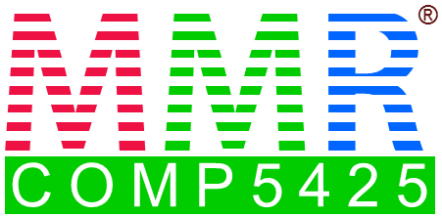


MULTIMEDIA RETRIEVAL

THE UNIVERSITY OF SYDNEY

Week08

Semester 1, 2025

Large Scale Retrieval

- Image/Video Annotation
  - Semantic gap
- Bag of Visual Words model
  - Video Google

Semantic Gap

- Content based retrieval
  - Use low level features
- Human understanding
  - Semantics: objects and meaningful attributes

CBIR: Semantic Gap

Query: “Find me pictures of tiger”

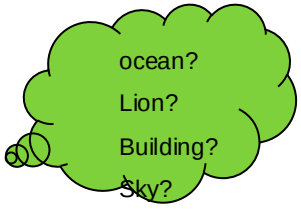
Annotation Task

- Training

tiger cat grass

hippo, bull, mouth, walk flower, coralberry, leaves, plant

- Testing



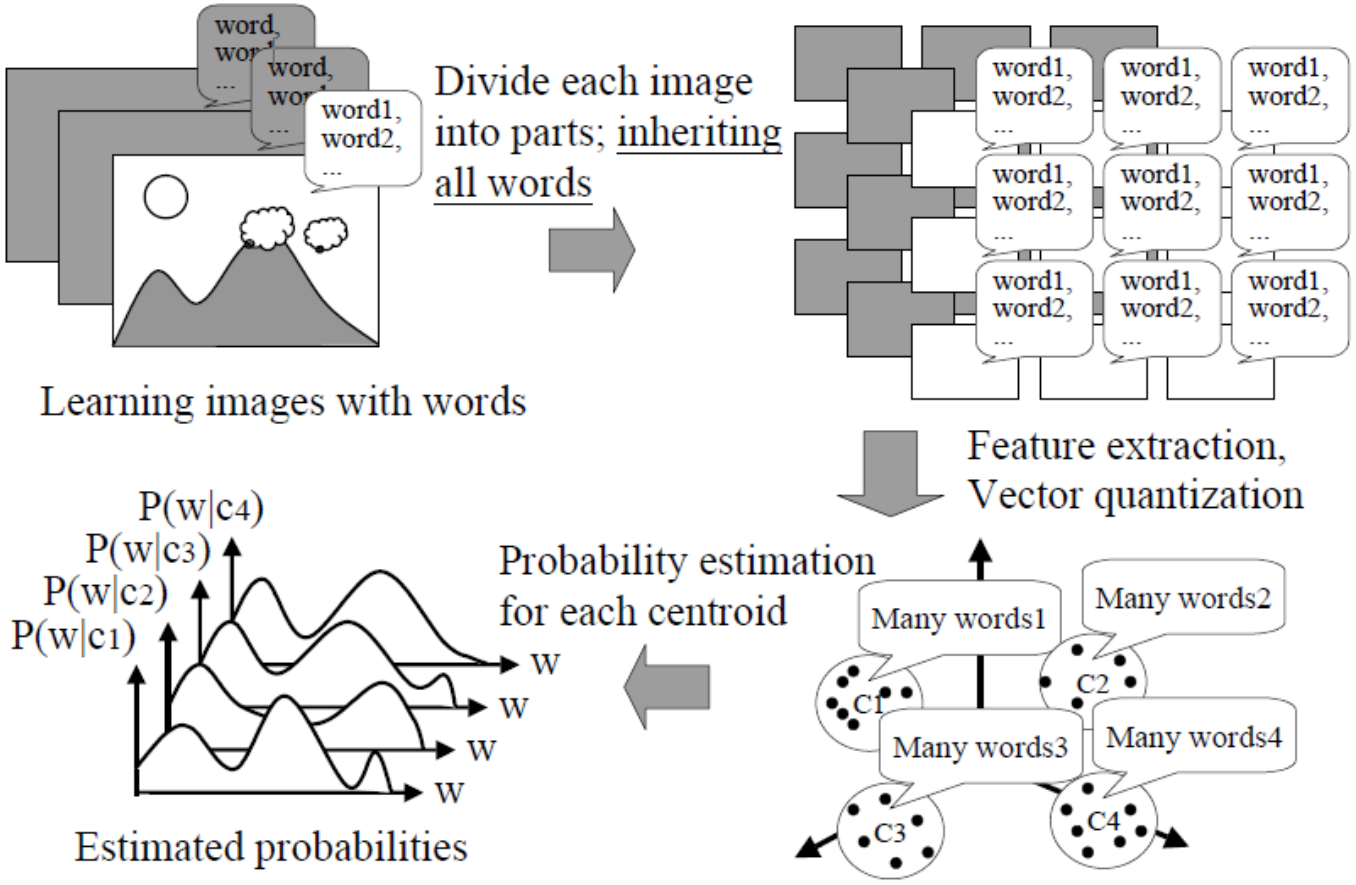
Annotation Approaches

- Word co-occurrence model
- Machine translation model
- Statistic models
- Refinement strategies
- … …

Co-occurrence Model

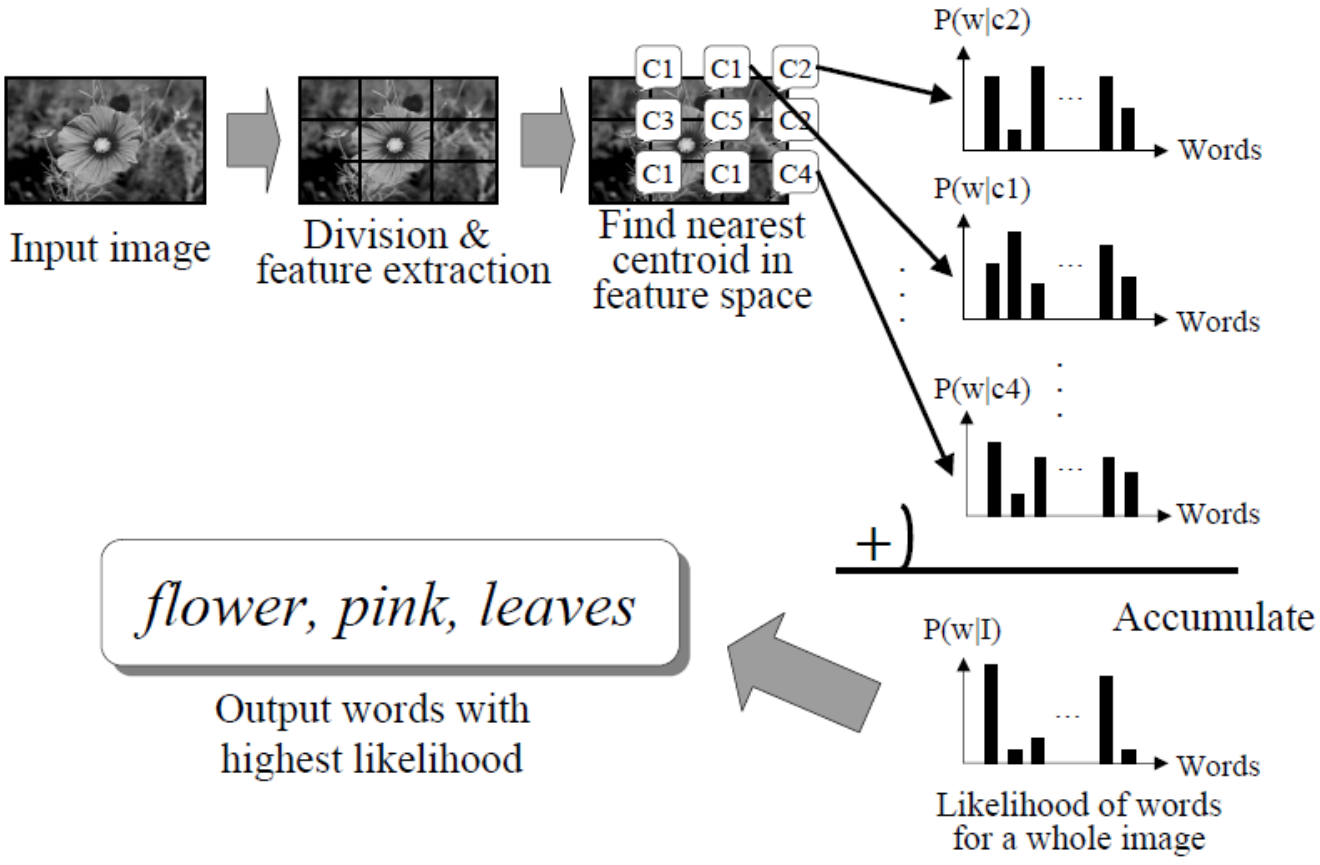
- An image is divided into parts
- Annotated words are inherited for each part
- Parts are vector quantized to make cluster
- P(word|cluster) for all word for all cluster is estimated statistically

Co-occurrence Model



Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999.

Co-occurrence Model

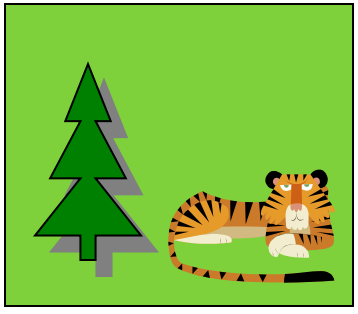


Input image	Output words (top 3)	Input image	Output words (top 3)
	**year**, **Japan**, family		year, **age**, white
	**year**, **many**, family		area, east, shore
	year, **park**, family		**park**, **national**, center
	year, **ten thousand**, **city**		**city**, god, layer

Machine Translation Model

I love multimedia computing technology very much.

我 非常 喜欢 多媒体 计算 技术.



Tree Tiger

P. Duygulu, K. Barnard, N. de Fretias, and D. Forsyth, Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, ECCV 2002.

Blob Representation

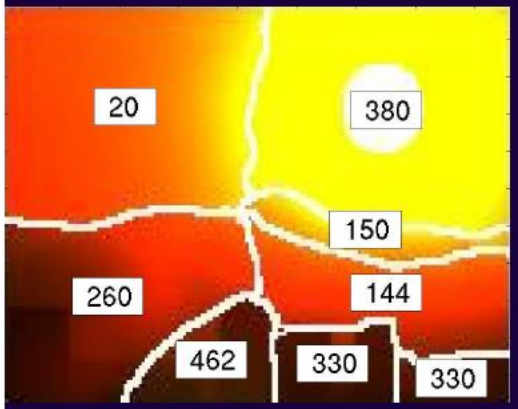
- Visual features
  - Region size
  - Position
  - Color
  - Oriented energy (12 filters)
  - Simple shape features

Tokenization

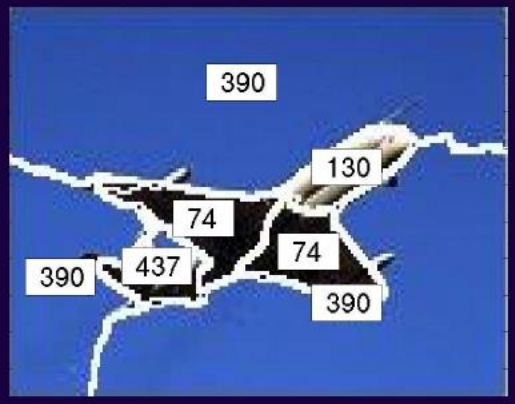
- Words → word tokens
- Image segments
  - represented by 30 features
    - (size, position, color, texture and shape)
  - k-means to cluster features
  - best cluster for the blob → blob tokens

Data

- 160 CD's from Corel Data Set
  - 100 images in each
- 10 sets, each :
  - randomly selected 80 CD's
  - ~6000 training
  - ~2000 test
  - 150-200 word tokens
  - 500 blob tokens
- Segmentation (using Ncuts)
  - about a month

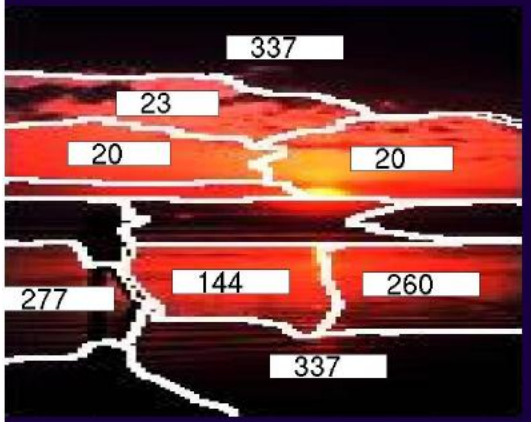


city mountain sky sun

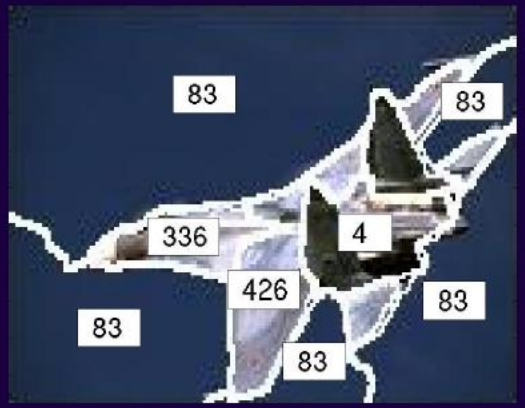


jet plane sky

cat forest grass tiger



beach people sun water



jet plane sky

cat grass tiger water

Annotation

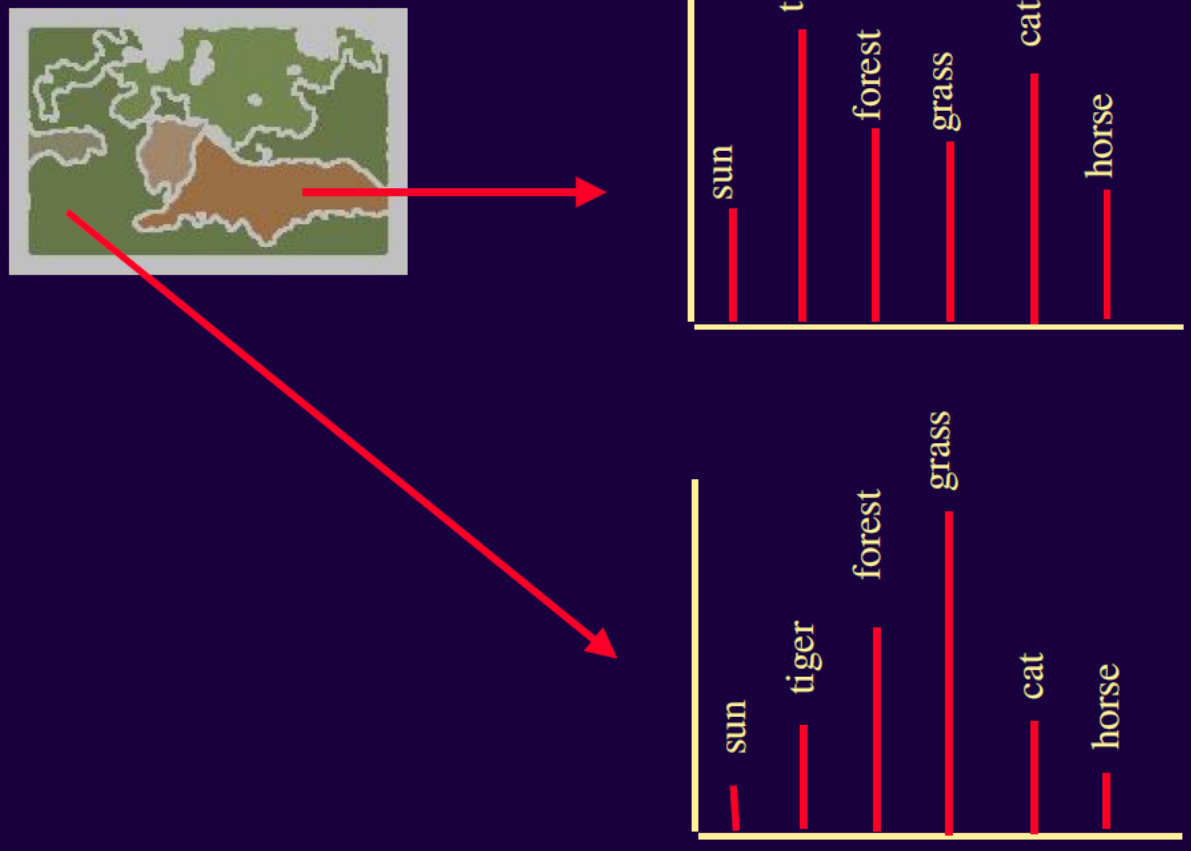
- Conditional probability p(w|b)
  - The likelihood function

$$P(w \mid b) = \sum_{a} P(w, a \mid b)$$

The event a_j=i means that the *j*th word in the possible translation translates the *i*th blob

