Activity analysis from video

David Hogg
University of Leeds

Mathematical Challenges in Computer Vision,
Warwick, March 2009

Introduction

Understanding our world requires knowledge about a huge variety of objects and activities.

How is this knowledge acquired and used?

Summary of talk:

- Learning about objects
- Learning about activities
- Dealing with visual ambiguity in recognising activities

Object recognition

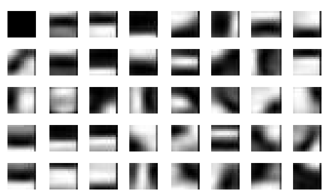
Challenge posed by Heinrich Bulthoff, MPI

How many chairs are there in the picture?

Learning from labelled examples – ‘bag of words’

corpus of labelled examples

$$\{(I_i, c_i)\}$$



set of ‘visual words’ $\{w_j\}$
(e.g. patches, texture descriptors)

$$\{(\mathbf{n}_i, c_i)\}$$

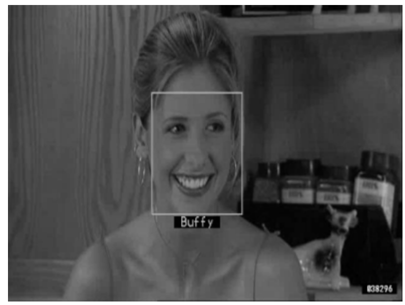
histograms of word occurrences

classifier

$$c = f(\mathbf{n})$$

Obtaining labelled examples

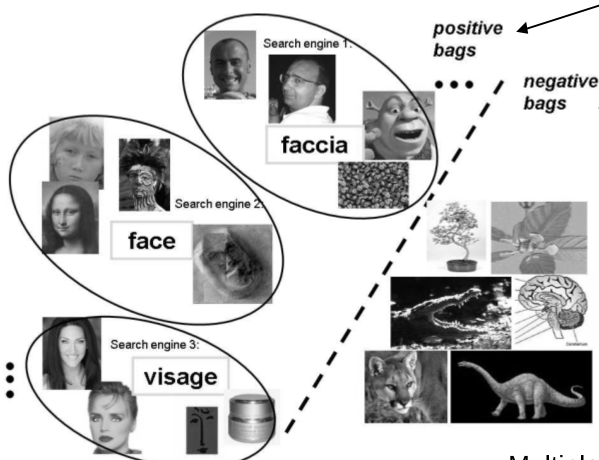
the world wide web
TV and radio shows
CCTV networks



Everingham et al., BMVC 2006

Dealing with 'mislabelled' data

each containing at least one positive instance



each containing only negative instances

Multiple Instance Learning

e.g. sparse-MIL (sMIL)
Bunescu and Mooney, ICML 2007

from *Vijayanarasimhan and Grauman, CVPR 2008*

Learning without labels

Given a set of unlabelled images $\{I_i\}$

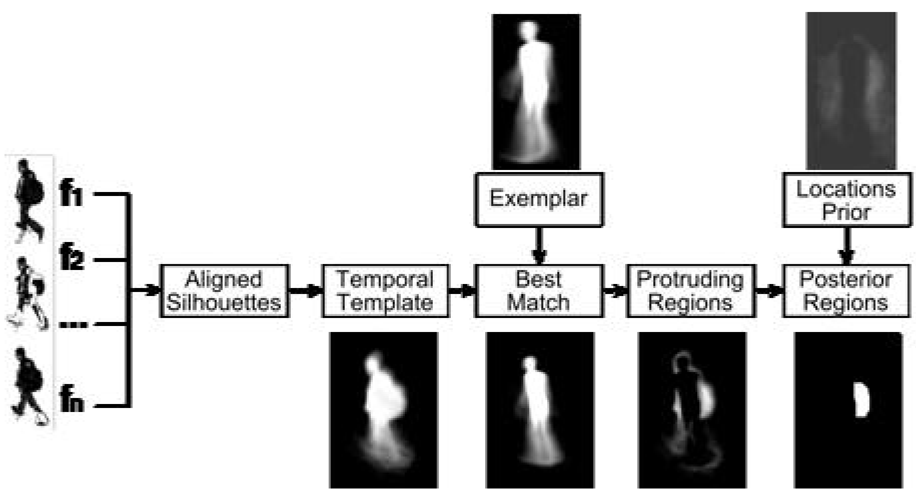
Cluster co-occurring visual words into classes, for example:

- using spectral clustering on an adjacency graph between images (*Graumann and Darrell, CVPR 2006*);

- through modelling each histogram as a mixture of 'category' histograms (Latent Dirichlet Allocation LDA, Latent Semantic Analysis LSA) (e.g. *Sivic, Russell, Efros, Zisserman, and Freeman., ICCV 2005*)

