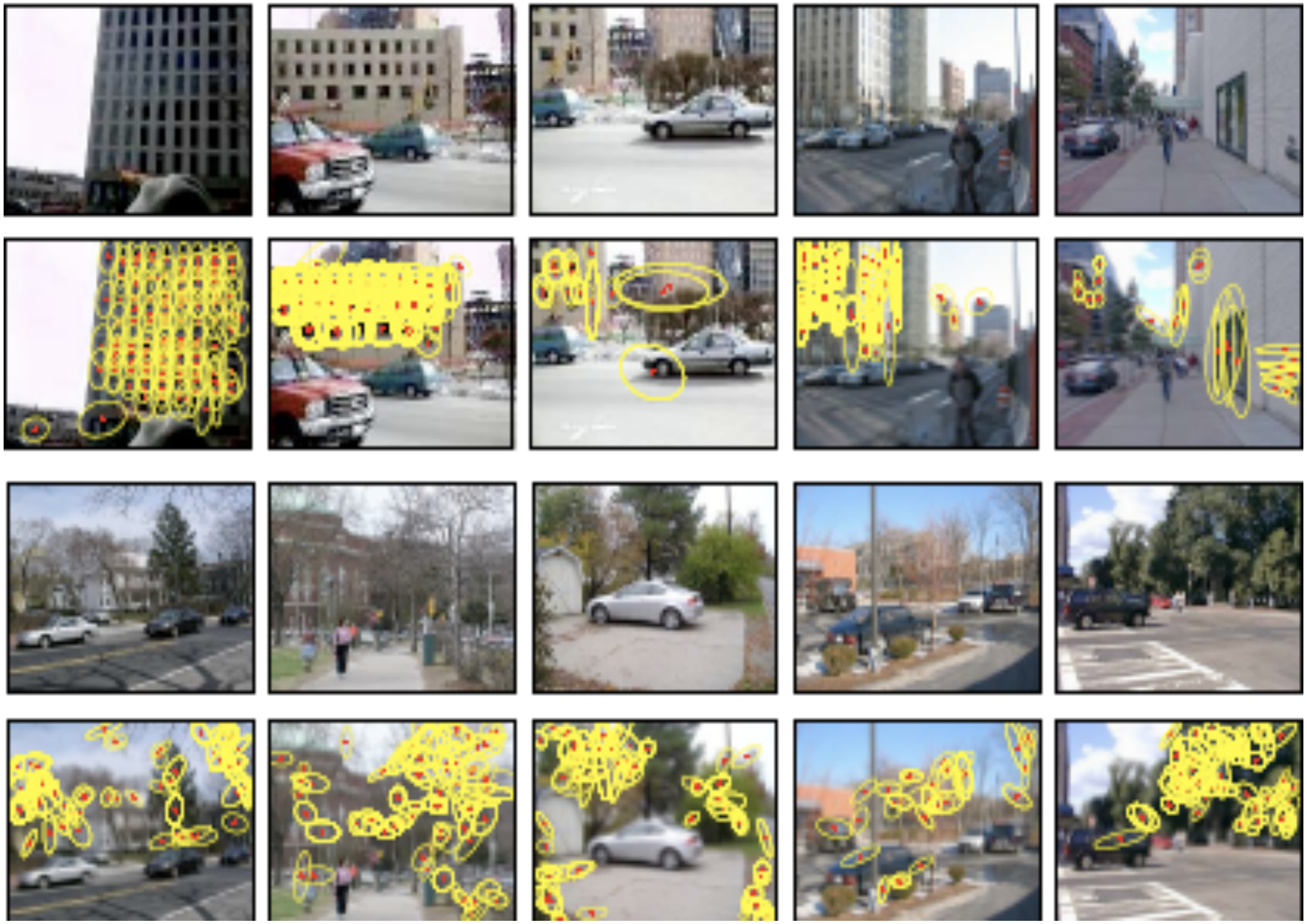
CS395T-Visual Recognition and Search

Gautam S. Muralidhar

Today's Theme

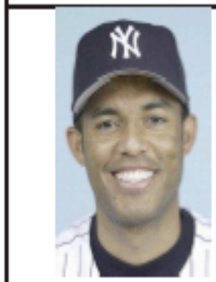
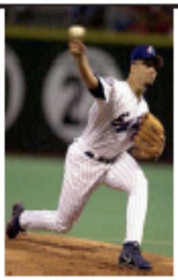
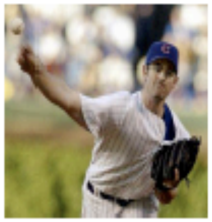
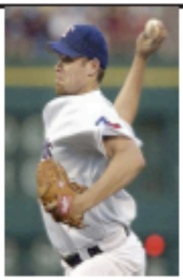
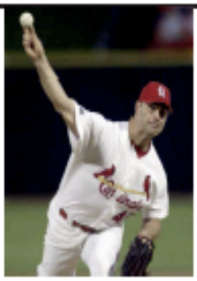
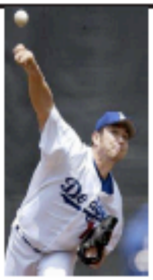
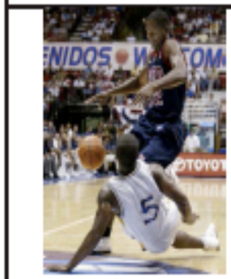
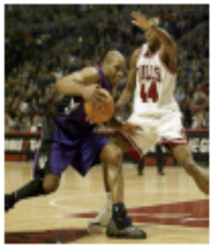
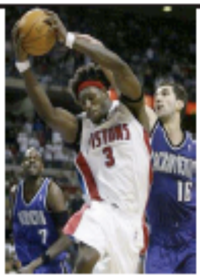
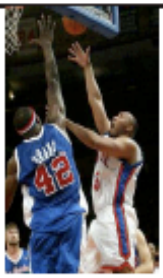
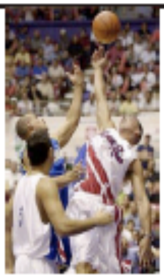
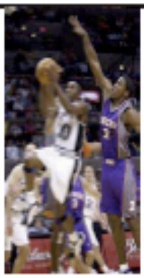
- Unsupervised discovery of images
- Main motivation behind unsupervised discovery is that supervision is expensive
- Common tasks include –
  - Detecting objects and their locations
  - Segmentation
  - Activity recognition
  - Irregularities in images and videos

Detecting Objects and Segmentation



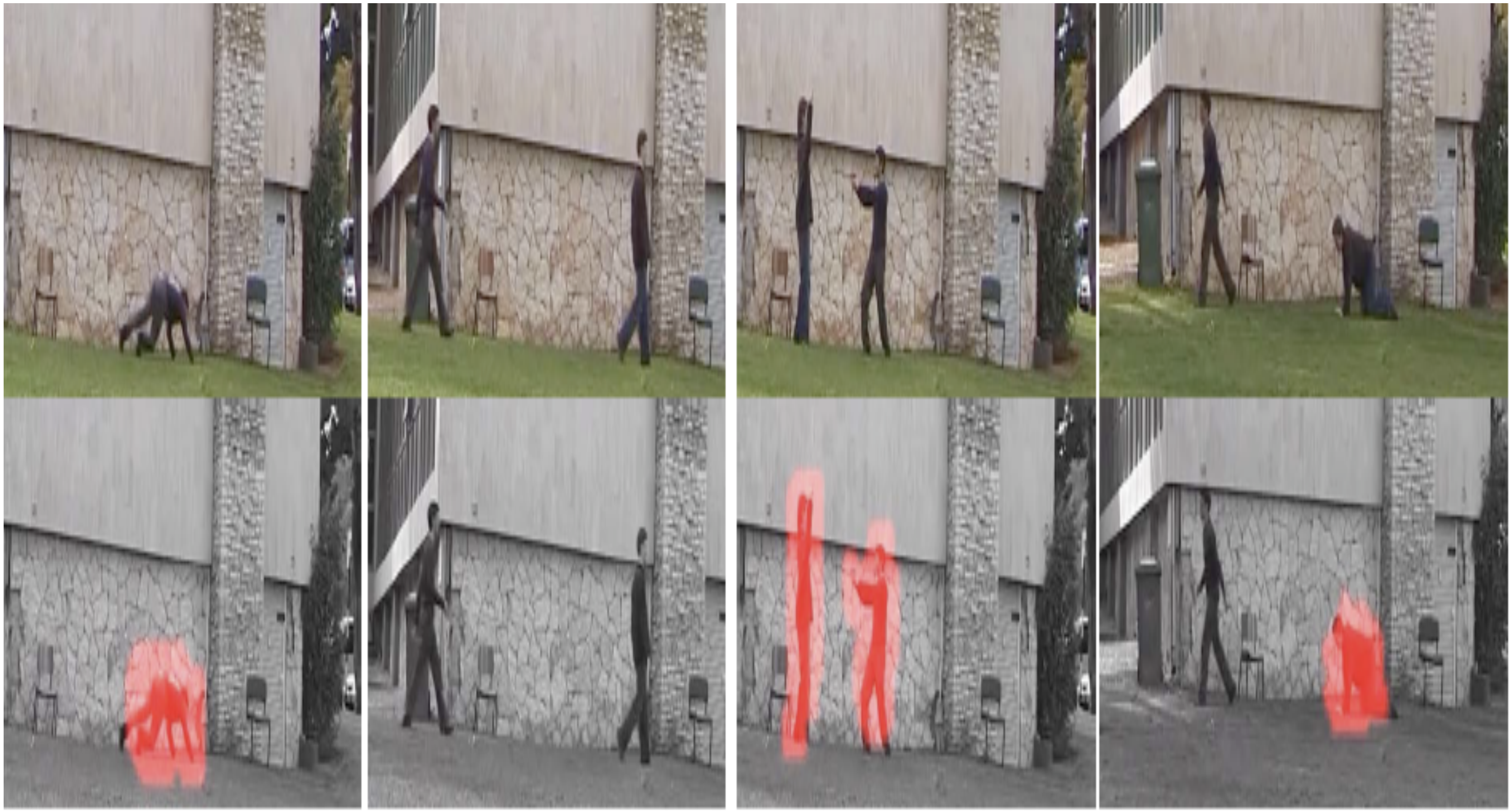
From Sivic et al

Action Class Recognition

					
face close-up picture	right-handed pitcher throws	right-handed pitcher throws	face close-up picture	right-handed pitcher throws	right-handed pitcher throws
					
a player drives past another	a player drives past another	a player has his shot blocked	a player goes for a lay-up against a defender	a player goes for a lay-up against a defender	a player goes for a lay-up against a defender

From Wang et al

Detecting Irregularities



From Boiman and Irani

Recipes

- Usually –
  - Process images and detect interest points
  - Extract low level features /descriptors (e.g., SIFT)
  - Cluster the image based on the descriptors
  - Learn statistical models to infer object categories / activity classes
- An alternative –
  - To make use of an existing database as evidence for a task, for e.g., detecting irregularities

Detecting objects and their locations in images

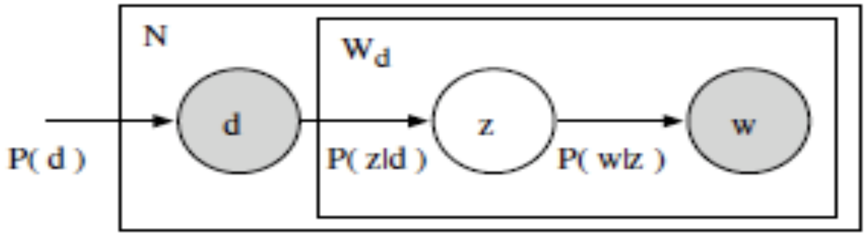
- Sivic et al

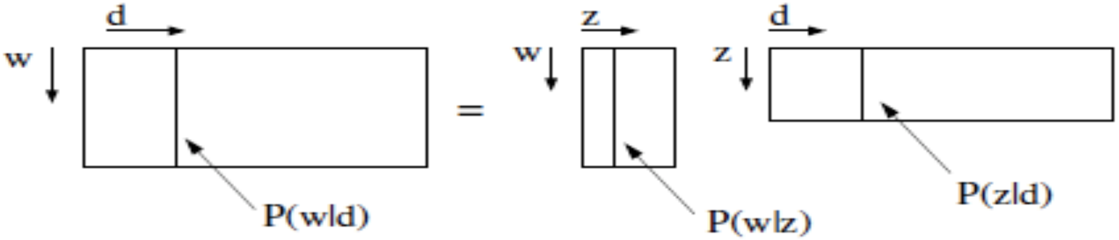
Analogy between text documents and images

