, et al, 2010)

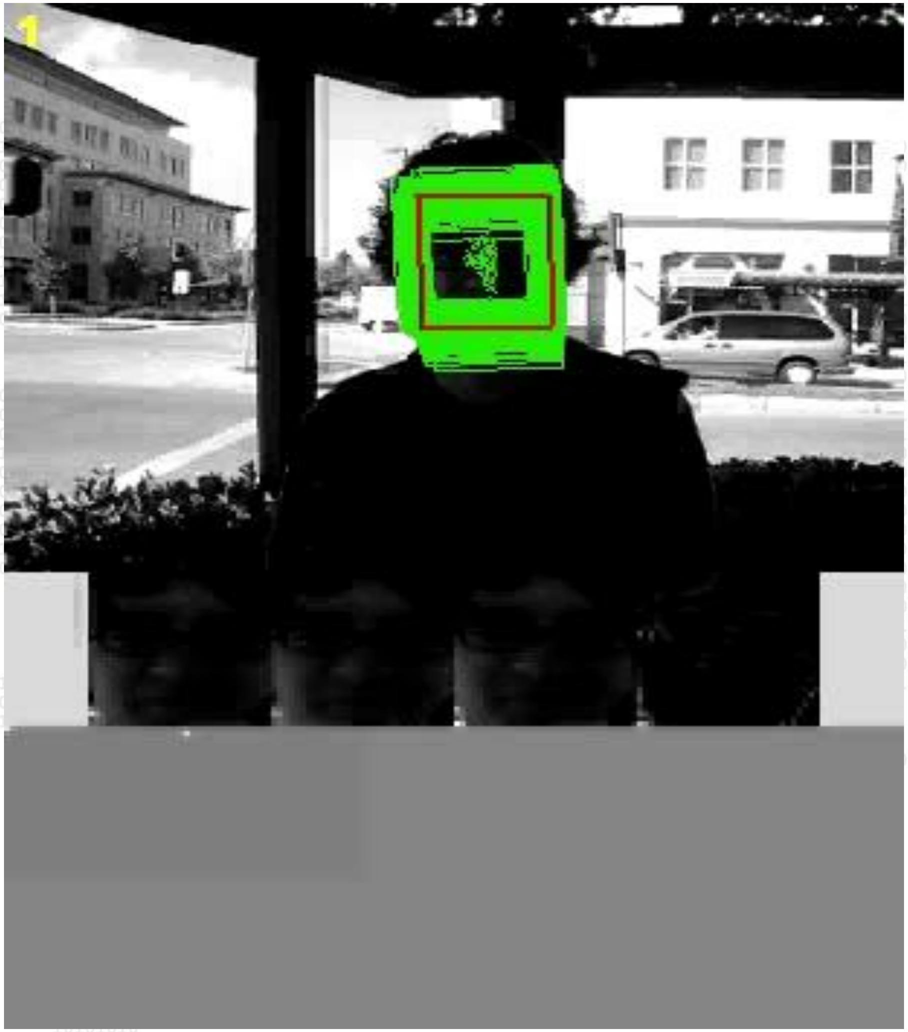
Incremental learning for visual tracking

- Issues of classical approaches:
  - Build an appearance model before tracking.
  - View based.
  - Complicated optimisation.
- Challenges:
  - Object appearance and the scene are dynamically changed.
  - Pose variations.
  - Drifts.

Incremental learning for visual tracking

- Algorithm (Lim et al, 2004):
  - Choose an initial location L_0.
  - Search for possible locations: $p(L_t | L_{t-1})$→dynamic model.
  - Predict a location: $p(L_t | F_t, L_{t-1}) \propto p(F_t | L_t) p(L_t | L_{t-1})$, where $p(F_t | L_t)$ is the observation model using Eigenbasis.
  - Use R-SVD algorithm to update Eigenbasis.

Incremental learning for visual tracking

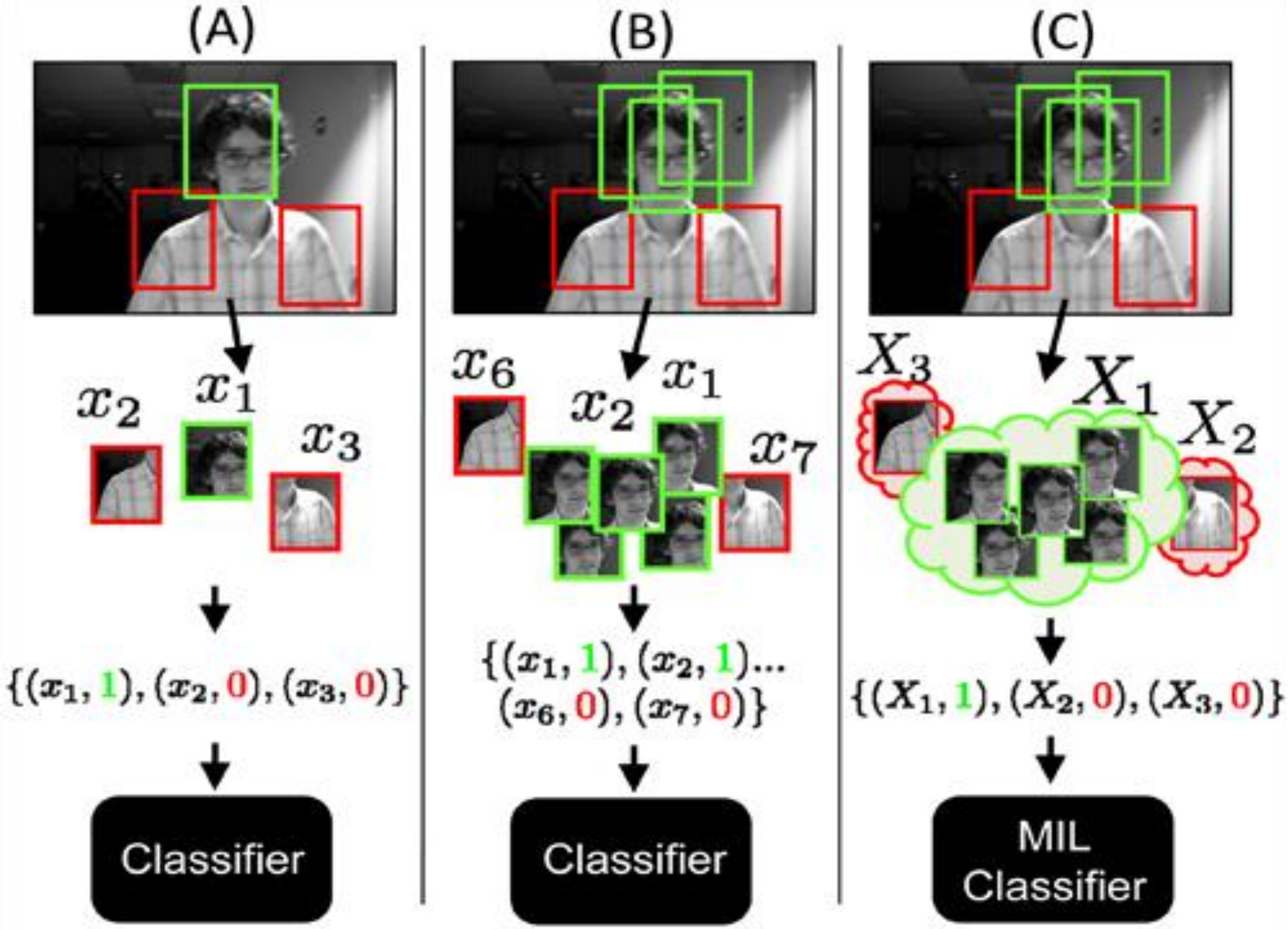


Courtesy of the algorithmic developer

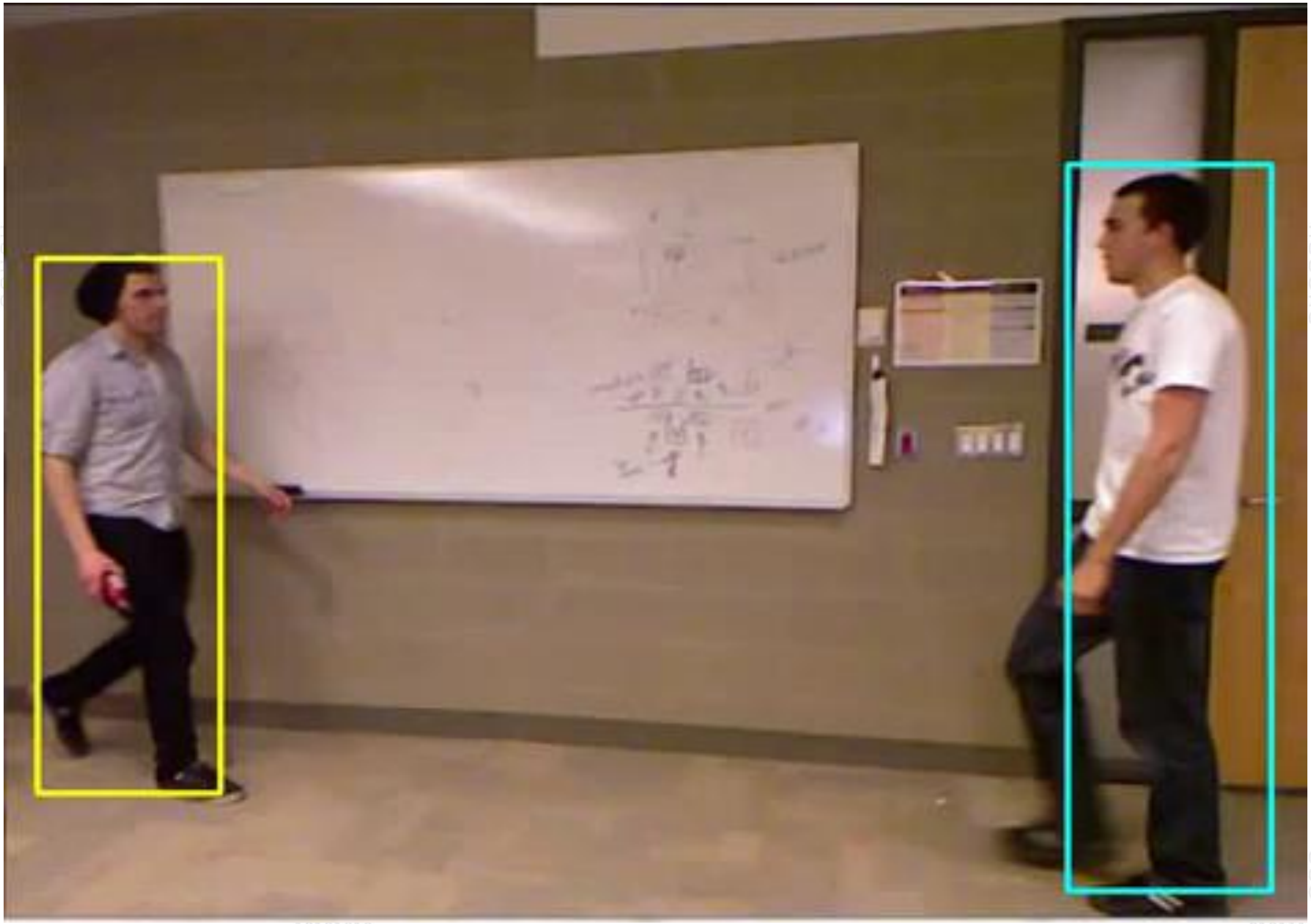
Tracking with online multiple instance learning

- Classical “tracking by detection”:
  - Train a discriminative classifier on-line to separate the object from the background.
  - The classifier uses the current state to extract positive/negative examples from the current frame.
  - Inaccurate tracks can lead to incorrectly labelled examples.
  - Drifts occur due to the poor examples.

MILTrack

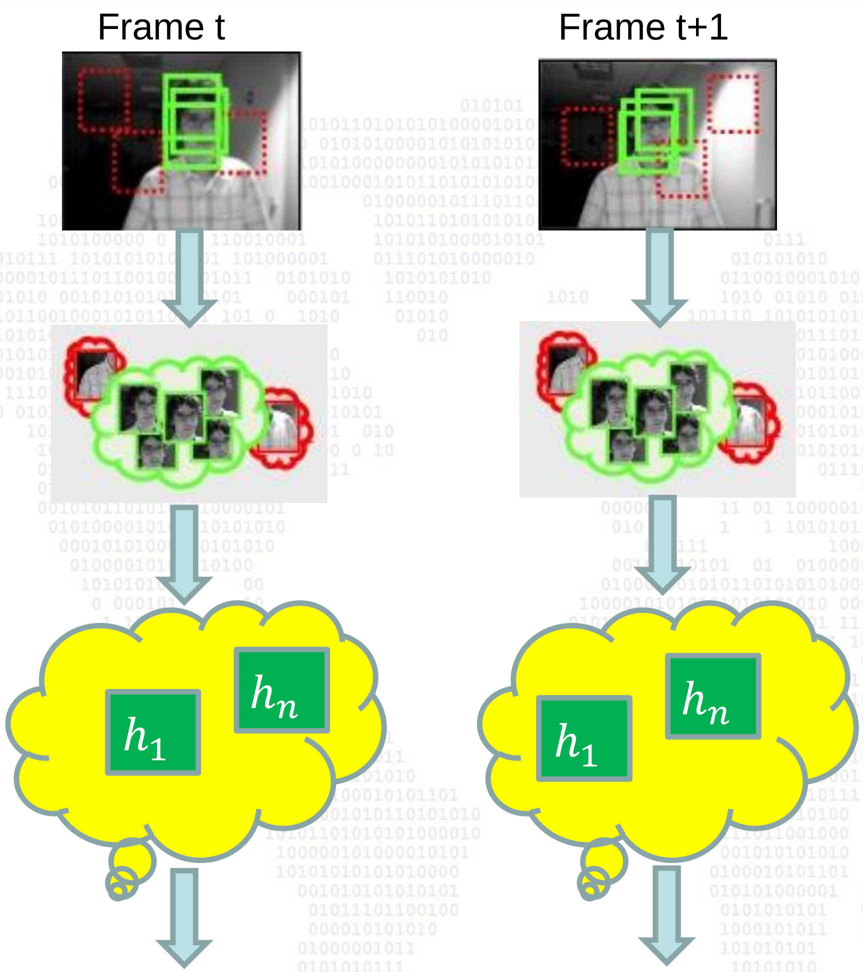

(A)
(B)
(C)
x_2 x_1 x_3
$\{(x_1, 1), (x_2, 0), (x_3, 0)\}$
Classifier
x_6 x_2 x_1 x_7
$\{(x_1, 1), (x_2, 1)\ldots (x_6, 0), (x_7, 0)\}$
Classifier
X_3 X_1 X_2
$\{(X_1, 1), (X_2, 0), (X_3, 0)\}$
MIL Classifier

Demo of MILTrack



Courtesy of the algorithmic developer

Online MIL Boost



$$H = \underset{h \in \{h_1, \ldots, h_n\}}{\operatorname{argmax}} \; L(H_k + h)$$

H

Obtain bags of words

Update classifiers in the pool

Greedily add best k candidates to the stronger classifier

Combining local features with kernel tracking

- Issues of classical mean-shift:
  - Less efficient in the presence of significant intensity or colour changes.
  - Lacks consistency in the case of occlusions.
  - Best works with colour features.
- SIFT features – scale invariant feature transform:
  - Keypoint localisation:
    - Interpolation of neighbouring data.
    - Discarding low-contrast keypoints.
    - Eliminating edge responses.
  - Orientation assignment.
  - Keypoint descriptor.

Zhou, et al, 2009

Combining local features with kernel tracking

- Entire algorithm:
  - Choose a region to track in the current frame.
  - Apply CamShift to find a possible match in the next frame.
  - Generate a set of windows around the centre of the match window.
  - Match the extracted SIFT features from the two frames.
  - Obtain the residuals of colour and SIFT based matching.
  - Establish a weighted cost function for the residuals.
  - Apply an EM algorithm to search for the window with the minimum residuals.

Comparisons

Mean Shift based

SIFT based

Proposed system

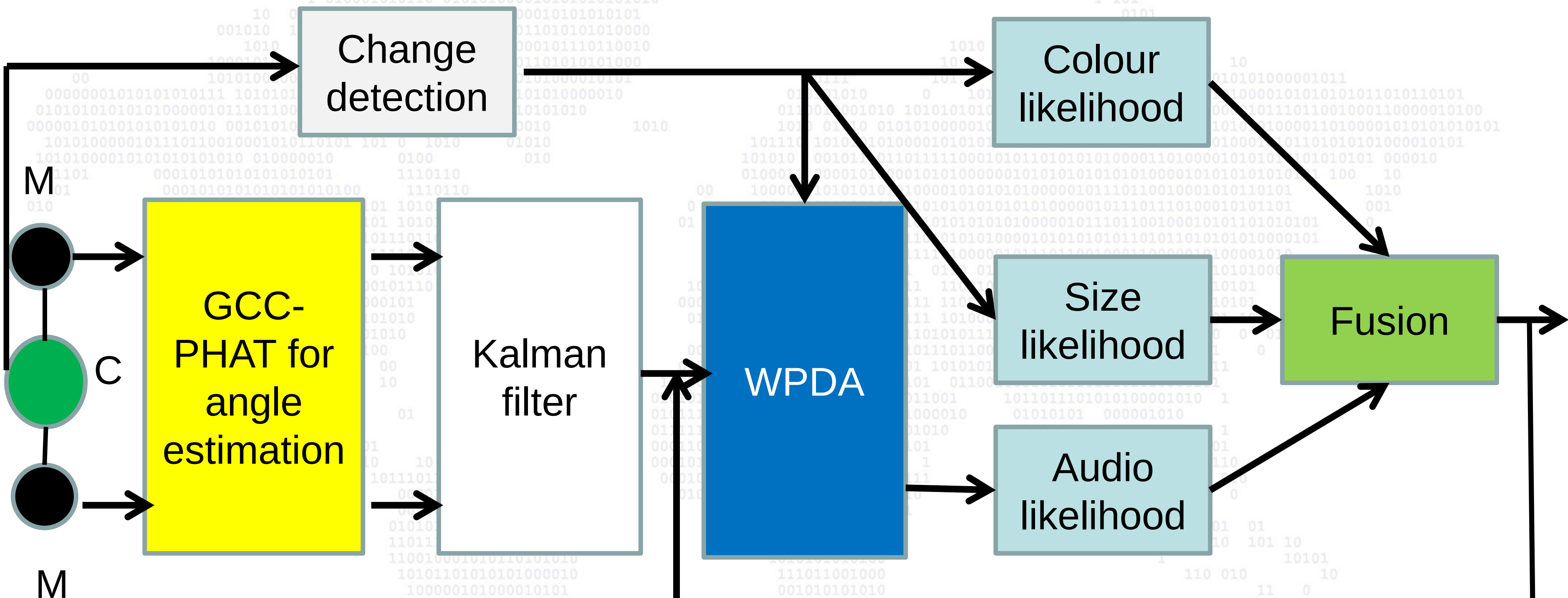
Audiovisual tracking

- Benefits of using multi-modality based systems:
  - Each modality may compensate for the weakness of the other.
  - Each modality can provide additional information.
- Challenges of audiovisual systems:
  - Unstable acoustic measurement.
  - Importance determination of audio and visual components.

Established approaches

References	Sensor types	Algorithms	Applications
Asoh, 2004	Stereo camera and circular microphone array	PF	Multimodal user interface
Checka et al, 2004	2 cameras and 4 microphone arrays	PF	Indoor multiple person tracking
Cevher et al, 2007	Camera and 10 element uniform circular array	PF	Outdoor surveillance
D. Gatica-Perez, et al, 2003	Wide-angle camera and a microphone array	I-PF	Meeting rooms
Rui and Chen, 2001	PTZ camera and 2 microphones	PF	Teleconferencing
Beal et al, 2002	Camera and 2 microphones	GM	Indoor environments

Target detection and tracking with heterogeneous sensors



Zhou et al, 2008

Target detection and tracking with heterogeneous sensors

- Exemplar results:

  Comparison of tracking results for “1-room” (Frame numbers: (a) 814,

  (b) 926, and (c) 1010).