- Using Expectation Maximization
  - Predicting correspondences from translation probabilities
  - Predicting translation probabilities from correspondences

Labeling Regions



Large Scale Annotation/Tagging

5+ billion (Sep 2010)

- 160 years to view all of them (1s per image)
- **3000+** uploads/minute
- 2% Internet users visit (2009)
- Daily time on site: 4.7 minutes (2009)

400 million (2010)

- 2,000 years to see all of them
- ~20 hours uploaded/minute (09)
- 20% Internet users visit (2009)
- Daily time on site: 23 minutes (2009)
- 2007 bandwidth = entire Internet in 2000
- 3B+ views per day (2010)

60 billion (Dec 2010)

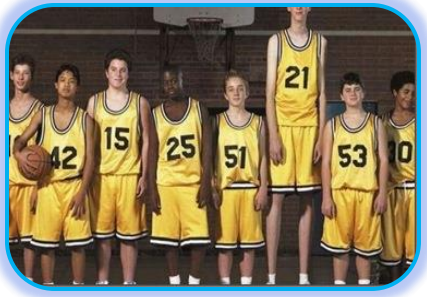
- 1,920 years to view all of them (1s per image)
- ~138M uploads/minute
- 24% Internet users visit (2009)
- Daily time on site: 30 minutes (2009)

Following slides are from Xian-Sheng Hua, Image and Video Tagging in the Internet Era

Characteristics of Internet Multimedia

Huge Amount of Data

Increasing Very Rapidly



Variances Are Very Large

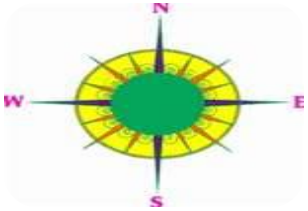
Be consumed Frequently

Affection Highly Involved

Connected to Each Other

Variety of Internet Multimedia Applications

Search



Browsing

Sharing

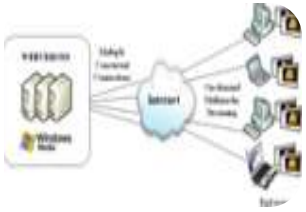
Authoring/Editing

Copy Detection

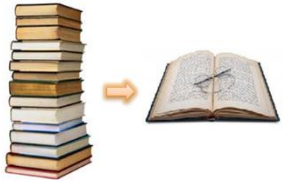
Recommendation

Tagging

Mining



Streaming



Summarization

Visualization

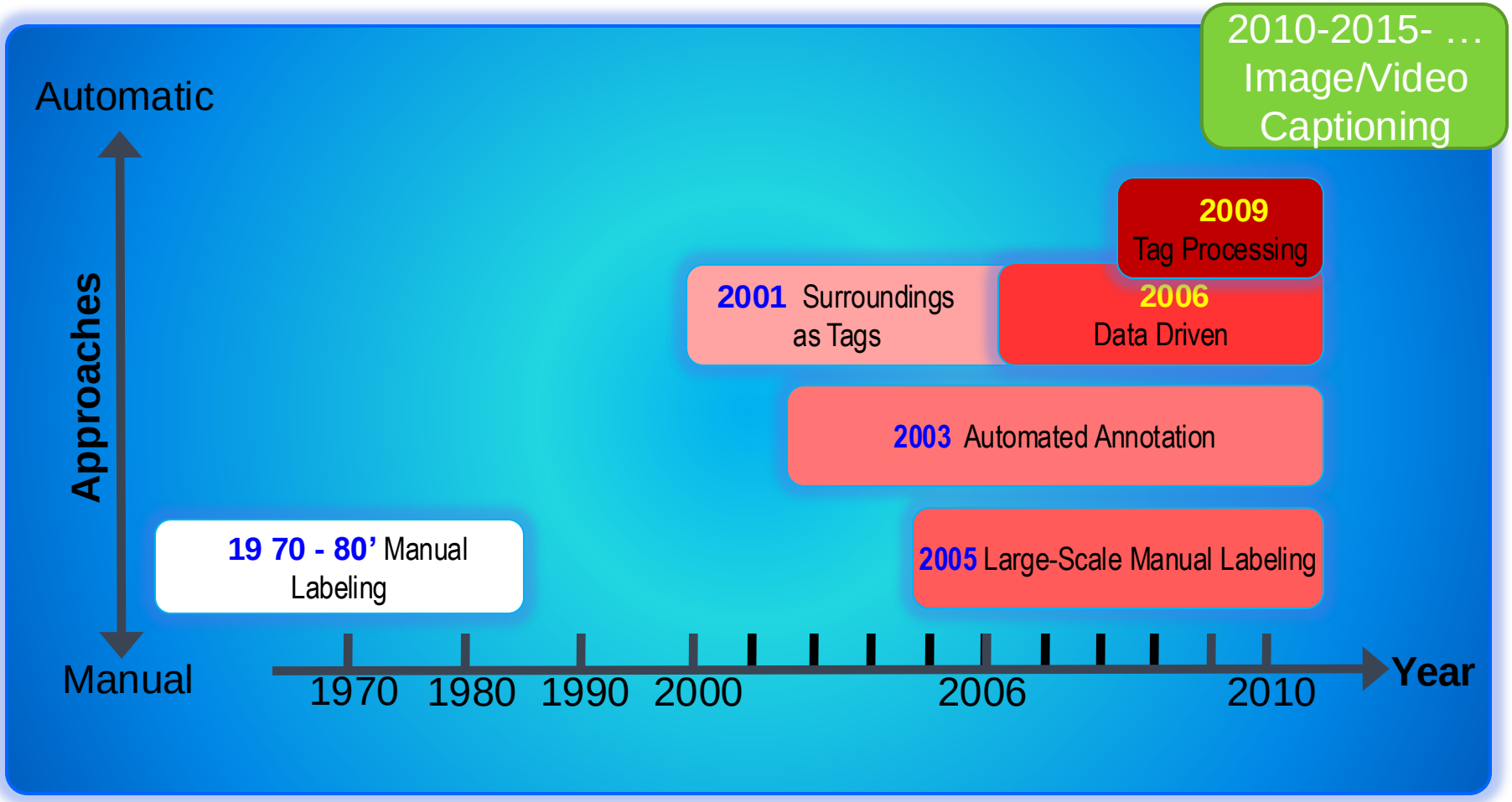
Advertising

Categorization

Forensics

Media on Mobiles

Evolution of Media Tagging



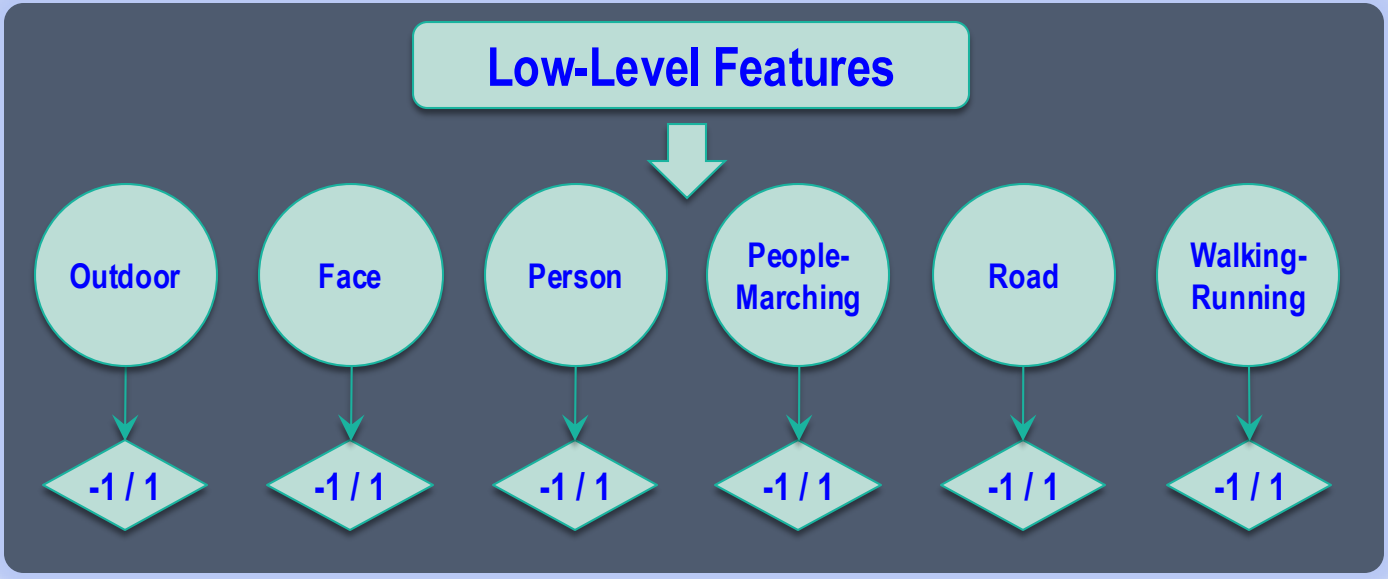
Synonyms

(Automatic) Annotation
≈ Concept Detection
≈ (Automatic) Tagging
≈ (Automatic) Labeling

Tags
≈ Labels
≈ Concepts

Automated Annotation – 1st Paradigm

- A typical strategy – Individual Concept Detection
  - Annotate multiple concepts separately