- Text documents - composed of words, Images – composed of visual words
- Both can be represented by a bag of words approach
- Associated with each (visual) word is an (object) topic category
- Text documents – mixture of topics, Images – mixture of object categories

pLSA

- The joint probability $P(w_i,d_j,z_k)$ is assumed to have the following graphical model:





- Goal of pLSA – find topic specific word distribution P(w|z), document specific mixing proportions P(z|d) and from these, the document specific word distribution P(w|d)

From Sivic et al

pLSA model

- Fitting the model involves determining the topics, which are common to all documents and mixture of coefficients, which are specific to each document

- Maximizing the objective function

$$L = \prod_{i=1}^{M} \prod_{j=1}^{N} P(w_i|d_j)^{n(w_i,d_j)}$$

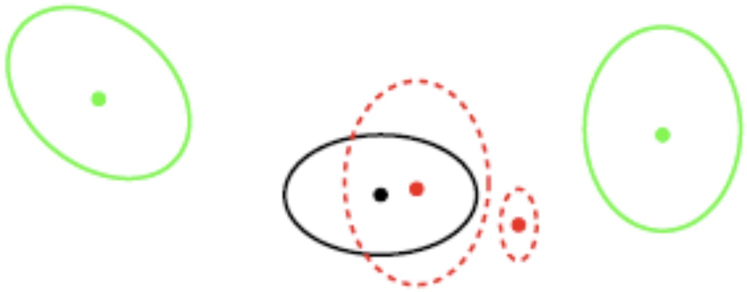
  yields the maximum likelihood estimate of the parameters of the model that gives high probability to words that appear in a corpus

From Sivic et al

Obtaining Visual Words

- SIFT descriptors extracted from ellipitical shape adaptation about an interest point and maximally stable extremal regions
- SIFT has all the nice properties ☺
- The SIFT descriptors are then vector quantized (k-means) into visual words
- Total vocabulary size = 2237 words

Doublets of Visual Words



- Black Ellipse represents the visual word whose doublets we want to estimate, ellipses that are red and green are candidate neighbors
- The large red ellipse significantly overlaps with the black ellipse and is discarded
- Likewise, the smaller red ellipse is 'too small' compared to the black ellipse and is discarded
- The Green ellipses are returned as doublets for the black ellipse

From Sivic et al

Model Learning and Baseline Method

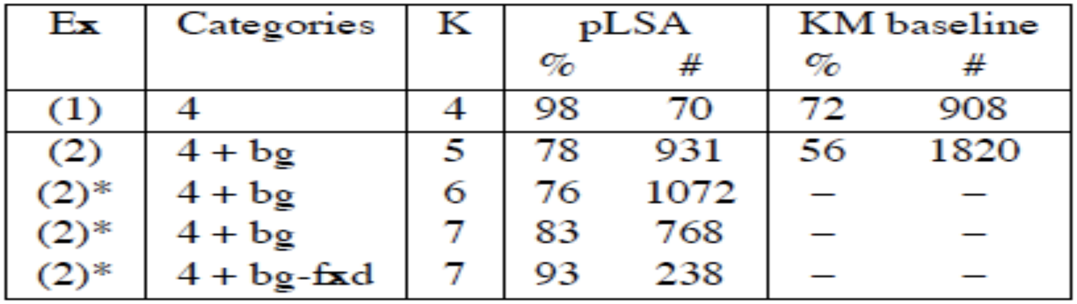
- EM algorithm for pLSA- converges in 40-100K iterations
- For the baseline method k-means was employed on the same features of the word frequency vectors for each image

Experiments and Datasets

- Three experiments:
  1. Topic discovery – categories are discovered by pLSA clustering on all available images
  2. Classification of unseen images – topics on one set of images are learnt to determine the topics in another set
  3. Object detection to determine the location and approximate segmentation of the objects
- Dataset – Caltech 101 (5 categories) and MIT

Topic Discovery Experiment

- Case 1 – Images of 4 object categories with cluttered background



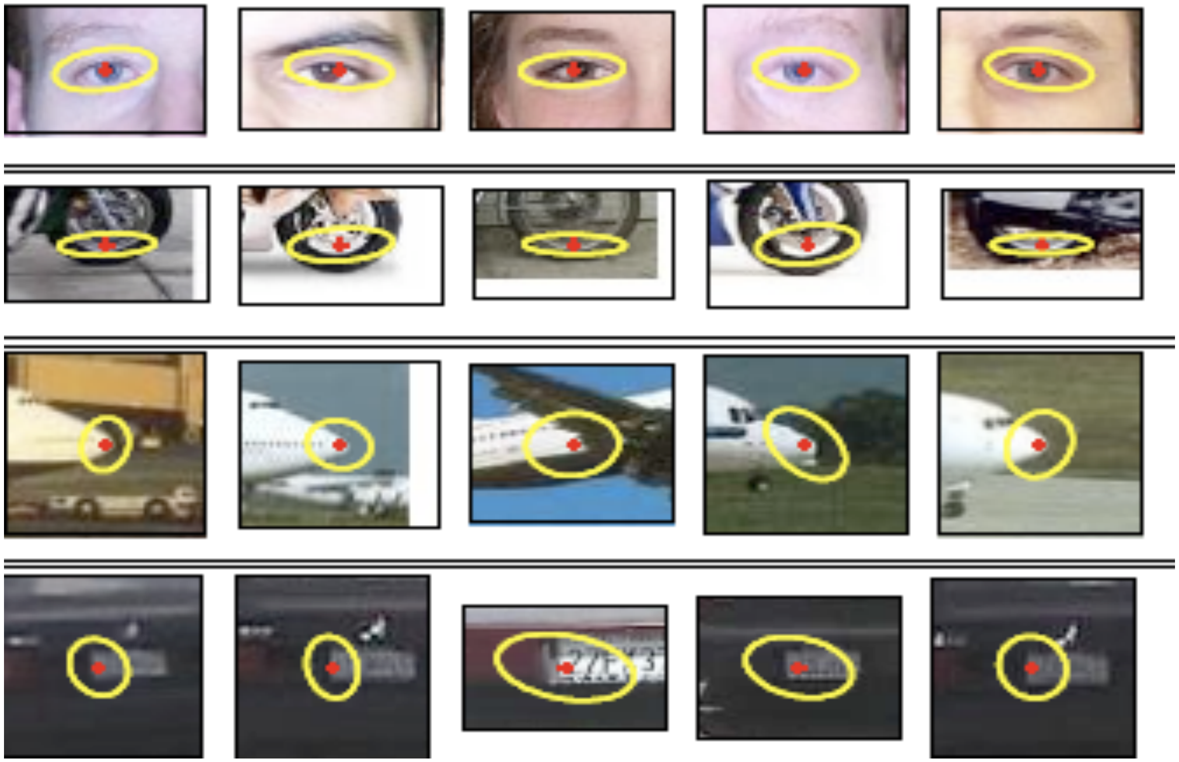
Ex	Categories	K	pLSA		KM baseline	
			%	#	%	#
(1)	4	4	98	70	72	908
(2)	4 + bg	5	78	931	56	1820
(2)*	4 + bg	6	76	1072	–	–
(2)*	4 + bg	7	83	768	–	–
(2)*	4 + bg-fxd	7	93	238	–	–

- When number of topics K = 4, 98% of the 4 different categories are accurately discovered
- K = 5 splits the car dataset into twp subtopics as the data consists of sets of many repeated images of the same car
- K = 6, splits the motorbike data into sets with plain and cluttered background
- K = 7 and 8, discovers two more sub-groups of the car data containing again other repeated images of the same/similar cars.

From Sivic et al

Most probable visual words

- Visual words with high topic specific probability - $P(w_i | z_k)$



From Sivic et al

Topic Discovery Experiment - Case 2, with Background Topics

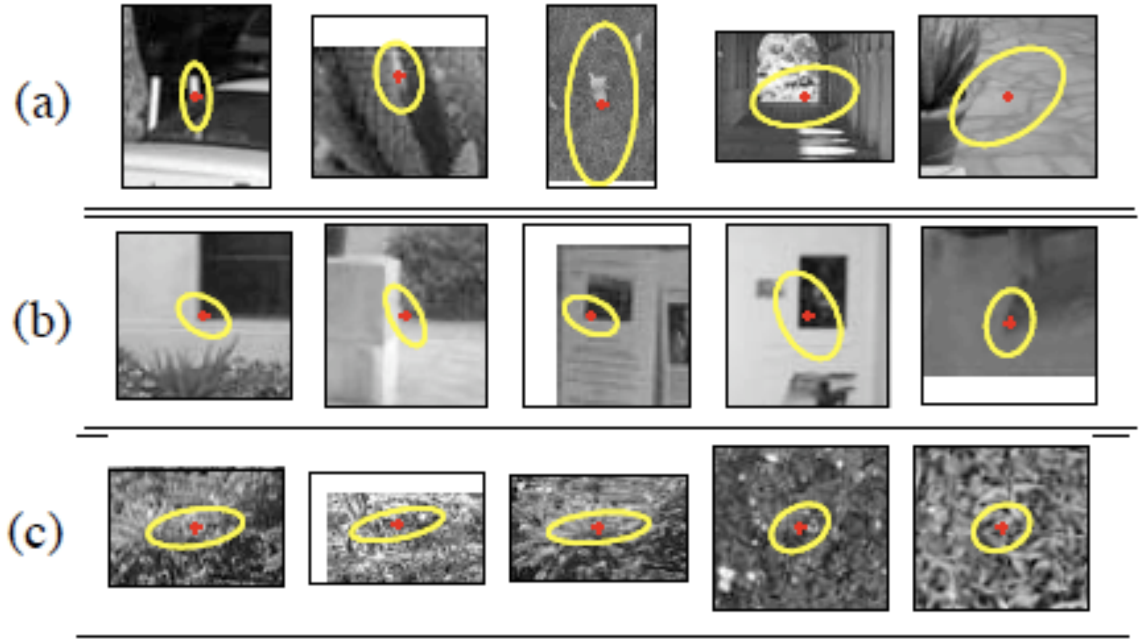


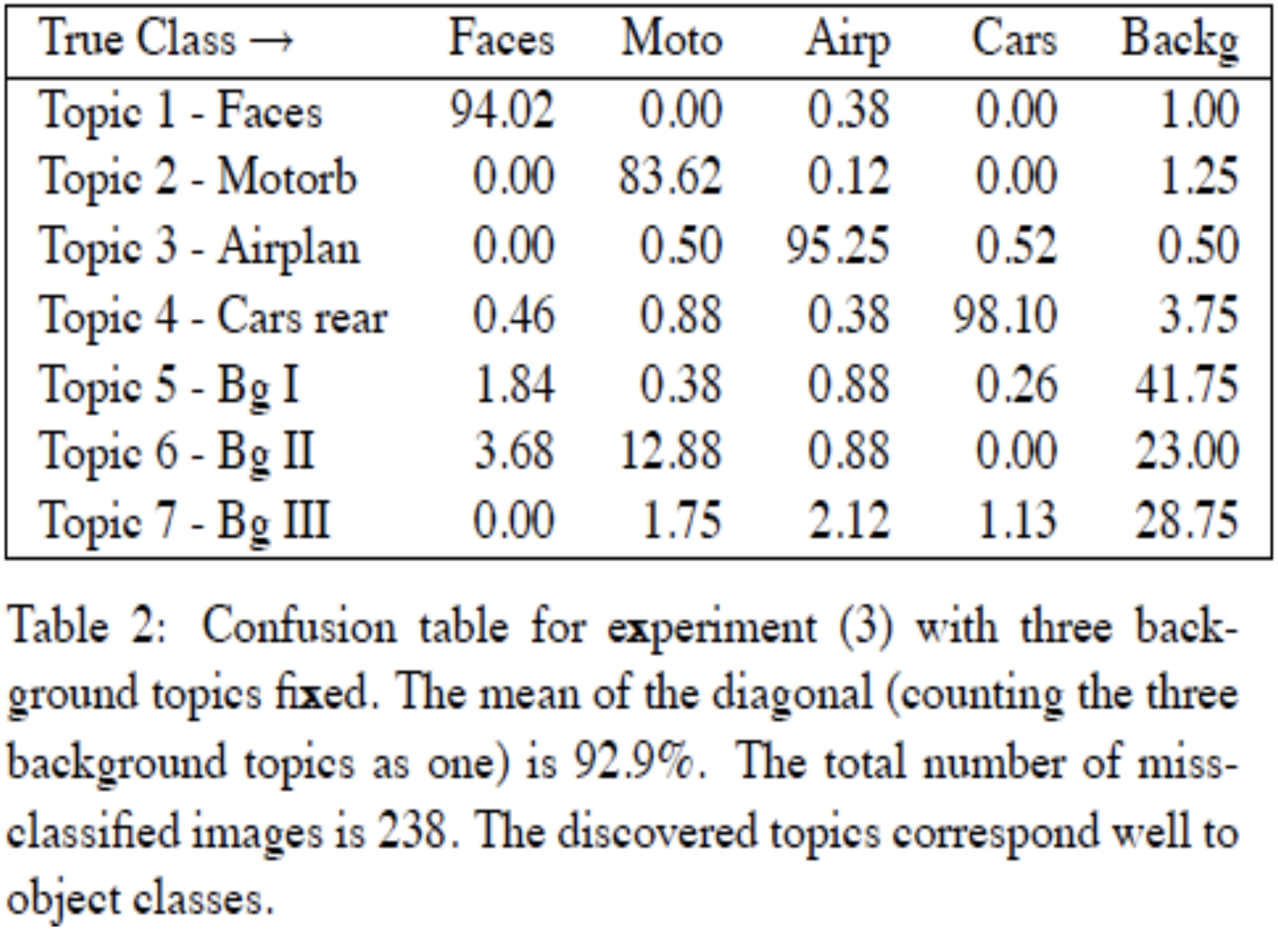
Figure 5: The most likely words (shown by 5 examples in a row) for the three background topics learned in experiment (2): (a) Background I, mainly local feature-like structure (b) Background II, mainly corners and edges coming from the office/building scenes, (c) Background III, mainly textured regions like grass and trees. For topic numbers refer to figure 6(c).