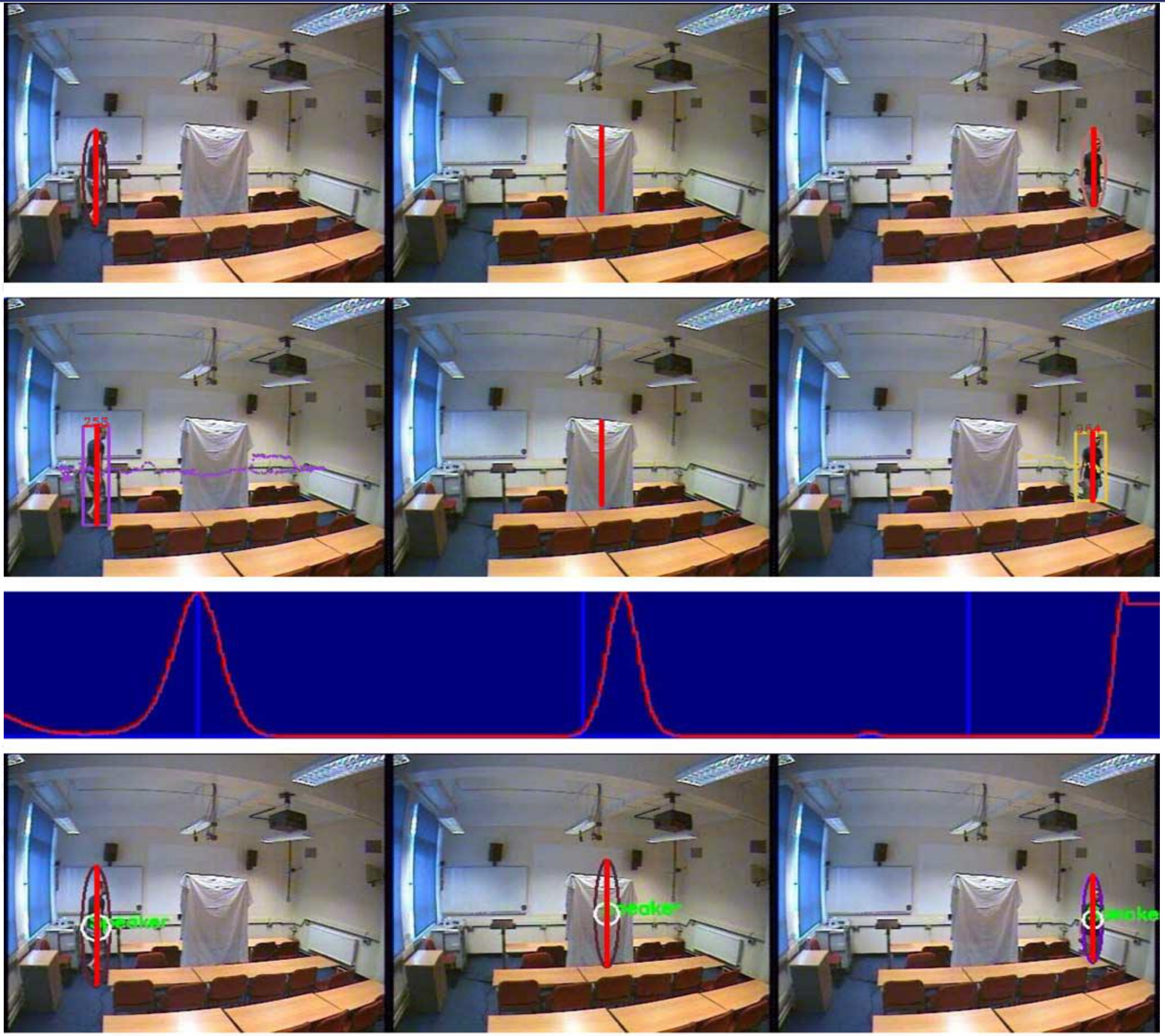
  Row 1: PF (Particle Filter);

  Row 2: GM (Graph Matching);

  Row 3: GCC (*generalized cross correlation*);

  Row 4: KF-PF-P (Kalman filtering audio detection and the particle filter-based audiovisual tracker with PDA).

  The red bar indicates the true target position.



(a) (b) (c)

- Introduction
- Human detection and tracking
- **Human profiling**
- Activity recognition
- Trajectory clustering
- Summary

Introduction

- Profiling
  - “Extrapolation of information about something, based on known qualities” (Wikipedia).
  - No explicit definition for “human profiling”.
- In homeland security
  - ~70% crime offender are young adolescent males in the UK.
  - There is a need to identify gender, age and ethnicity of a pedestrian through facial or body images.
- This is a classification problem
  - Separate pedestrians into different groups.

Overview

- Challenges
- General approaches
- State of the art techniques
- Exemplar systems for age/gender/ethnicity classification:
  - Age classification using Radon transform and scaling SVM.
  - Ethnicity classification based on gait using multi-view fusion.
  - Ethnicity- and gender-based subject retrieval using 3-D face-recognition techniques.

General approaches

(1) Principle Component Analysis (PCA).
(2) Scale Invariant Feature Transform (SIFT).
(3) Histogram of Oriented Gradients (HOG) and variants.
(4) Gabor.
(5) Local Binary Patterns (LBP).
(6) Speeded Up Robust Feature (SURF).

Feature extraction

(1) Support Vector Machine (SVM).
(2) Nearest Neighbor.
(3) Linear Discriminant Analysis (LDA).
(4) Boosting.
(5) Bayesian.
(6) Neural Networks.
(7) Hidden Markov Model (HMM).
(8) Active Appearance Model (AAM) with a classifier.

Feature classification

Previous techniques

- Age classification
  - **Kwon and Lobo, 1999**: Geometrical ratios from the distance and size of facial characteristics and wrinkles detected by snakes.

Previous techniques

- Age classification
  - **Kwon and Lobo, 1999**: Geometrical ratios from the distance and size of facial characteristics and wrinkles detected by snakes.
  - **Lanitis et al, 2002**: Active Appearance Model based coding for dimensional reduction.

Previous techniques

- Age classification
  - **Kwon and Lobo, 1999**: Geometrical ratios from the distance and size of facial characteristics and wrinkles detected by snakes.
  - **Lanitis et al, 2002**: Active Appearance Model based coding for dimensional reduction.
  - **Geng et al, 2007**: Generated aging patterns for each person in a dataset, where face images show each subject at different ages.

Previous techniques

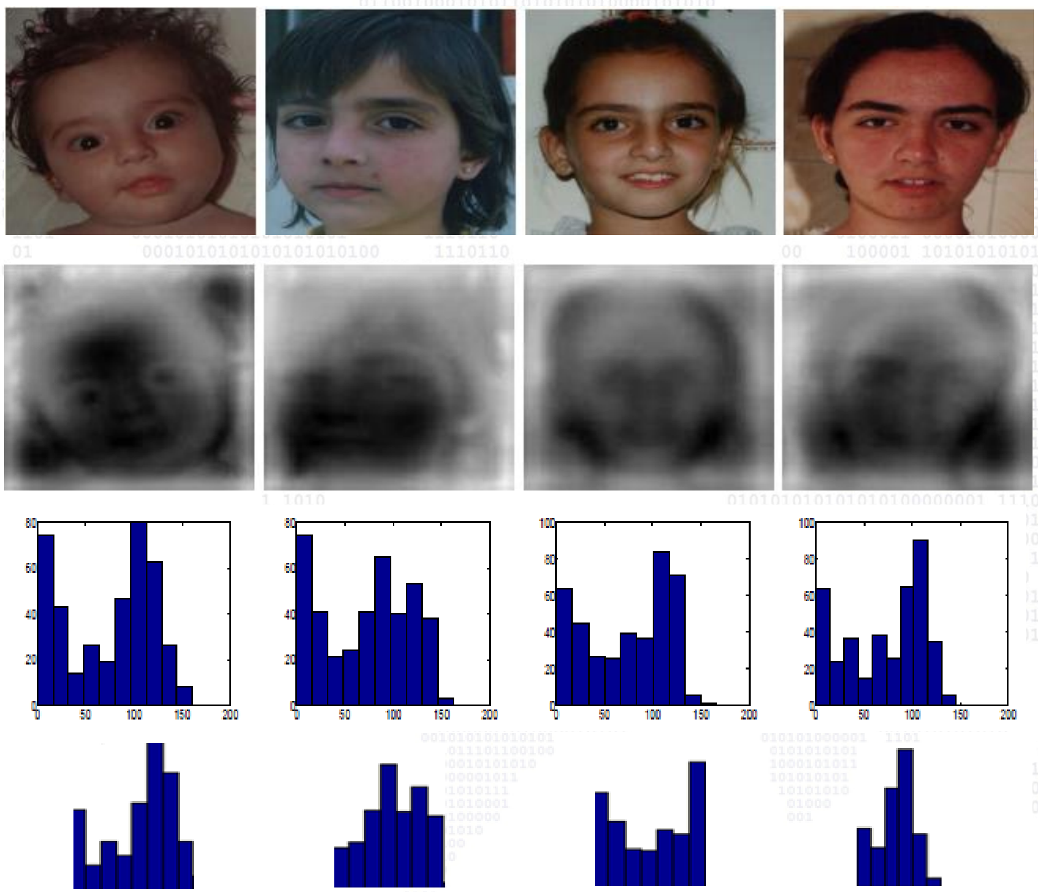
- Age classification
  - **Kwon and Lobo, 1999**: Geometrical ratios from the distance and size of facial characteristics and wrinkles detected by snakes.
  - **Lanitis et al, 2002**: Active Appearance Model based coding for dimensional reduction.
  - **Geng et al, 2007**: Generated aging patterns for each person in a dataset, where face images show each subject at different ages.
  - **Fu and Huang, 2008**: Represent aging patterns using manifold learning.

Previous techniques

- Age classification
  - **Kwon and Lobo, 1999**: Geometrical ratios from the distance and size of facial characteristics and wrinkles detected by snakes.
  - **Lanitis et al, 2002**: Active Appearance Model based coding for dimensional reduction.
  - **Geng et al, 2007**: Generated aging patterns for each person in a dataset, where face images show each subject at different ages.
  - **Fu and Huang, 2008**: Represent aging patterns using manifold learning.
  - **Wang et al, 2009**: Applied Error-Correcting Output Codes (ECOC) to the fused Gabor and Local Binary Patterns (LBP) features.

Age classification using Radon transform and scaling SVM

Months 4 years 7 years 14 years



- Original images
- Adaptive Difference of Gaussian (DoG)
- Radon Transform (RT): x – intensity, y – bins
- Feature selection/Support Vector Machine classification

Previous techniques

- Gender classification
  - **Moghaddam and Yang, 2002**: Applied Support Vector Machine (SVM) with Radian Basis Function to thumbnail facial images.

Previous techniques

- Gender classification
  - **Moghaddam and Yang, 2002**: Applied Support Vector Machine (SVM) with Radian Basis Function to thumbnail facial images.
  - **BenAbdelkader and Griffin, 2005**: Combined local region matching and holistic features with Linear Discriminant Analysis (LDA) and SVM.

Previous techniques

- Gender classification
  - **Moghaddam and Yang, 2002**: Applied Support Vector Machine (SVM) with Radian Basis Function to thumbnail facial images.
  - **BenAbdelkader and Griffin, 2005**: Combined local region matching and holistic features with Linear Discriminant Analysis (LDA) and SVM.
  - **Lapedriza et al, 2006**: Compared facial features from internal zone (eyes, nose and mouth) and external zone (hair, chin, and ears).

Previous techniques

- Gender classification
  - **Moghaddam and Yang, 2002**: Applied Support Vector Machine (SVM) with Radian Basis Function to thumbnail facial images.
  - **BenAbdelkader and Griffin, 2005**: Combined local region matching and holistic features with Linear Discriminant Analysis (LDA) and SVM.
  - **Lapedriza et al, 2006**: Compared facial features from internal zone (eyes, nose and mouth) and external zone (hair, chin, and ears).
  - **Gao and Ai, 2009**: Adopted the probabilistic boosting tree with Harr-like features.

Previous techniques

- Gender classification
  - **Moghaddam and Yang, 2002**: Applied Support Vector Machine (SVM) with Radian Basis Function to thumbnail facial images.
  - **BenAbdelkader and Griffin, 2005**: Combined local region matching and holistic features with Linear Discriminant Analysis (LDA) and SVM.
  - **Lapedriza et al, 2006**: Compared facial features from internal zone (eyes, nose and mouth) and external zone (hair, chin, and ears).
  - **Gao and Ai, 2009**: Adopted the probabilistic boosting tree with Harr-like features.
  - **Shan, 2012**: LBP was employed to describe faces and AdaBoost was used to select the discriminative LBP features.

Ethnicity- and gender-based subject retrieval using 3-D face-recognition techniques

- 3-D images
  - Distance between two geometries: an L_1 measure on the Haar wavelets and the complex wavelet structural similarity measure on the pyramid coefficients.
- Fusion techniques:
  - K-Nearest-Neighbors.
  - Kernelised k-Nearest-Neighbors.
  - Learning from the Face-Similarity Space.
  - Learning from Algorithm-Specific Features.

Toderici et al, 2010

(a) -743.12 (b) -580.63 (c) -522.05 (d) -451.01
(e) 28.413 (f) 66.656 (g) 101.54 (h) 137.46
(i) 316.63 (j) 363.72 (k) 396.51 (l) 453.62

Photographs of subjects sampled along the dimension most discriminative of race in the data.

Previous techniques

- Ethnicity classification
  - **Gutta et al, 2000**: Applied the mixture of experts using radial basis functions networks with inductive decision trees and SVM.

Previous techniques

- Ethnicity classification
  - **Gutta et al, 2000**: Applied the mixture of experts using radial basis functions networks with inductive decision trees and SVM.
  - **Lu and Jain, 2004**: An ensemble framework that integrated the Linear Discriminant Analysis (LDA) was deployed for classifying the face images at different scales.

Previous techniques

- Ethnicity classification
  - **Gutta et al, 2000**: Applied the mixture of experts using radial basis functions networks with inductive decision trees and SVM.
  - **Lu and Jain, 2004**: An ensemble framework that integrated the Linear Discriminant Analysis (LDA) was deployed for classifying the face images at different scales.
  - **Zhang et al, 2010**: A multi-linear principal component analysis (MPCA) was used to extract features.

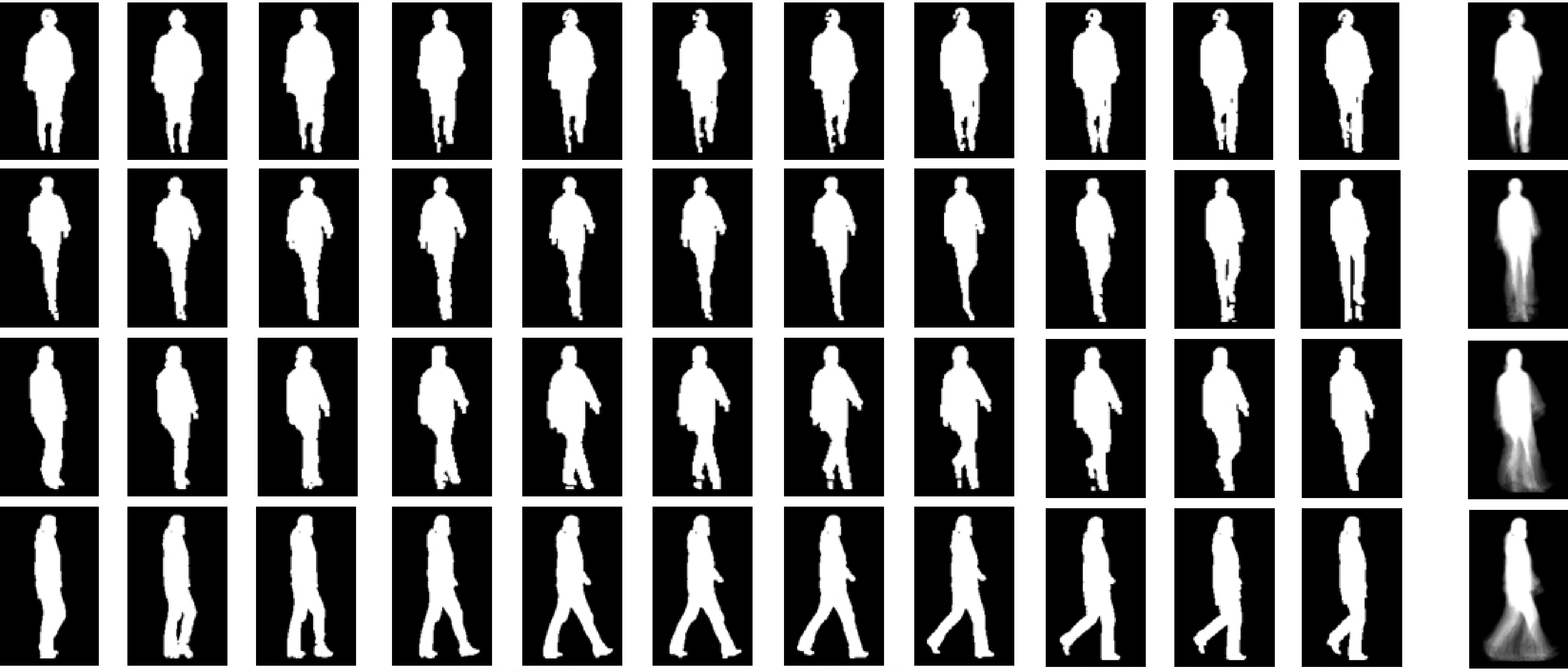
Previous techniques

- Ethnicity classification
  - **Gutta et al, 2000**: Applied the mixture of experts using radial basis functions networks with inductive decision trees and SVM.
  - **Lu and Jain, 2004**: An ensemble framework that integrated the Linear Discriminant Analysis (LDA) was deployed for classifying the face images at different scales.
  - **Zhang et al, 2010**: A multi-linear principal component analysis (MPCA) was used to extract features.
  - **Hosoi et al, 2004**: Gabor wavelet transform and retina sampling were combined to extract features, followed by SVM.

Previous techniques

- Ethnicity classification
  - **Gutta et al, 2000:** Applied the mixture of experts using radial basis functions networks with inductive decision trees and SVM.
  - **Lu and Jain, 2004**: An ensemble framework that integrated the Linear Discriminant Analysis (LDA) was deployed for classifying the face images at different scales.
  - **Zhang et al, 2010**: A multi-linear principal component analysis (MPCA) was used to extract features.
  - **Hosoi et al, 2004**: Gabor wavelet transform and retina sampling were combined to extract features, followed by SVM.
  - **Zhang et al, 2012**: Fused LBP features of face and gait using canonical correlation analysis (CCA).

Ethnicity classification based on gait using multi-view fusion



Feature extraction: Examples of normalized and centered silhouette frames from different views for one walk. From the top row to bottom row, the view angles are 0, 30, 60, 90, 120, 150 and 180 degrees respectively. The rightmost image in each row is the corresponding gait energy image (GEI).

Coffee break!

- Introduction
- Human detection and tracking
- Human profiling
- **Activity recognition**
- Trajectory clustering
- Summary

What is activity recognition?

- “It aims to recognise the actions and goals of one or more agents from a series of observations on the agents’ action and the environment conditions” – Wikipedia

Levels of human activity

- Gesture – atomic movements
- Actions – single actor
- Interactions – human-human and human to computer
- Group activities – physical/mental

Research challenges

- Environmental changes:
  - Changing backgrounds.
  - Changing view points.
- Human movement variations:
  - Same activity but different styles.
- Unconstrained activities.
- Needs of robust learning algorithms.