To Exploit Label Correlations

√ Person
√ Street
√ Building

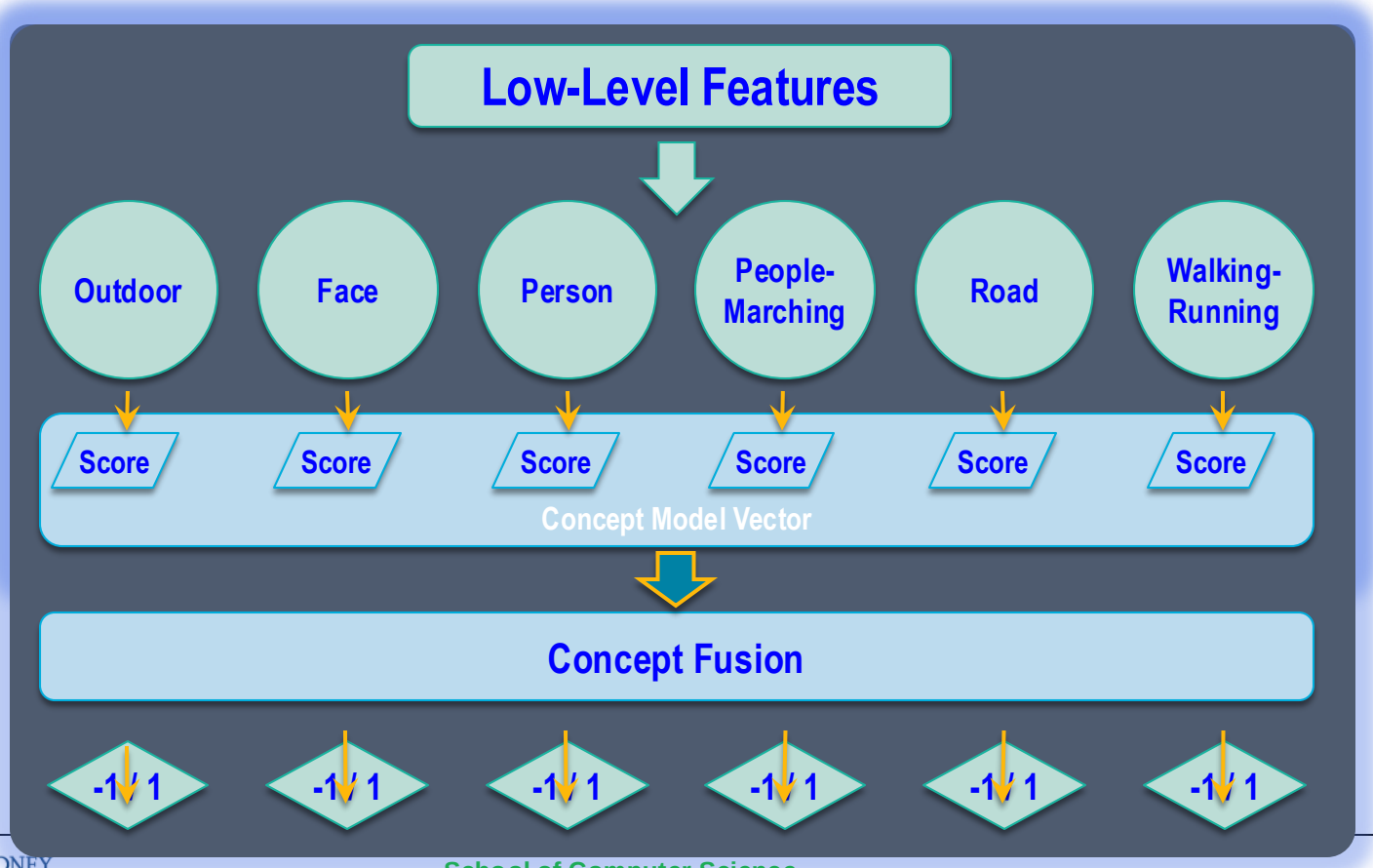
× Beach
× Mountain

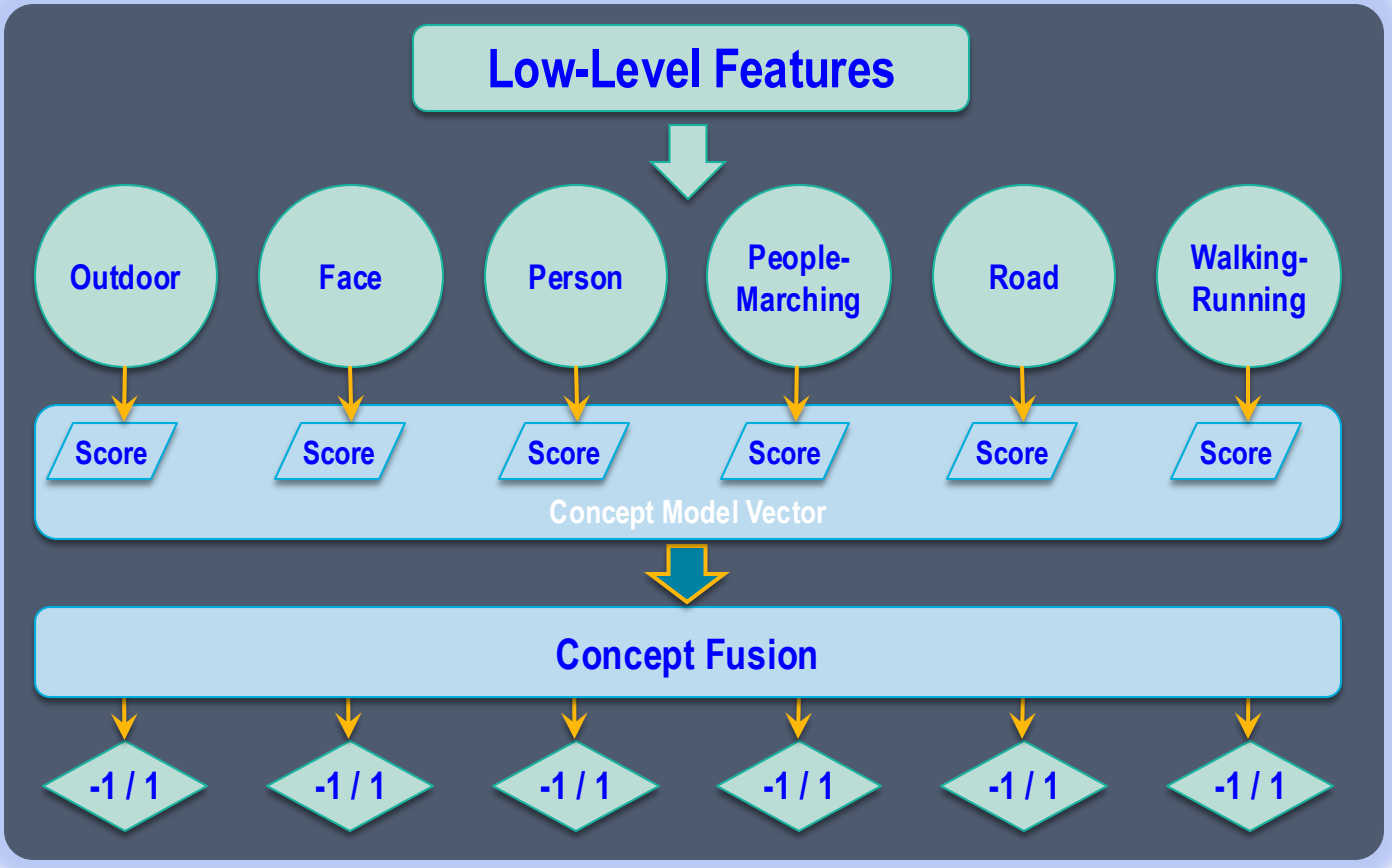
√ Crowd
√ Outdoor
√ Walking/Running

√ Marching ? Marching

Automated Annotation – 2nd Paradigm

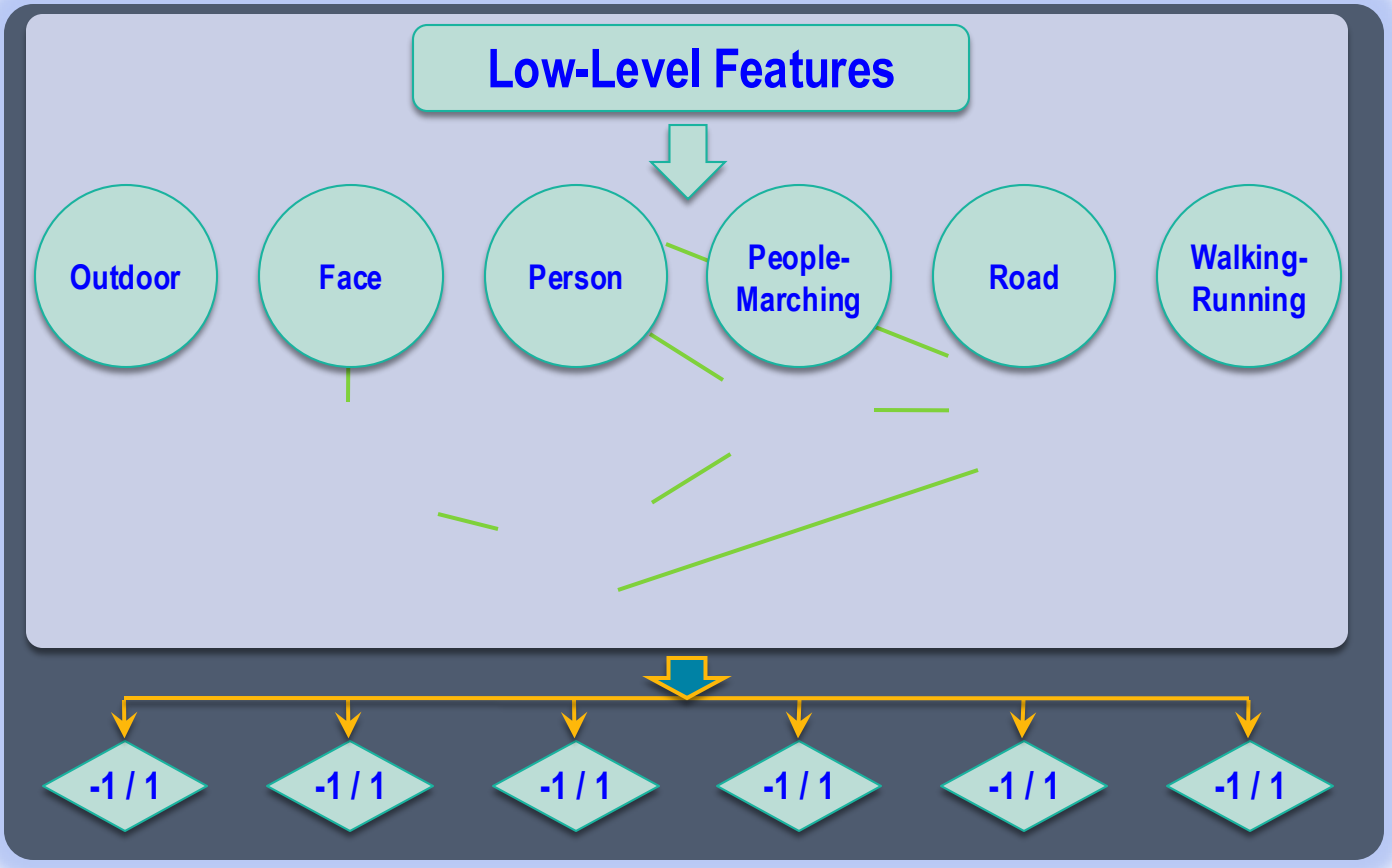
- Another typical strategy – Fusion-Based
  - Context Based Concept fusion (CBCF)



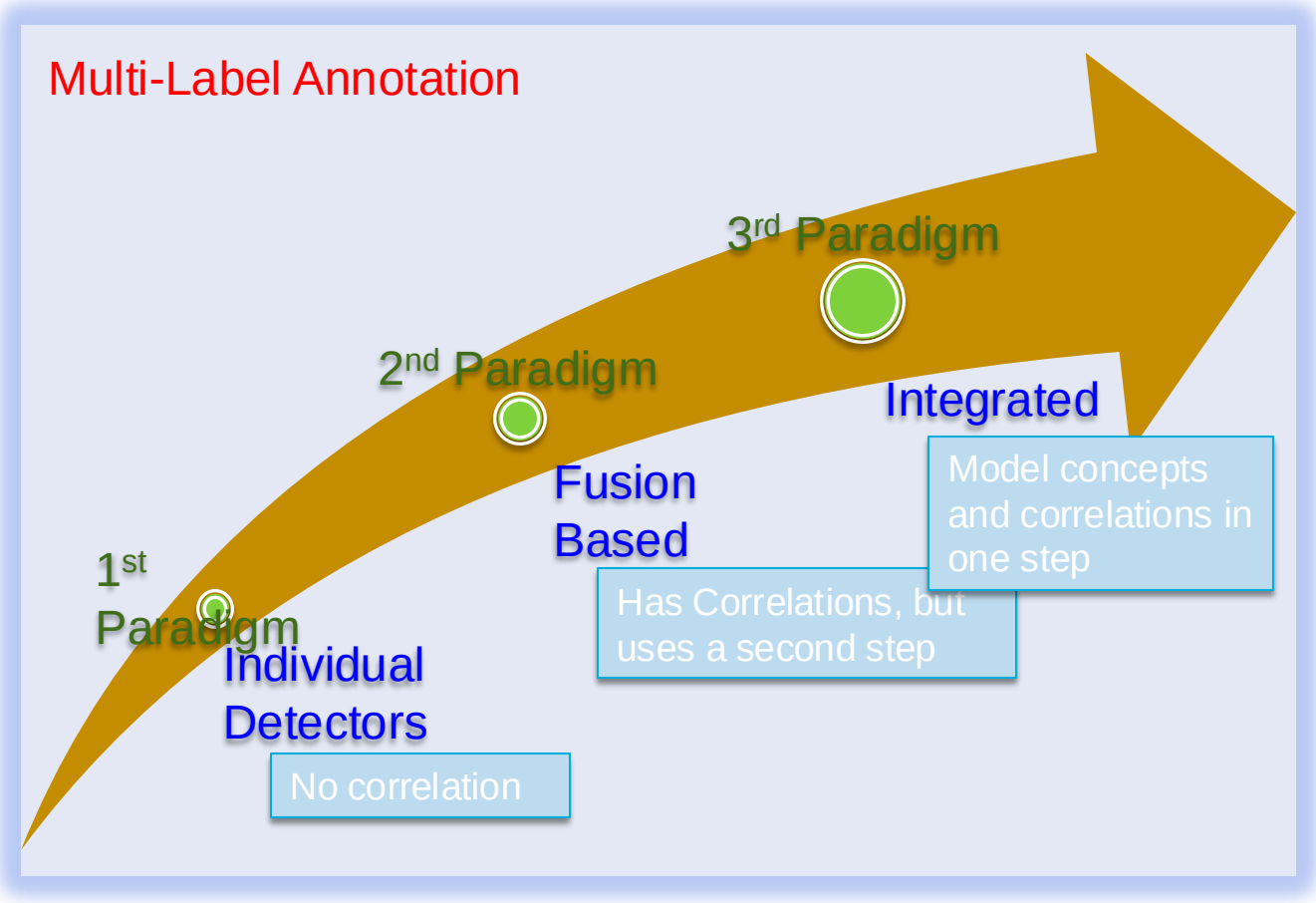

Low-Level Features
Outdoor
Face
Person
People-Marching
Road
Walking-Running
Score
Score
Score
Score
Score
Score
Concept Model Vector
Concept Fusion
-1 / 1
-1 / 1
-1 / 1
-1 / 1
-1 / 1
-1 / 1

Automated Annotation – 3rd Paradigm

- Integrated Concept Detection
  - Correlative Multi-Label Learning (CML)



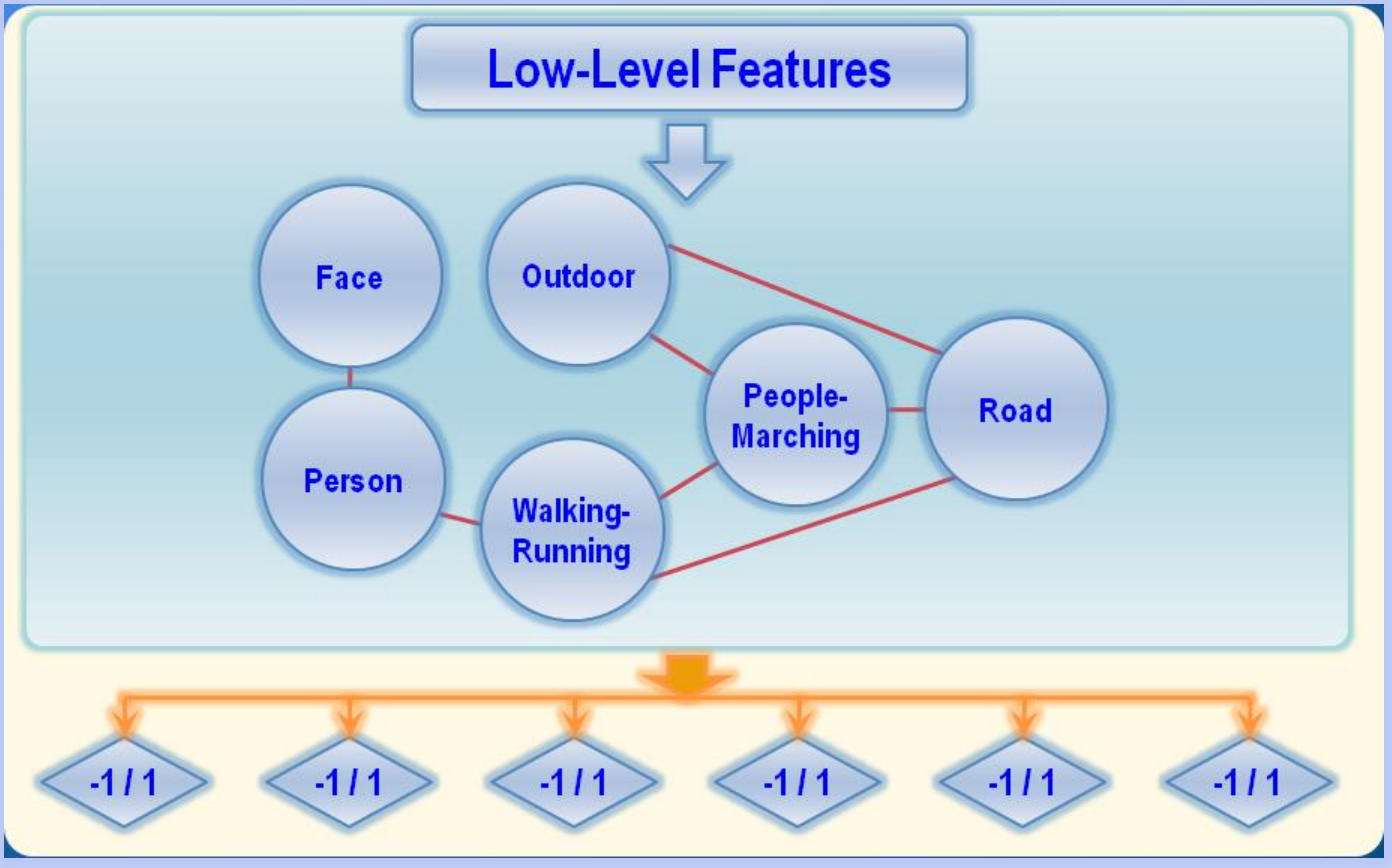
CML Roadmap



G.-J. Qi, et al., Correlative Multi-Label Video Annotation, ACM Multimedia 2007.