From Sivic et al

Classifying New Images Experiment

- P(w|z) – topic specific distributions are learned from a separate set of training images
- When observing a new, previously unseen test image, the document specific mixing coefficients P(z|test) are computed
- Achieved by EM with only coefficients P(z|test) updated in each M-step and the learned P(w|z) are kept fixed

Classification Results

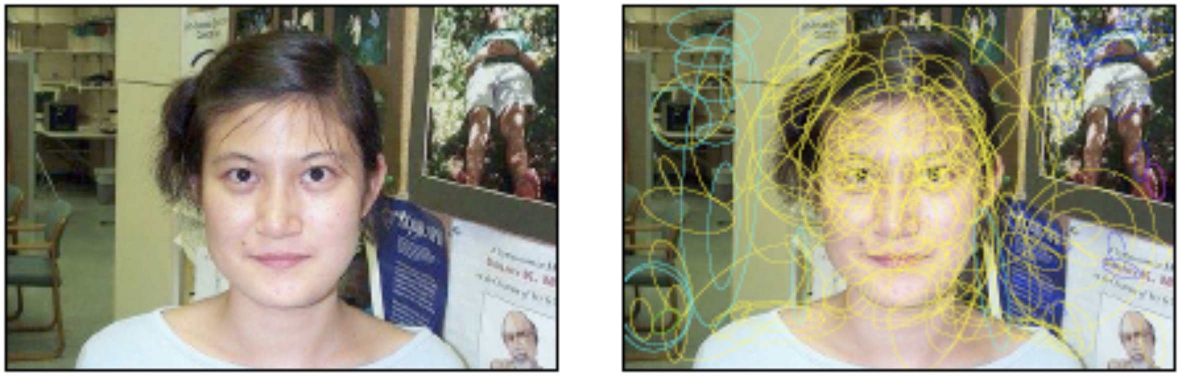


True Class →	Faces	Moto	Airp	Cars	Backg
Topic 1 - Faces	94.02	0.00	0.38	0.00	1.00
Topic 2 - Motorb	0.00	83.62	0.12	0.00	1.25
Topic 3 - Airplan	0.00	0.50	95.25	0.52	0.50
Topic 4 - Cars rear	0.46	0.88	0.38	98.10	3.75
Topic 5 - Bg I	1.84	0.38	0.88	0.26	41.75
Topic 6 - Bg II	3.68	12.88	0.88	0.00	23.00
Topic 7 - Bg III	0.00	1.75	2.12	1.13	28.75

Table 2: Confusion table for experiment (3) with three background topics fixed. The mean of the diagonal (counting the three background topics as one) is 92.9%. The total number of missclassified images is 238. The discovered topics correspond well to object classes.

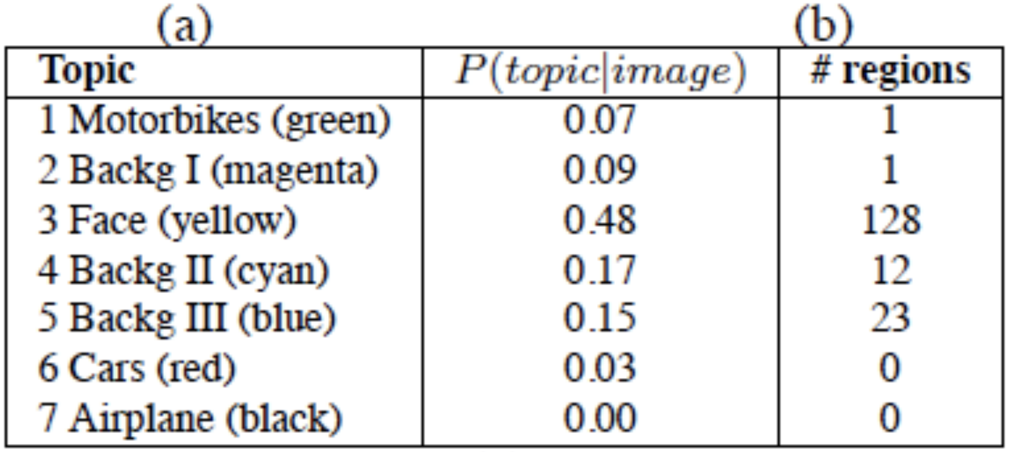
From Sivic et al

Segmentation Results from the Posteriors



(a)

(b)



Topic	$P(topic\|image)$	# regions
1 Motorbikes (green)	0.07	1
2 Backg I (magenta)	0.09	1
3 Face (yellow)	0.48	128
4 Backg II (cyan)	0.17	12
5 Backg III (blue)	0.15	23
6 Cars (red)	0.03	0
7 Airplane (black)	0.00	0

(c)

Mixing coefficients

From Sivic et al

Segmentation Results – with Doublets

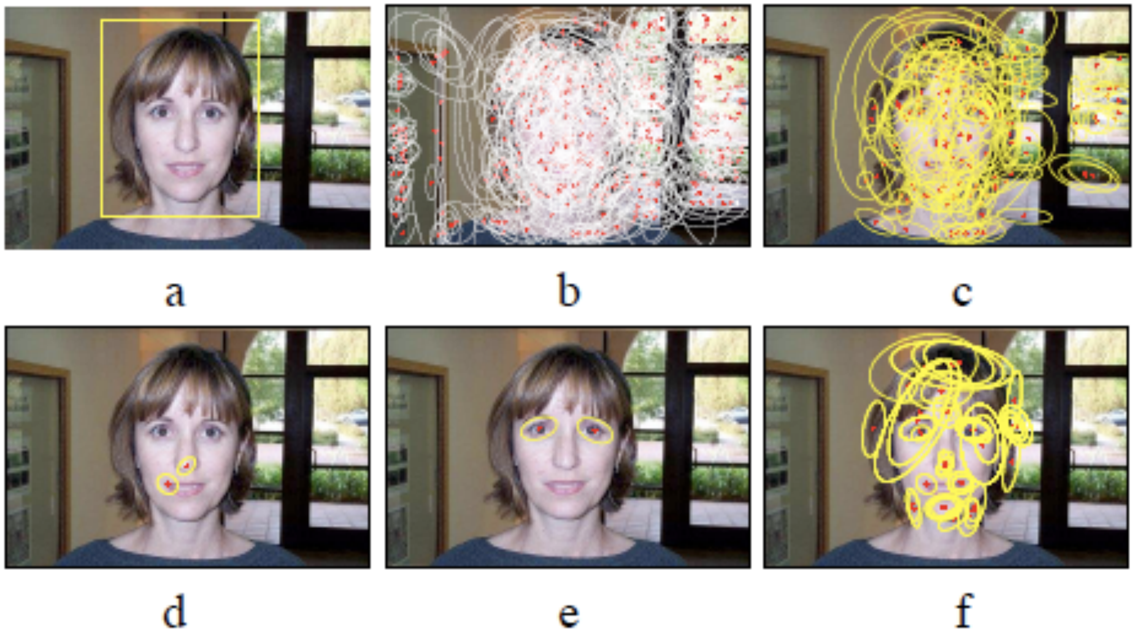
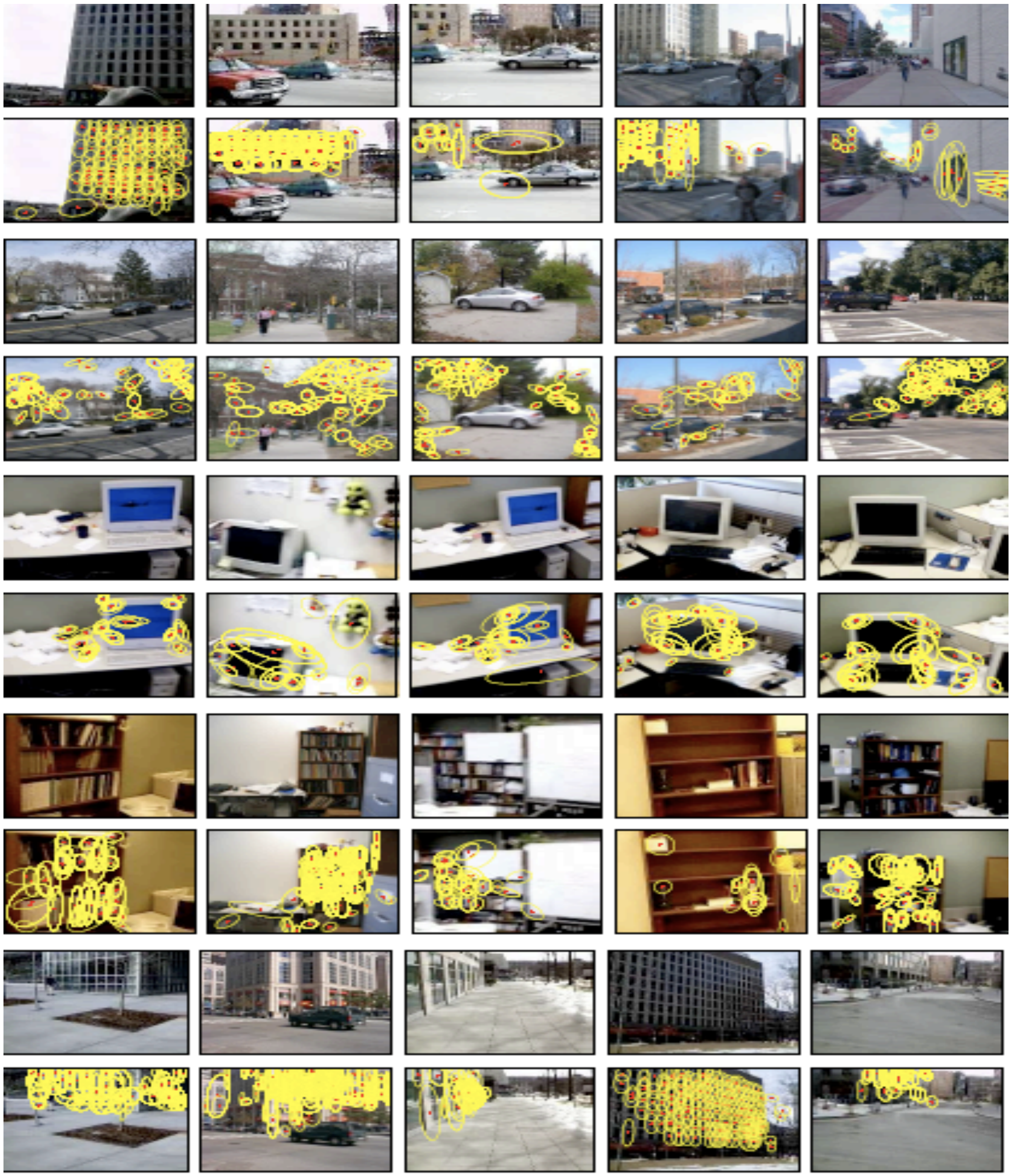
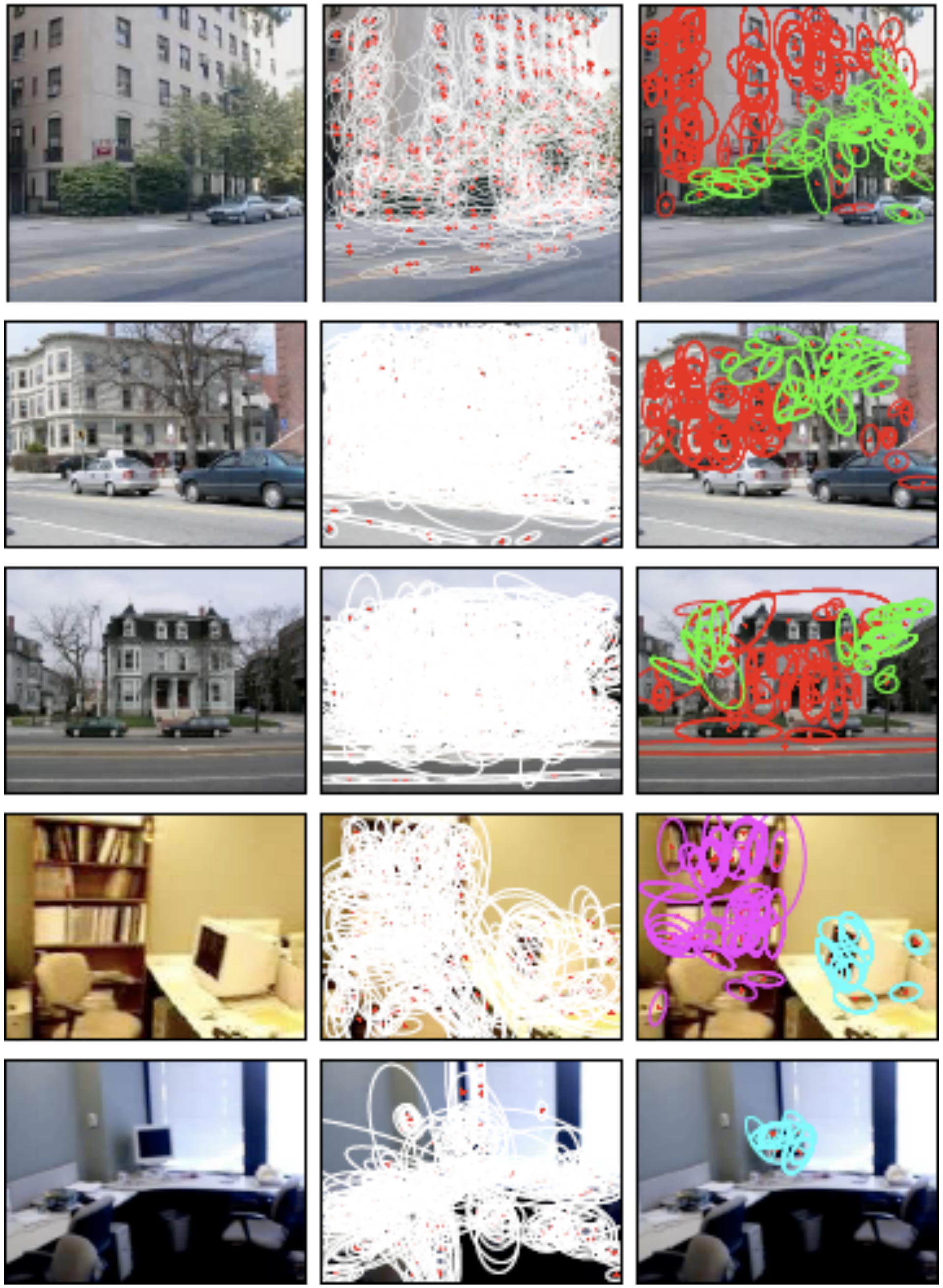