Categorisation

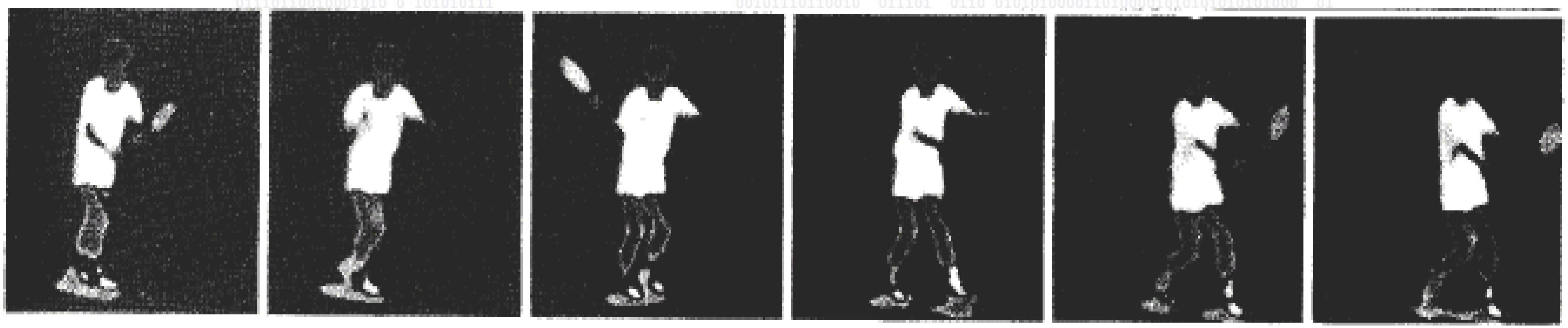
- Sequential approaches
  - Data based.
  - State model based.
- Hierarchical approaches
  - Statistical.
  - Description based.

Sequential approaches

- Data based, for example,
  - Darrel and Pentland, 1993.
  - Yacoob and Black, 1998.
  - Ali and Aggarwal, 2001.
  - Lublinerman et al, 2006.
  - Jiang et al, 2006.
- State model based, for example,
  - Yamato et al, 1992.
  - Starner and Pentland, 1995.
  - Bregler, 1997.
  - Bobick and Wilson, 1997.
  - Park and Aggarwal, 2004.
  - Natarajan and Nevacia, 2007
  - Gupta and Davis, 2007.

Action recognition using HMMs

- Concepts:
  - Each HMM is related to a specific sequence of features.
  - Match the observed features with the model.
  - An action refers to a set of sequences of features.

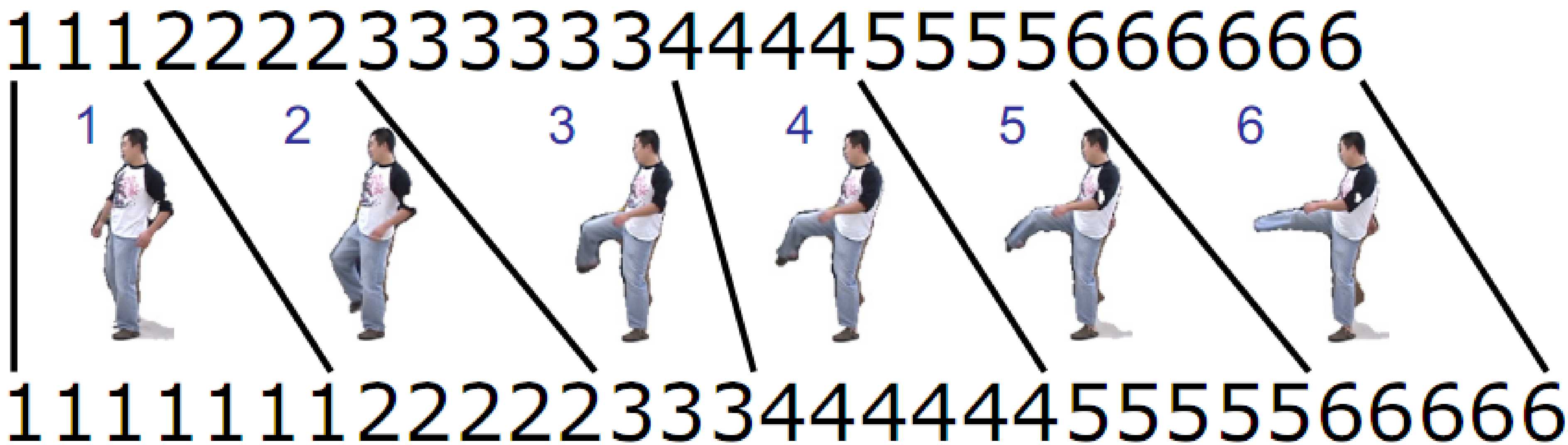


Symbol sequence	60 61 61 62 62 62 63 63 64 64 65 66 66 66 67 68 68 69 69 70 70 70 71 71

Yamato et al, 1992

Dynamic time warping

- Applied dynamic programming to match two strings/sequences.
- Each image frame generates a symbol or a feature vector.

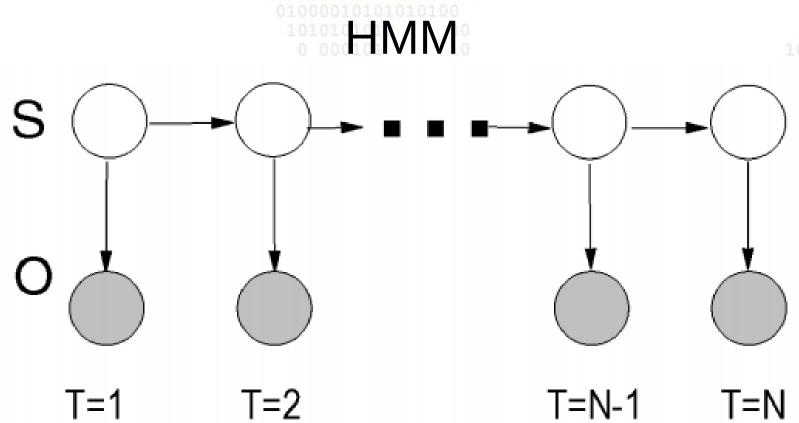


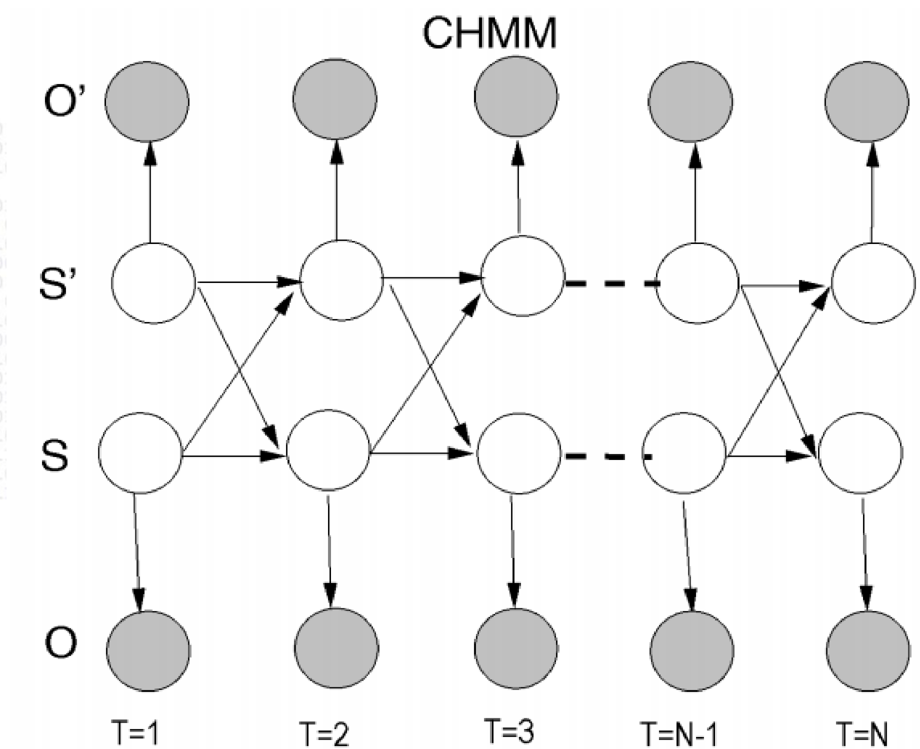
Gavrila and Davies, 1995

Coupled HMMs

- Set up two types of states for two different agents.
- Synthetic agents for training HMMs.

What is the difference between these left and right structures?





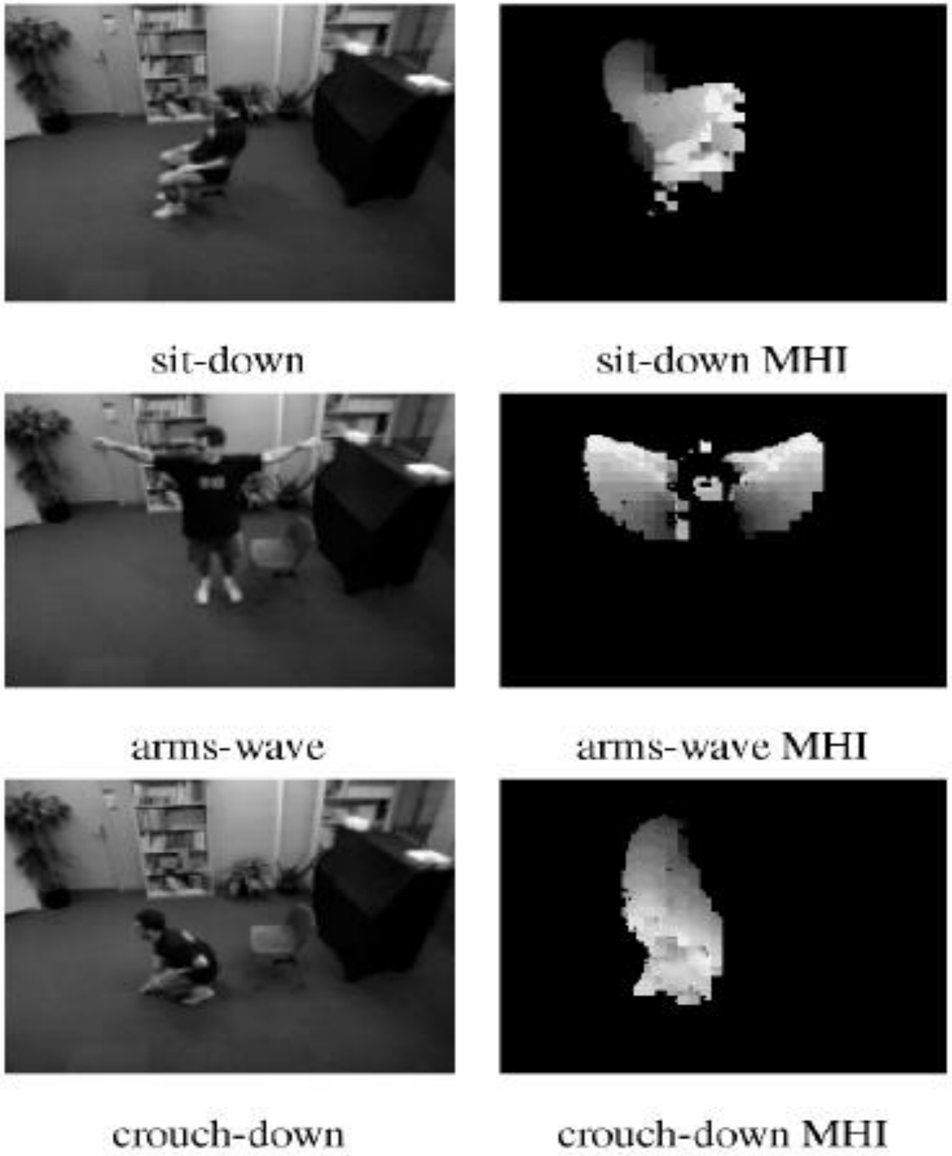
Oliver et al, 2000

What we observed from sequential approaches?

- Common approaches
  - Markovian process.
  - Motion features are required of each frame.
- Advantages
  - Straightforward.
  - Quick process.
- Weaknesses
  - Need good features from valid observations.
  - Large training data.

Motion history images

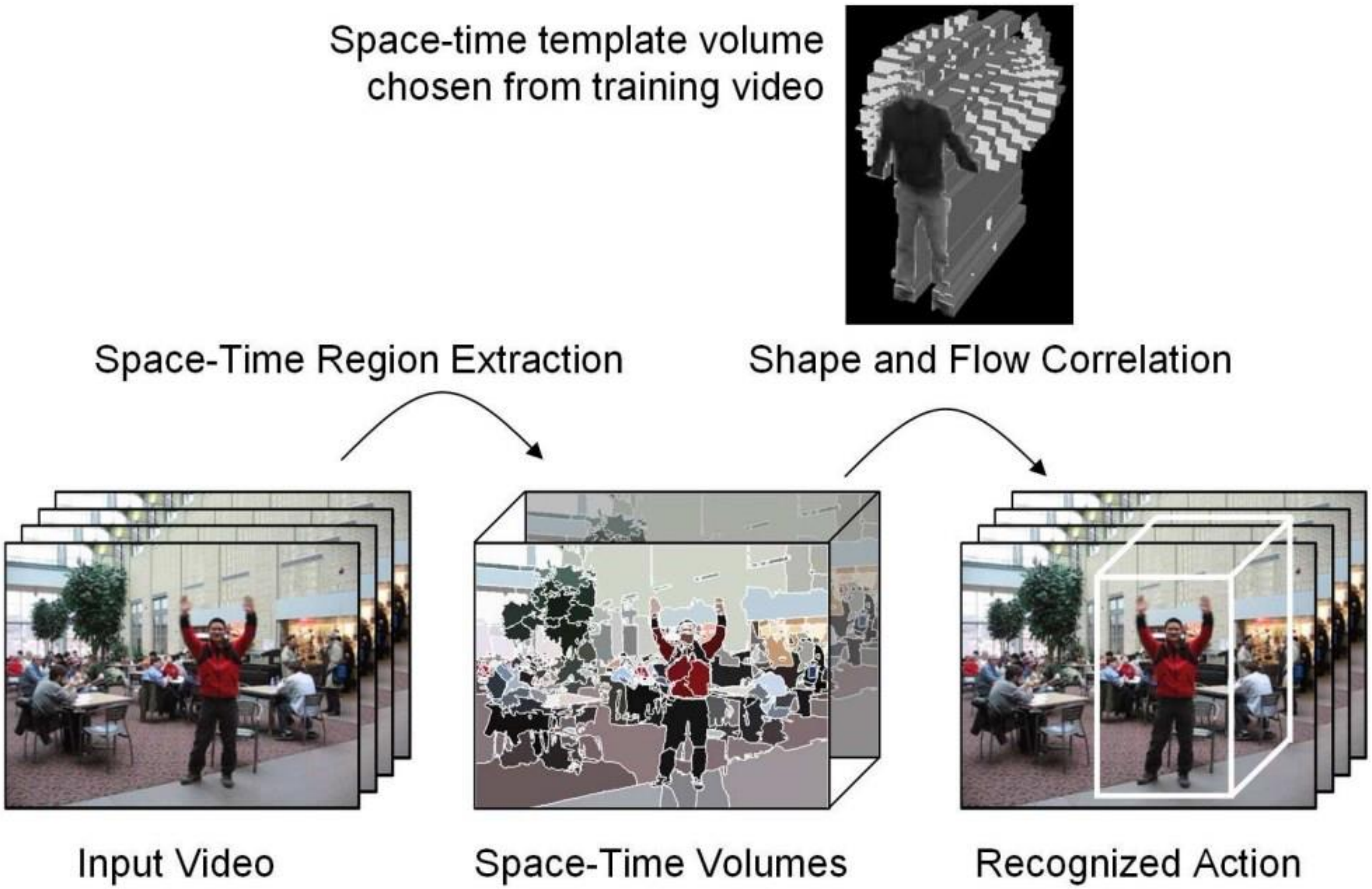
- Motion history images (MHIs).
- Weighted projection of a x-y-t foreground volume.
- Template matching.



Bobick and Davis, 2001

3-D Volume Matching

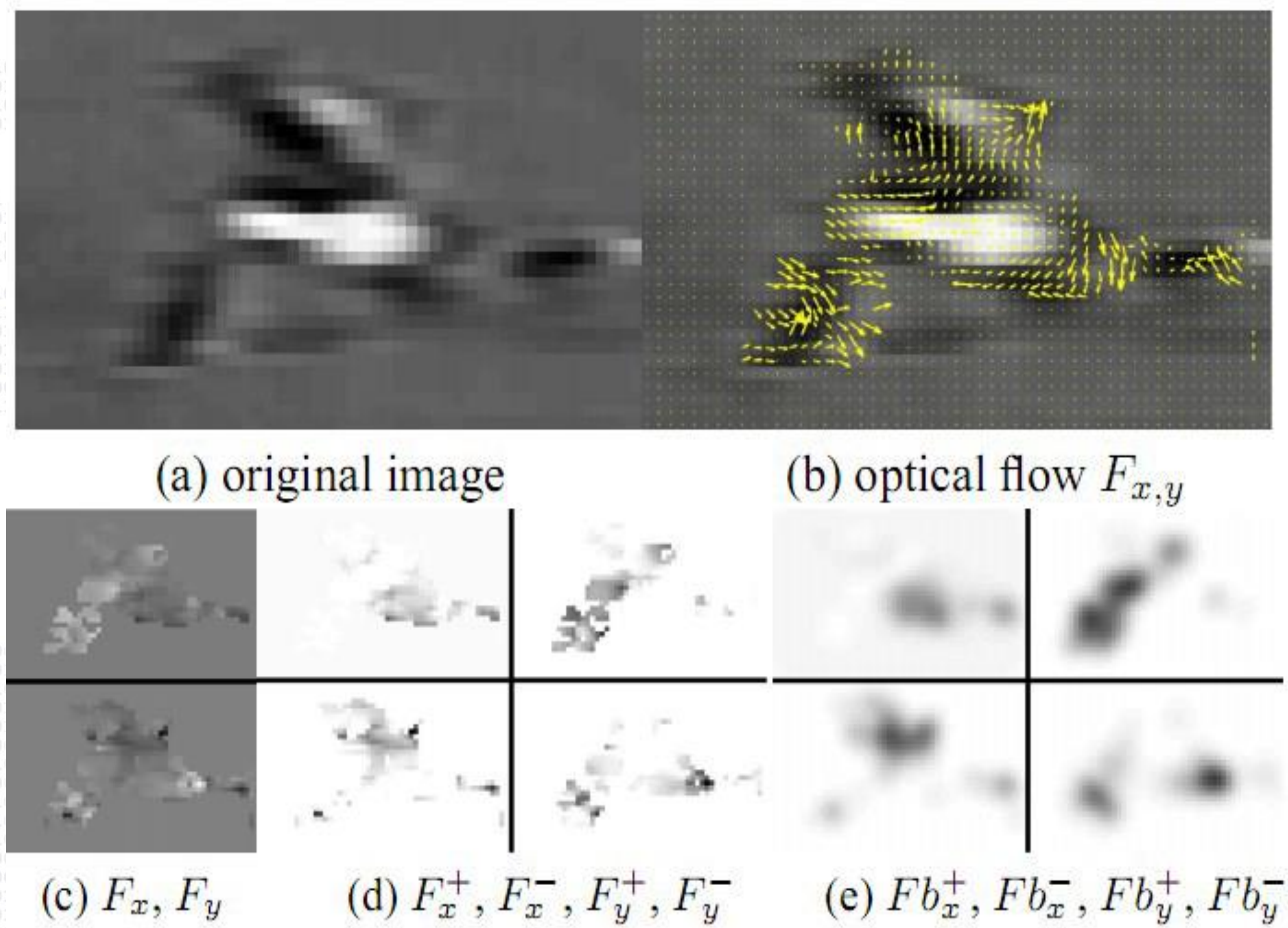
- Perform volume matching for segments.
- Combine scores of segment matching.



Ke, et al, 2007

Global features from volumes

- Concatenate optical flow features from x-y-t volumes.
- Good performance in low resolution videos.