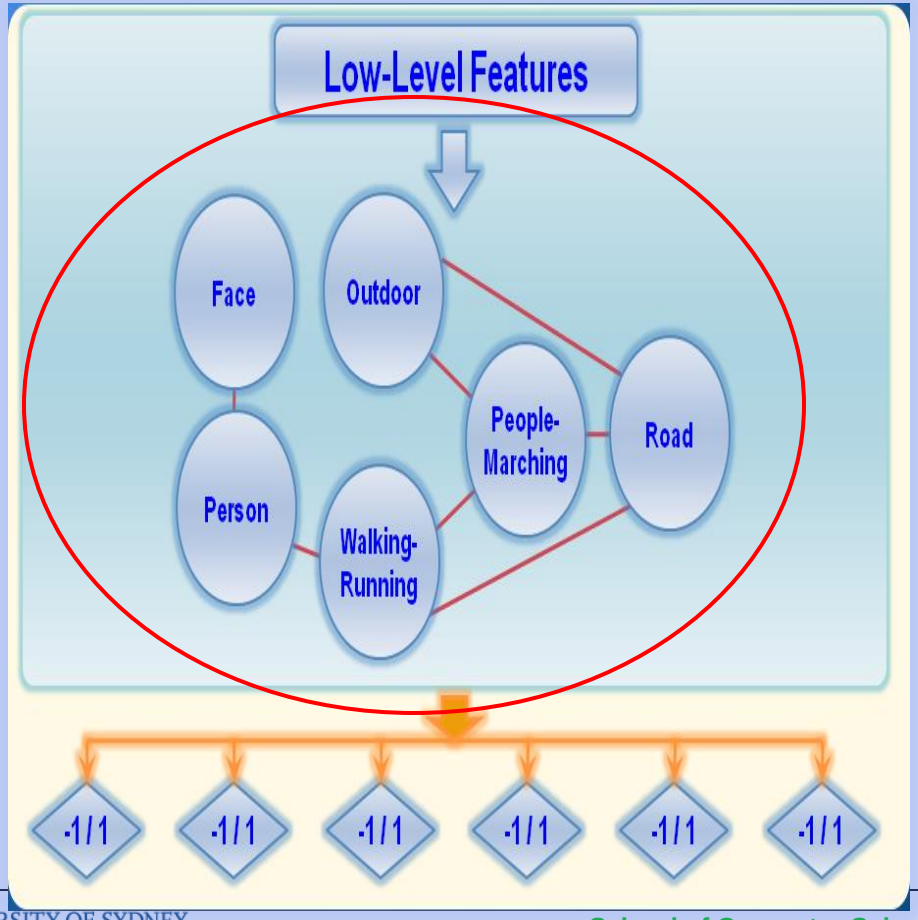
Automated Annotation – 3rd Paradigm

- Integrated Concept Detection
  - Correlative Multi-Label Learning (CML)



How to Model Concept Correlations

- How to model concepts and the correlations among concept in a single step



Strategy

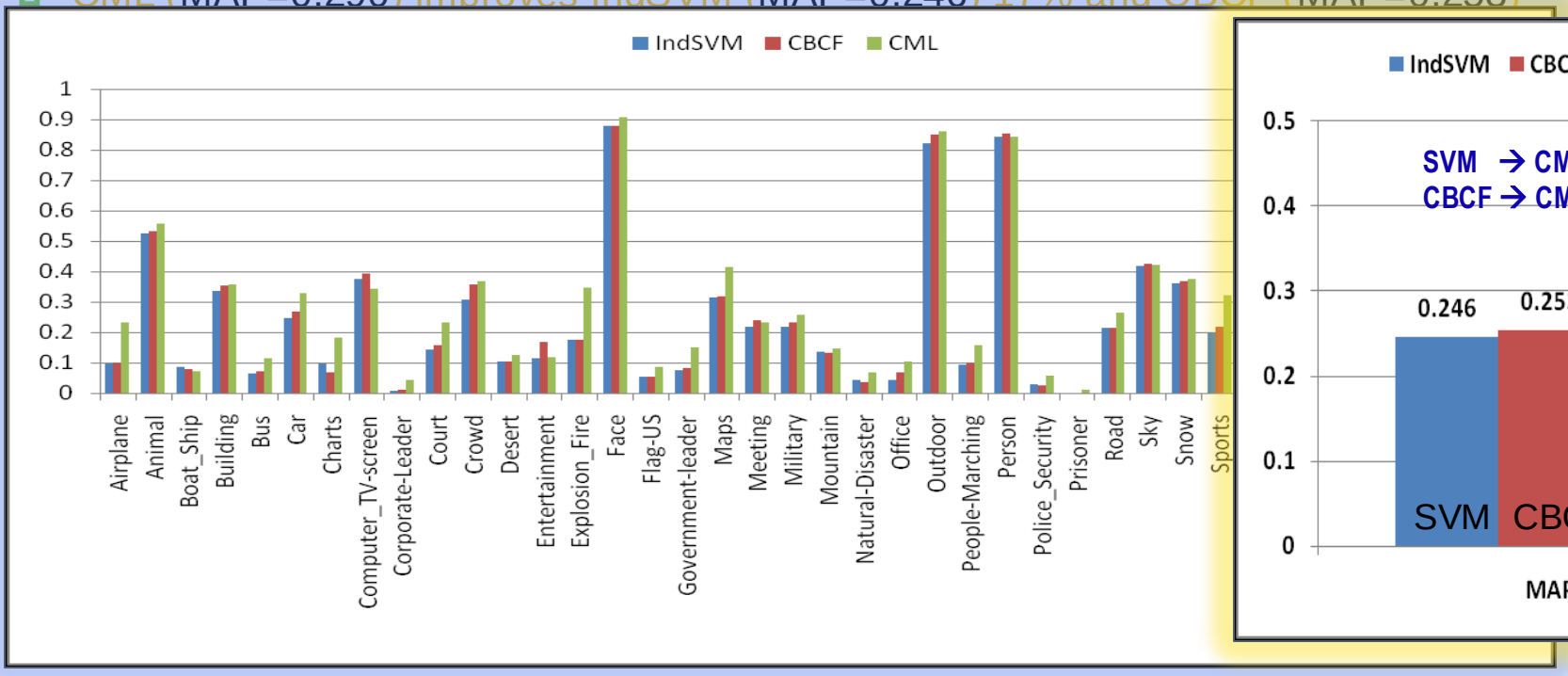
Converting correlations into features.

Constructing a new feature vector that captures both

- The characteristics of concepts, and
- The correlations among concepts

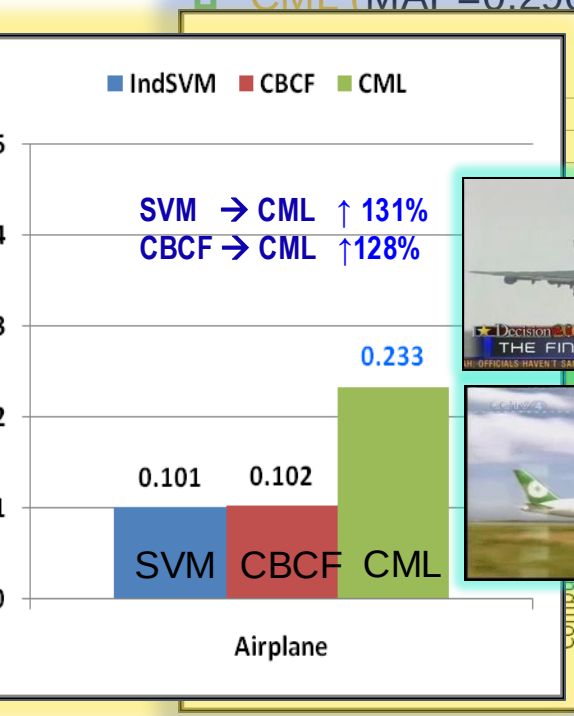
Correlative Multi-Label Video Annotation

- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)
  - CML (MAP=0.290) improves IndSVM (MAP=0.246) 17% and CBCF (MAP=0.253)



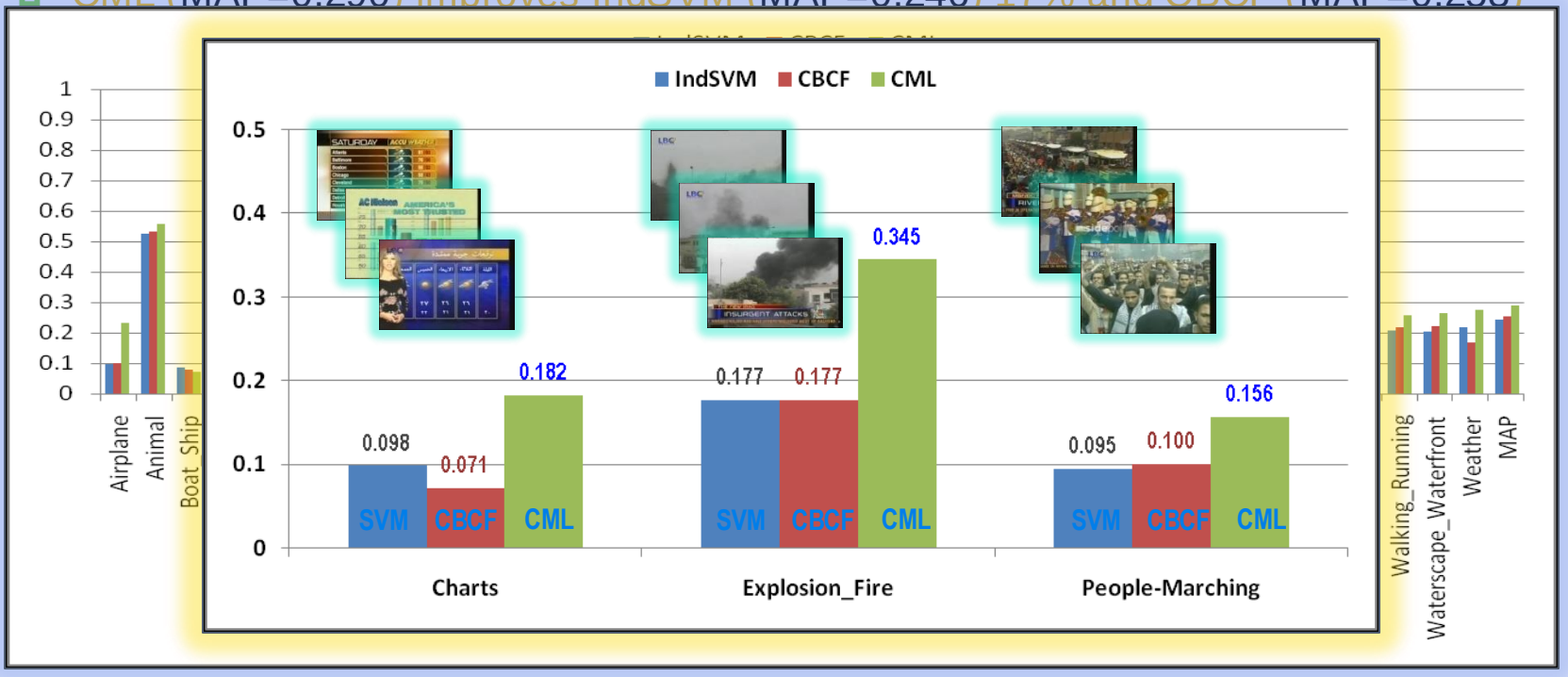
Correlative Multi-Label Video Annotation

- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)
  - CML (MAP=0.290) improves IndSVM (MAP=0.246) 17% and CBCF (MAP=0.253)



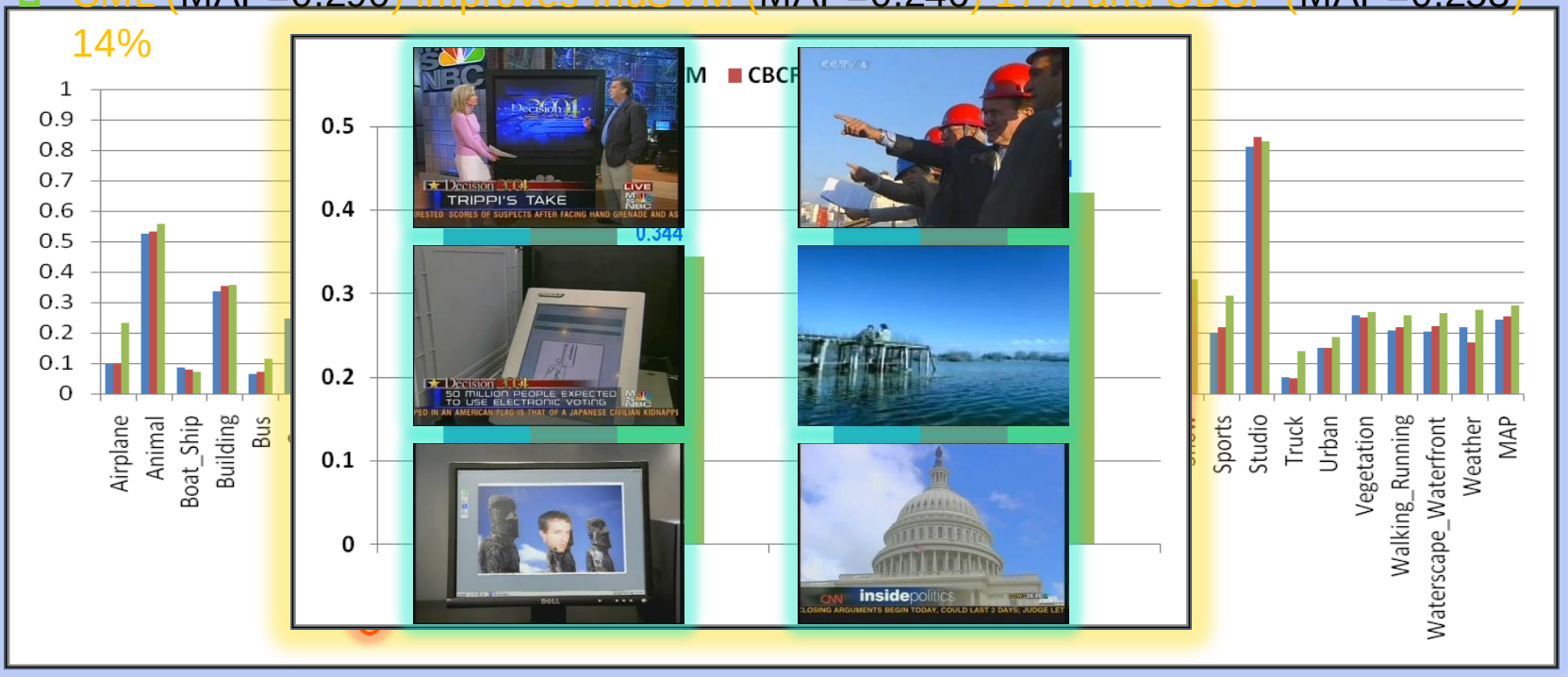
Correlative Multi-Label Video Annotation

- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)
  - CML (MAP=0.290) improves IndSVM (MAP=0.246) 17% and CBCF (MAP=0.253)



Correlative Multi-Label Video Annotation

- Experiments
  - TRECVID 2005 dataset (170 hours)
  - 39 concepts (LSCOM-Lite)
  - Training (65%), Validation (16%), Testing (19%)
  - CML (MAP=0.290) improves IndSVM (MAP=0.246) 17% and CBCF (MAP=0.253) 14%



Internet Media

flickr

www, www2009, madrid, spain
w3c, Don Quixote, Don, Quixote
cervantes, Sancho, ...

www2009, w3c, futuro, future, workshop, congreso
palacio, municipal, Madrid, consortium, consorcio
20, aniversario, España, Spain, Vinton, ...

YouTube

Social tags

- Good, but
- Noisy
- Ambiguous
- Incomplete
- No relevance information

Two directions to improve tag quality

- During tagging – Tag Recommendation
- After tagging – Tag Refinement/Ranking

The most relevant tag is NOT at the top position in the tag list of the following social image.

Social tags for online images are better than automatic annotation in terms of both scalability and accuracy.

This phenomenon is widespread on social media websites such as Flickr.