Figure 7: **Improving object segmentation.** (a) The original frame with ground truth bounding box. (b) All 601 detected elliptical regions superimposed on the image. (c) Segmentation obtained by pLSA on single regions. (d) and (e) show examples of 'doublets' —locally co-occurring regions. (f) Segmentation obtained using doublets. Note the extra regions on the right-hand side background of (c) are removed in (f).

From Sivic et al



MIT dataset results
from Sivic et al



MIT dataset results
from Sivic et al

Conclusion

- Visual object categories can be discovered using an unsupervised approach
- Tasks such as segmentation can be performed using simple bag of features combined with statistical models
- However, bag of features does not take into account semantic context – Can models from statistical text literature handle context?
- Are models from text really appropriate? – Unlike text, Images have a strong spatial structure

Moving on….

Unsupervised Discovery of Action Classes

By Wang et al.

Basic Idea

- Cluster images that depict similar actions together and label these clusters (action classes)
- Assign a new image to an action class based on its distance from the centroids of the clusters

Approach

- Human shape as a cue to determine the action
- The similarity measure for clustering has to take into account the deformations when comparing two images of different people performing different actions

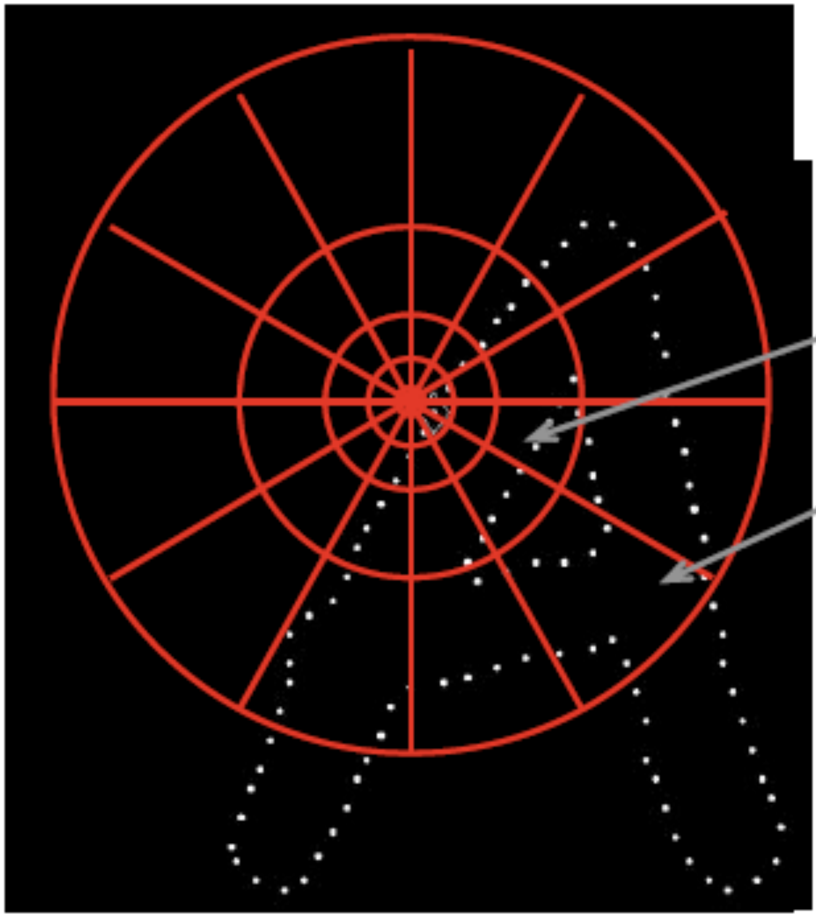
Similarity Measure

- Requirement - yield a high value on a pair of images when similar poses are depicted and a low value on dissimilar poses
- Spectral Clustering- Affinity Matrix W (n x n) where W_{ij} = affinity between images i and j, $W_{ij} = \exp(-(d_{ij}^2 + d_{ji}^2)/2)$

Deformable Template Matching

- Algorithm to match actions in images by measuring affinity (similarity)
- Posed as an Integer Linear Programming Problem
- Computationally not feasible when required to compute n x n affinity measures (required to do so as the affinity measure is not symmetric)
- A fast pruning algorithm based on shape contexts is used to address this issue

Fast Pruning using Representative Shape Contexts



Count the number of points inside each bin, e.g.:

Count = 4

⋮

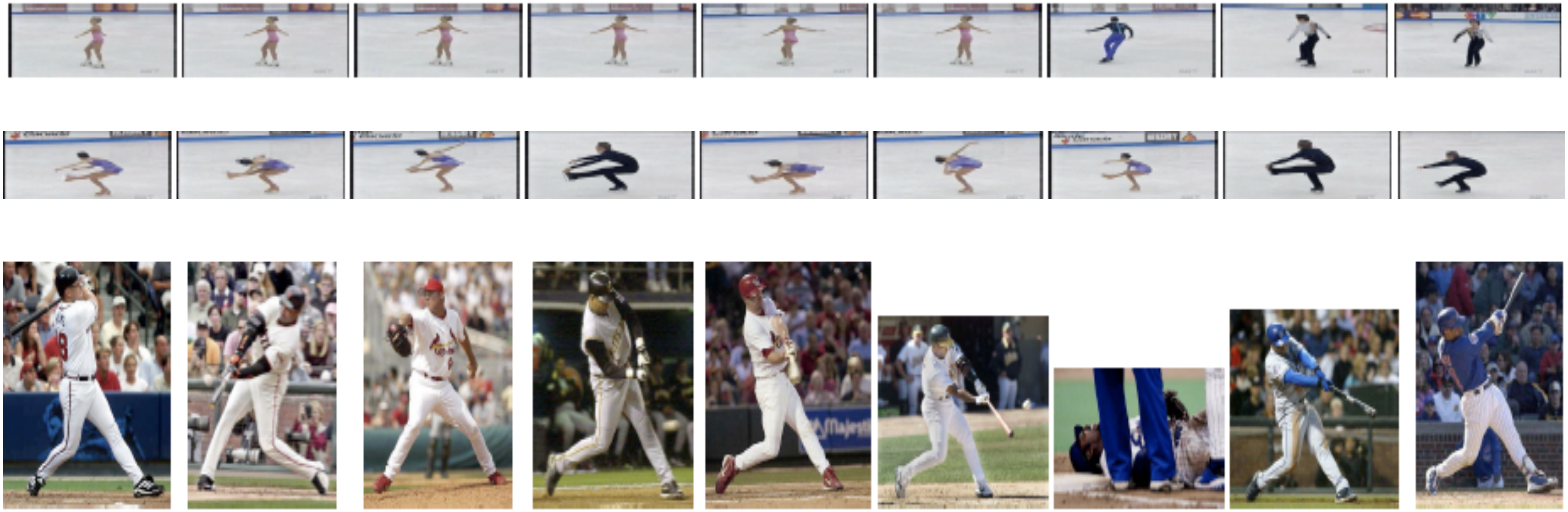
Count = 10

Log-polar binning: more precision for nearby points, more flexibility for farther points.

Local descriptor

Slide From Grauman,
Original Source from Belongie, Malik and Puzicham ICCV 2001

Pruning results



- A criticism of the RSC pruning algorithm is that it ignores spatial structure and this leads to some errors in matching.
- The deformable template matching algorithms tries to rectify these errors
- For the pruning algorithm, a shortlist of length 50 was used.