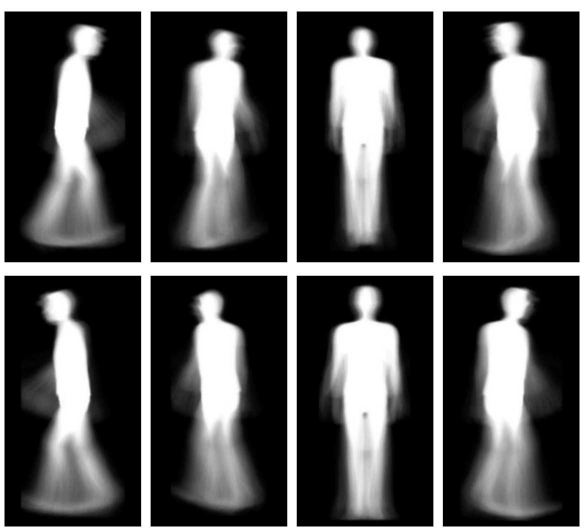
Detecting carried objects

Damen and Hogg, ECCV 08



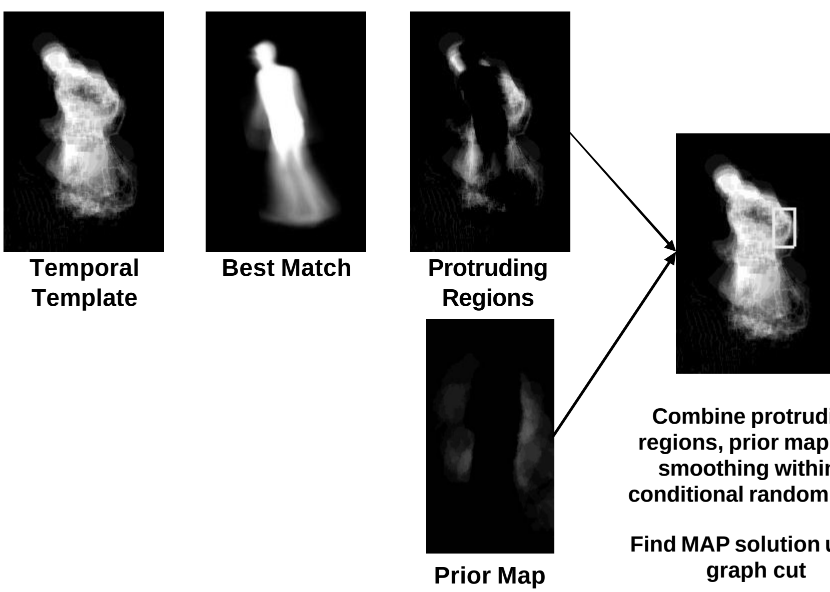
Building the temporal template

View-specific model templates for an average person



Aligning to the model template

Combining with the prior map and smoothness



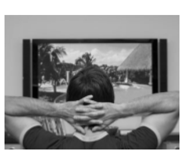
Combine protruding regions, prior map and smoothing within a conditional random field.

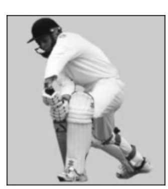
Find MAP solution using graph cut

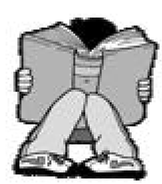
Demo

Activity discovery challenge

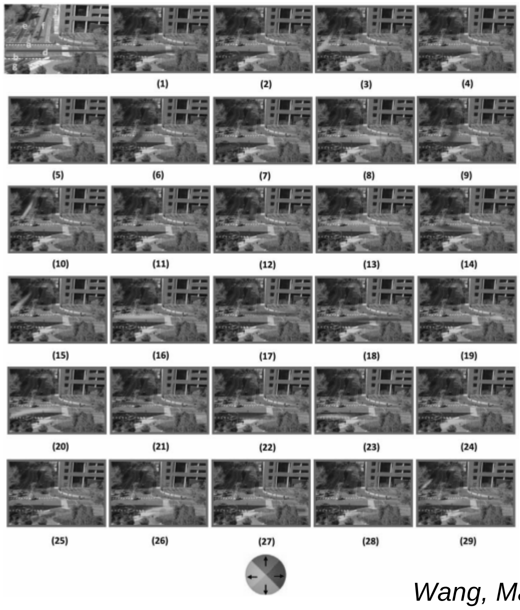
To infer the principles of the game of cricket by visual observation alone







Learning a layered model of activities without labels



Visual words: Quantised position and flow direction of 'change pixels'

Atomic activities: Co-occurring visual words (in short clips)

Interactions: Co-occurring atomic activities (in short clips)