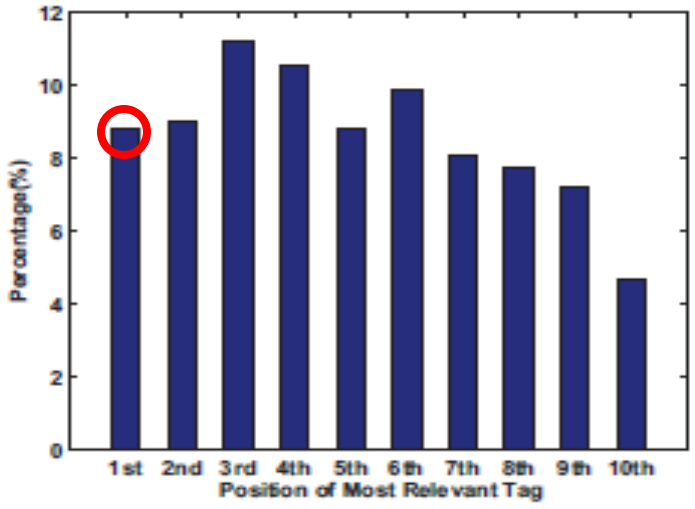
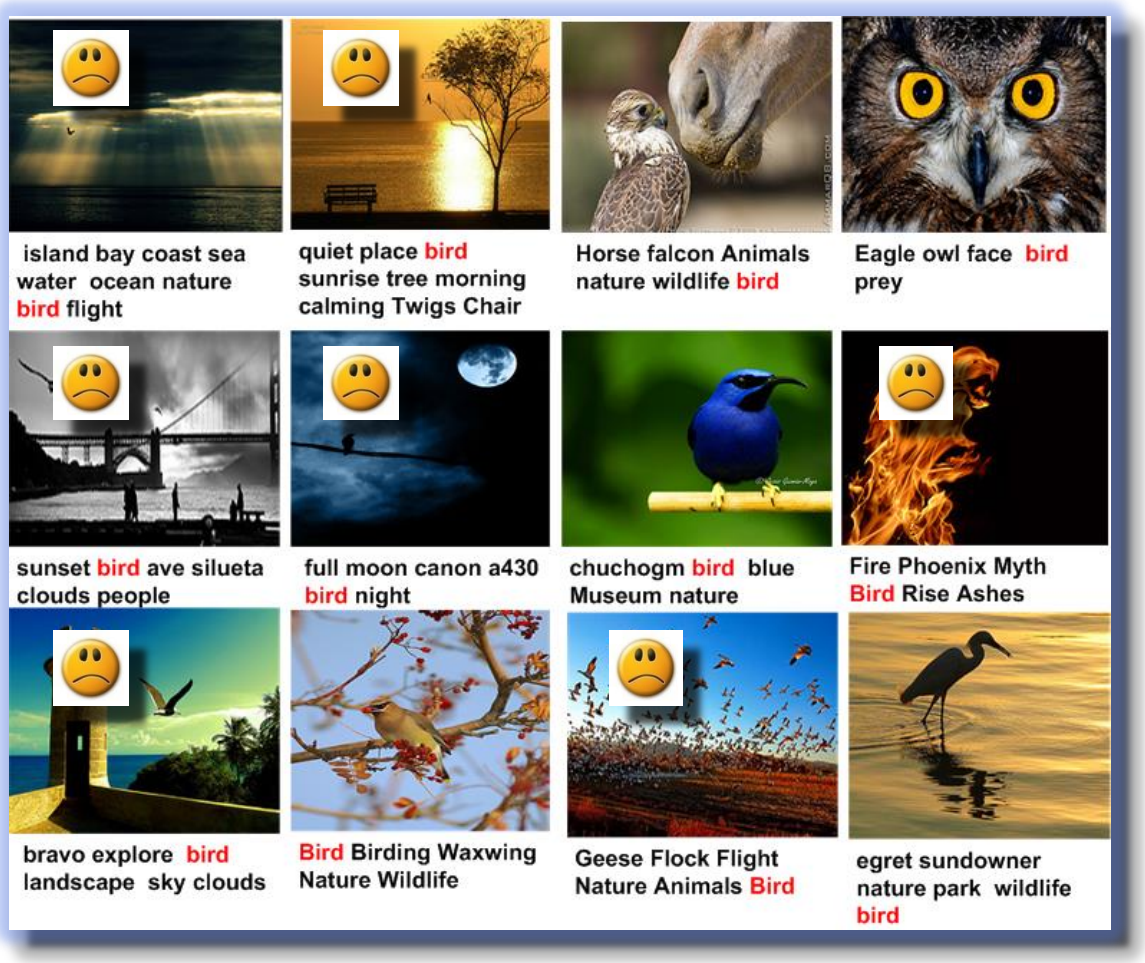


Figure 2: Percentage of images that have their most relevant tag at the n-th position in the associated tag list, where $n = 1, 2, 3 \ldots, 10$.

Only less than 10% images have their most relevant tag at the top position in their tag list.

This has significantly limited the performance of tag-based image search and other applications.

For example, when we search for “bird” on Flickr.



What we are going to do:

Rank the tags according to their relevance to the image.

Towards Storytelling

- Image and video captioning

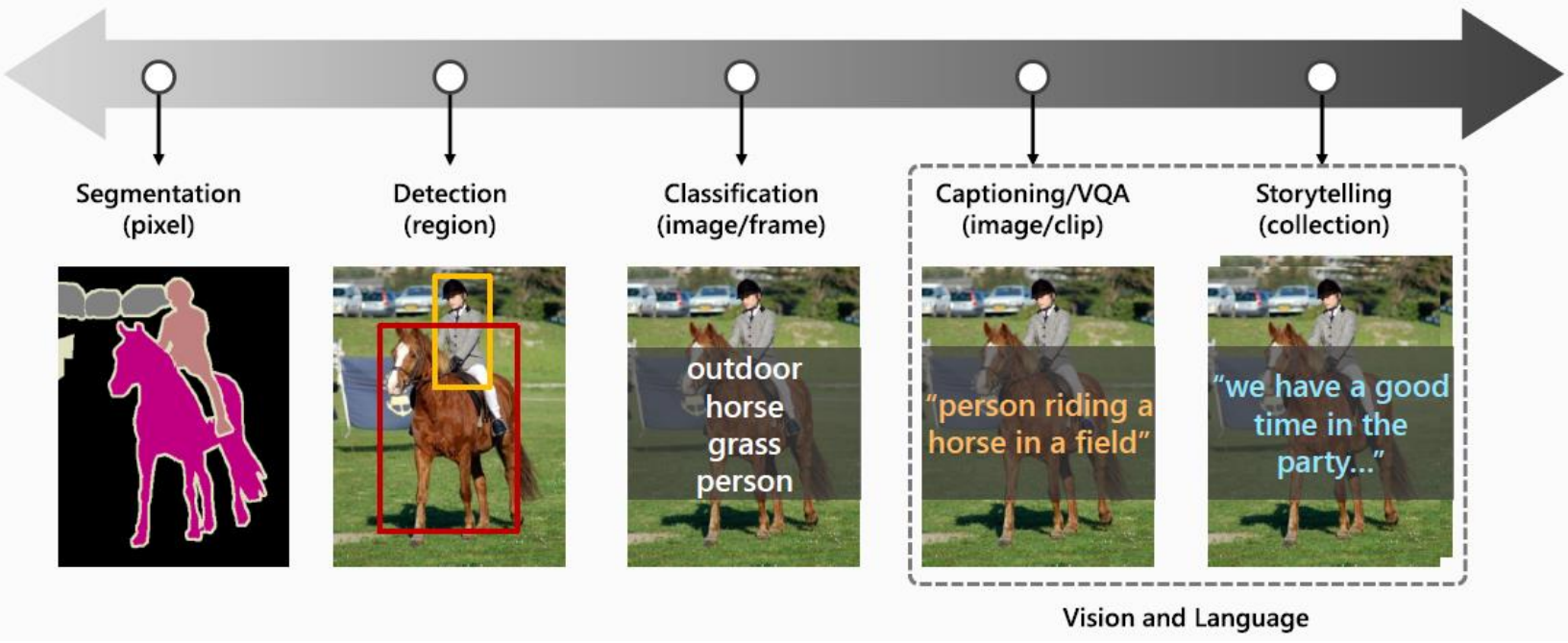


Figure from Tao Mei, ACM MM 2016.

Bag-of-words (BoW) Model

A document

A collection of the words in the document

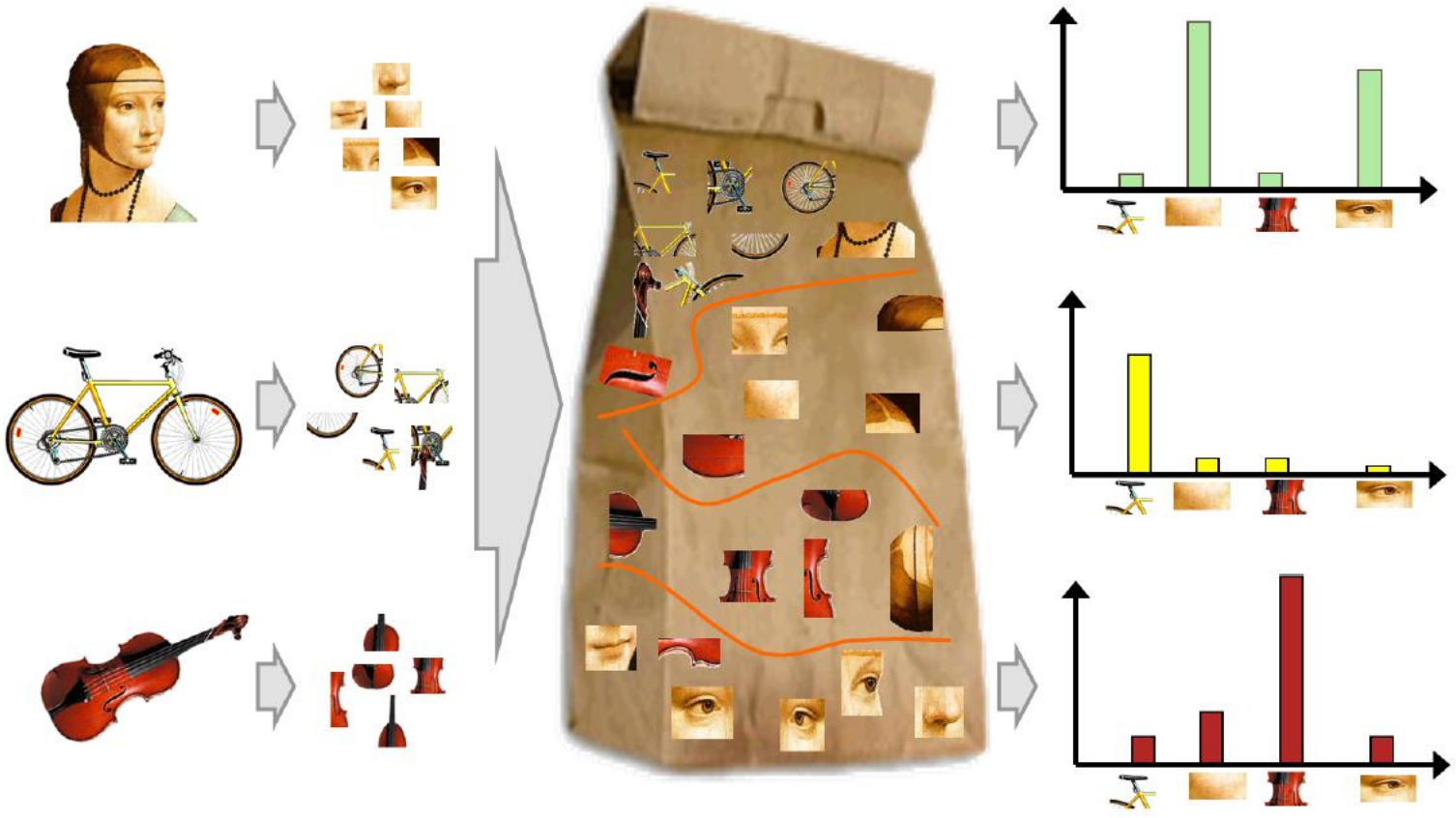
An image

A collection of the objects in the image

What features could characterize an object well?

R. Jin, Content based Image Retrieval.

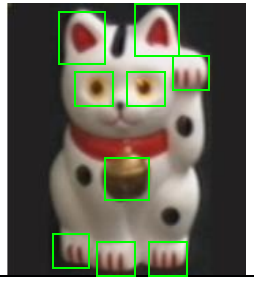
Bag-of-Visual-Words (BoVW)



Slides credit: M. Bressan and L. Fei-Fei

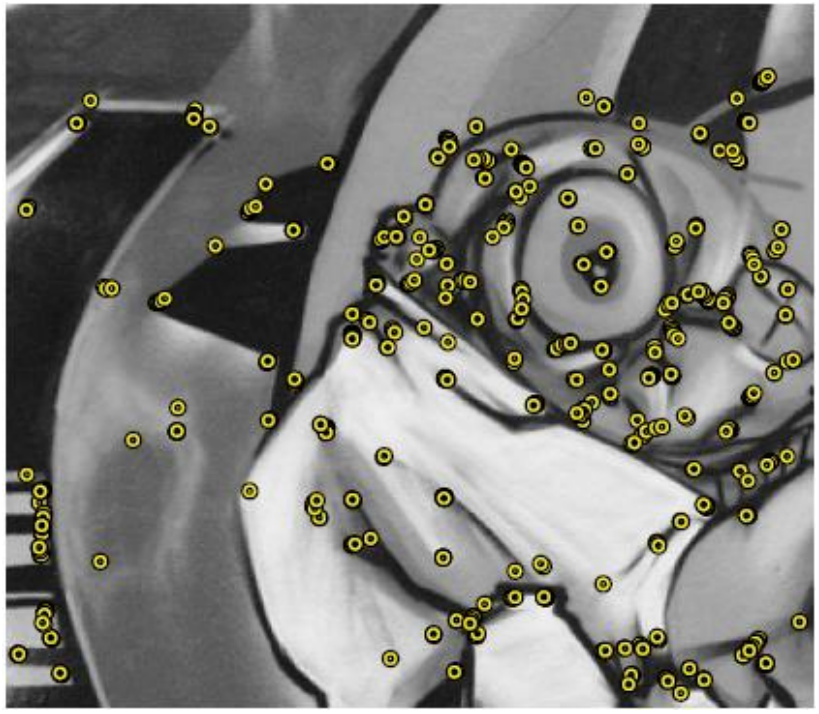
Interest Point Detection

- *Local features* have been shown to be effective for representing images
- They are image patterns which differ from their immediate neighborhood.
- They could be points, edges, small patches.
- We call local features *key points* or *interest points* of an image



Interest Point Detection

- An image example with key points detected by a corner detector.



Interest Point Detection

- The detection of interest point needs to be robust to various geometric transformations

(1)

Original

(2)

Scaling+Rotation+Translation

(3)

Projection

Interest Point Detection

- The detection of interest point needs to be robust to imaging conditions, e.g. lighting, blurring.




Descriptor

- Representing each detected key point
- Take measurements from a region centered on an interest point
  - E.g., texture, shape, …
- Each descriptor is a vector with fixed length
  - E.g. SIFT descriptor is a vector of 128 dimension

Descriptor

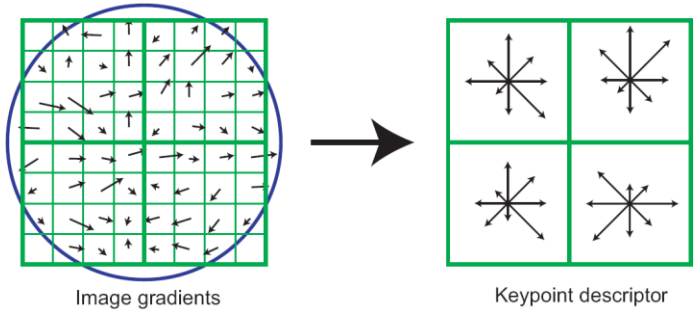
- The descriptor should also be robust under different image transformation.