Back to the Deformable Template Matching Algorithm

- Goal : To find the optimal matching of sample points from one image to another
- The distance between two images is the minimum of the following objective function over all possible candidate matches:

$$\min_{\mathbf{f}} E : \sum_{\mathbf{s}\in S} c(\mathbf{s}, \mathbf{f}_{\mathbf{s}}) + \sum_{\{\mathbf{p},\mathbf{q}\}\in\mathcal{N}} \lambda_{\mathbf{p},\mathbf{q}} ||(\mathbf{f}_{\mathbf{p}} - \mathbf{p}) - (\mathbf{f}_{\mathbf{q}} - \mathbf{p})||$$

From Wang et al

More on the Objective Function

$$\min_{\mathbf{f}} E : \sum_{\mathbf{s} \in S} c(\mathbf{s}, \mathbf{f_s}) + \sum_{\{\mathbf{p},\mathbf{q}\} \in \mathcal{N}} \lambda_{\mathbf{p},\mathbf{q}} ||(\mathbf{f_p} - \mathbf{p}) - (\mathbf{f_q} - \mathbf{p})||$$

- The objective function clearly shows that a set of points {f} is sought such that:
  - the cost of matching points {f} with the {s} is minimum
  - the points in the neighborhood of every point in {f} matches closely to the points in the neighborhood of every point in {s}

Cost Function

- The cost function c(s,f) is computed as follows:
  - Convert the edge maps to gray scale images by taking the distance transform
  - Consider small neighborhood of 9x9 pixels around each feature point in the two images
  - Compute the normalized sum of absolute differences between the 2 patches to get c(s,f)

Linearizing the Objective Function

$$\min_{\mathbf{f}} E : \sum_{\mathbf{s}\in S} c(\mathbf{s}, \mathbf{f}_{\mathbf{s}}) + \sum_{\{\mathbf{p},\mathbf{q}\}\in\mathcal{N}} \lambda_{\mathbf{p},\mathbf{q}} ||(\mathbf{f}_{\mathbf{p}} - \mathbf{p}) - (\mathbf{f}_{\mathbf{q}} - \mathbf{p})||$$

- The first term in the objective function measures similarity between the 2 sets of points, while the second term is a mesure of relative spatial deformation
- The objective function is neither linear nor convex – a hard optimization problem

Linearizing the Objective Function contd.

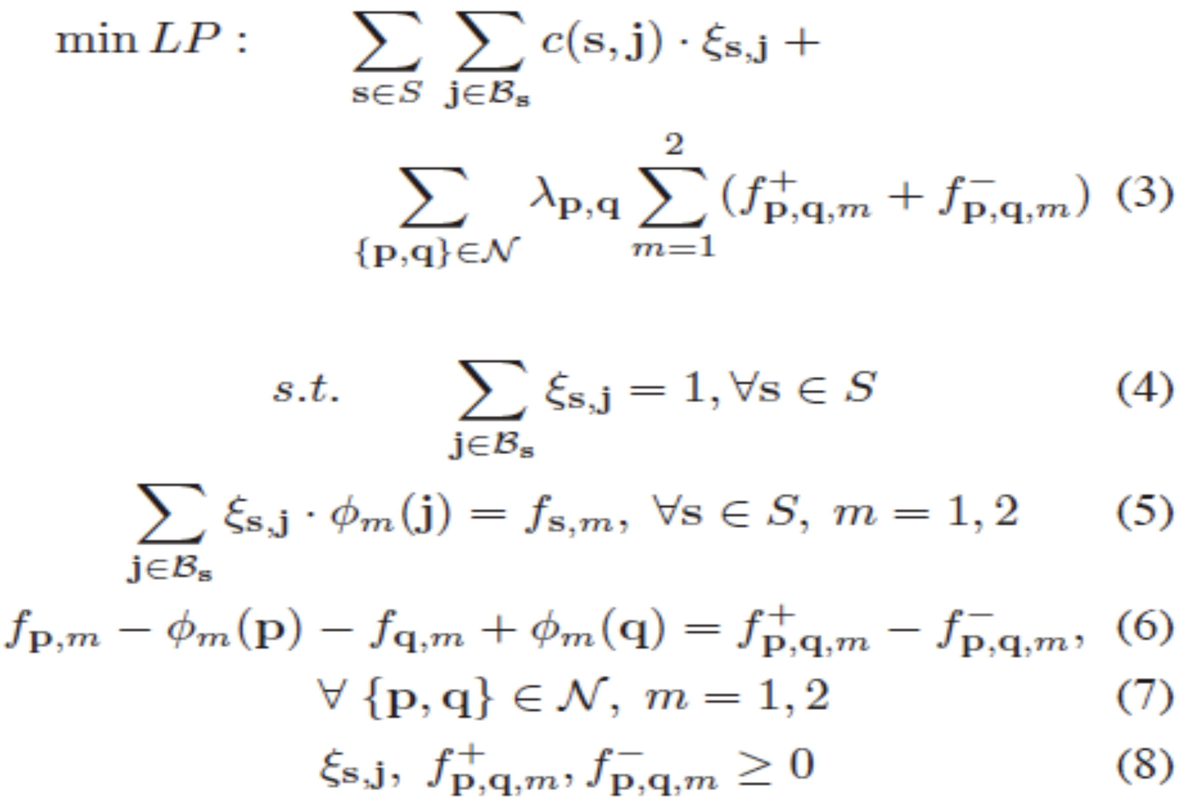
- $\min_{\mathbf{f}} E : \sum_{\mathbf{s} \in S} c(\mathbf{s}, \mathbf{f_s}) + \sum_{\{\mathbf{p},\mathbf{q}\} \in \mathcal{N}} \lambda_{\mathbf{p},\mathbf{q}} ||(\mathbf{f_p} - \mathbf{p}) - (\mathbf{f_q} - \mathbf{p})||$
- Lets focus on the 2nd term first (it's easy!)
- Recall that the L1 norm is just the absolute value (or to be formal the sum of absolute values of all elements in the vector)
- An absolute value can always be replaced by the difference of two non-negative values

Linearizing the objective function

$$\min_{\mathbf{f}} E : \sum_{\mathbf{s}\in S} c(\mathbf{s}, \mathbf{f_s}) + \sum_{\{\mathbf{p},\mathbf{q}\}\in\mathcal{N}} \lambda_{\mathbf{p},\mathbf{q}} ||(\mathbf{f_p} - \mathbf{p}) - (\mathbf{f_q} - \mathbf{p})||$$

- To linearize the first term:
  - For the set of coordinates f, find its basis vectors
  - Represent each fs as a linear combination of the basis vectors
  - Approximate the cost function as the linear combination of the cost between the sample point s and the basis vectors

The final LP



$$\min LP: \quad \sum_{\mathbf{s}\in S}\sum_{\mathbf{j}\in\mathcal{B}_\mathbf{s}} c(\mathbf{s},\mathbf{j})\cdot\xi_{\mathbf{s},\mathbf{j}} + \sum_{\{\mathbf{p},\mathbf{q}\}\in\mathcal{N}} \lambda_{\mathbf{p},\mathbf{q}} \sum_{m=1}^{2} (f^{+}_{\mathbf{p},\mathbf{q},m} + f^{-}_{\mathbf{p},\mathbf{q},m}) \quad (3)$$

$$s.t. \quad \sum_{\mathbf{j}\in\mathcal{B}_\mathbf{s}} \xi_{\mathbf{s},\mathbf{j}} = 1, \forall \mathbf{s}\in S \quad (4)$$

$$\sum_{\mathbf{j}\in\mathcal{B}_\mathbf{s}} \xi_{\mathbf{s},\mathbf{j}}\cdot\phi_m(\mathbf{j}) = f_{\mathbf{s},m},\ \forall\mathbf{s}\in S,\ m=1,2 \quad (5)$$

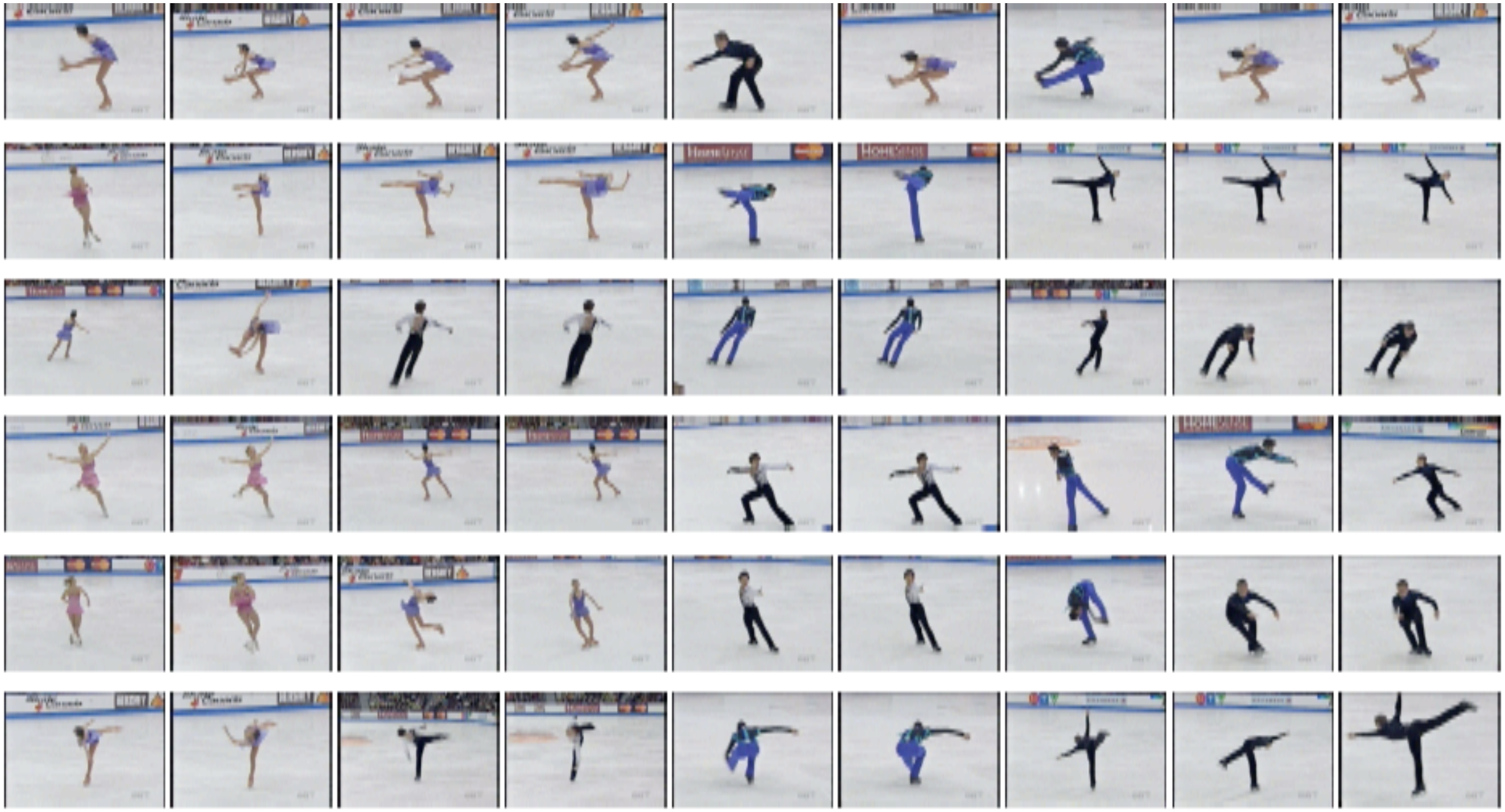
$$f_{\mathbf{p},m} - \phi_m(\mathbf{p}) - f_{\mathbf{q},m} + \phi_m(\mathbf{q}) = f^{+}_{\mathbf{p},\mathbf{q},m} - f^{-}_{\mathbf{p},\mathbf{q},m}, \quad (6)$$

$$\forall\,\{\mathbf{p},\mathbf{q}\}\in\mathcal{N},\ m=1,2 \quad (7)$$

$$\xi_{\mathbf{s},\mathbf{j}},\ f^{+}_{\mathbf{p},\mathbf{q},m}, f^{-}_{\mathbf{p},\mathbf{q},m} \geq 0 \quad (8)$$

The LP can be solved using the Simplex method

Clustering Results



- Each row corresponds to a cluster
- Number of clusters fixed between 100 and 200