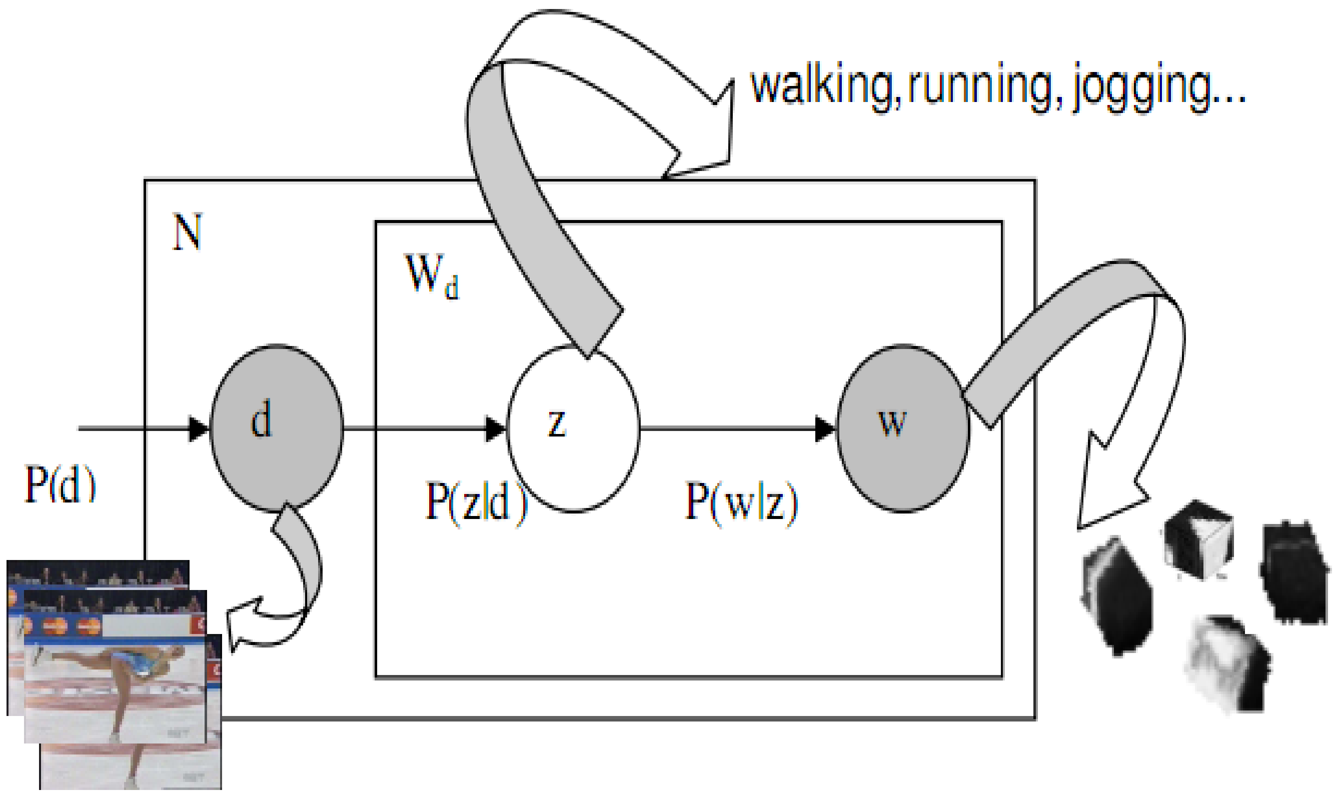
(a) original image (b) optical flow $F_{x,y}$

(c) F_x, F_y (d) $F_x^+, F_x^-, F_y^+, F_y^-$ (e) $Fb_x^+, Fb_x^-, Fb_y^+, Fb_y^-$

Efros et al, 2003

pLSA models for actions

- Probabilistic Latent Semantic Analysis (pLSA).
- Estimate the probability of features from an action video.



Niebles et al, 2006

What did we learn from these examples?

- Use of local spatio-temporal features
  – Bag of words, cuboid, grouping, etc.
- Incorporating standard classifiers.
- Any extension?
  – Structural information.
  – Hybrid features.

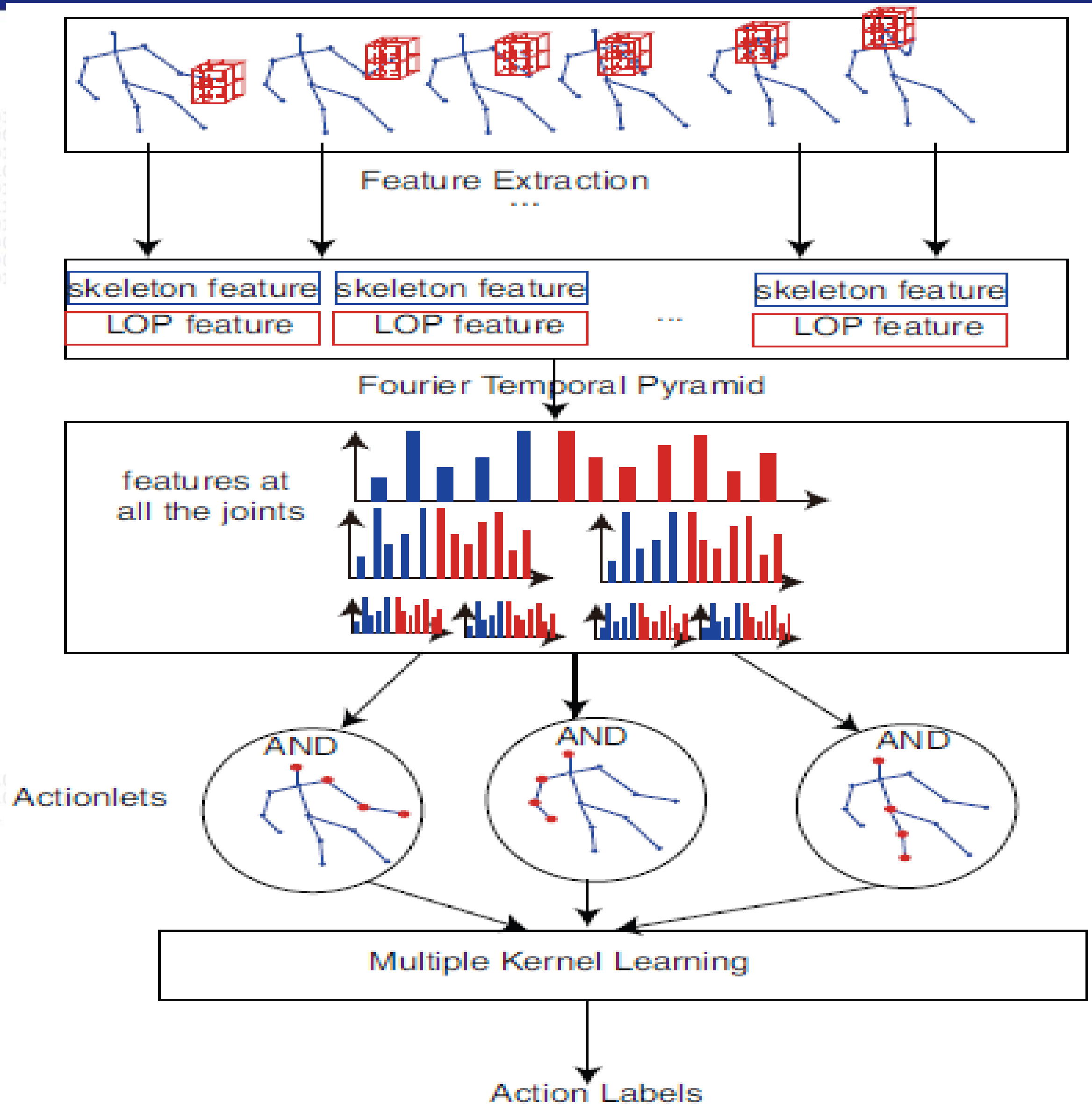
Structural information

- Previously introduced methods:
  - There is no structure in local features.
- Exemplar approaches considering structures:
  - pLSA-ISM: takes into account the locations of features (Wong et al, 2007).
  - Feature correlation: pair-wise proximity (Savarese et al, 2008).

Mining actionlet ensemble with depth cameras

- Actionlet: a conjunction of the features for a subset of the joints.
- A linear combination of actionlet was obtained with learnt weights.

Wang et al, 2012



Hybrid features

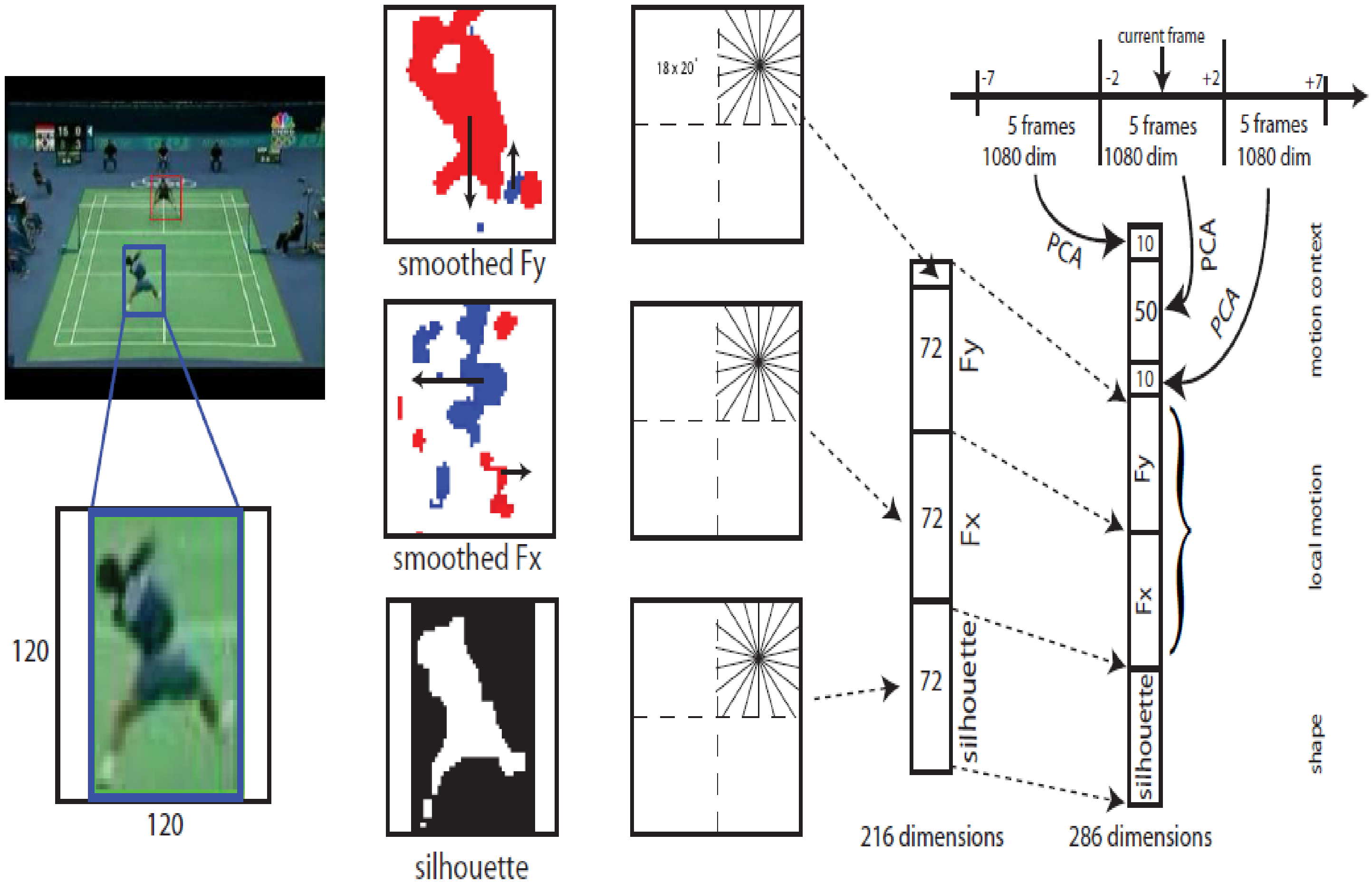
- Challenges: scenes with camera movements.

- Features: gradients + optical flows.

Laptev et al, 2009

More examples

- To reject unseen activities and learn with few examples.
- Features: silhouettes + optical flows.

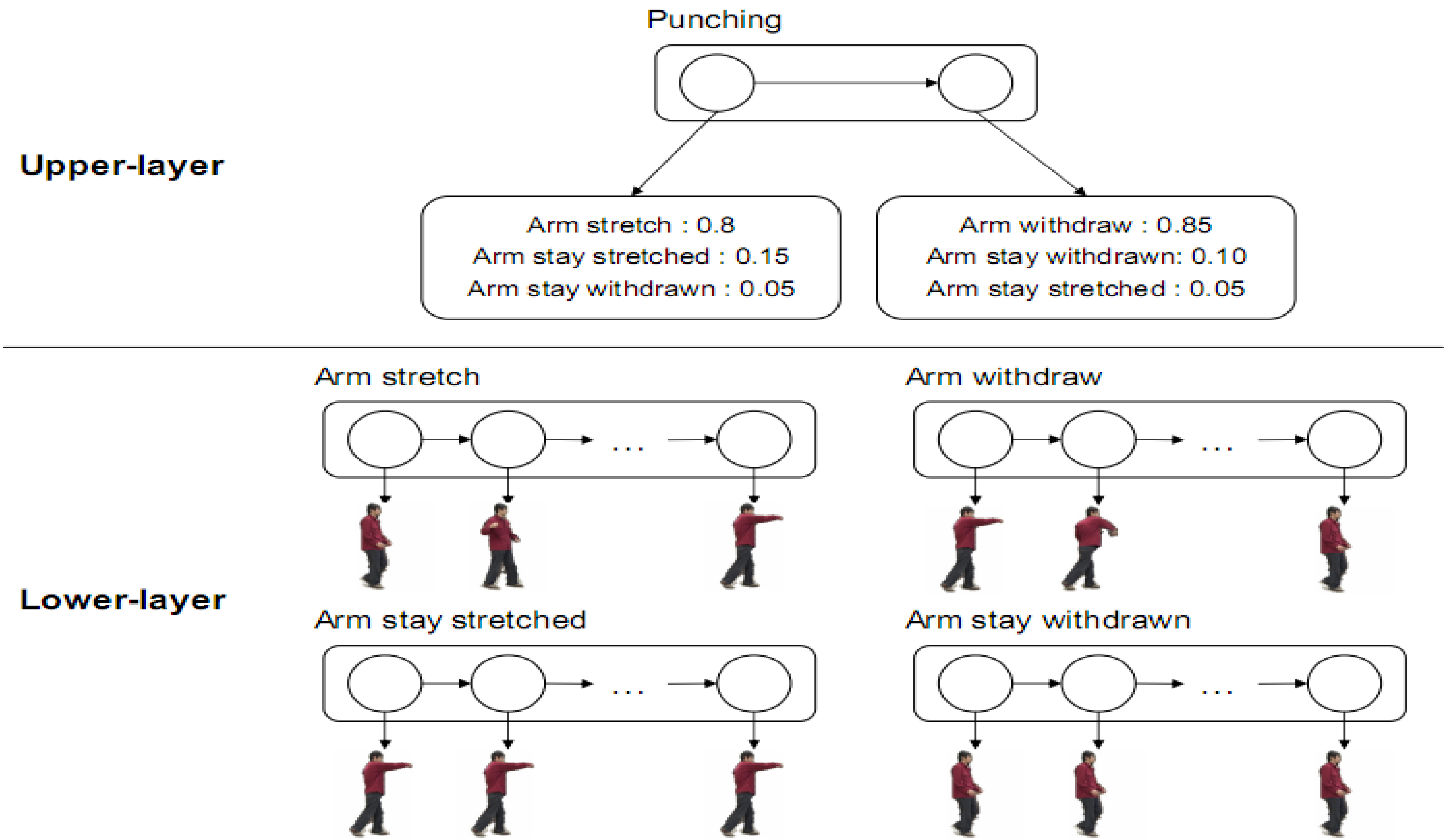


Tran and Sorokin, 2008

Hierarchical approaches

- Why do this research?
  - Sequential approaches cannot effectively handle complicated activities.
- How is it working?

Hierarchy – an illustration



Aggarwal et al, 2011

Category

- Statistical
- Syntactic
- Descriptive

Syntactic approaches

- Use of context free grammar.
- A grammar is described: G = <S, T, N, P>.

Generic language	Natural language
Start symbol (S)	Sentences
Terminal symbols (T)	Words
Non-terminal symbols (N)	Speech
Production rules (P)	Syntax rules

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

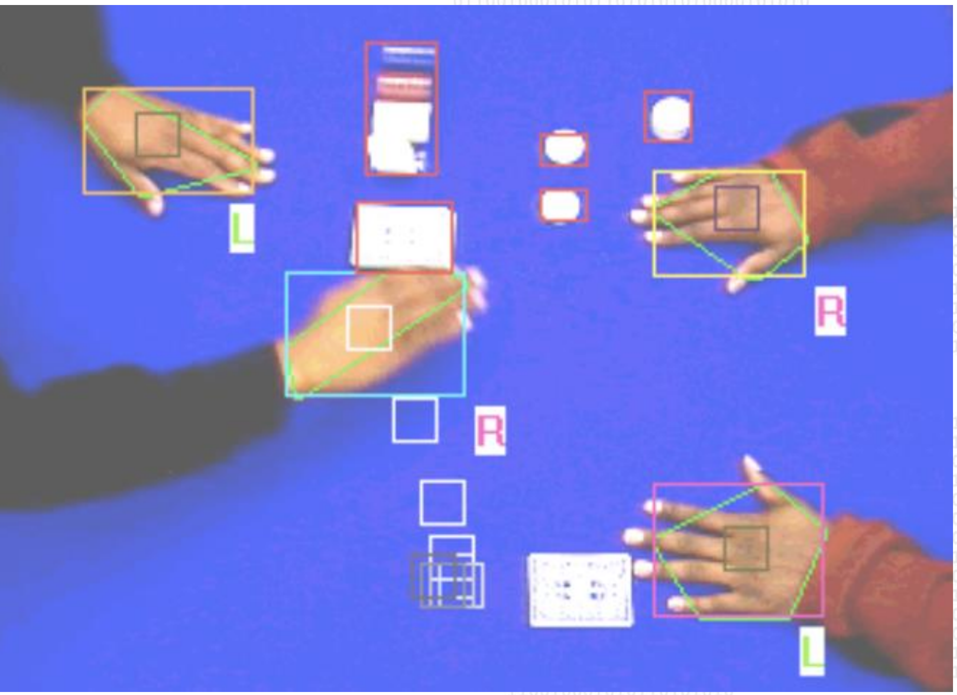
Action recognition using probabilistic parsing

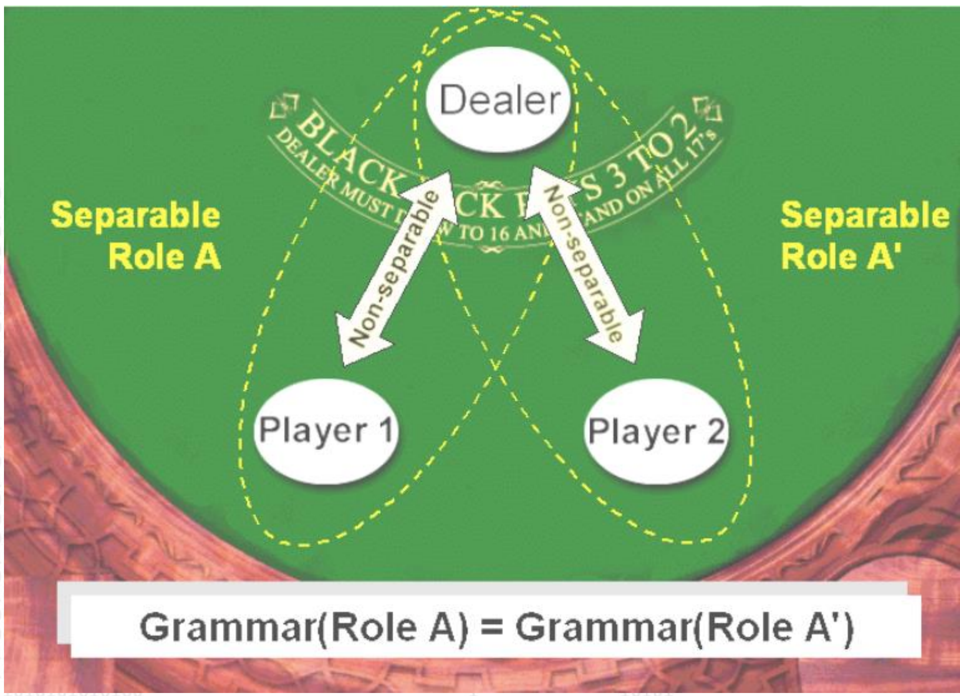
G_{square} :