Wang, Ma and Grimson, TPAMI 31(3) 2009

Learning activities

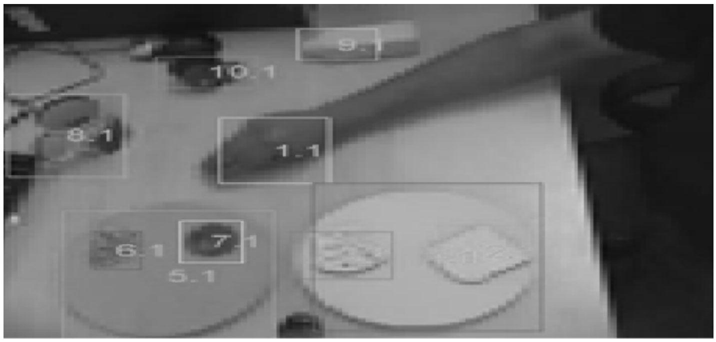
Sridhar, et al., ECAI 2008

Overview of method

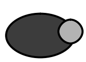
- Detect and track objects in video
- Represent spatio-temporal relationships as a labelled graph
- Find maximal frequent isomorphic subgraphs – output as discovered activities

Object detection and tracking

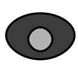
Colour-based segmentation
Careful to minimise (but not eliminate) occlusions
Lots of spurious blobs



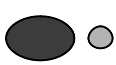
Spatial relations



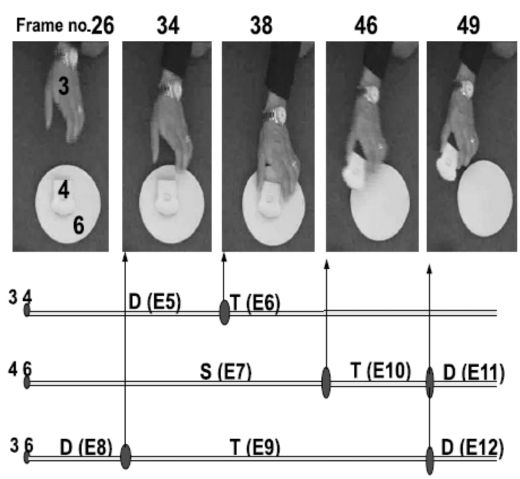
T: touches



S: surrounds

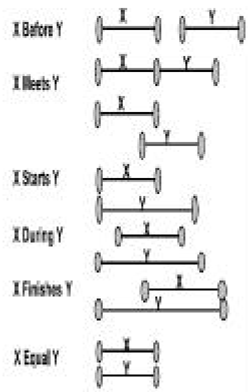


D: disconnected

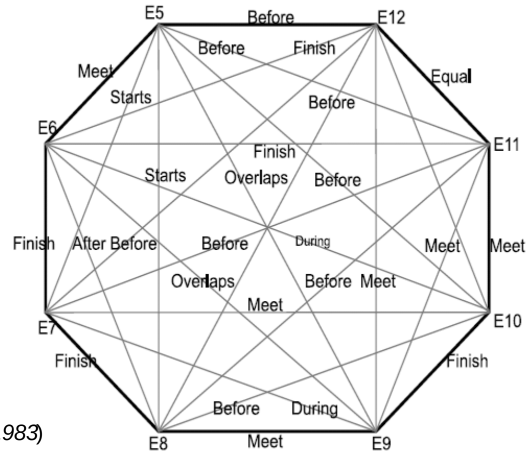


Temporal relations and the spatio-temporal graph

Dealing with parallel activities and varying sequential order of the spatial configurations that make up an activity

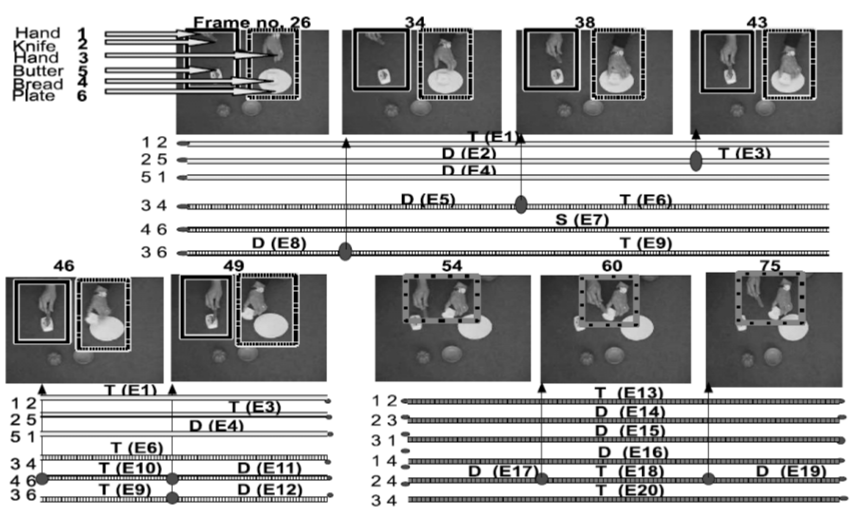


Allen's relations (*Allen, CACM 1983*)