From Wang et al

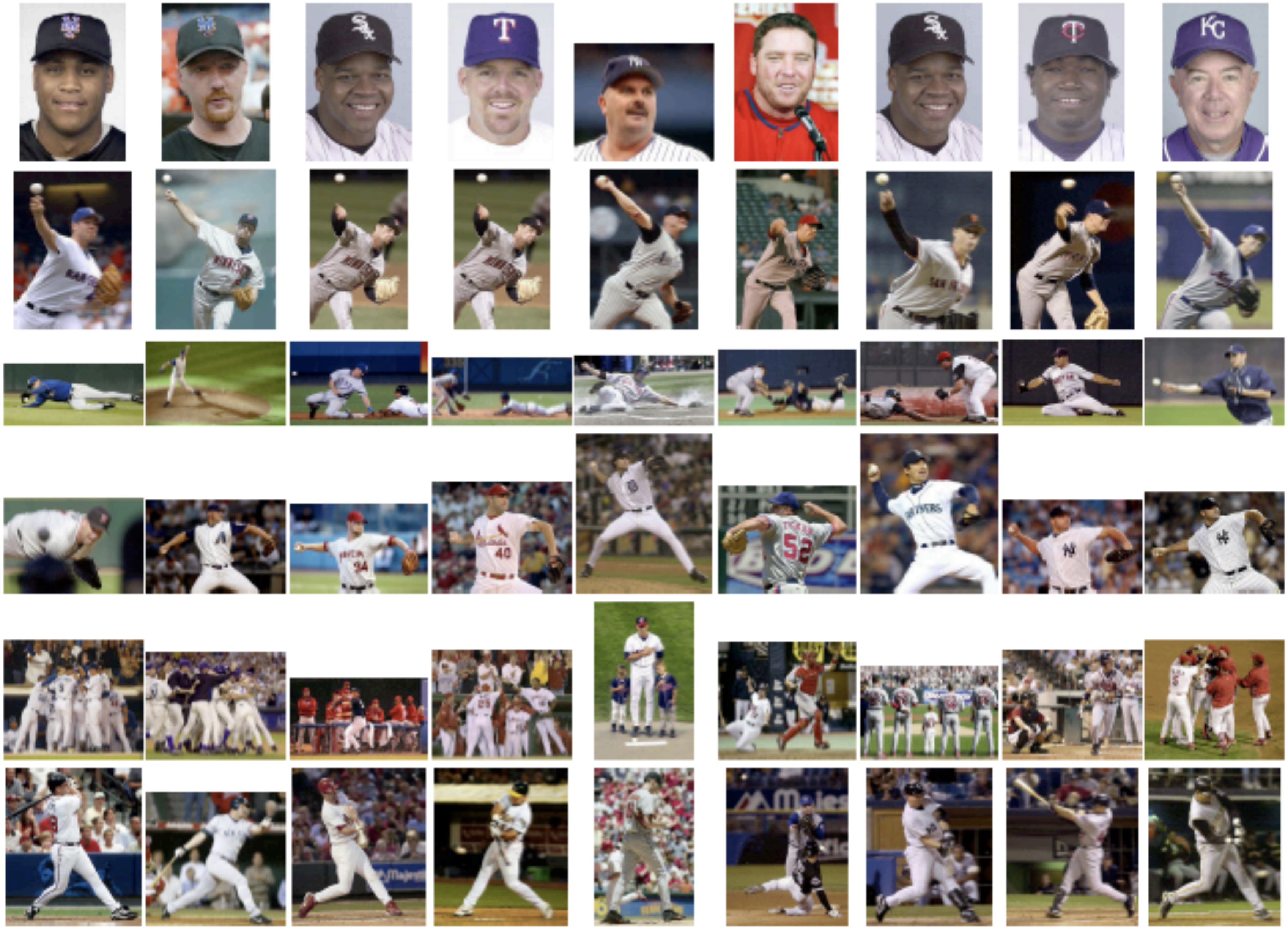


Figure 5. Examples of clusters in baseball images. Each row corresponds to a cluster.

From Wang et al

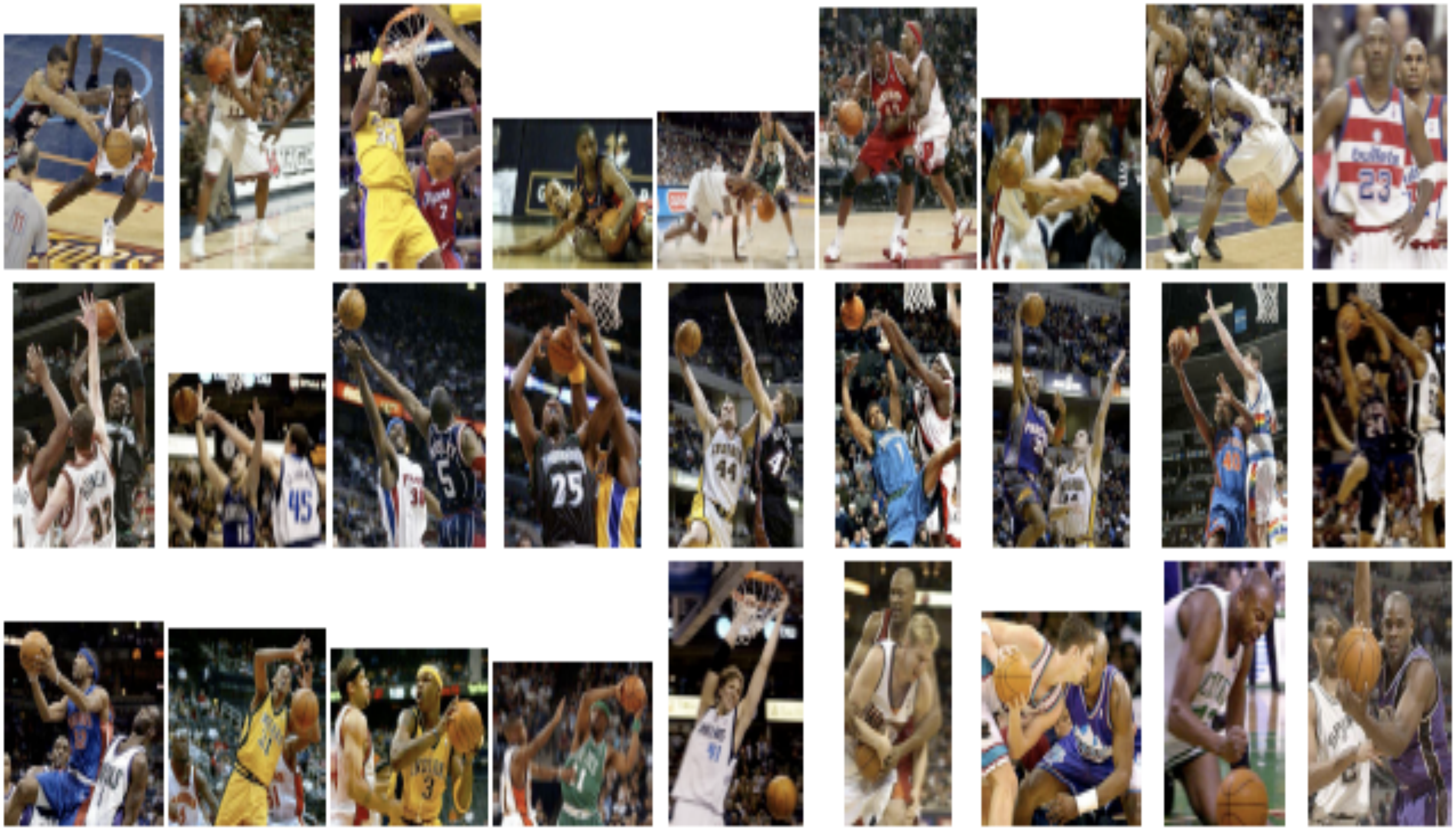
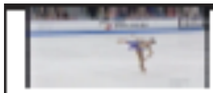
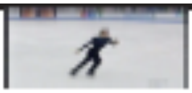


Figure 6. Examples of clusters in basketball images. Each row corresponds to a cluster.

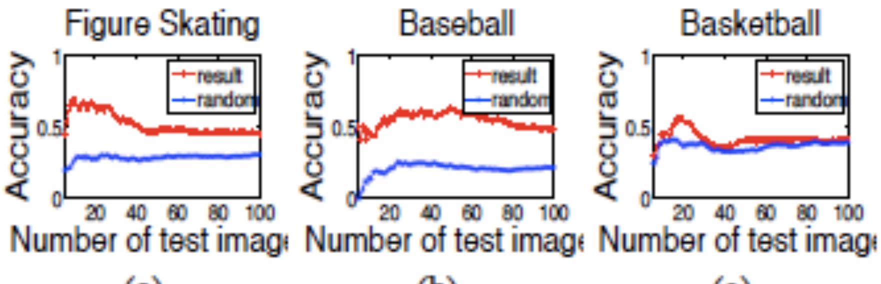
From Wang et al

Image Labeling Experiment

					
camel spin, leg to left of image	skates with arms down	sit spin, leg to right if image	sin spin, leg to left of image	skates on one leg	sit spin, leg to left of image
face close-up picture	right-handed pitcher throws	right-handed pitcher throws	face close-up picture	right-handed pitcher throws	right-handed pitcher throws
a player drives past another	a player drives past another	a player has his shot blocked	a player goes for a lay-up against a defender	a player goes for a lay-up against a defender	a player goes for a lay-up against a defender

From Wang et al

Image Labeling Experiment



(a) (b) (c)

Figure 8. Quantitative results on image labeling.

Accuracy of automatic labeling as evaluated subjectively by 4 naïve observers

From Wang et al

Conclusions

- An unsupervised recipe for action class recognition
- Spectral clustering vs. Topic discovery – comparing pairs of images for deciding on common actions vs. learning the visual codewords associated with each action class
- A few points to note –
  - The action classes considered are substantially different from one another
  - What happens when actions depicted on still images are similar – for e.g., brisk walking vs. jogging / sleeping vs. just lying down! Would there be ties ?

Moving On…

Detecting Irregularities in Images and in Video

- Boiman and Irani

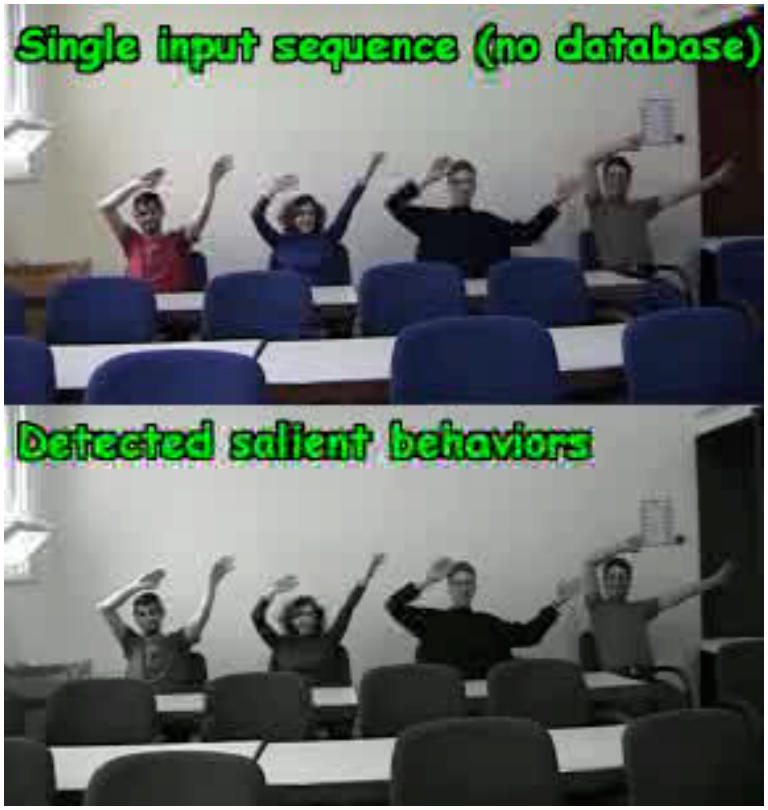
Recap

- We just saw-
  - Unsupervised approach to learn codewords specific to object categories to detect objects in images
  - Spectral clustering to compare images and infer about commonality in actions
- We will now see –
  - An unsupervised approach that makes use of a database of ‘evidence’ to infer what is regular and what is ‘not’ in actions!