SQUARE	$\rightarrow$	RH	[0.5]
	$\mid$	LH	[0.5]
RH	$\rightarrow$	TOP UD BOT DU	[1.0]
LH	$\rightarrow$	BOT DU TOP UD	[1.0]
TOP	$\rightarrow$	LR	[0.5]
	$\mid$	RL	[0.5]
BOT	$\rightarrow$	RL	[0.5]
	$\mid$	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]

Bobick and Ivanov, 1998

Recognising multitasked activities from video using stochastic context free grammar

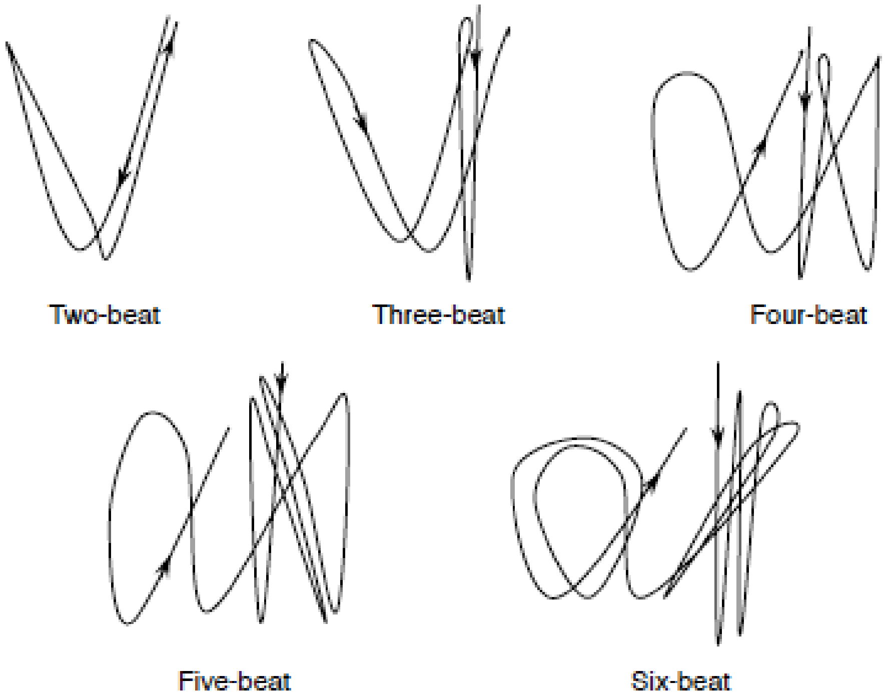




Moore and Essa, 2001

Heuristic grammatical induction

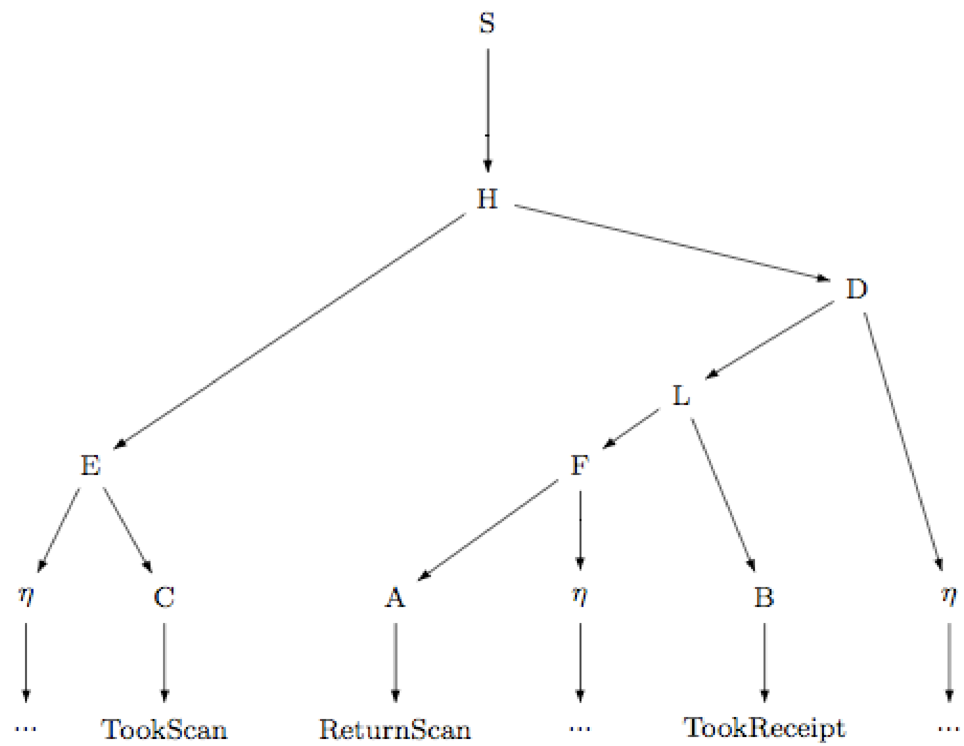
- Lexicon learning
  - Learning by HMMs.
  - Clustering by HMMs.
- Convert a video to a string.
- Learn grammar(s).



Wang et al, 2001

Learning to handle noise

Example: Learn the process of transactions.



Kitani et al, 2008

Review of syntactic approaches

- Robust against errors.
- Accurate detail descriptions.
- But, need quite a lot training sets.
- Computationally complex.

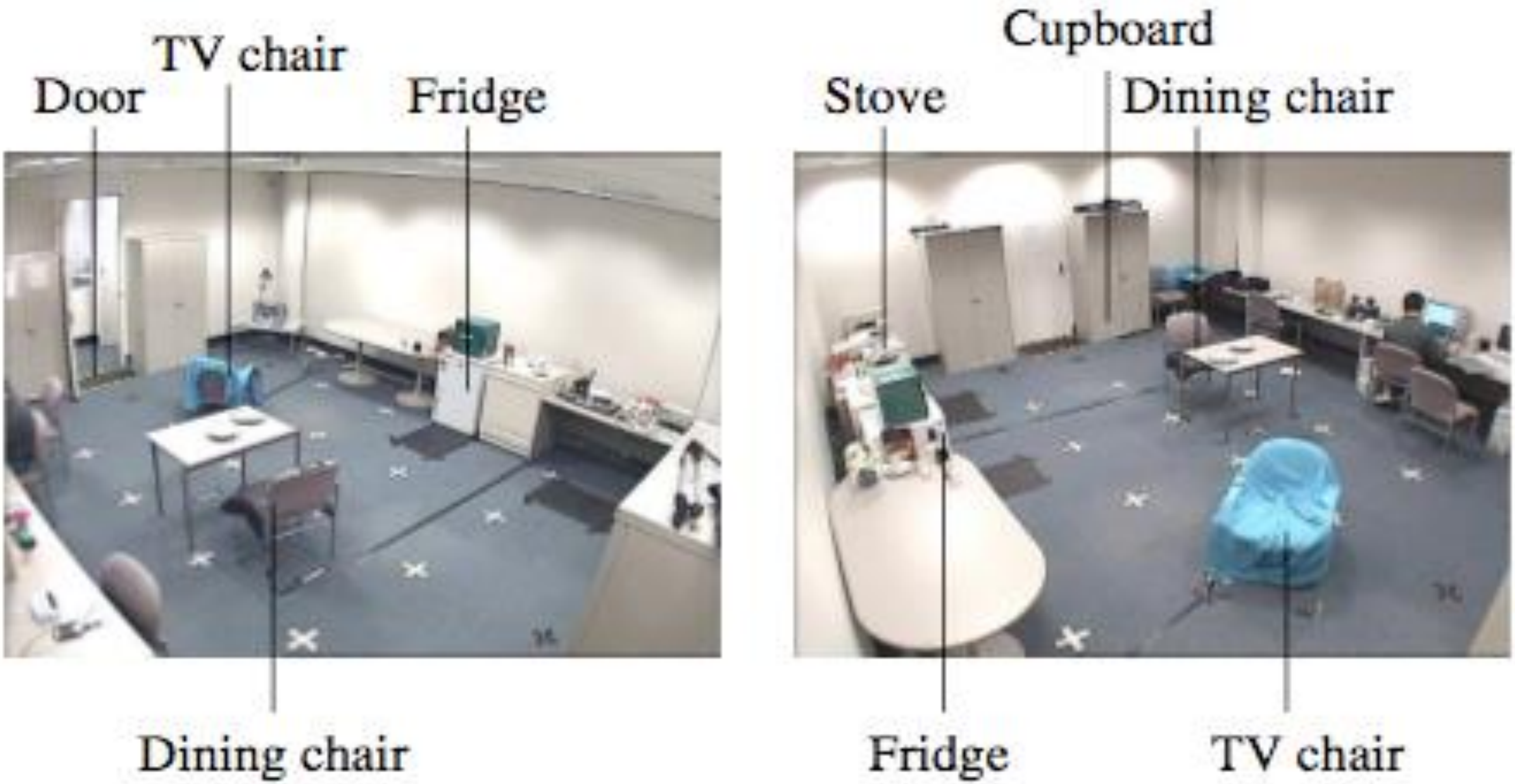
Statistical approaches

- When we apply these approaches:
  - Few features extracted from videos are “noisy”.
  - Activity structure is not complicated.
  - Rich and clear video dynamics.

Characteristics

- Strong Markovian assumption.
- Known priors of dynamics.
- We can reason certain ambiguity/uncertainty.

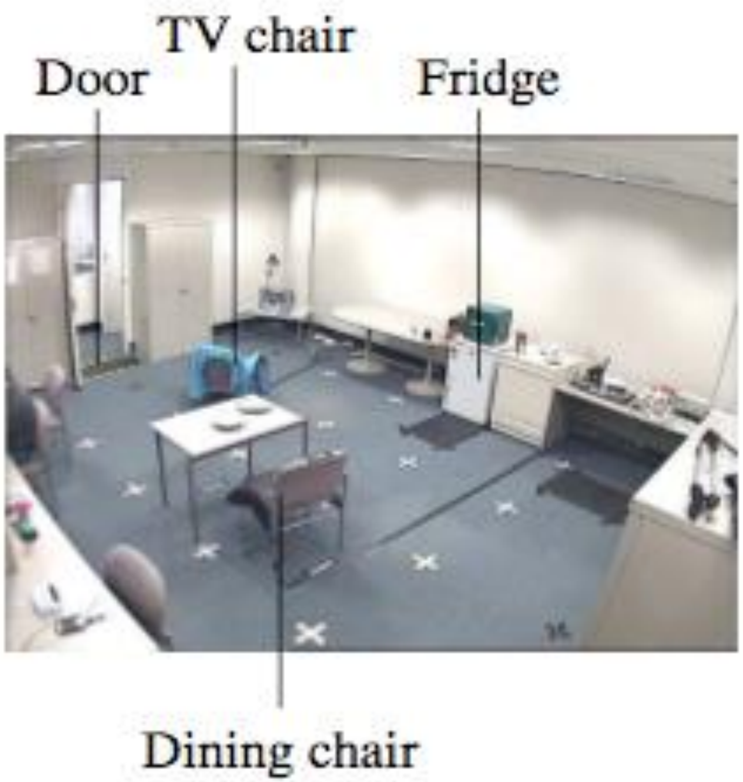
Context free activity grammar

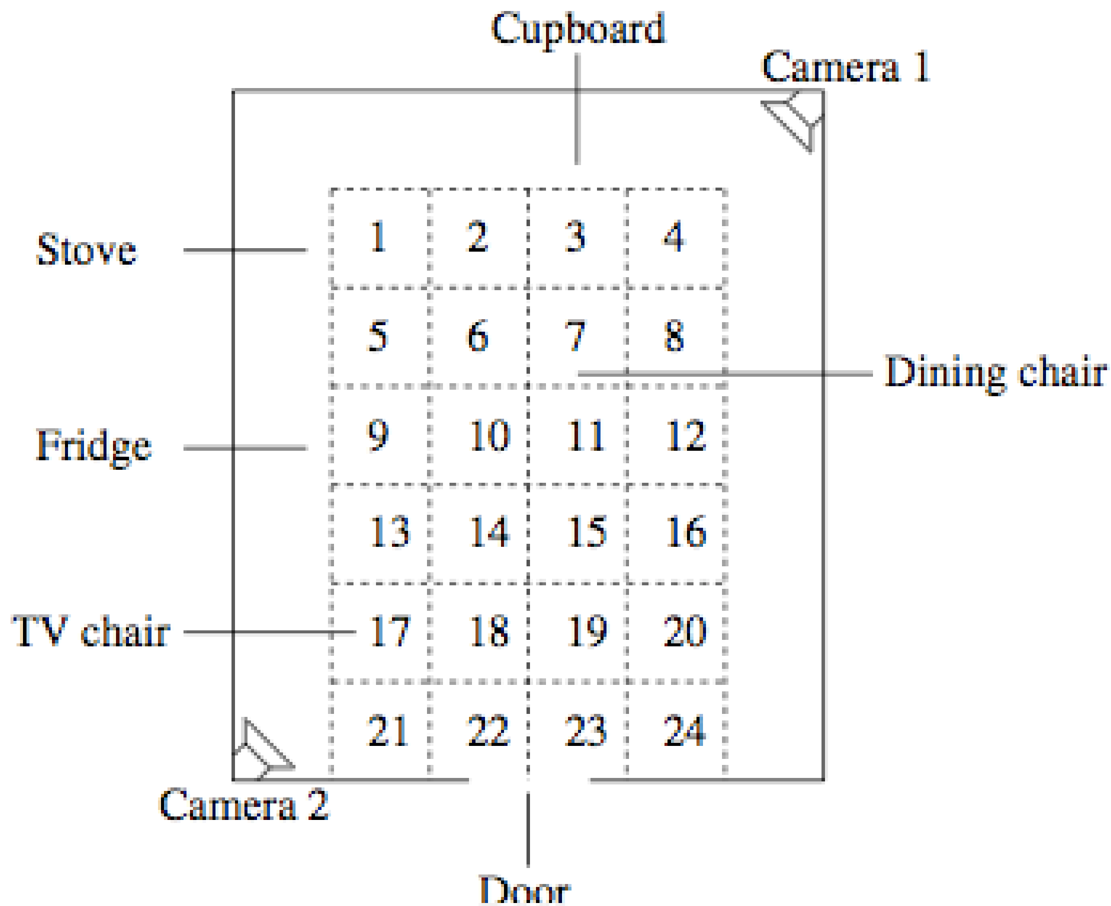


Nguyen et al, 2005

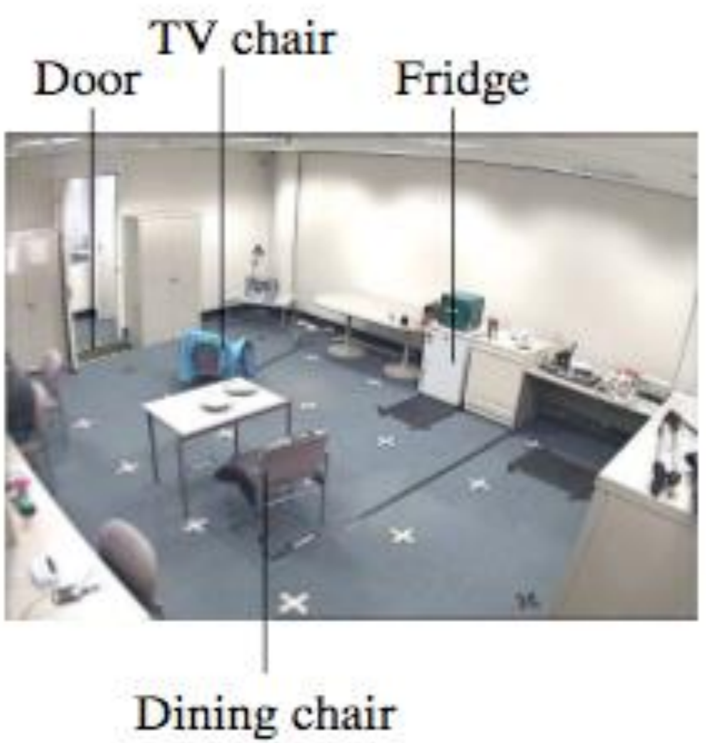
CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