Keypoint + Local Descriptors

- SIFT (Scale Invariant Feature Transform)

Feature detector

SIFT descriptor [Lowe'04]

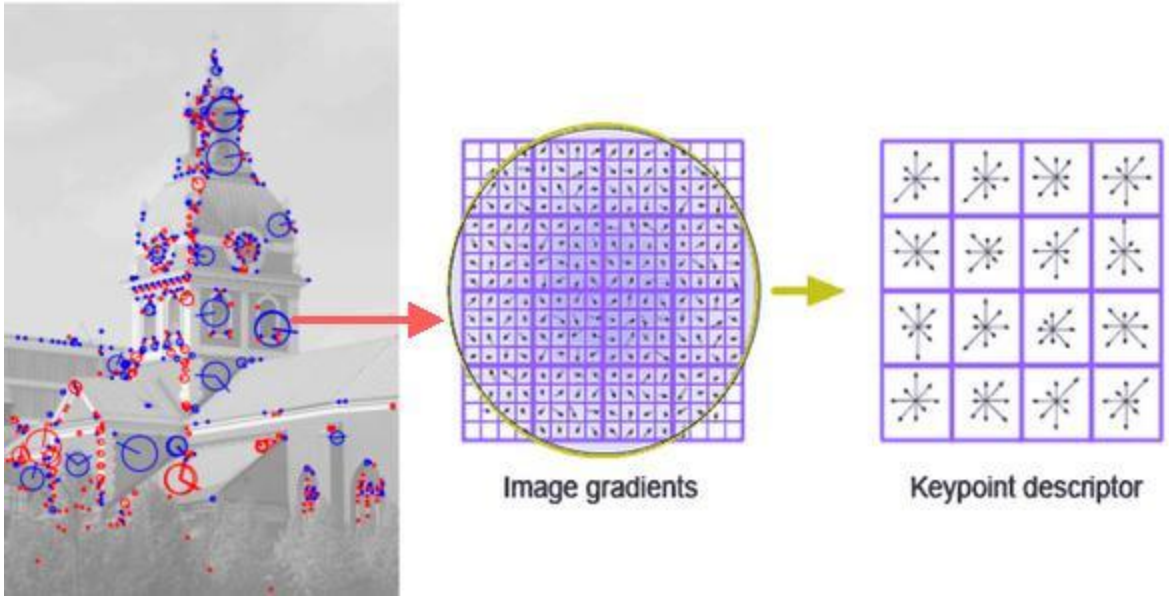


Calculated in a 16x16 window

Difference of Gaussian
Eliminating edge response

128 dimension = (4 x 4) * 8

SIFT

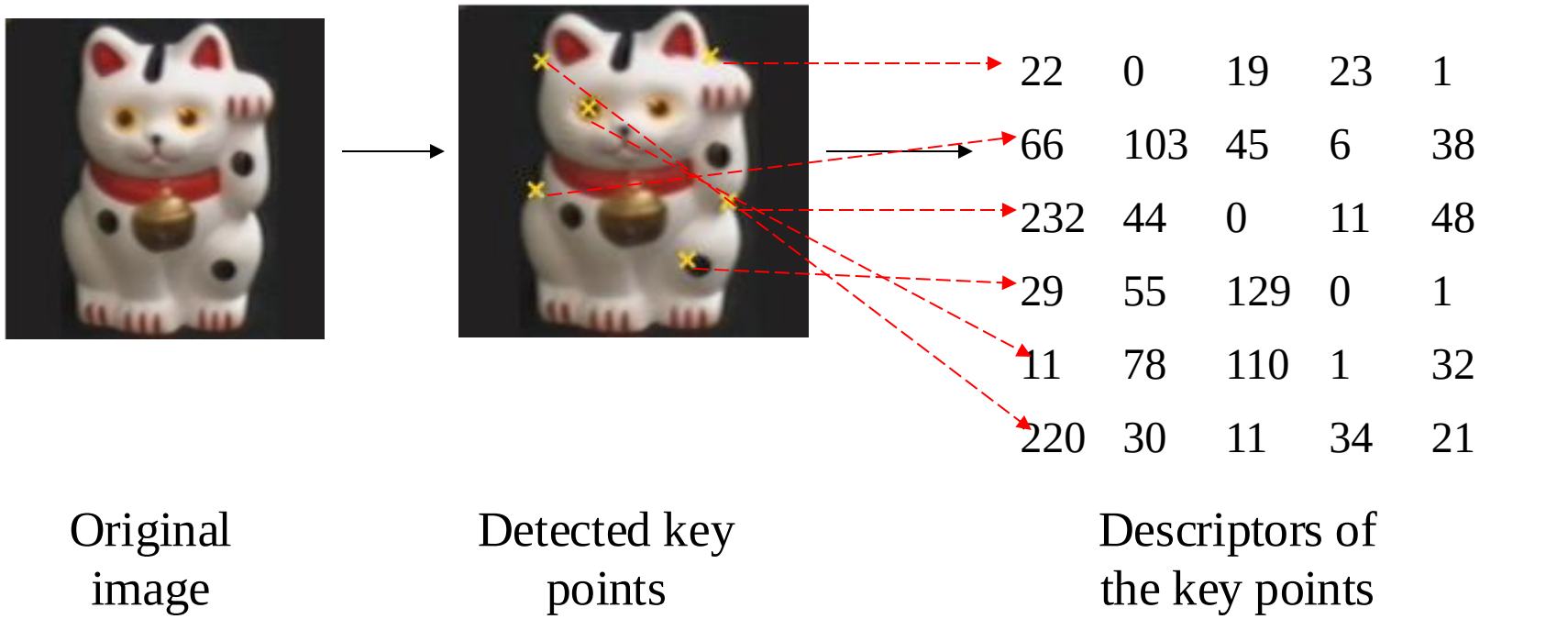


128 dimension = (4 x 4) * 8

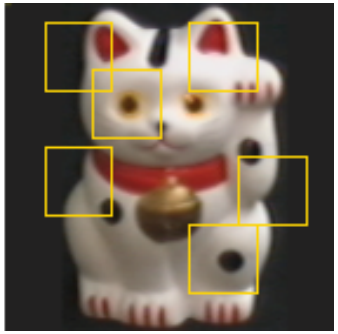
http://www.codeproject.com/Articles/619039/Bag-of-Features-Descriptor-on-SIFT-Features-with-O

Image Representation

- Bag-of-features representation: an example



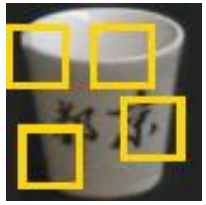
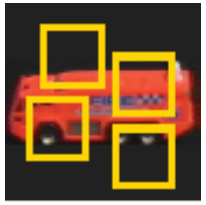
Retrieval



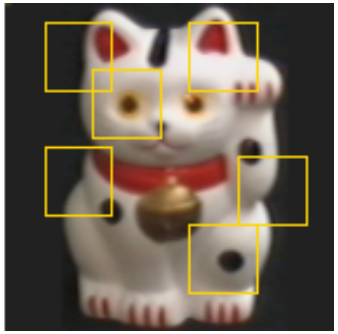
22	0	19	23	1
66	103	45	6	38
232	44	0	11	48
29	55	129	0	1
...				

How to measure similarity?



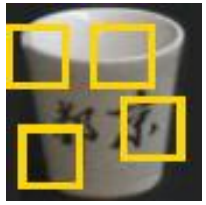




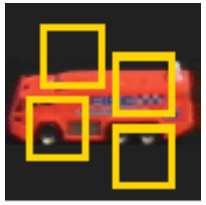
Retrieval



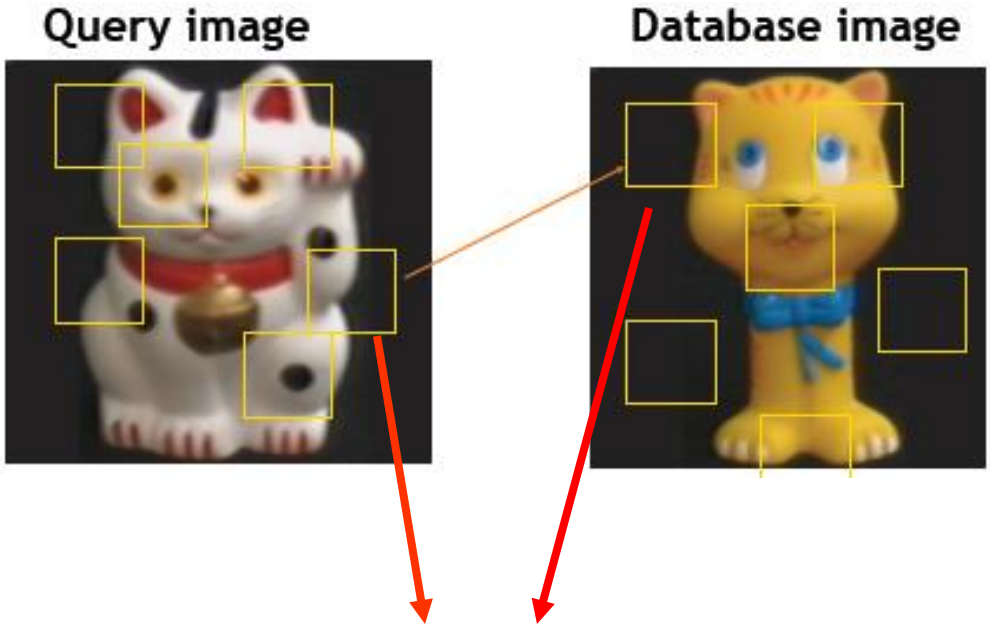
22	0	19	23	1
66	103	45	6	38
232	44	0	11	48
29	55	129	0	1
...				

Count number of matches !



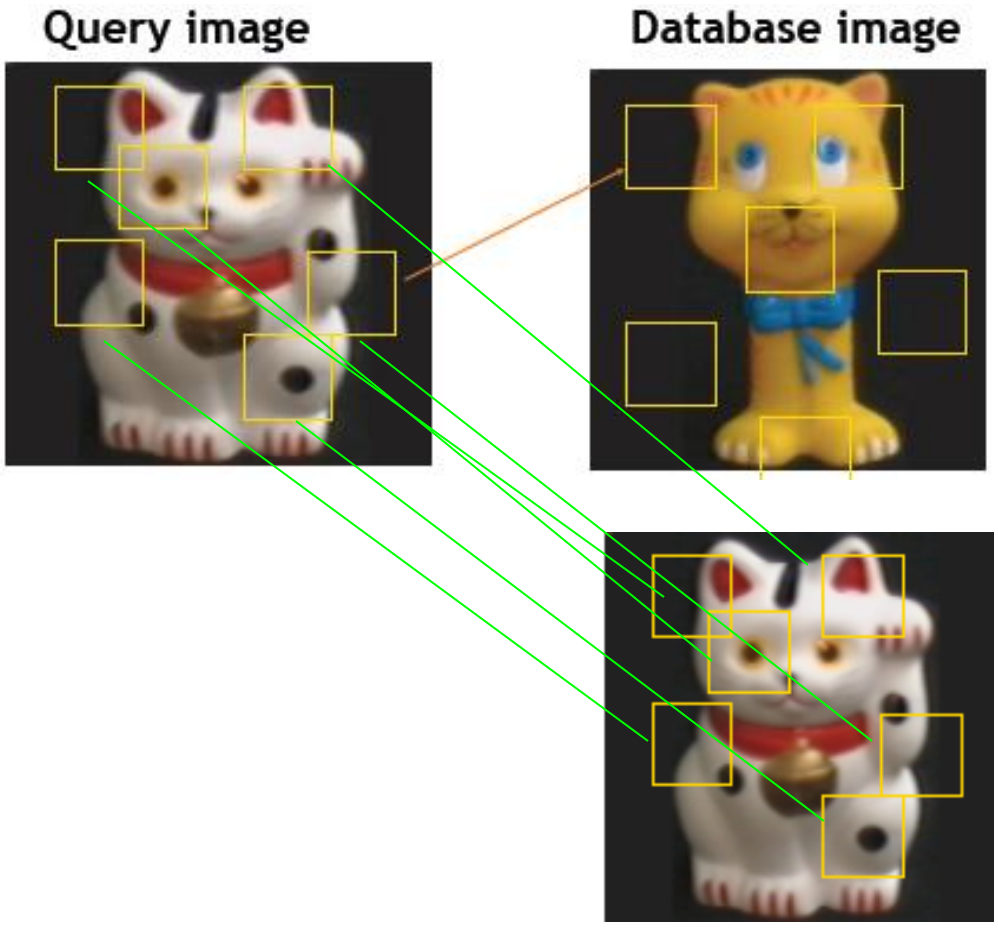


Retrieval



If the distance between two vectors is smaller than the threshold, we get one match

Retrieval



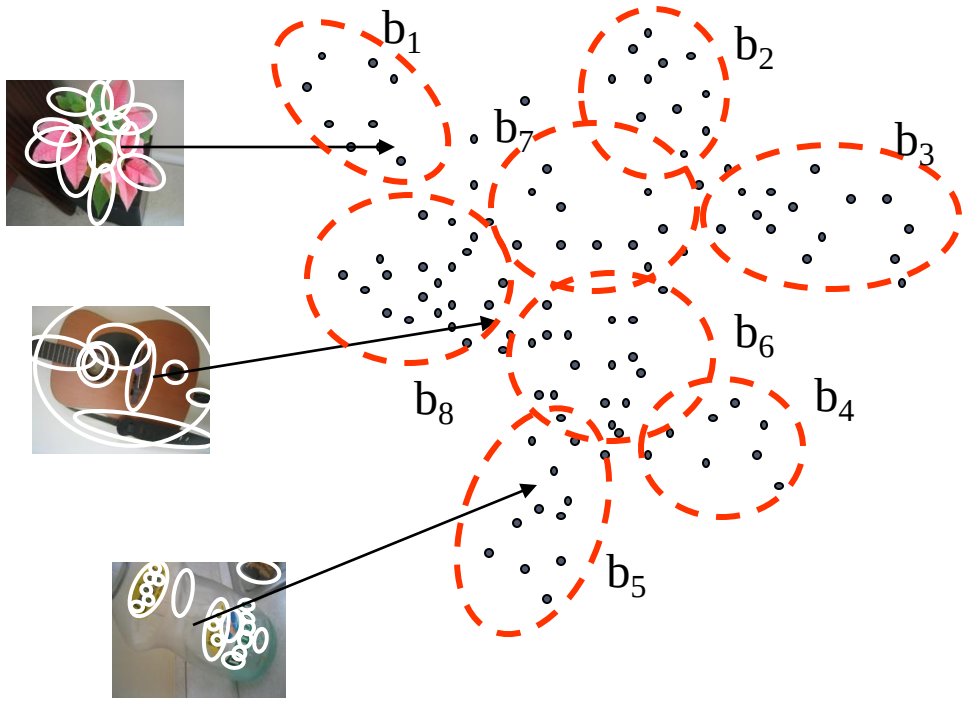
Matched points: 1

Matched points: 5

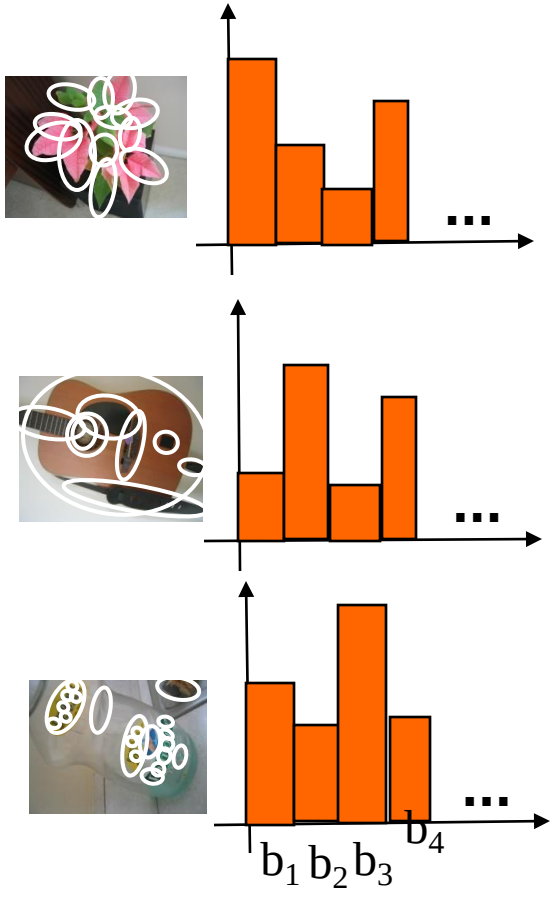
Problems

- Computationally expensive
  - Requiring linear scan of the entire data base
- Example: match a query image to a database of 1 million images
  - 0.1 second for computing the match between two images
  - Take more than one day to answer a single query