Detecting Irregularities

From - http://www.wisdom.weizmann.ac.il/~vision/Irregularities.html#Video%20Examples

Spatio Temporal Saliency

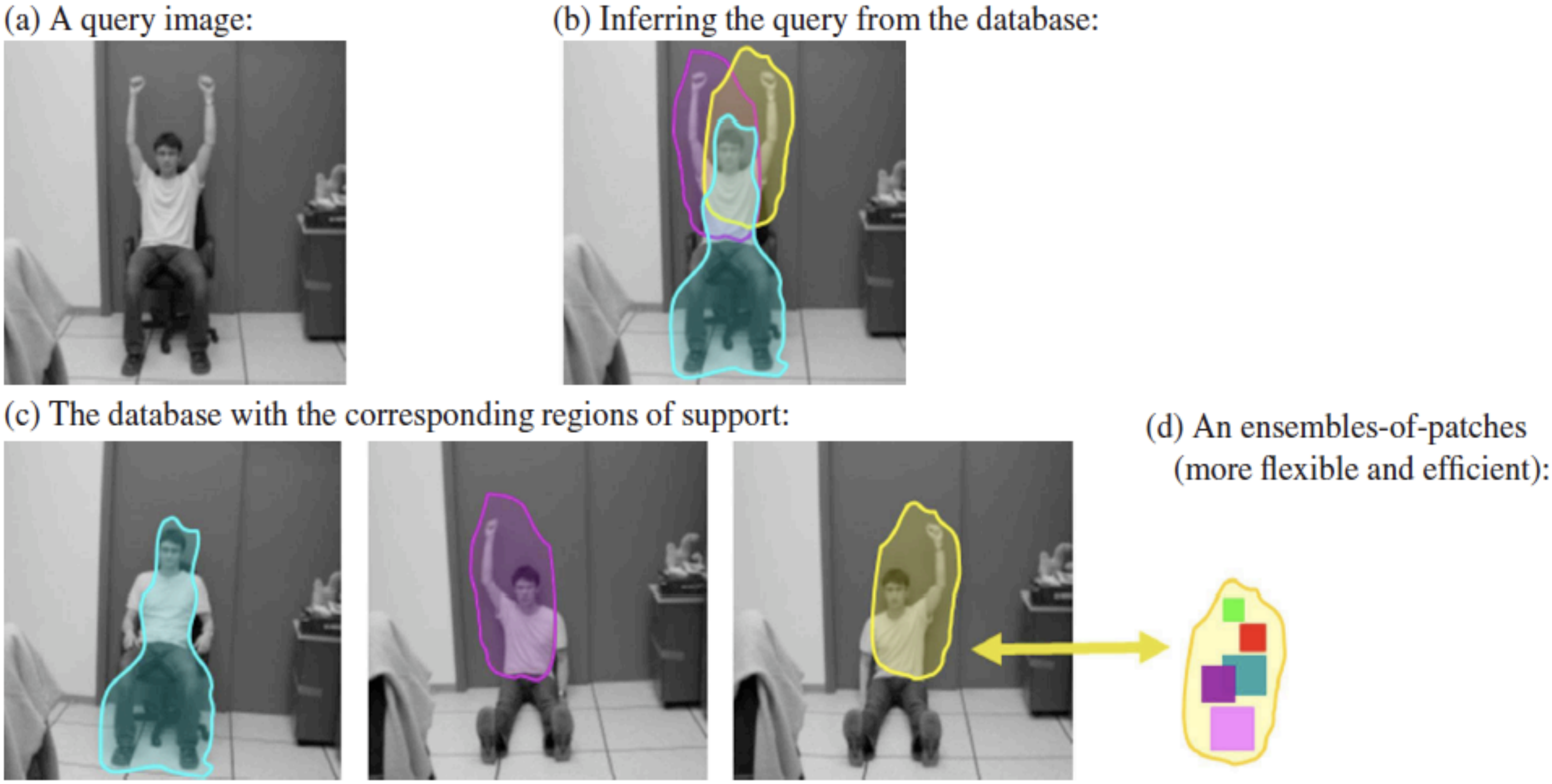

Single input sequence (no database)
Detected salient behaviors

From - http://www.wisdom.weizmann.ac.il/~vision/
Irregularities.html#Video%20Examples

Approach

- Analyze patches in an image and compare their relative location and appearance with ensemble of patches in a ‘database’
- Regions in the image that have a contiguous region of support in the database are considered likely
- A graphical model to capture the dependencies between the query patch and a patch from the ensemble

Inference by Composition



(a) A query image:

(b) Inferring the query from the database:

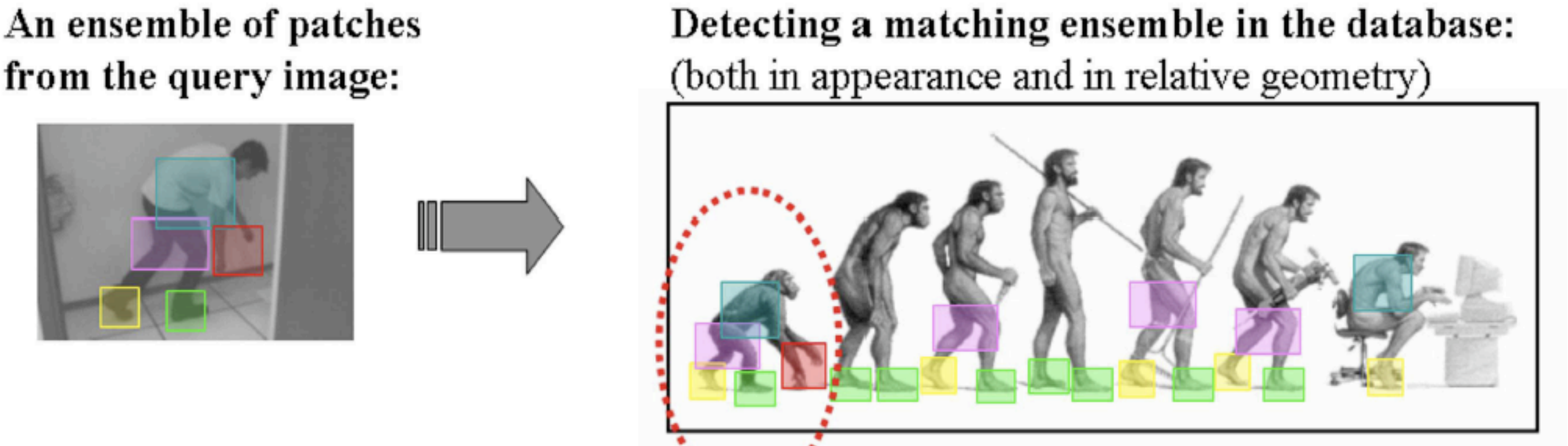
(c) The database with the corresponding regions of support:

(d) An ensembles-of-patches (more flexible and efficient):

Given a new image, can each region be explained by a large enough contiguous region of support by a database of evidence ?

From Boiman and Irani

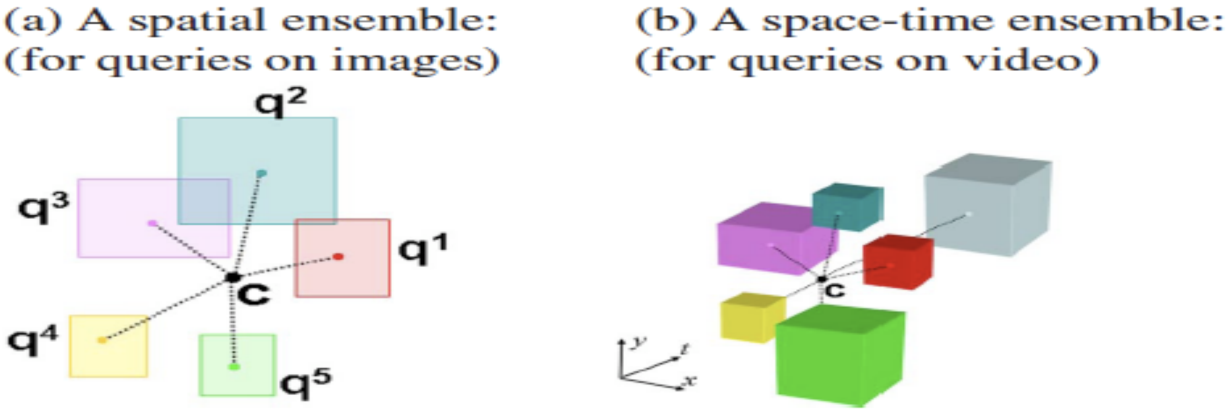
Ensemble of Patches



An ensemble of patches from the query image:

Detecting a matching ensemble in the database: (both in appearance and in relative geometry)

Figure 2. Detecting a matching ensemble of patches.



(a) A spatial ensemble: (for queries on images)

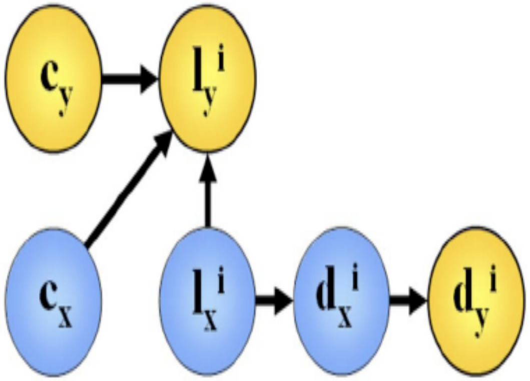
(b) A space-time ensemble: (for queries on video)

From Boiman and Irani

Ensemble of Patches

- Local patches at multiple scales to allow for local, non-rigid deformations
- Patches of similar properties (appearance, or behavior) and having similar geometric configuration are searched
- An ensemble typically consists of hundreds of patches at multiple scales
- Patch defined in terms of a simple descriptor vector and absolute location

Statistical Formulation



Goal – to estimate likelihood at every pixel ;The similarity between a pair of ensembles y and x is given by the likelihood:

$$P(c_x, d_x^1, \ldots, l_x^1, \ldots, c_y, d_y^1, \ldots, l_y^1) = \alpha \prod_i P(l_y^i | l_x^i, c_x, c_y) P(d_y^i | d_x^i) P(d_x^i | l_x^i)$$

$$P\left(d_x | l_x\right) = \begin{cases} 1 & (d_x, l_x) \in DB \\ 0 & otherwise \end{cases}$$

$$P\left(d_y^i | d_x^i\right) = \alpha_1 \exp\left(-(d_y^i - d_x^i)^T S_D^{-1} (d_y^i - d_x^i)\right)$$

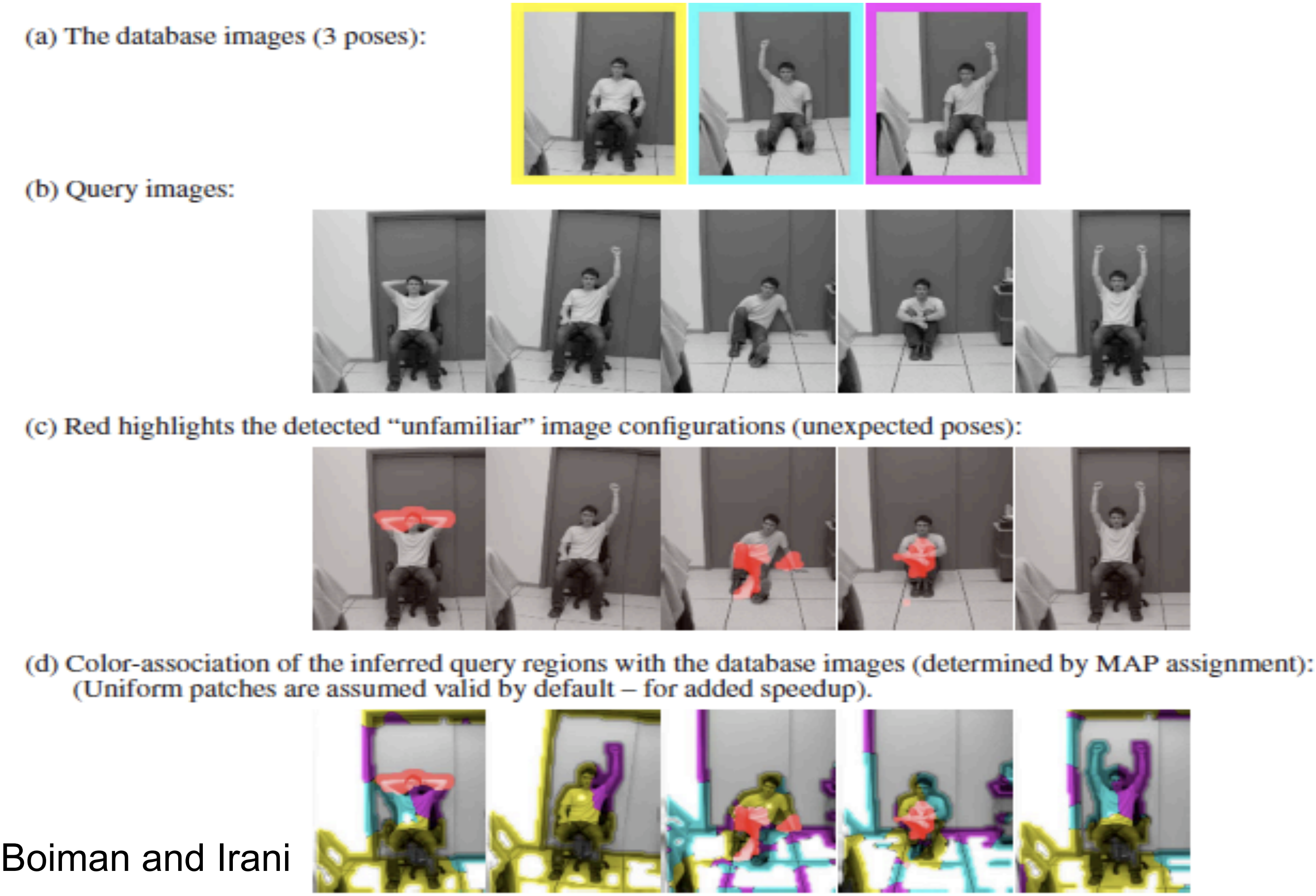
From Boiman and Irani

$$P\left(l_y^i | l_x^i, c_x, c_y\right) = \alpha_2 \cdot \\ \exp\left(-((l_y^i - c_y) - (l_x^i - c_x))^T S_L^{-1} ((l_y^i - c_y) - (l_x^i - c_x))\right)$$

Given an observed ensemble, a set of ensembles from the database that maximizes the MAP probability assignment is sought. This is solved by implementing an efficient message passing algorithm.