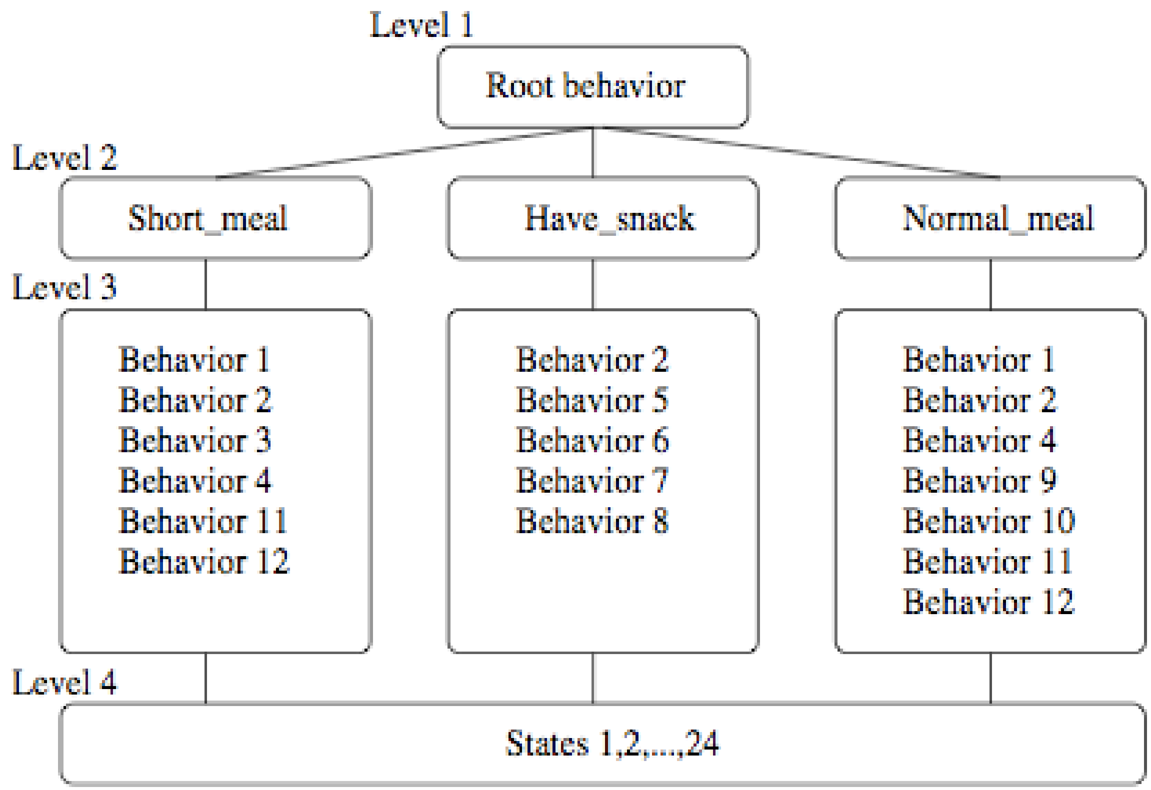
Context free activity grammar



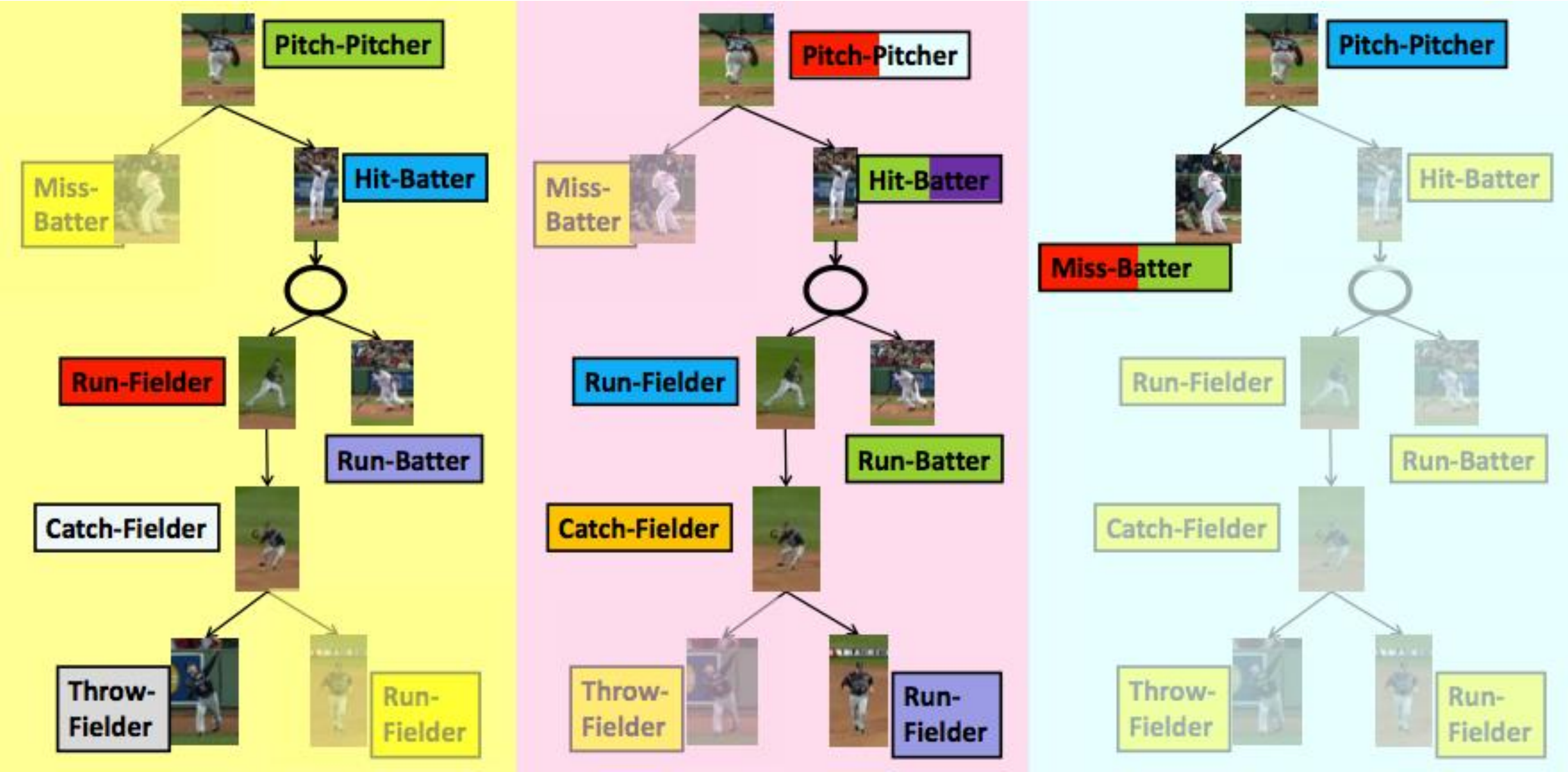


Context free activity grammar



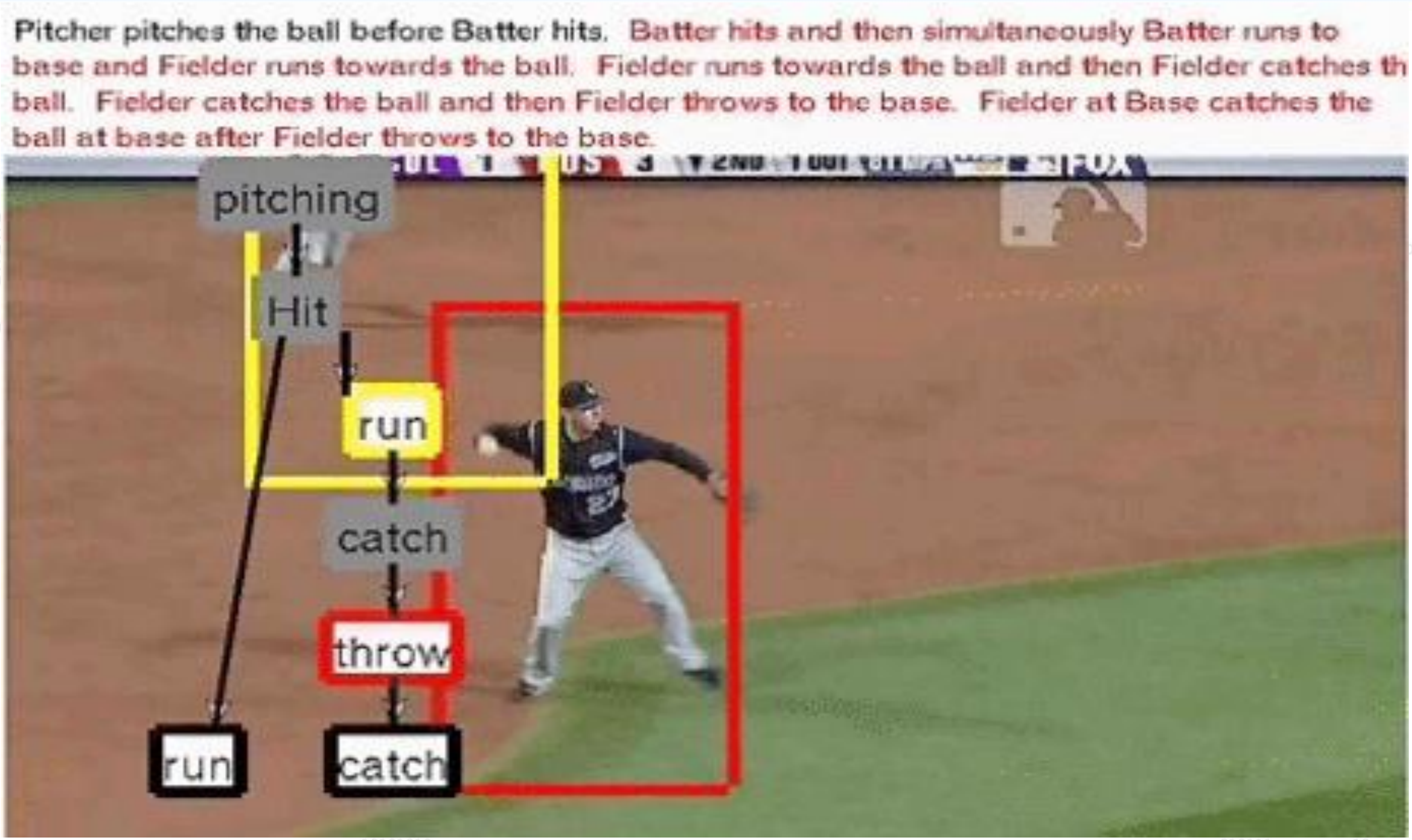


Learning storylines



Gupta et al, 2009

Scripts used to describe



Gupta et al, 2009

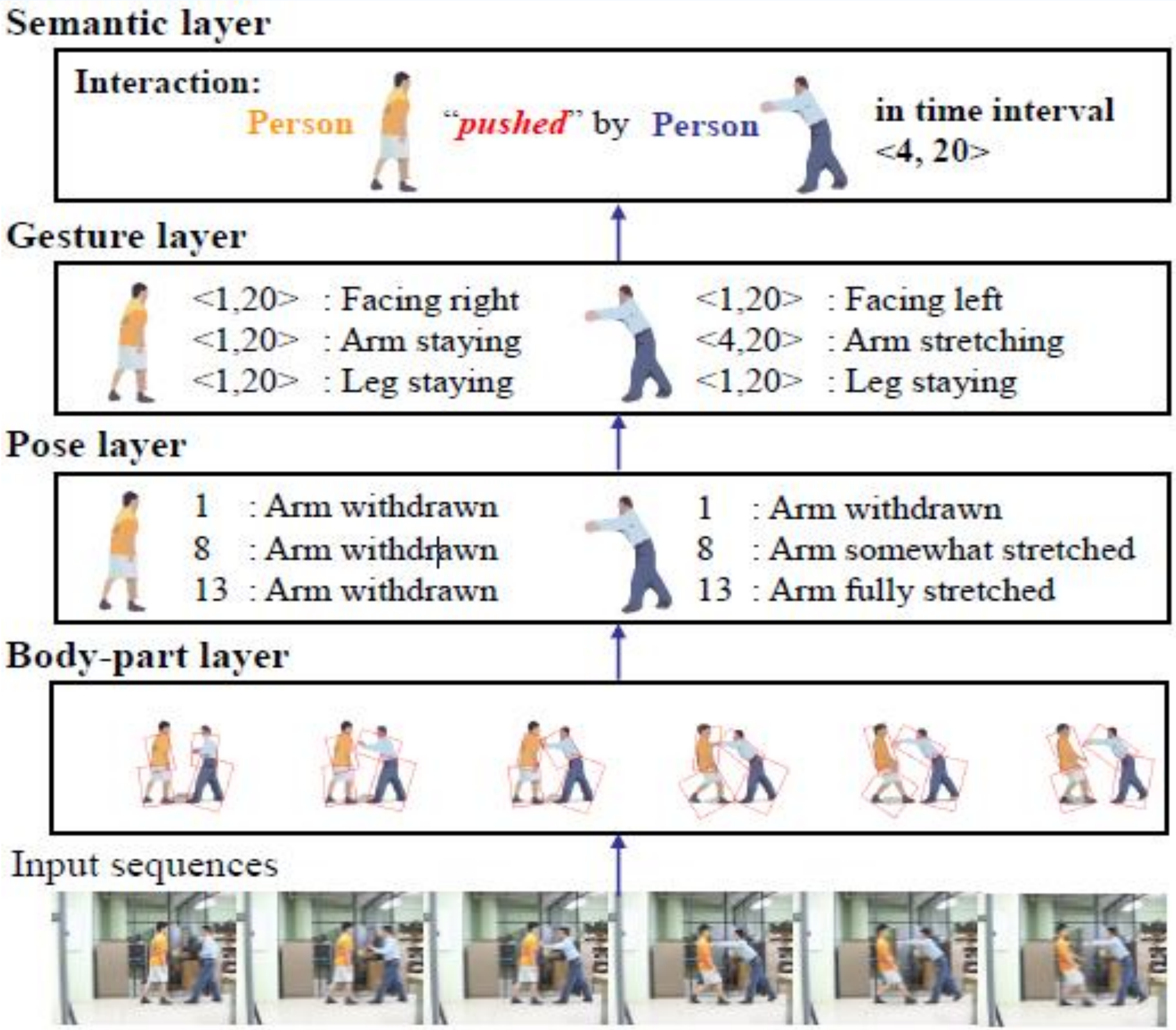
Unsuitable for statistical approaches

- Activities: too many structures to build.
- Activities: too complicated temporal correlation.

Descriptive approaches

- Using semantic matching for recognising activities
  - Football kick = “a person touches a football using her/his foot”.
  - Recognition is achieved by matching the components to the definition.

Ryoo and Aggarwal, 2006



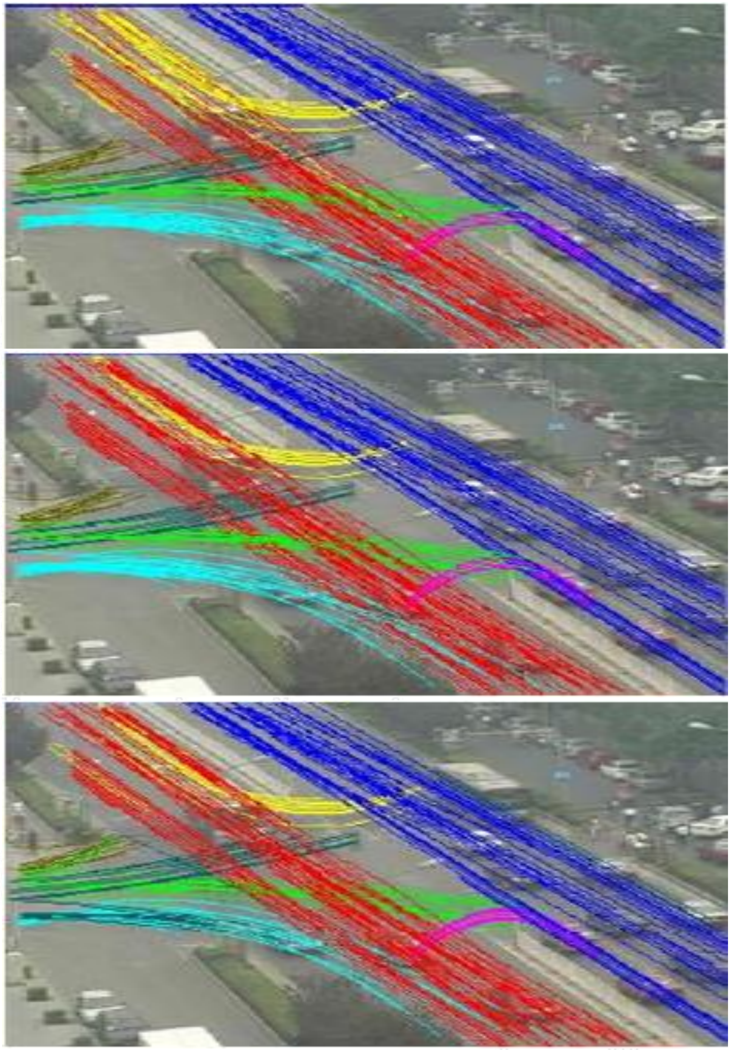
- Interaction
- Gesture
- Pose
- Body part feature

- Introduction
- Human detection and tracking
- Human profiling
- Activity recognition
- **Trajectory clustering**
- Summary

Trajectory

- Trajectories describe the movement behaviours of objects.

- Challenges in clustering:
  - Fast changes in routes.
  - Intersection of different routes.
  - Similar route but different direction or speeds.



Trajectory analysis

- Trajectory analysis is part of behaviour understanding from videos.
- It aims to extract relevant visual information with proper representation and interpretation for behaviour learning and recognition.
- Trajectory clustering provides a tool to implement the learning and analysis of human activities.

Clustering procedure

- To define a distance (or similarity) measure.
- To propose a cluster update methodology.
- To perform cluster validation.

Distance – examples

- Euclidean distance (two routes)

$$d_E(F_i, F_j) = \sqrt{(F_i - F_j)^T(F_i - F_j)}$$

- Mahalanobis

$$D_M(x) = \sqrt{(x - \mu)^T S^{-1}(x - \mu)}$$

- Hausdorff

$$D_H(F_i, F_j) = \max\left(D_h(F_i, F_j), D_h(F_j, F_i)\right)$$

- Bhattacharyya

$$D_B(p, q) = -\ln\left(BC(p, q)\right)$$

Clustering methods

- Iterative optimisation
- On-line adaptive
- Hierarchical
- Neural networks
- Co-occurrence decomposition
- …

Iterative optimisation

- Advantages:
  - – Simple.
  - – Tractable.
  - – Closed form solutions.
- Weaknesses:
  - – Need to specify cluster number.
- Examples:
  - – K-means.
  - – Fuzzy C-means and variants.

On-line adaption

- Advantages:
  - No need to specify cluster number.
  - Does not require training datasets.
- Weaknesses:
  - Hard to obtain a good cluster initialisation.
- Examples:
  - Similarity threshold.
  - Iterative K-means.

Hierarchical

- Advantages:
  - Allowing an intelligent choice of cluster number.
  - Well suited for graphic models (max-flow/min-cut, dominant set).
- Weaknesses
  - Usually do not re-evaluate decisions.
- Examples:
  - Agglomerative.
  - Divisive.

Neural networks

- Advantages
  - Describing linear and non-linear relationship.
  - Trained to update unseen scenes.
- Weaknesses
  - A large training set.
  - Complex parameterisation.
- Examples:
  - SOM (self-organising map).
  - Fuzzy SOM.

Co-occurrence

- Trajectories: a bag of words; use of a co-occurrence matrix.
- Advantages
  - Independent of trajectory length.
- Weaknesses
  - Limited vocabulary size.
  - Unpreserved time order.
- Examples:
  - Document keyword.

Comparisons of different methods

Table 1. Trajectory Distance Measures

Technique	Publication
HU	Hu 2007 [7]
PCA	Bashir 2007 [8]
DTW	Keogh 2000 [9]
LCSS	Buzan 2004 [10]
PF	Piciarelli 2006 [3]
MODH	Atev 2006 [4]

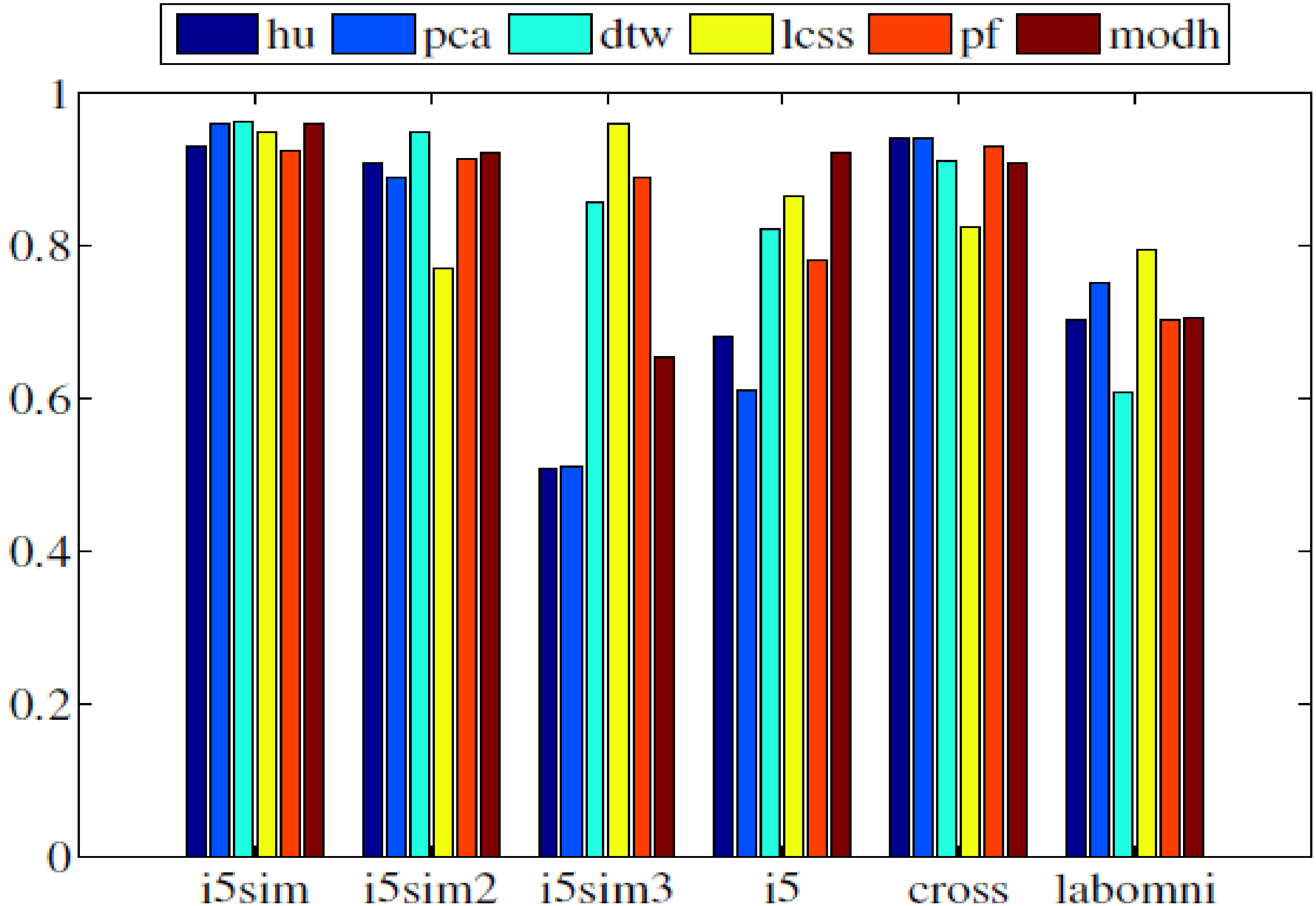
Table 2. Clustering Techniques

Technique	Publication
Direct	Morris 2008 [11]
Divisive (rb,rbr)	Billotti 2005 [12]
Agglomerative	Buzan 2004 [10]
Hybrid (cham)	Karypris 1999 [13]
Graph	Li 2006 [14]
Spectral	Hu 2007 [7]