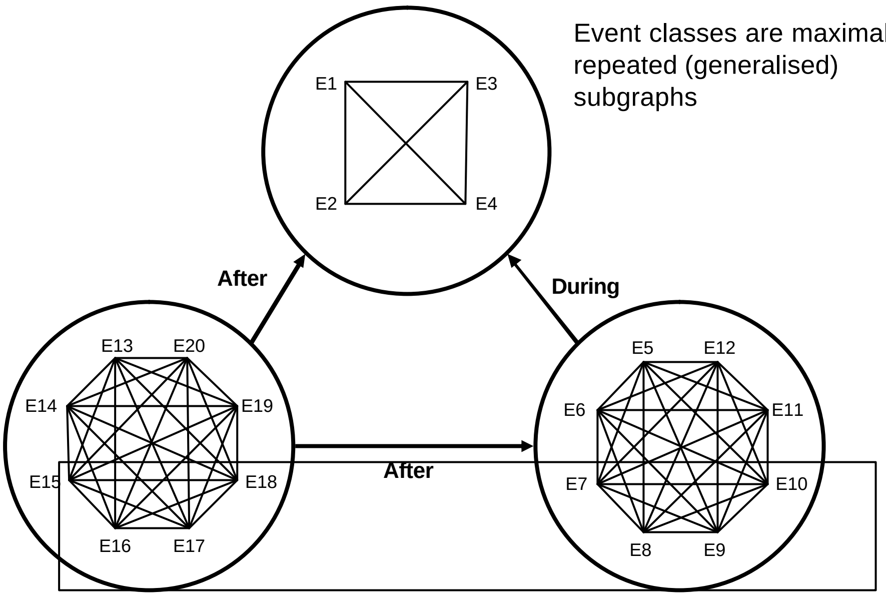


Attention

Focus on *atomic events*: maximal sub-graphs involving a constant set of connected (S,T) objects, at least one of which must move



Replace full graph by a graph over atomic events



Inducing a functional object taxonomy

Form Boolean matrix of the role played by objects in each event class (+ partially generalised classes)

Event classes

$$
\begin{array}{c|ccccccccccc}
 & & E_1 & & & E_2 & & & & E_m & \\
 & (\bullet & \bullet & \bullet) & (\bullet & \bullet) & \cdots & (\bullet & \bullet & \bullet & \bullet) \\
\hline
o_1 & 0 & 1 & 0 & 0 & 0 & & 0 & 0 & 1 & 0 \\
o_2 & 1 & 0 & 0 & 0 & 1 & & 0 & 0 & 0 & 0 \\
\vdots & & & & & & & & & & \\
o_n & 1 & 0 & 0 & 0 & 0 & & 0 & 0 & 0 & 0
\end{array}
$$

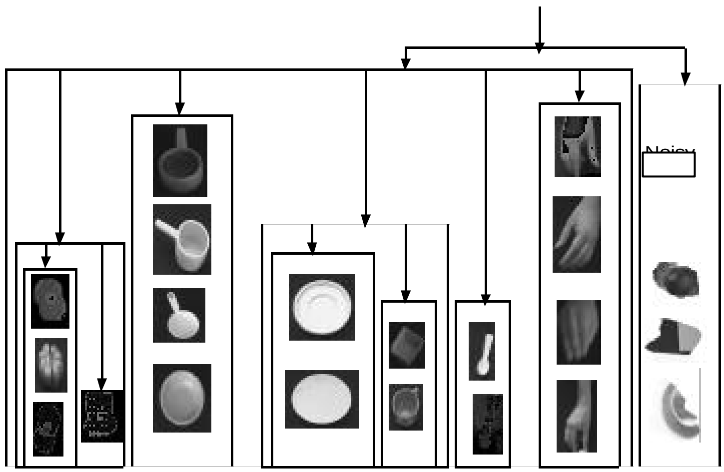
Objects

Compress the rows (pattern for each object) using PCA

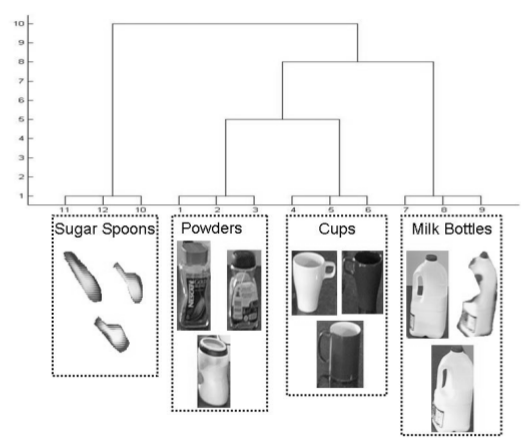
Obtain object taxonomy by hierarchical-clustering of the compressed rows

Experimental Results

Video of 5 minutes: preparing breakfast with tea, and a simple vegetable curry
50 objects, ~ 3000 roles. PCA reduces to 50 x 20