BoVW Model

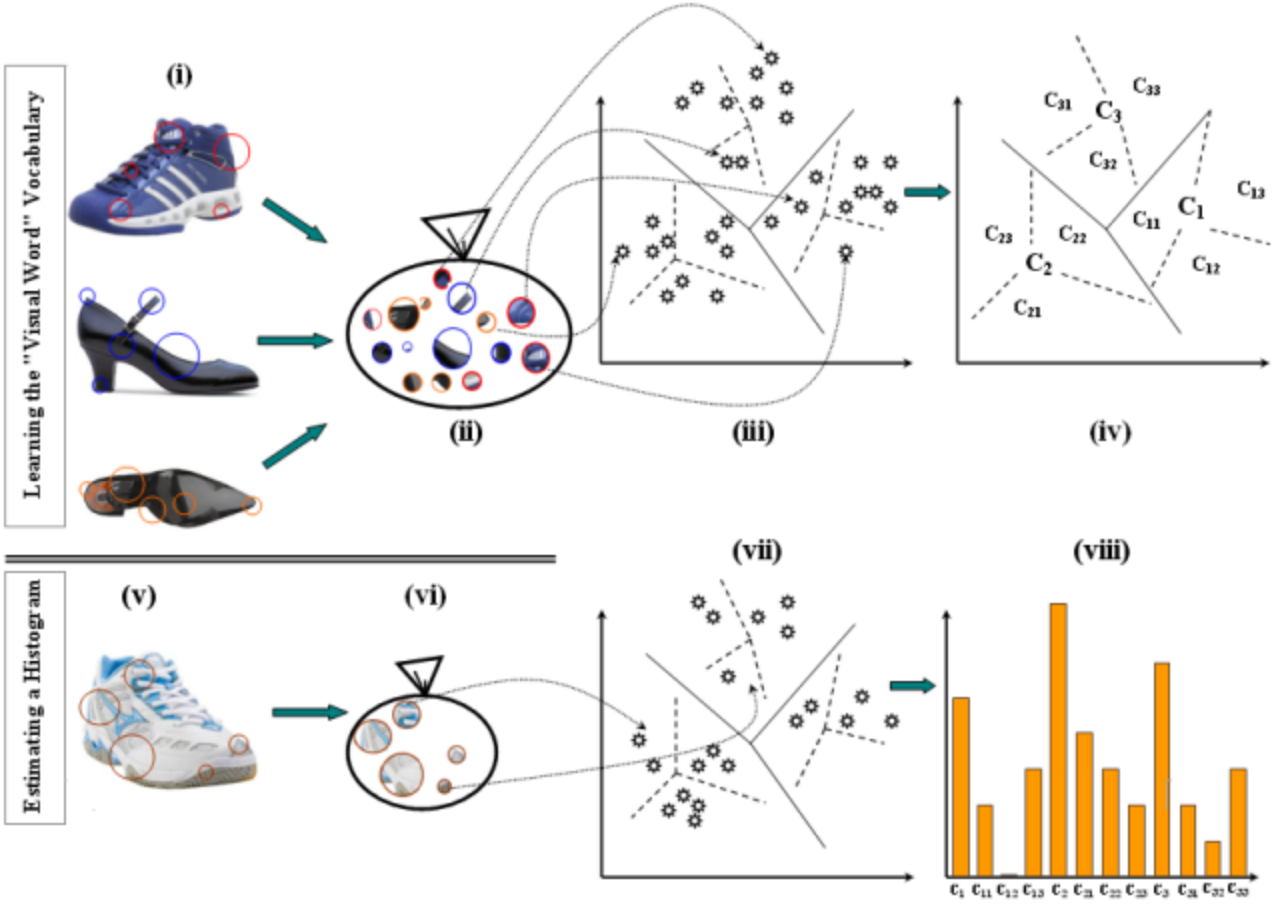


Group key points into visual words



Represent images by histograms of visual words

BoVW for Product Tagging



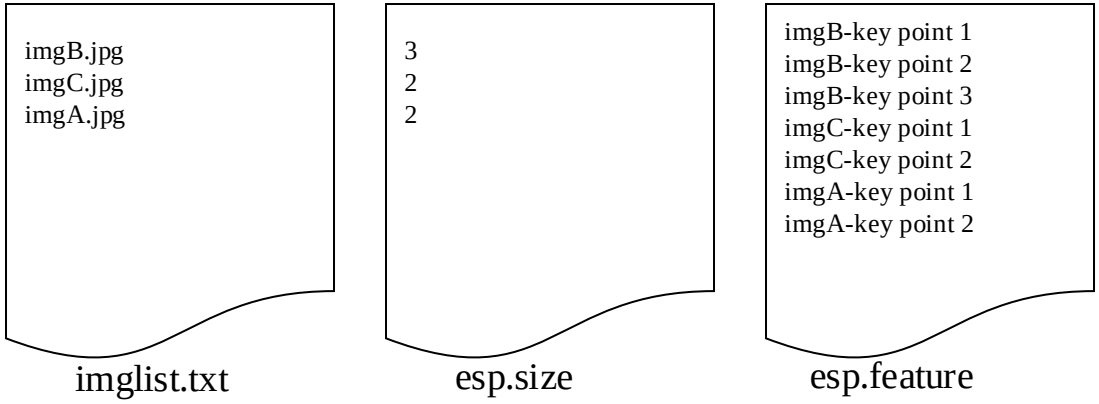
http://www.sccs.swarthmore.edu/users/09/btomasi1/tagging-products.html

BoVW Model

- Generate “visual vocabulary”
- Represent each key point by its nearest “visual word”
- Represent an image by “a bag of visual words”
- Text retrieval technique can be applied directly.

Step 1: Dataset

- Example: Three images imgA, imgB, imgC.
- imgA : 2 key points; imgB: 3 key points; imgC: 2 key points.



Step 2: Key Point Quantization

- Represent each image by a bag of visual words:
  - Construct the visual vocabulary
    - Clustering all the key points into 10,000 clusters
    - Each cluster center is a visual word
  - Map each key point to a visual word
    - Find the nearest cluster center for each key point (*nearest neighbor search*)

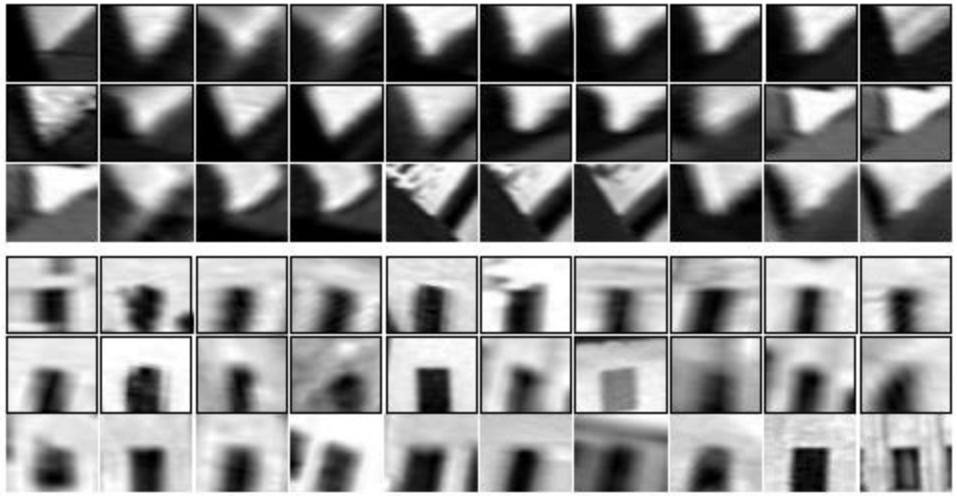
K-Mean Clustering

- Input
- E.g. Minimize square distance of vectors to centroids

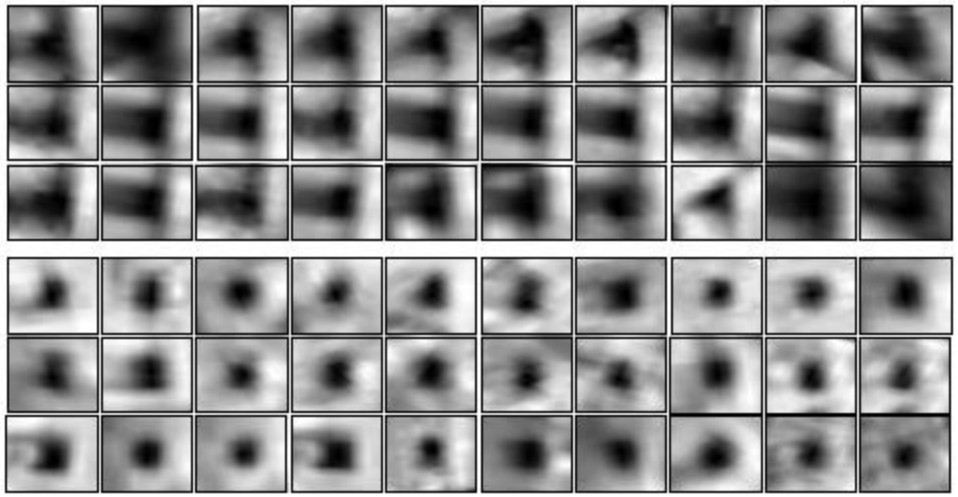
$$\sum_{j=1}^{K} \sum_{x \in D_j} \left(x - m_j\right)^2 \quad ; \quad m_j = \frac{1}{|D_j|} \sum_{x \in D_j} x$$

$$D = \bigcup_{j=1}^{K} D_j \qquad D_i \bigcap D_j = \Phi \qquad i \neq j$$

  - So that the points in each subset are coherent according to certain criterion



(a)

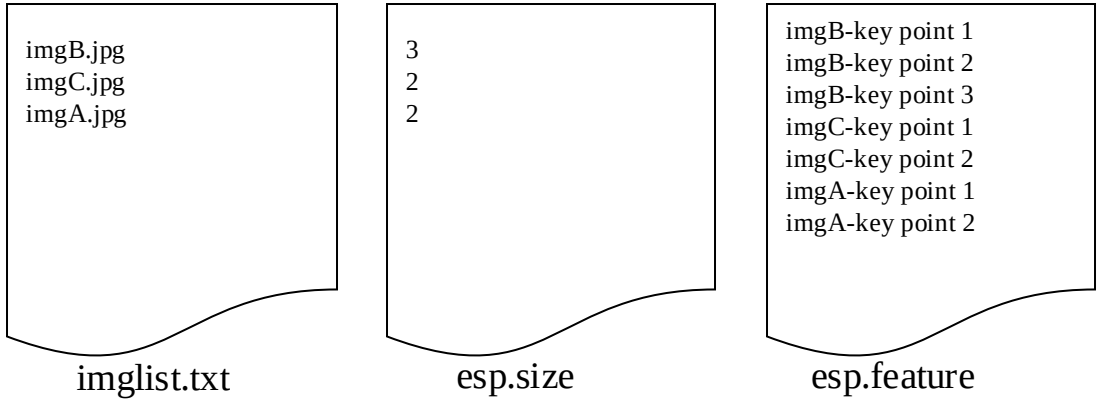


(b)

Figure 2: Samples from the clusters corresponding to a single visual word. (a) Two examples of clusters of Shape Adapted regions. (b) Two examples of clusters of Maximally Stable regions.

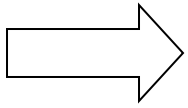
Step 2: Key Point Quantization

- Clustering 7 key points into 3 clusters
  - The cluster centers are: cnt1, cnt2, cnt3
  - Each center is a visual word: w1, w2, w3
- Find the nearest center to each key point



Step 2: Key Point Quantization

- imgA.jpg
  - 1st key point → w2
  - 2nd key point → w1
- imgB.jpg
  - 1st key point → w3
  - 2nd key point → w3
  - 3rd key point → w2
- imgC.jpg
  - 1st key point → w3
  - 2nd key point → w2



Bag-of-words Rep.

imgA.jpg: w2 w1

imgB.jpg: w3 w3 w2

imgC.jpg: w3 w2

Step 2: Key Point Quantization

- In this step, you need to save:
  - the cluster centers to a file. You will use this later on for quantizing key points of query images
  - bag-of-words representation of each image in “trec” format.

Bag-of-words Rep.

imgA.jpg: w2 w1

imgB.jpg: w3 w3 w2

imgC.jpg: w3 w2

```
<DOC>
<DOCNO>imgB</DOCNO>
<TEXT>
w3 w3 w2
</TEXT>
</DOC>
```

```
<DOC>
<DOCNO>imgA</DOCNO>
<TEXT>
w2 w1
</TEXT>
</DOC>
```

```
<DOC>
<DOCNO>imgC</DOCNO>
<TEXT>
w3 w2
</TEXT>
</DOC>
```

Step 3: Build index

- Refer to the indexing process of textual information retrieval
- IR libraries:
  - Lucene: http://lucene.apache.org/
  - Lemur: http://www.lemurproject.org/
- LIRE
  - http://www.semanticmetadata.net/lire/
  - http://www.lire-project.net/

Step 4: Extract key points for a query

Step 5: Generate BoVW for a query

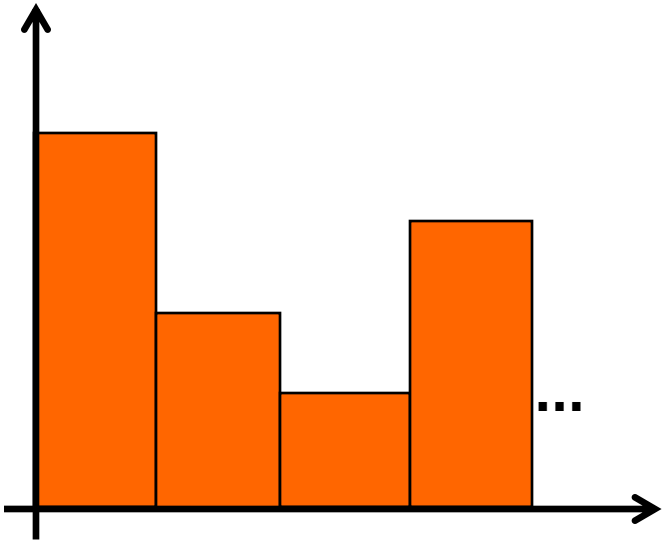
The mapped cluster ID for the 1st key point

The mapped cluster ID for the 2nd key point

…

The mapped cluster ID for the 1st key point

Step 5: Generate BoVW for a query



Step 6: Retrieval

- Refer to the indexing process of textual information retrieval

Video Google

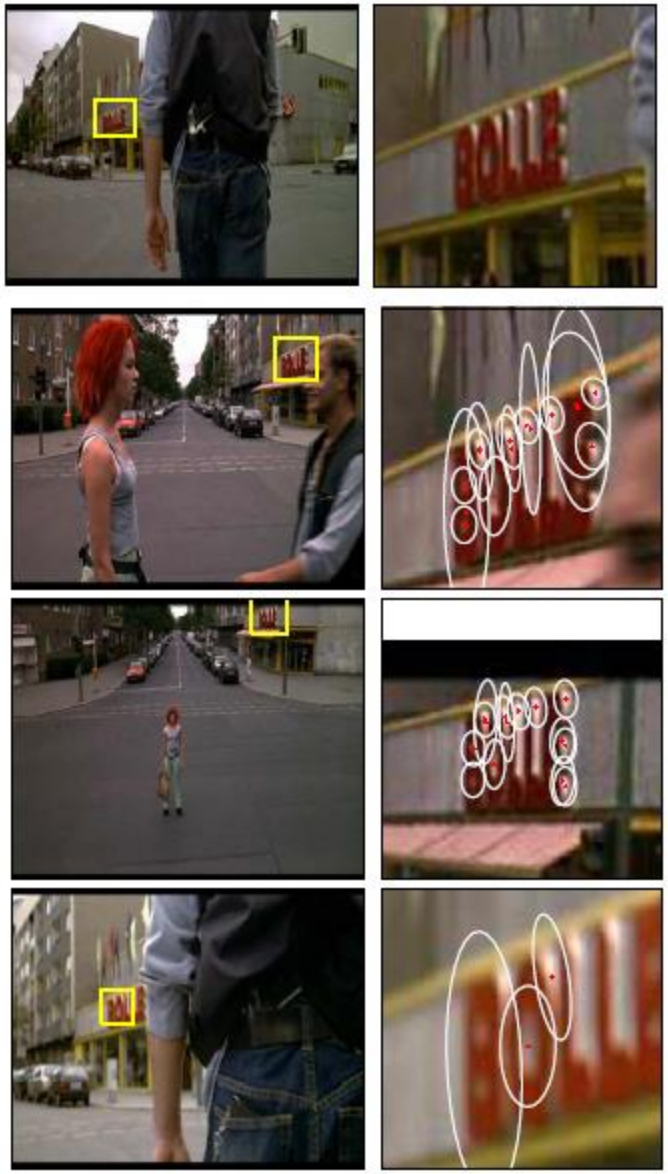
An example search in the movie 'Groundhog Day'

Query frame and query region (yellow). Click images to enlarge.