Morris and Trivedi, 2008

LCSS: longest common subsequence; MODH: modified Hausdorff

Outcomes of comparison

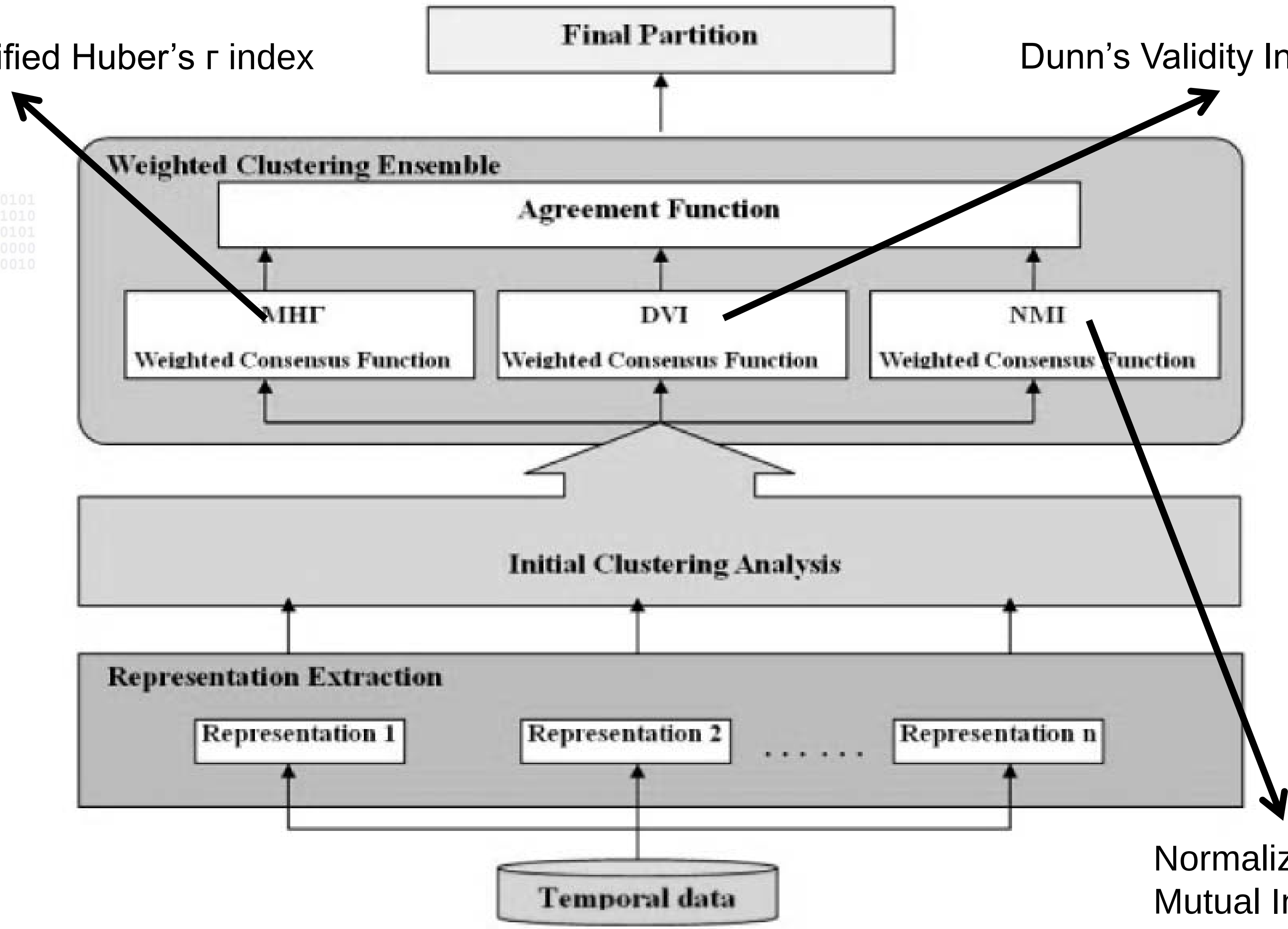


Average performance for the different similarity measures for each dataset.

Temporal data clustering with different representation

Modified Huber’s Γ index

Dunn’s Validity Index

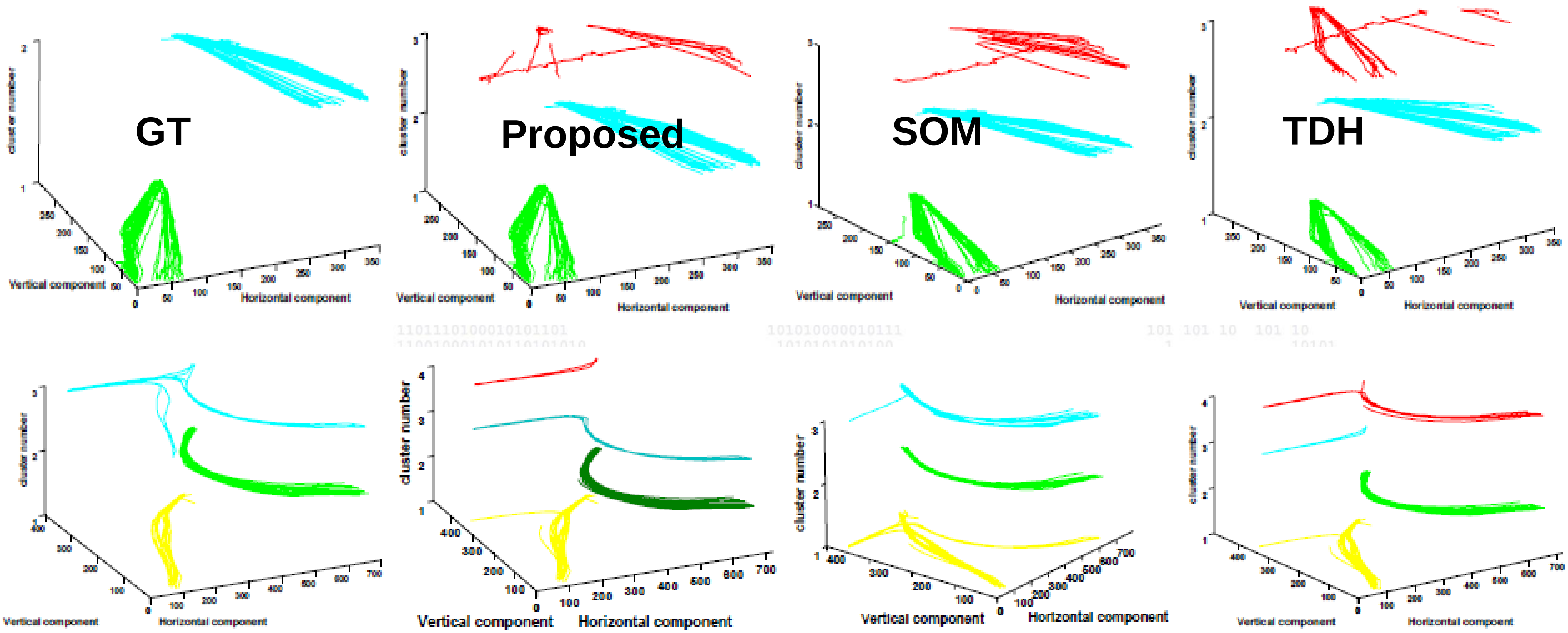


Normalized
Mutual Information

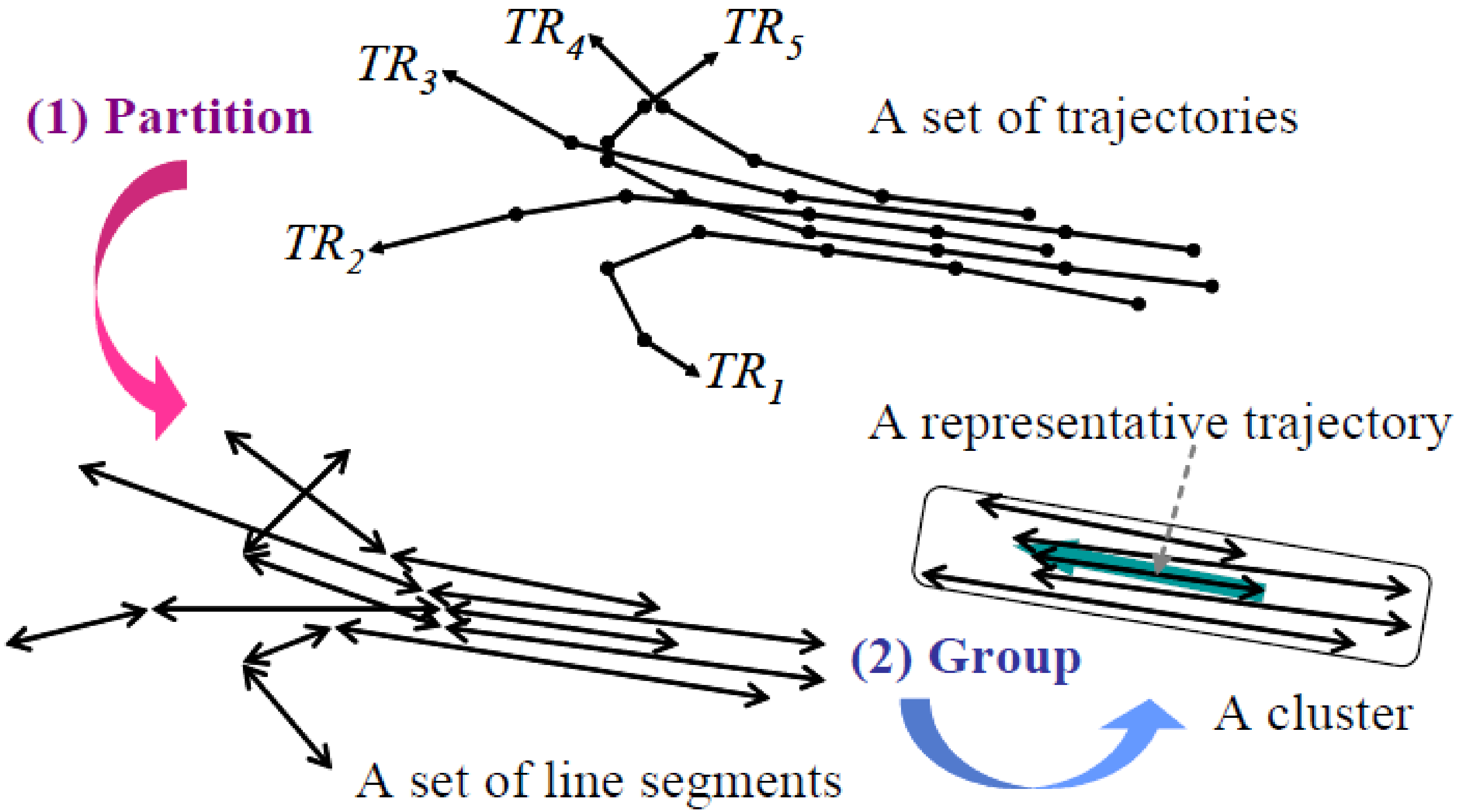
CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Multi-feature object trajectory clustering (Anjum and Cavallaro, 2008)

- Transform trajectories to a set of feature spaces using mean-shift.
- A merging procedure is devised to refine the features.

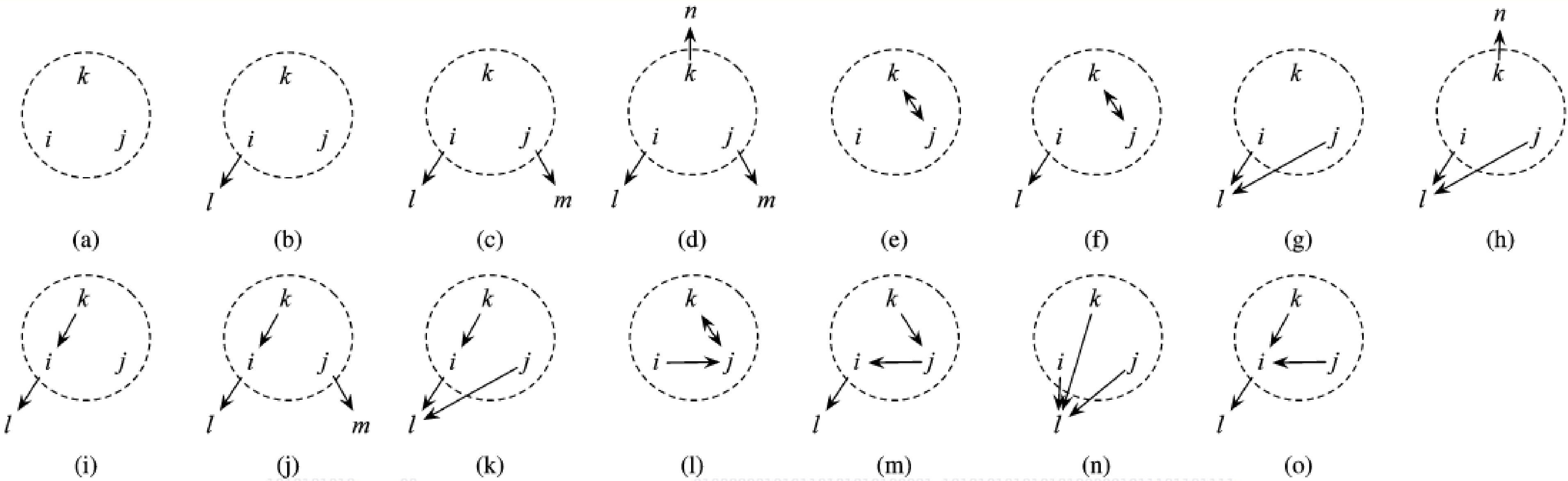


Trajectory Clustering: A Partition-and-Group Framework



Using minimum description length (MDL), Lee et al, 2007

Dynamic hierarchical clustering for trajectories



Fifteen categories of any three trajectory groups according to different nearest neighbours

Jiang et al, 2009

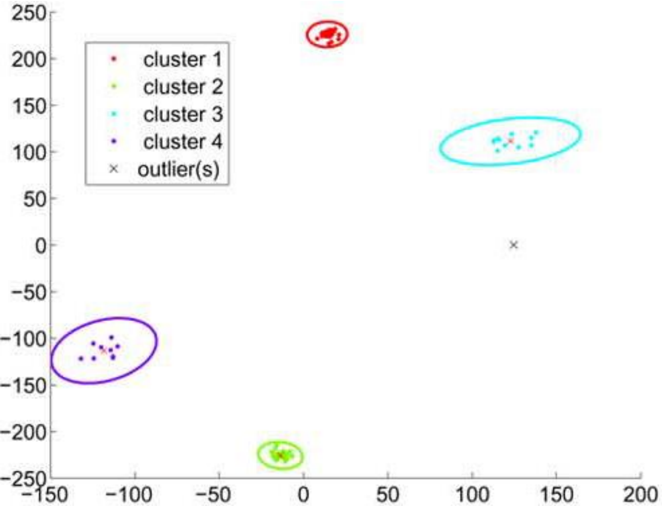
(1) HMMs are applied for events.
(2) Bayesian information criterion (BIC) is used for event clustering.
(3) An EM algorithm is deployed.

Trajectory Clustering using 4-D Histograms

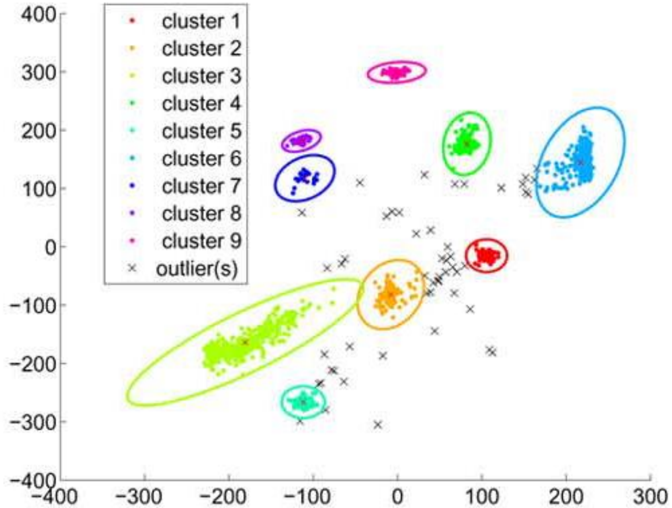
- Group trajectories into clusters of “main coherent motion”.
- Position/velocity over time are used to form 4-D histogram.
- Spatial proximity is applied.

Jung et al, 2008

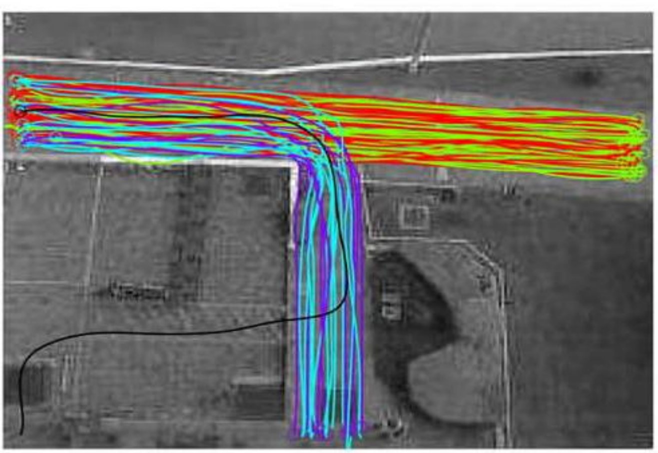
Trajectory Clustering using 4-D Histograms



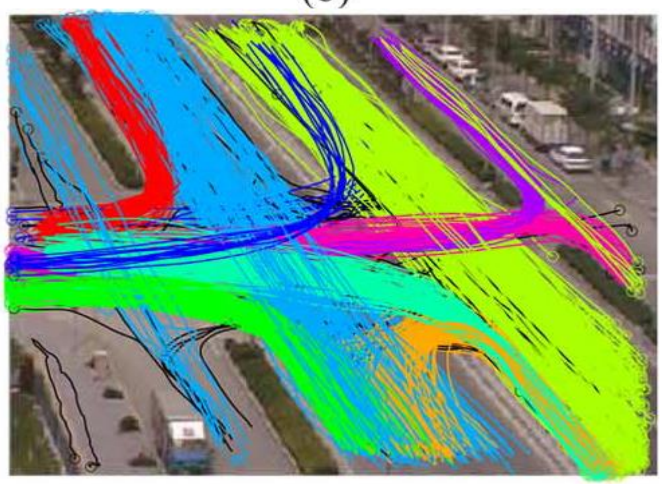
(a)



(b)



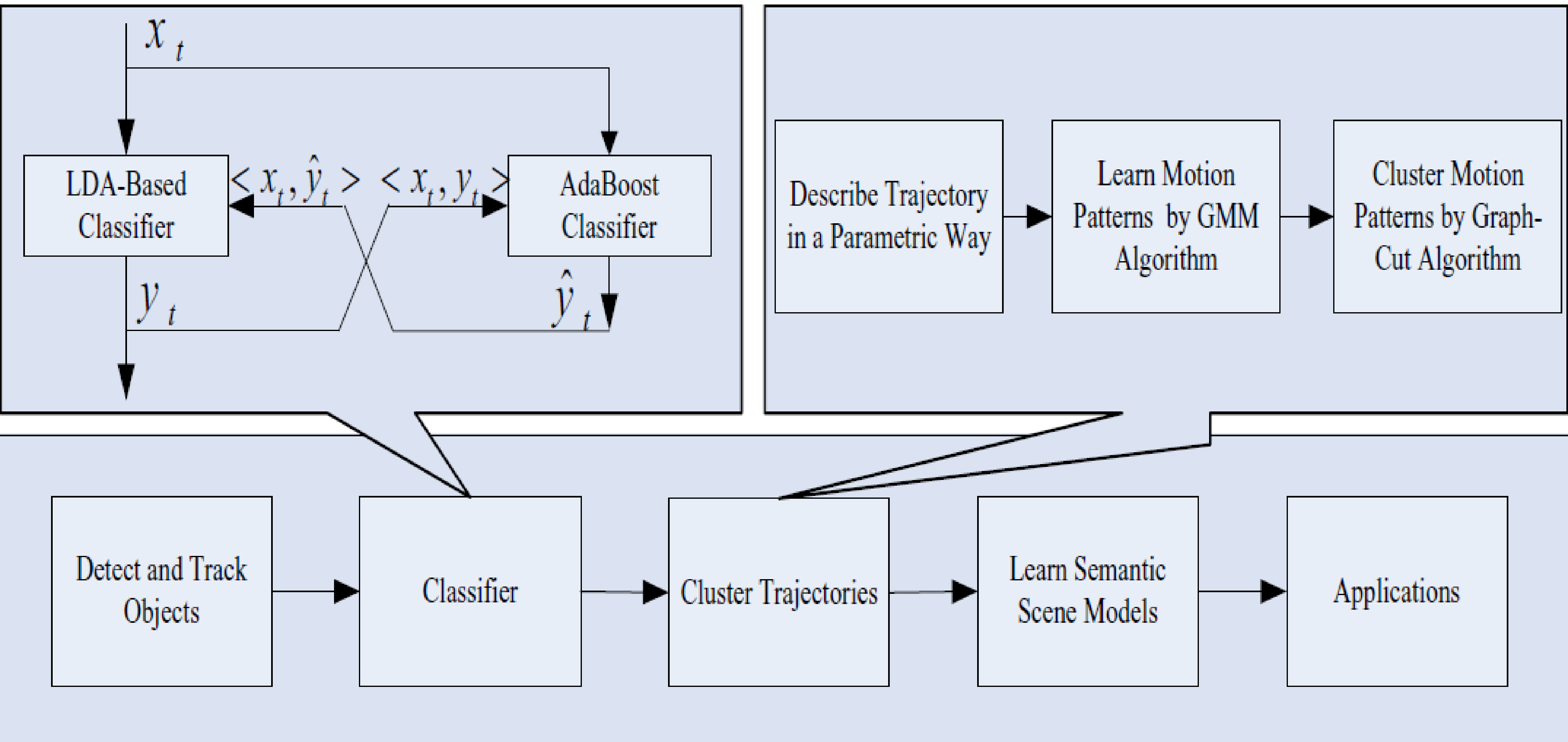
(c)



(d)

(a)–(b) Final clustering result with outlier removal.
(c)–(d) Trajectories used in the training stage shown in different colors for each cluster, and black ones
were classified as outliers.