Experimental results – a more challenging dataset

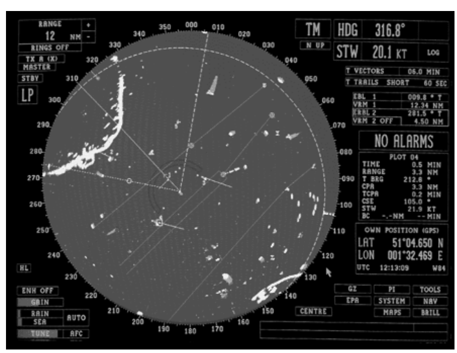


Dealing with detection errors and ambiguity

Radar tracking

Dealing with

- missed detections
- spurious detections



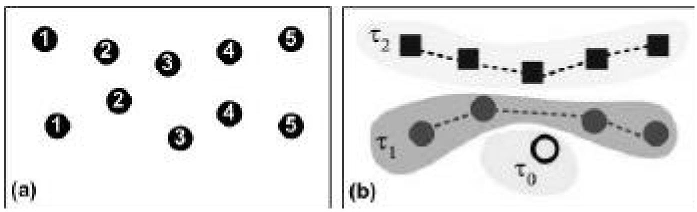
Long history from radar literature and elsewhere:

Ingemar Cox, A Review of Statistical Data Association Techniques for Motion Correspondence, International Journal of Computer Vision, vol. 10, pp. 53-66, 1993.

Standard approach

Find the optimal global explanation:

Given a set of noisy observations Y over a period of time.
An explanation is a partition of these observations $w = \{t_0, t_1, \ldots t_K\}$ where each part defines a track and t_0 contains all spurious observations (false alarms)



Seek $\underset{w \in \Omega}{\operatorname{argmax}}(p(w \mid Y))$

Formulation from Oh, Russell and Sastry, CDC-04

Defining $p(w \mid Y)$

Assumptions:

(1) each track behaves as a stochastic linear system:

$$x_{t_{i+1}} = Ax_{t_i} + h$$
$$y_{t_i} = Cx_{t_i} + u$$

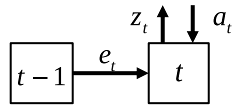
(note that matrix A and noise term scaled according to the width of interval

(2) new objects and false alarms occur as Poisson processes

(3) objects disappear and are undetected with fixed probability at each time-step

For a given w at time-step t, assume:

- e_t objects persist from t-1
- a_t new objects appear
- z_t objects disappear
- d_t objects detected
- f_t false alarms
- $u_t = e_t - z_t + a_t - d_t$ objects undetected

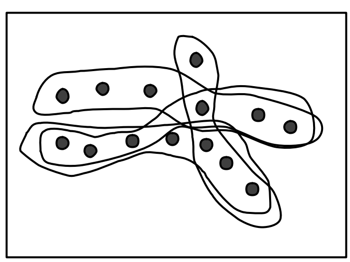


$$P(w \mid Y) = \frac{1}{Z} \prod_{t=1}^{T} \underbrace{p_Z^{z_t} (1-p_z)^{e_t - z_t}}_{\text{track terminations}} \underbrace{p_d^{d_t} (1-p_d)^{u_t}}_{\text{missing observations}} \underbrace{\frac{l_b^{a_t}}{a_t!} \frac{l_f^{f_t}}{f_t!}}_{\text{new objects and false alarms}} \prod_{t \in w \setminus \{t_0\}} \prod_{i=1}^{|t|-1} \underbrace{?\,(t(t_{i+1}) \mid C\bar{x}_{t_i}(t), B_{t_{i+1}}(t))}_{\text{stochastic linear system}}$$

Integer Programming

Morefield, IEEE-TAC 1977

- Create a large set of feasible tracks F (a covering), many of which will be inconsistent with one another.



- Seek the optimal partition from a subset of these tracks + false alarms

$$\operatorname*{argmax}_{\substack{w \subset F \\ w \in \Omega}} (p(w \mid Y))$$

Example
from *Leibe, Schindler, and Van Gool, ICCV 2007*

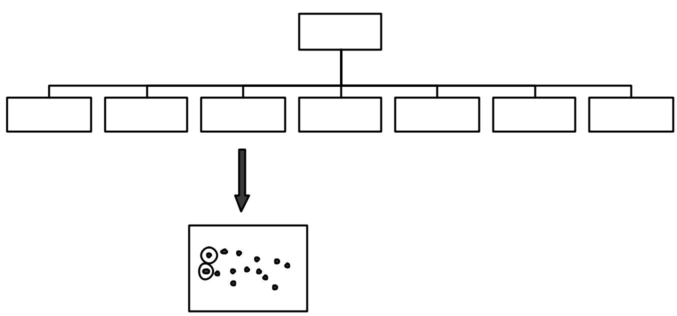
Uses a trained pedestrian detector operating on each frame

from http://www.vision.ee.ethz.ch/~bleibe/index.html

Multiple-Hypothesis Tree (MHT)
Reid, IEEE-TAC 1979

- Iteratively extend partial tracks at each time-step
- Pursue multiple hypotheses where there is ambiguity
- Prune unlikely hypotheses to keep search tractable