Applications

- Detecting Unusual Image Configurations



(a) The database images (3 poses):

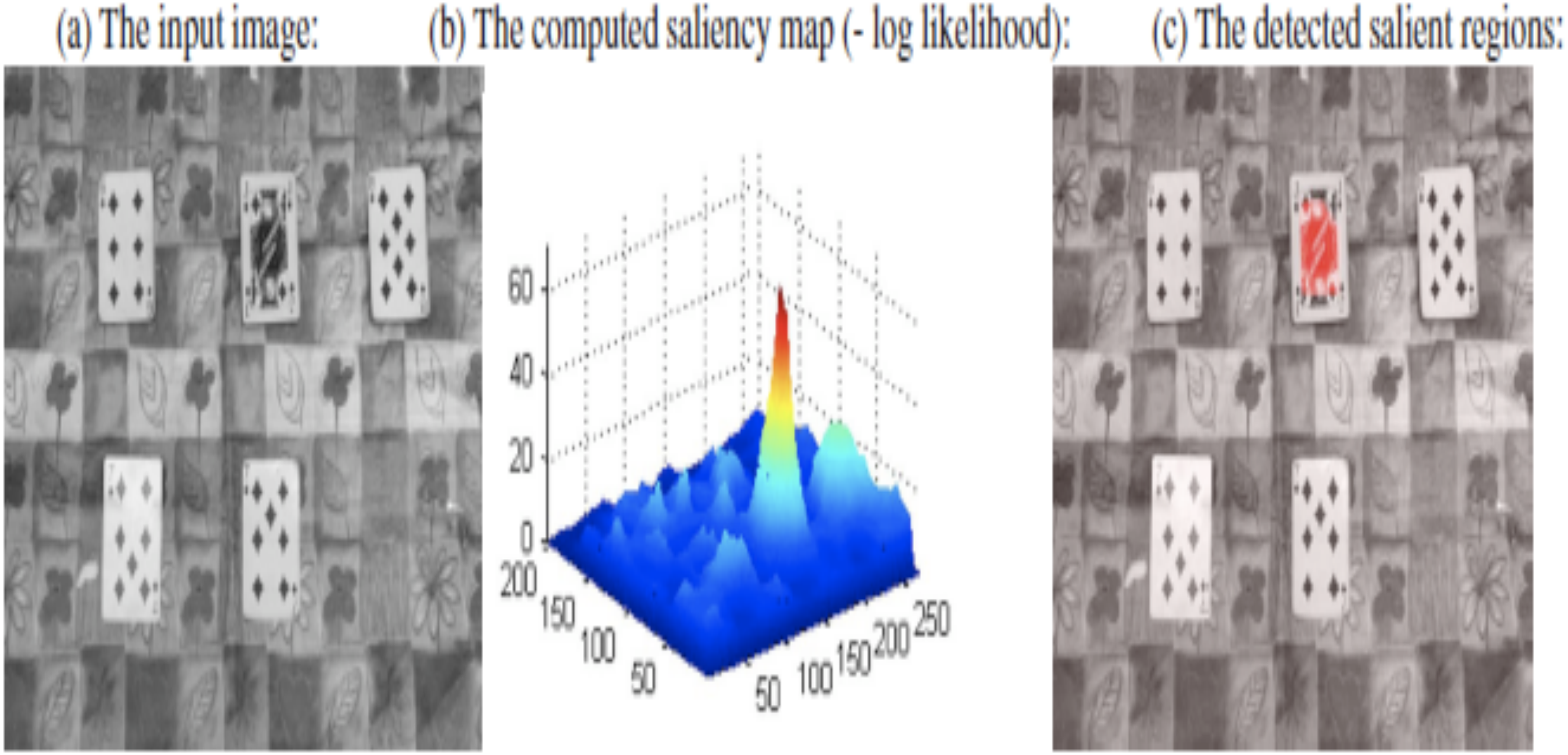
(b) Query images:

(c) Red highlights the detected "unfamiliar" image configurations (unexpected poses):

(d) Color-association of the inferred query regions with the database images (determined by MAP assignment): (Uniform patches are assumed valid by default – for added speedup).

From Boiman and Irani

Spatial Saliency in an Image



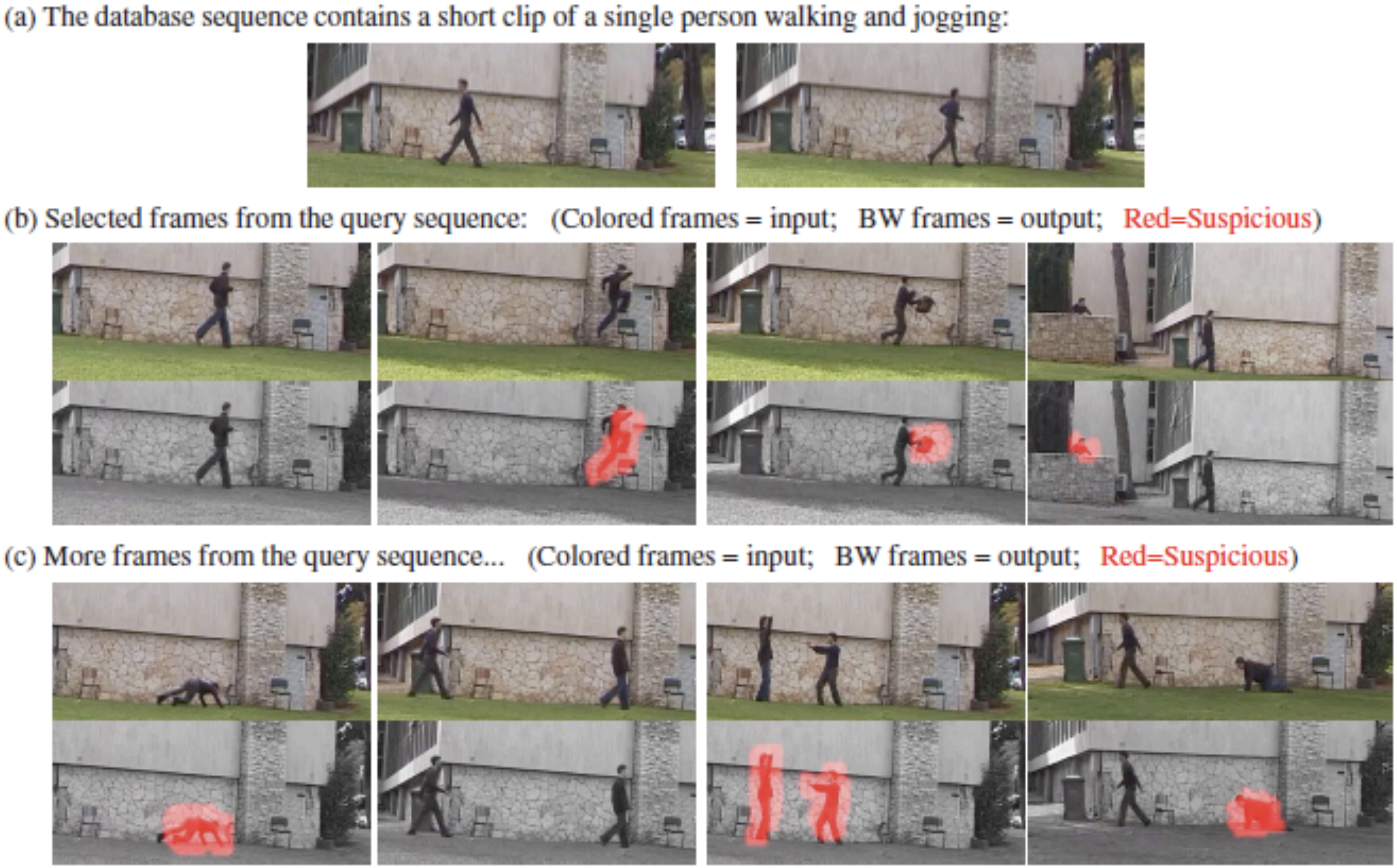
(a) The input image:

(b) The computed saliency map (- log likelihood):

(c) The detected salient regions:

From Boiman and Irani

Detecting Suspicious Behavior



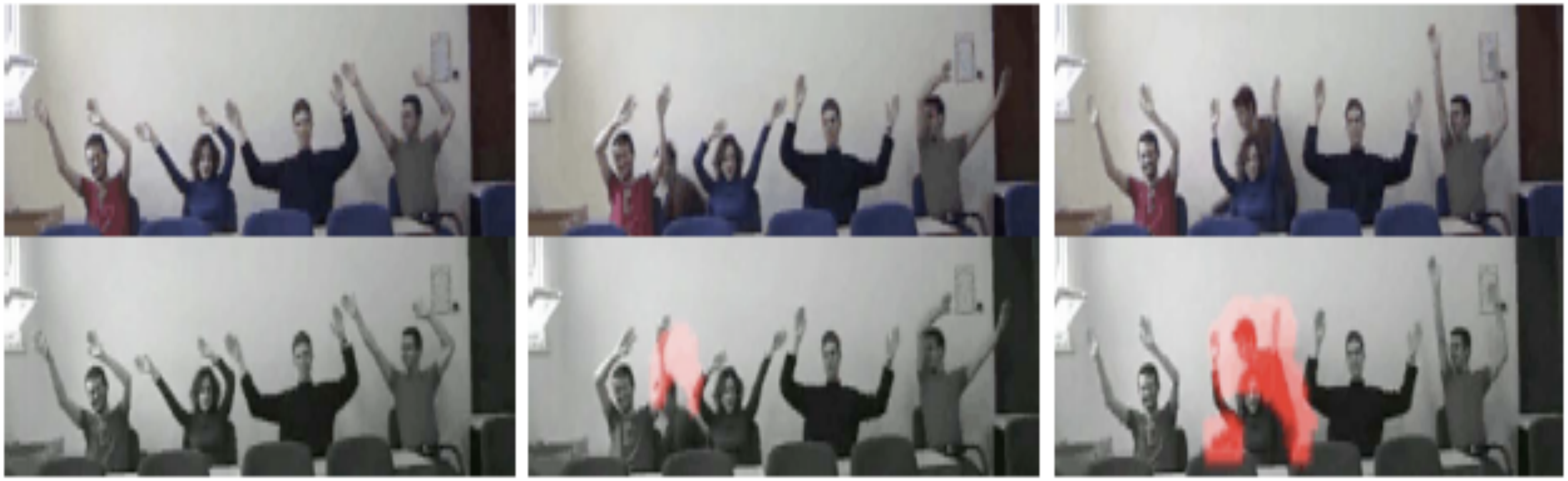
(a) The database sequence contains a short clip of a single person walking and jogging:

(b) Selected frames from the query sequence: (Colored frames = input; BW frames = output; Red=Suspicious)

(c) More frames from the query sequence... (Colored frames = input; BW frames = output; Red=Suspicious)

From Boiman and Irani

Detecting Salient Behaviors



From Boiman and Irani

To Conclude

- We saw applications of unsupervised discovery
- Approaches include
  - Topic models that learn codewords specific to object categories to detect objects in images
  - Spectral clustering to compare pairs of images to infer common action classes
  - Evidence based detection of irregularities

Final Thoughts

- How can we improve on unsupervised techniques ?
  - Semi-supervised ?
  - Active Annotation ?
  - Any other thoughts ?