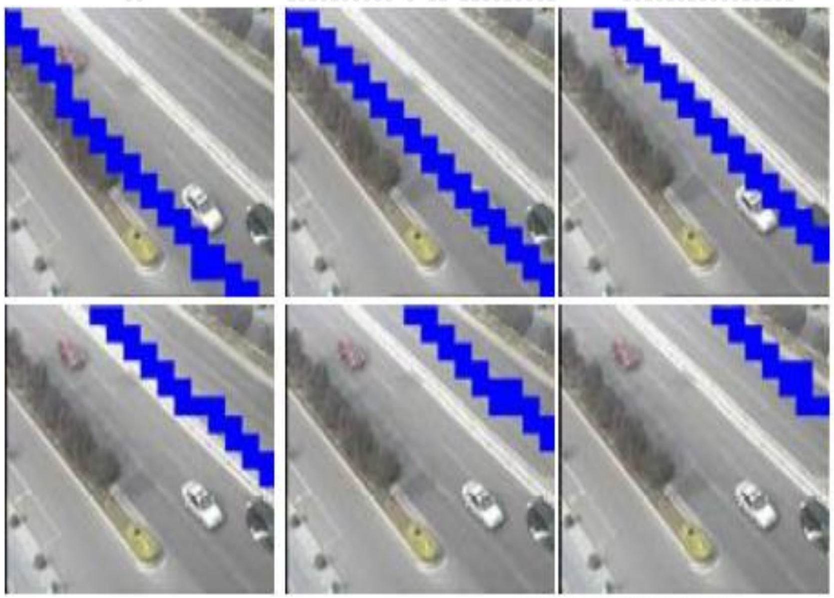
Learning Semantic Scene Models by Trajectory Clustering

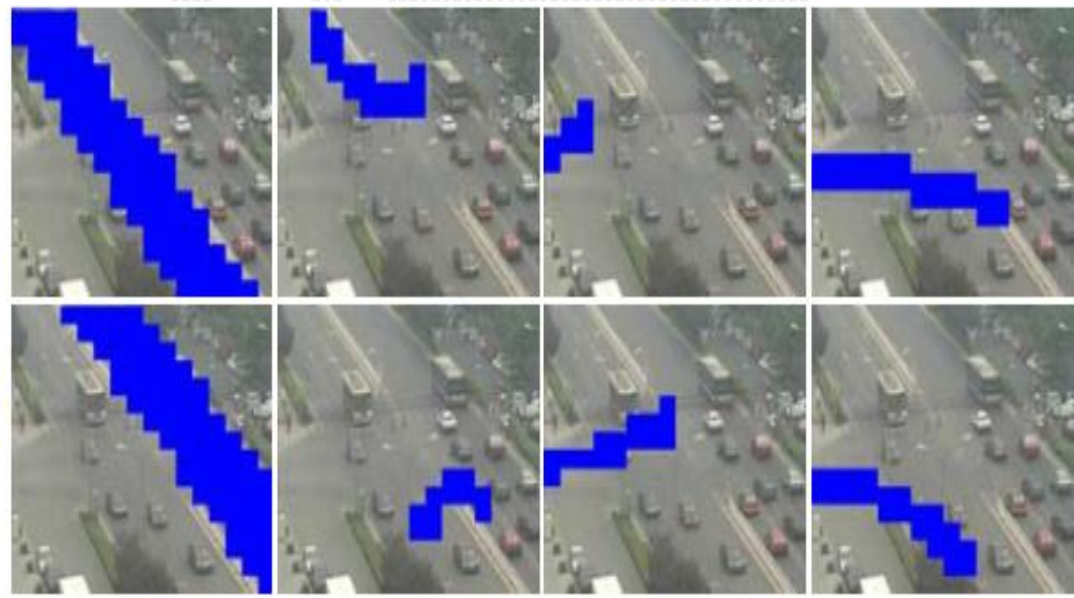


Zhang et al, 2009

Learning Semantic Scene Models by Trajectory Clustering

Exemplar results





Cluster validation

- No ground-truth for clusters.
- To minimise or maximise criteria for obtaining correct clusters and numbers:
  - Change initial number of clusters.
  - Use criteria such as “tightness and separation”.
  - Measure the distance between clusters.

- Introduction
- Human detection and tracking
- Human profiling
- Activity recognition
- Trajectory clustering
- Summary

Starting

- What is video surveillance?
- Why is it important?
- Challenges?

Human detection/tracking

- Human detection
  - Background subtraction
  - Mixture of Gaussian
  - Viola-Jones method
  - HoG
  - Shape context
- Human tracking
  - Incremental learning for visual tracking
  - Tracking with online multiple instance learning
  - Combining local features with kernel tracking
  - Audiovisual tracking

Human profiling

- State of the art techniques
  - Age classification using Radon transform and scaling SVM
  - Ethnicity classification based on gait using multi-view fusion
  - Ethnicity- and gender-based subject retrieval using 3-D face-recognition techniques

Human profiling

- State of the art techniques
  - Age classification using Radon transform and scaling SVM
  - Ethnicity classification based on gait using multi-view fusion
  - Ethnicity- and gender-based subject retrieval using 3-D face-recognition techniques

Activity recognition

- Sequential approaches
  - Data based
  - State model based
- Hierarchical approaches
  - Statistical
  - Description based

Trajectory clustering

- Distance (or similarity) measure
- Cluster update methodology
- Cluster validation

Thank you very much!

Q & A

References

- J.K. Aggarwal, M.S. Ryoo and K. Kitani, “CVPR 2011 Tutorial on Human Activity Recognition – Frontiers of human activity analysis”.
- N. Anjum and A. Cavallaro, “Multi-feature object trajectory clustering for video analysis”, IEEE Trans. CSVT, 18(11): 1555-1564, 2008.
- H. Asoh *et al.*, “An application of a particle filter to Bayesian multiple sound source tracking with audio and video information fusion”, *Proc. of Int. Conf. Inf. Fusion*, Stockholm, Sweden, Jun. 2004, pp. 805–812.
- B. Babenko, M.-H. Yang and S. Belongie, “Viusla tracking with online multiple instance learning”, *Proc. Of CVPR*, 2009.
- A. Bobick and J. Davis, “The recognition of human movement using temporal templates”. IEEE T PAMI 23(3), 2001.
- Bar-Shalom, Y. and Foreman, T. 1988. Tracking and Data Association. Academic Press Inc.
- M. J. Beal, H. Attias, and N. Jojic, “Audio–video sensor fusion with probabilistic graphical models,” *Proc. Of Eur. Conf. Computer Vision*, Copenhagen, Denmark, Jun. 2002, pp. 736–752.
- C. BenAbdelkader and P. Griffin. “A local region-based approach to gender classification from face images”. In: IEEE Conf. on Computer Vision and Pattern Recognition Workshop, 2005, pp. 52–52.
- T. Broida and R. Chellapa, “Estimation of object motion parameters from noisy images”. IEEE Trans. Patt. Analy. Mach. Intell. 8, 1986, pp. 90–99.
- Z. Cai, M. Saberian and N. Vasconcelos, “Learning complexity-aware cascades for dep pedestrian detection”, Proc. Of ICCV, 2015.
- V. Cevher,A. C. Sankaranarayanan, J. H. McClellan, and R. Chellappa, “Target tracking using a joint acoustic video system,” *IEEE Trans. Multimedia*, vol. 9, no. 4, pp. 715–727, Jun. 2007.
- N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darrell, “Multiple person and speaker activity tracking with a particle filter,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Cambridge, MA, May 2004, vol. 5, pp. V-881–V-884.
- S.-H. Cho and H.-B. Kang, “Panoramic background generation using mean-shift in moving camera environment”, Proc. Of IPCV, 2011.
- R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, “Detecting moving objects, ghosts and shadows in video streams”, IEEE Trans. on Patt. Anal. and Machine Intell., vol. 25, no. 10, Oct. 2003, pp. 1337-1342.
- N. Dalal and B. Triggs. “Histograms of oriented gradients for human detection”. Proc. CVPR, 2005.
- T. Dekel, S. Oron, M. Rubinstain, S. Avidan and W.T. Freeman, “Best-buddies similarity for robust template matching”, Proc. Of CVPR, 2015.
- A. Efros, A. Berg, G. Mori, and J. Malik, “Recognizing action at a distance”, ICCV, 2003.
- A. Elgammal, D. Harwood, and L.S. Davis, “Non-parametric Model for Background Subtraction”, Proc. of ICCV '99 FRAME-RATE Workshop, 1999.
- Y. Fu and T.S. Huang (2008). ‘Human Age Estimation with Regression on Discriminative Aging Manifold’, IEEE Transactions on Multimedia, 10(4), pp. 578-584.
- W. Gao and H, Ai, “Face gender classification on consumer images in a multiethnic environment”. In: Internat. Conf. on Biometrics (ICB), pp. 169–178, 2009.

- D. Gatica-Perez, G. Lathoud, I. McCowan, J. Odobez, and D. Moore, “Audio–visual speaker tracking with importance particle filters,” *Proc. IEEE Int. Conf. Image Process.*, 2003, pp. 25–28.
- D. Gavrila and L. Davis, Towards 3-D model-based tracking and recognition of human movement. In International Workshop on Face and Gesture Recognition 1995.
- D. M. Gavrila and V. Philomin. “Real-time object detection for smart vehicles”. Proc. Of ICCV, 1999.
- X. Geng, Z.H. Zhou and K. Smith-Miles, “Automatic Age Estimation Based on Facial Aging Patterns”. IEEE Transactions of Pattern Analysis and Machine Intelligence, 29(12), 2234-2240, 2007.
- A. Gupta, N. Srinivasan, J.Shi and L.Davis. “Understanding Videos, Constructing Plots - Learning a Visually Grounded Storyline Model from Annotated Videos”. Proc. Of CVPR, 2009
- S. Gutta, J.R.J. Huang, P. Jonathon and H. Wechsler, “Mixture of Experts for classification of gender, ethnic origin, and pose of human faces”, IEEE Trans. NN, 11(4), 2000, pp. 948-960.
- G. H. Golub and C. F. Van Loan. Matrix Computations. The Johns Hopkins University Press, 1996.
- B. Han, D. Comaniciu, and L. Davis, "Sequential kernel density approximation through mode propagation: applications to background modeling”, Proc. Asian Conf. on Computer Vision, 2004.
- S. Hosoi, E. Takikawa and M. Kawade, “Ethnicity estimation with facial images”, Proc. Of IEEE FGR, 2004, pp. 195-200.
- Z. Hu, Y. Wang, Y. Tian and T. Huang, “Selective eigenbackgrounds method for background subtraction in crowed scenes”, Proc. Of ICIP, 2011.
- F. Jiang, Y. Wu and A.K. Katsaggelos, “A dynamic hierarchical clustering method for trajectory-based unusual video event detection”, IEEE Tran.s IM, 18(4), 2009, pp. 907-913.
- C.R. Jung, L. Hennemann and S.R. Musse, “Event detection using trajectory clustering and 4-D histograms”, IEEE Trans. CSVT, 18(11), 2008, pp. 1565-1575.
- Y. Ke, R. Sukthankar, and M. Hebert, “Spatio-temporal shape and flow correlation for action recognition”. Proc. Of CVPR 2007.
- Y. H. Kwon and N. da Vitoria Lobo. “Age Classification from Facial Images”, Computer Vision and Image Understanding, 74(1), pp. 1–21, 1999.
- A. Lanitis, C.J. Taylor and T.F. Cootes. “Toward Automatic Simulation Of Aging Effects on Face Images”. IEEE Transactions of Pattern Analysis and Machine Intelligence, 24 (4), pp. 442-455, 2002.
- A. Lapedriza, M.J. Marin-Jimenez, J. Vitria, “Gender recognition in non-controlled environments, ” In: Internat. Conf. on Pattern Recognition, pp. 834–837, 2006.
- I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, Learning realistic human actions from movies, Proc. Of CVPR, 2008.
- J. Lim, D. Ross, R.-S. Lin and M.H. Yang, “Incremental learning for visual tracking”, Proc. NIPS, 2004.

CENTRE FOR SECURE INFORMATION TECHNOLOGIES

- B.P.L. Lo and S.A. Velastin, “Automatic congestion detection system for underground platforms,” Proc. of 2001 Int. Symp. on Intell. Multimedia, Video and Speech Processing, pp. 158-161, 2000.
- X. Lu and A.K. Jain, “Ethnicity identification from face images”, Proc. Of SPIE ISDS: Biometric Technology for human identification, 2004, pp. 114-123.
- L. Ma, J. Liu, J. Wang, J. Cheng and H. Lu, “An improved silhouette tracking integrating particle filler with graph cuts”, Proc. Of ICASSP, 2010.
- B. Moghaddam and M.-H. Yang, “Learning gender with support faces”, IEEE Trans. PAMI, vol. 24, No. 5, pp.707-711, 2002.
- G. Mori, X. Ren, A. A. Efros, and J. Malik. “Recovering human body configurations: Combining segmentation and recognition”. Proc. CVPR, 2:326–333, 2004.
- B. Morris and M.M. Trivedi: Learning trajectory patterns by clustering: Experimental studies and comparative evaluation. Proc. Of CVPR 2009: 312-319.
- M. Narayana, A. Hanson and E.G. Learned-Miller, “Background subtraction – separating the modelling and the inference”, MVA, Vol. 25, Issues 5, pp. 1163-1174, 2014.
- N.T. Nguyen, D.Q. Phung, S. Venkatesh and H. Bui,” Learning and Detecting Activities from Movement Trajectories Using the Hierarchical Hidden Markov Models”, Proc. Of CVPR, 2005.
- J.C. Niebles, H. Wang, and L. Fei-Fei, “Unsupervised learning of human action categories using spatial-temporal words”, Proc. Of BMVC, 2006.
- N. M. Oliver, B. Rosario, and A. P. Pentland, “A Bayesian Computer Vision System for Modeling Human Interactions,” IEEE Trans. on Patt. Anal. and Machine Intell., vol. 22, no. 8, pp. 831-843, 2000.
- S. Park and J.K.Aggarwal, A hierarchical Bayesian network for event recognition of human actions and interactions. Multimedia Systems, 2004.
- N.-J. Pyun, H. Sayah and N. Vicent, “Adaptive Harr-like features for head pose estimation”, Proc. Of ICIAR, 2014.
- R. Ronfard, C. Schmid, and B. Triggs. “Learning to parse pictures of people”. Proc. ECCV, 2002.
- Y. Rui and Y. Chen, “Better proposal distributions: Object tracking using unscented particle filter,” in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognition*, Dec. 2001, pp. 786–793.
- S. Savarese, A. DelPozo, J. Niebles, and L. Fei-Fei, “Spatial-temporal correlatons for unsupervised action classification”, Proc. Of WMVC, 2008.
- C. Shan, “Learning local binary patterns for gender classification on real =world face images”, Pattern Recognition Letters, 33(2012), 431-437.
- L. Shen, T.W. Chua and K. Leman, “Shadow optimisation from structured deep edge detection”, Proc. Of CVPR, 2015.
- C. Stauffer, W.E.L. Grimson, “Adaptive background mixture models for real-time tracking”, Proc. of CVPR 1999, pp. 246-252.
- R.L. Streit and T.E. Luginbuhl, “Maximum likelihood method for probabilistic multi-hypothesis tracking”. In Proc. of the International Society for Optical Engineering (SPIE.) vol. 2235. 394–405, 1994.
- G. Toderici, S. M. O’Malley, G. Passalis, T. Theoharis, I.A. Kakadiaris, “Ethnicity- and gender-based subject retrieval using 3-D face-recognition techniques”, IJCV, 2010, 89: 382-391.

CENTRE FOR SECURE INFORMATION TECHNOLOGIES

- D. Tran and A. Sorokin, “Human activity recognition with metric learning”, Proc .of ECCV, 2008.
- G, Tzimiropoulos, S. Zafeirious and M. Pantic, “Subspace learning from image gradient orientation”, TPAMI,34(12), p. 2454-2466, 2012.
- P.Viola, M.Jones, and D.Snow. “Detecting pedestrians using patterns of motion and appearance”. Proc. ICCV, 2003.
- P. Viola and M. Jones, “Robust Real-time Object Detection”, IJCV, 2001.
- J. Wang, Z. Liu, Y. Wu and J. Yuan, “Mining Actionlet Ensemble for Action Recognition with Depth Cameras”, Proc. Of CVPR, 2012
- J.-G. Wang, W.-Y. Yau and H. L. Wang, “Age Categorization via ECOC with Fused Gabor and LBP Features”. Procs. of the IEEE Workshop on Applications of Computer Vision (WACV), pp.313-318, 2009.
- S.-F. Wong, T.-K. Kim, and R. Cipolla, “Learning motion categories using both semantic and structural information”, Proc. CVPR, 2007.
- J. Yamato, J. Ohya, and K. Ishii, Recognizing human action in time-sequential images using hidden Markov model. Proc. CVPR, 1992.
- Y. Yang and K. Chen, “Temporal data clustering via weighted clustering ensemble with different representations”, IEEE Trans. KDE, 23(2), 2011, pp. 307-320.
- A. Yilmaz, O. Javed and M. Shah, “Object tracking: a survey”, ACM Computing Surveys, Vol. 38, No. 4, 2006.
- D. Zhang, Y. Wang and B. Bhanu, “Ethnicity classification based on gait using multi-view fusion”, Workshop of CVPR, 2010.
- D. Zhang, Y. Wang, Z. Zhang and M. Hu, “Ethnicity classification based on fusion of face and gait”, Proc. Of ICB, 2012.
- T. Zhang, H. Lu and S.Z. Li, “Learning Semantic Scene Models by Object Classification and Trajectory Clustering “, Proc. Of CVPR, 2009.
- H. Zhou, P. Miller and J. Zhang "Age classification using Radon transform and entropy based scaling SVM". *Proc. of British Machine Vision Conference*, 2011.
- H. Zhou, Y. Yuan and C. Shi, “Object tracking using SIFT features and mean shift”, CVIU, 113(3), 345-352, 2009.
- H. Zhou, M. Taj and A. Cavallaro, “Target detection and tracking with heterogeneous sensors”, IEEE Journal of Selected Topics in Signal Processing, 113(3), 345-352, 2009.
- H. Zhou, A. Wallace and P. Green, “Efficient tracking and ego-motion recovery using gait analysis”, Signal Processing, 89(12), p. 2367-2384, 2009.