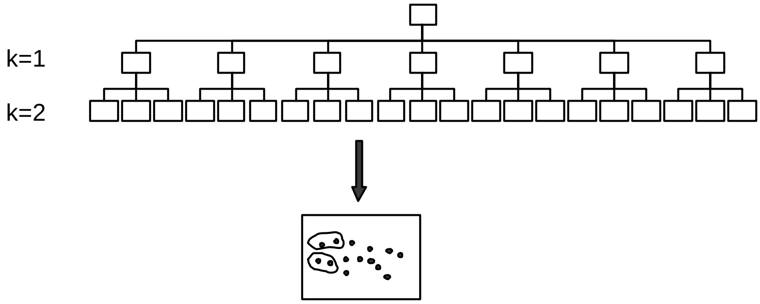
k=1



Multiple-Hypothesis Tree (MHT)

Reid, IEEE-TAC 1979

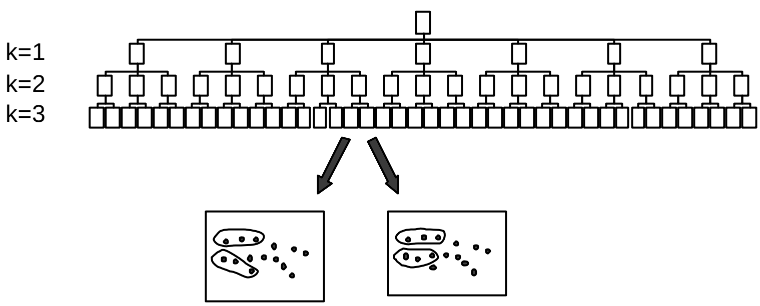
- Iteratively extend partial tracks at each time-step
- Pursue multiple hypotheses where there is ambiguity
- Prune unlikely hypotheses to keep search tractable



Multiple-Hypothesis Tree (MHT)

Reid, IEEE-TAC 1979

- Iteratively extend partial tracks at each time-step
- Pursue multiple hypotheses where there is ambiguity
- Prune unlikely hypotheses to keep search tractable



Markov Chain Monte Carlo Data Association

Oh, Russell, and Sastry, CDC-04, 2004

- Draw samples from posterior $p(w \mid Y)$ and select the maximum. Use Markov Chain Monte Carlo (MCMC) to do this efficiently.

initialise w

repeat many times

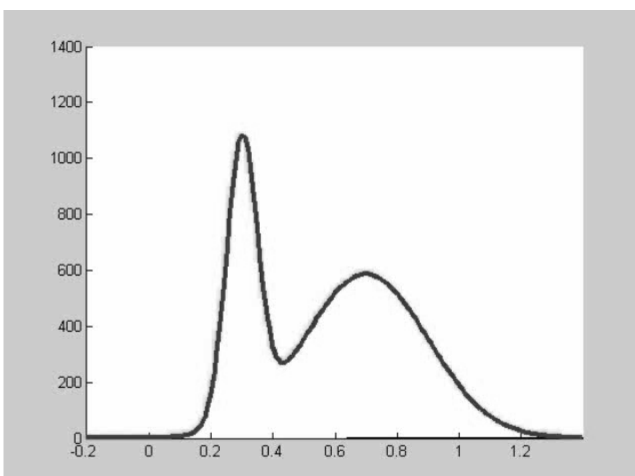
    Sample w' from proposal distribution $q(w, w')$

    Replace w by w' with (acceptance) probability:

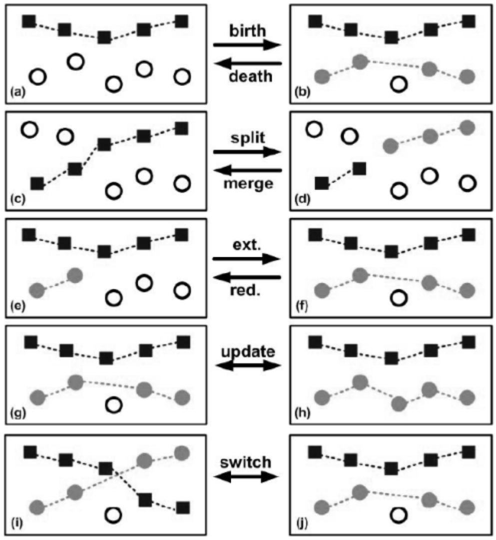
$$A(w,w') = \min\left(1, \frac{p(w' \mid Y) q(w', w)}{p(w \mid Y) q(w, w')}\right)$$