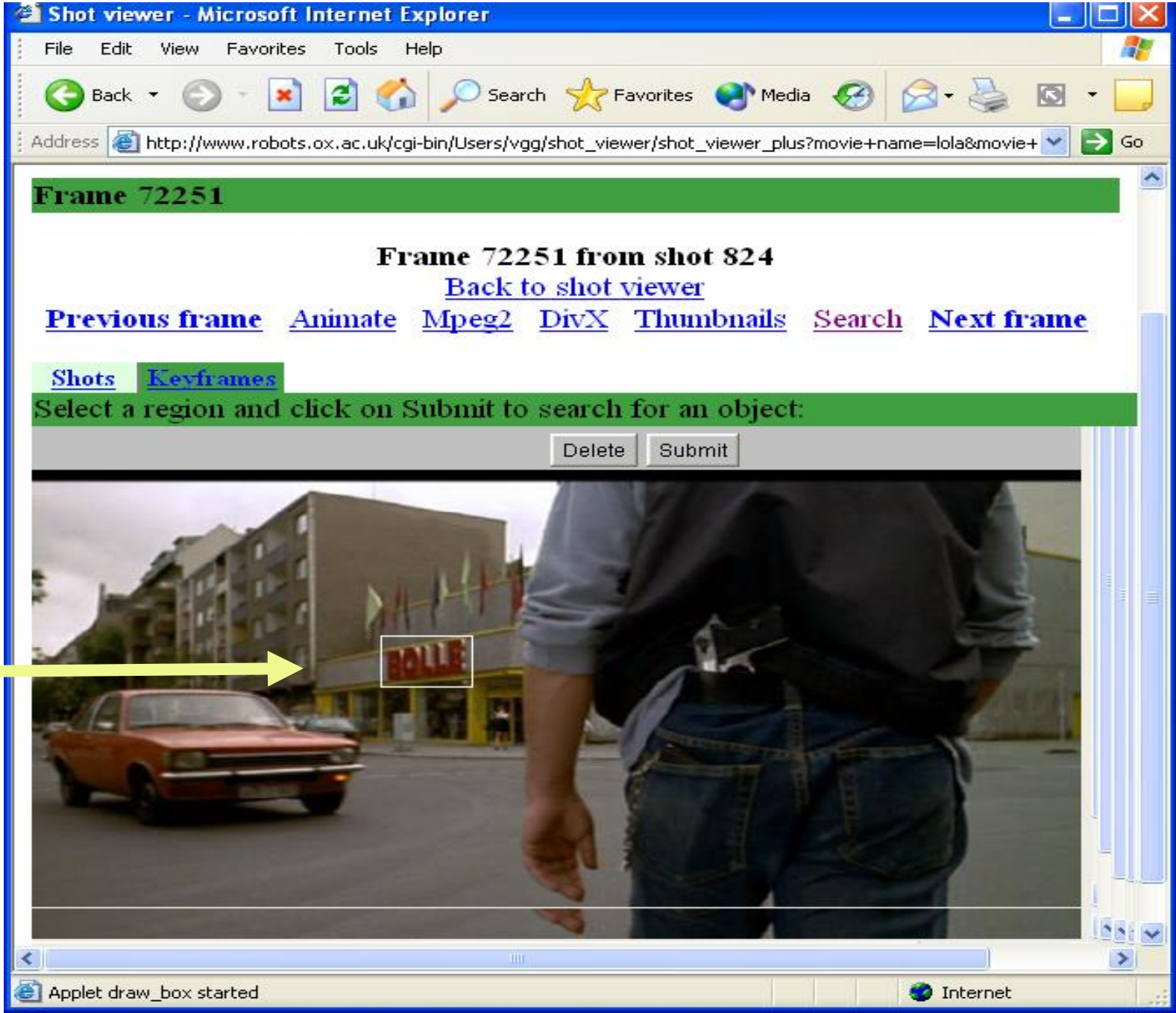
Frames (and region close-ups) from the first three retrieved shots. The query took 0.34 seconds. See more retrieved shots.

http://www.robots.ox.ac.uk/~vgg/research/vgoogle/index.html

J. Sivic and A. Zisserman, Video Google: A Text Retrieval Approach to Object Matching in Videos, ICCV 2003.


BOLLE

Query


Shot viewer - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.robots.ox.ac.uk/cgi-bin/Users/vgg/shot_viewer/shot_viewer_plus?movie+name=lola&movie+
Go
Frame 72251
Frame 72251 from shot 824
Back to shot viewer
Previous frame Animate Mpeg2 DivX Thumbnails Search Next frame
Shots Keyframes
Select a region and click on Submit to search for an object:
Delete Submit
ROLLE
Applet draw_box started
Internet

Object matches for frame 72251

Back to object search

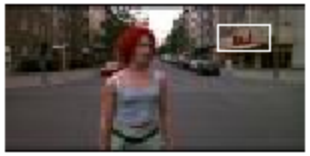
esults 1 to 20 of approximately 32 Time taken 0.180000 seconds

Frame 72251
Score 143

Frame 106426
Score 64

Frame 108401
Score 49

Frame 72226
Score 43



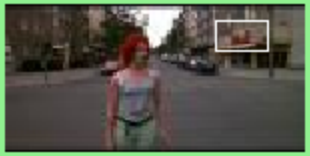
Frame 106601
Score 41

Frame 72301
Score 41

Frame 72326
Score 38

Frame 108351
Score 36

Frame 106676
Score 36



Frame 106576
Score 36

Frame 106401
Score 35

Frame 106451
Score 30

Object matches for frame 72251

Back to object search

esults 21 to 32 of approximately 32 Time taken 0.170000 seconds

Frame 106551
Score 17

Frame 106526
Score 14

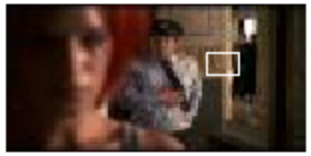
Frame 106476
Score 13

Frame 37126
Score 9

Frame 37101
Score 9

Frame 55051
Score 7

Frame 37201
Score 3



Frame 66176
Score 1

Frame 16001
Score 1

Frame 37076
Score 1

Frame 37001
Score 1

Frame 37051
Score 1

Matched image, frame 37101

Occluded !!!

Shots Keyframes

Select a region and click on Submit to search for an object:

Delete Submit

Edit View Favorites Tools Help

Back Search Favorites Media

http://www.robots.ox.ac.uk/cgi-bin/Users/vgg/shot_viewer/shot_viewer_plus?movie+name=lola&quadx1=69&quady1=88&quadx2=147&quady2=88&quadx3=147&quady3=137&

Original selection from frame 72251

Matched region from frame 55051

Matched image, frame 55051

Shots Keyframes

Select a region and click on Submit to search for an object:

Delete Submit

plet quad_outline started

Internet

Edit View Favorites Tools Help

Back Search Favorites Media

http://www.robots.ox.ac.uk/cgi-bin/Users/vgg/shot_viewer/shot_viewer_plus?frames+list=53001&Goto+frames=Go&movie+name=groundhog_day

rame 53001

Shot numbers 1 Go **Frame numbers** 0 Go

ame 53001

Frame 53001 from shot 294

Back to shot viewer

Previous frame Animate Mpeg2 DivX Thumbnails Search **Next frame**

Shots Keyframes

Select a region and click on Submit to search for an object:

Delete Submit

Back to object search

esults 1 to 10 of approximately 22 **Time taken 0.220000 seconds**

Shot number 294
Animate Thumbnails Mpeg2 DivX Search
Score 39.64


Start frame 52907


Key frame 53001


End frame 53028

Shot number 360
Animate Thumbnails Mpeg2 DivX Search
Score 22.27


Start frame 66136


Key frame 66451


End frame 66508

Shot number 140
Animate Thumbnails Mpeg2 DivX Search
Score 20.7


Start frame 28125


Key frame 28226


End frame 28228

Shot number 395
Animate Thumbnails Mpeg2 DivX Search
Score 17.41


Start frame 76139


Key frame 76576


End frame 77605

Shot number 372

Matched image, frame 28226

Shots Keyframes

Select a region and click on Submit to search for an object:

Delete Submit

Problem with bag-of-features

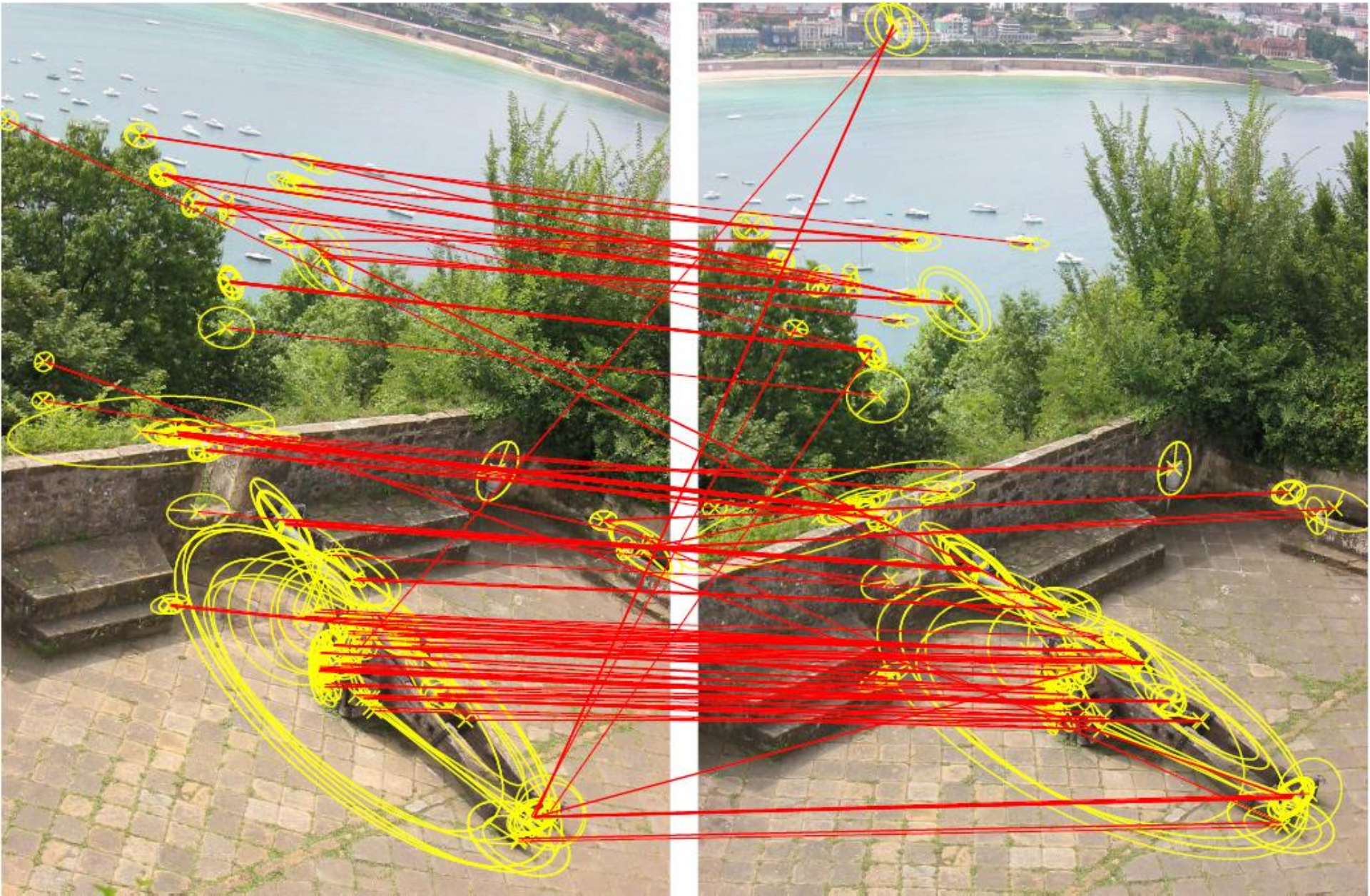
- The intrinsic matching scheme performed by BOF is weak
  - for a “small” visual dictionary: too many false matches
  - for a “large” visual dictionary: many true matches are missed

- No good trade-off between “small” and “large” !
  - either the Voronoi cells are too big
  - or these cells can’t absorb the descriptor noise

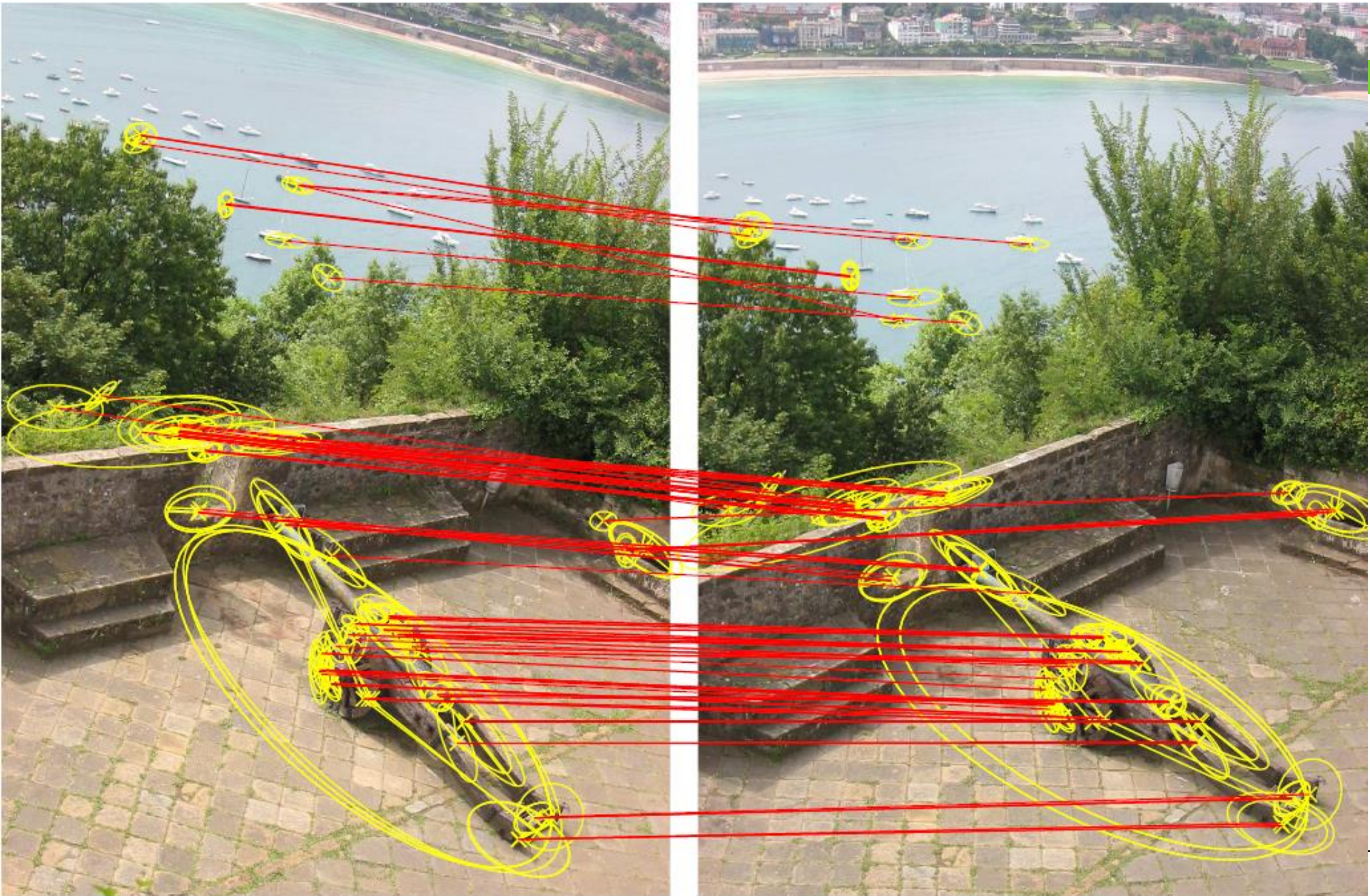
  → intrinsic approximate nearest neighbor search of BOF is not sufficient

www.ens-lyon.fr/LIP/Arenaire/ERVision/search_large.ppt

20K visual word: false matches



200K visual word: good matches missed



Need to Know

- Large scale retrieval
- Semantic gap
- Image/Video annotation/tagging/captioning
- Bag of visual words model