end

Introduction to MCMC



MCMC moves



From Oh, Russell and Sastry, CDC-04, 2004

Detecting people parking and collecting bikes

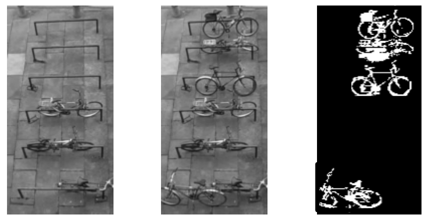
Damen & Hogg, BMVC 2007

Task: linking people dropping-off and picking-up bikes

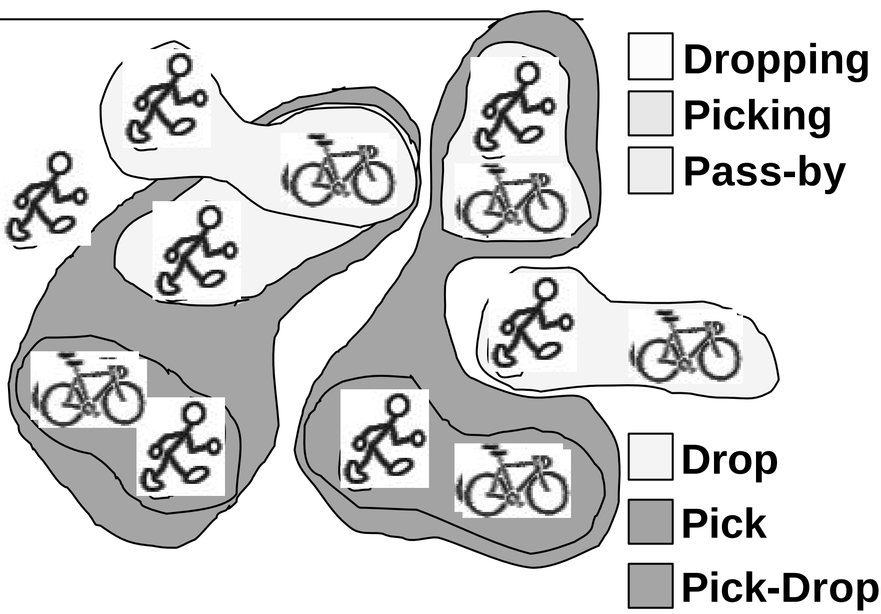
Method

- Track people (+/- bikes) entering and leaving the rack area
- Detect new clusters of dropped & picked bikes each time the rack area becomes empty
- List the possible new drop, pick and pass-through events, assuming people entering the rack, drop or pick no more than one bike
- Find optimal set of linked drop and pick events

$$\arg\max_{w}(p(w \mid Y))$$



Interpretation w

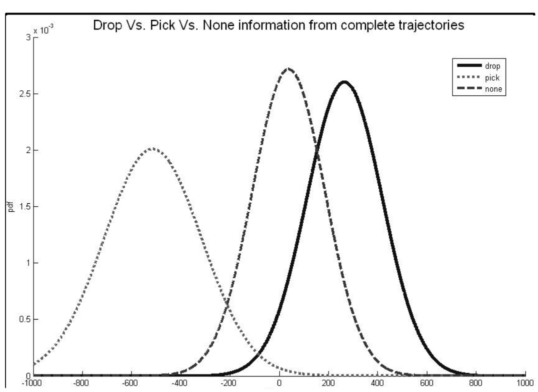


Defining $p(w \mid Y)$

Based on:

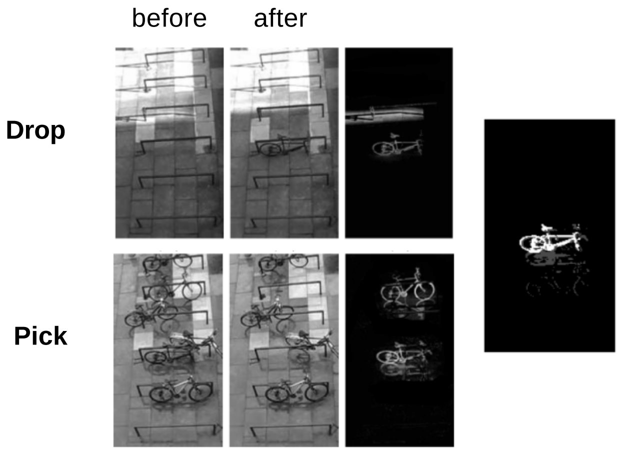
- Change in the area of person-blobs between entering and leaving rack
- Proximity of people to bike clusters
- Similarity of bike clusters between drop and pick
- Prior probabilities for the different events:

Likelihood of a person dropping, picking or passing through

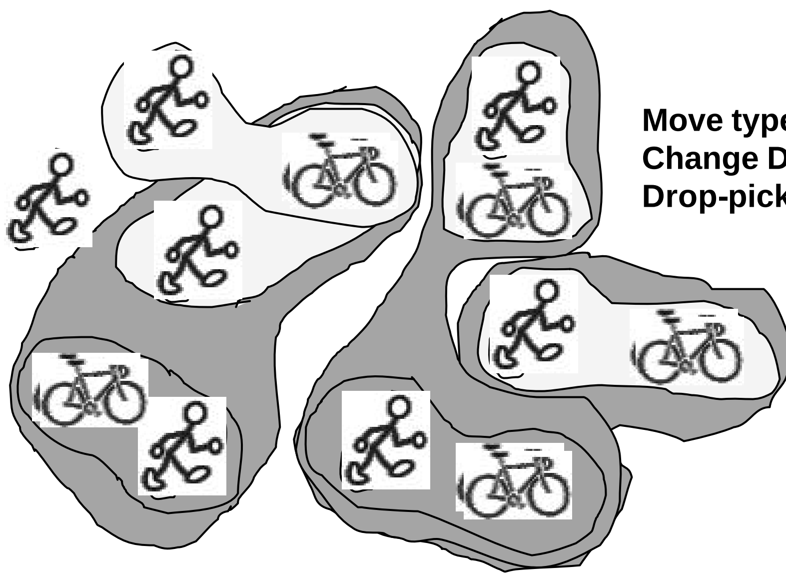


Area difference of person-blob entering and leaving the rack

Likelihood of a drop/pick linkage

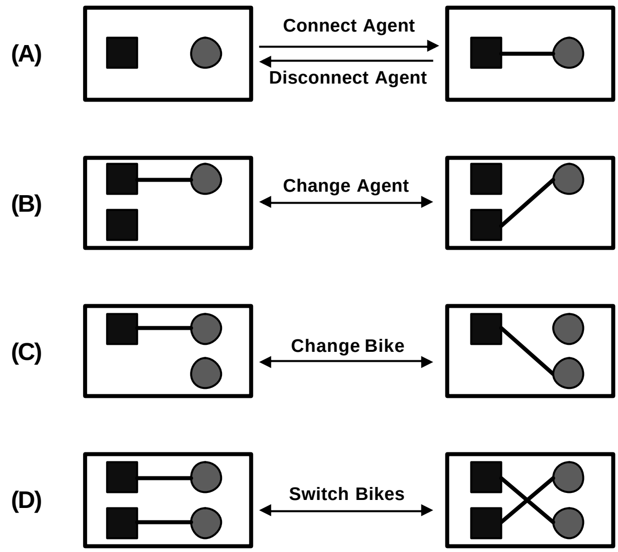


Find the MAP solution using MCMC

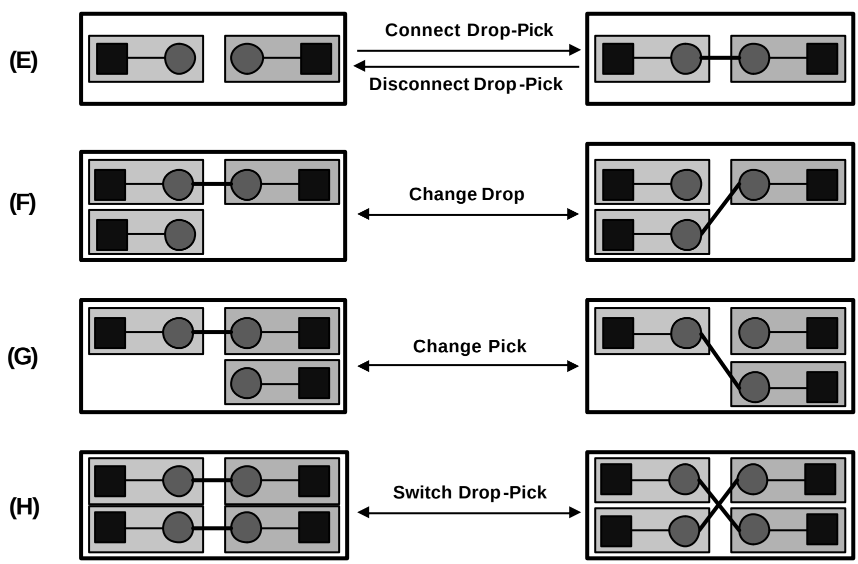


Move type: Change Drop in Drop-pick

Possible moves - 1



Possible moves - 2



Dataset

Results

Results

Number of drops, picks, drop-picks	Greedy		MCMC		SA-MCMC	
	log(p)	Accuracy	log(p)	Accuracy	log(p)	Accuracy
24,20,20	102.3	**72.41**	57.9	**91.38**	57.9	**91.38**
11,12,11	23.5	**85.19**	4.6	**100.00**	4.6	**100.00**
20,19,18	609.7	**58.59**	429.0	**88.28**	422.3	**89.84**
20,10,20	6272.7	**73.81**	6077.3	**83.33**	6083.7	**87.30**
14,13,14	5034.5	**89.05**	4944.7	**94.89**	4937.1	**94.16**
28,17,14	860.4	**66.07**	815.8	**71.43**	808.4	**76.79**
39,41,22	934.4	**45.69**	681.2	**48.22**	658.23	**51.78**