



UNIVERSITY OF
OXFORD

Towards On-the-fly Large Scale Video Search

Andrew Zisserman

Relja Arandjelović, Ken Chatfield, Omkar Parkhi

Visual Geometry Group,
Dept of Engineering Science,
University of Oxford

The Vision

All visual material (images, video) should be searchable for anything

- people, object categories, scene categories, particular objects, human actions and interactions, activities ...

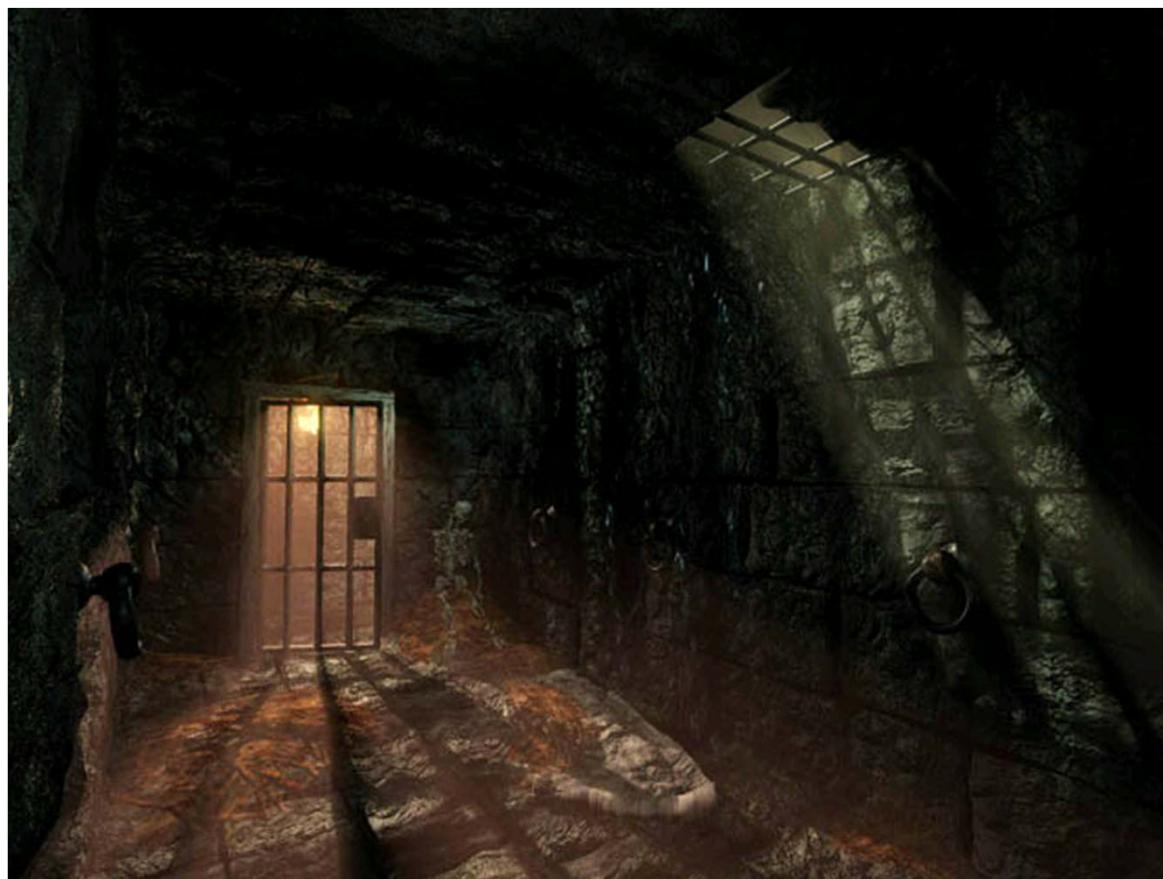
and retrieved with high precision and high recall



The Problem

There exist large data sets of images and videos lacking almost any annotation (apart from the date), e.g.

- archive datasets
- personal photo and video collections



The Problem

There exist large data sets of images and videos with sufficient annotations to retrieve thousands of examples of a query, e.g.

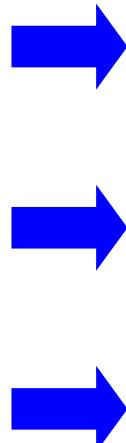
- Photos on web pages – Google Image Search
- Flickr, Facebook



The Solution

To harness some of the information from annotation rich sources and use it to enable searching of annotation starved datasets:

- efficiently, and
- in a scalable manner



On-the-fly search for faces

[Web](#) [Images](#) [Groups](#) [News](#) [Froogle](#) [more »](#)

Boris Johnson



Download images and detect faces



Person X
classifier



ranked frames
(BBC corpus)

Can search for anyone

Video dataset: BBC TV

- 4372 broadcasts from BBC 1, 2, 3 & 4
- Programmes from late 2011 to early 2012 from prime time slot (7pm-12pm) over five months
- 3007 hours of video represented by 1 frame per second
- 11M seconds of data, 3M keyframes
- Frames are 480 x 270 pixels



Instance Search – Example ‘Mona Lisa’

VISOR

mona lisa

+

BBCb

Search

Leonardo

Leonardo

Leonardo

Leonardo

Leonardo

bbc one

bbc one

bbc one

bbc one

bbc one

The National Lottery...

The National Lottery...

Leonardo

World News Today

World News Today

Leonardo

Strictly Come Dancin...

Strictly Come Dancin...

Leonardo

FREDIE NOTHROP
SOPHIE
ARTURO REIGALD
Zoe Walker
Matt Groombridge
JORDAN

Leonardo

Outline

On-the-fly *instance* search

- Specific places/scenes/objects e.g. White house, Mona Lisa, HSBC logo



On-the-fly *category* search

- Object and scene categories, e.g. cars, crowds, forest



On-the-fly *face* search

- Particular people and attributes, e.g. Obama, moustache



Long history of learning from Google images

- Berg & Forsyth, CVPR 06
- Fergus *et al.*, ICCV 05
- Li *et al.*, CVPR 07
- Liu *et al.*, ACM MM 09
- Schroff *et al.*, ICCV 07
- Sivic & Zisserman, ICCV 03, Proc. IEEE 08
- Torresani *et al.*, ECCV 10, NIPS 11

1. On-the-fly Instance Search

Instance Search

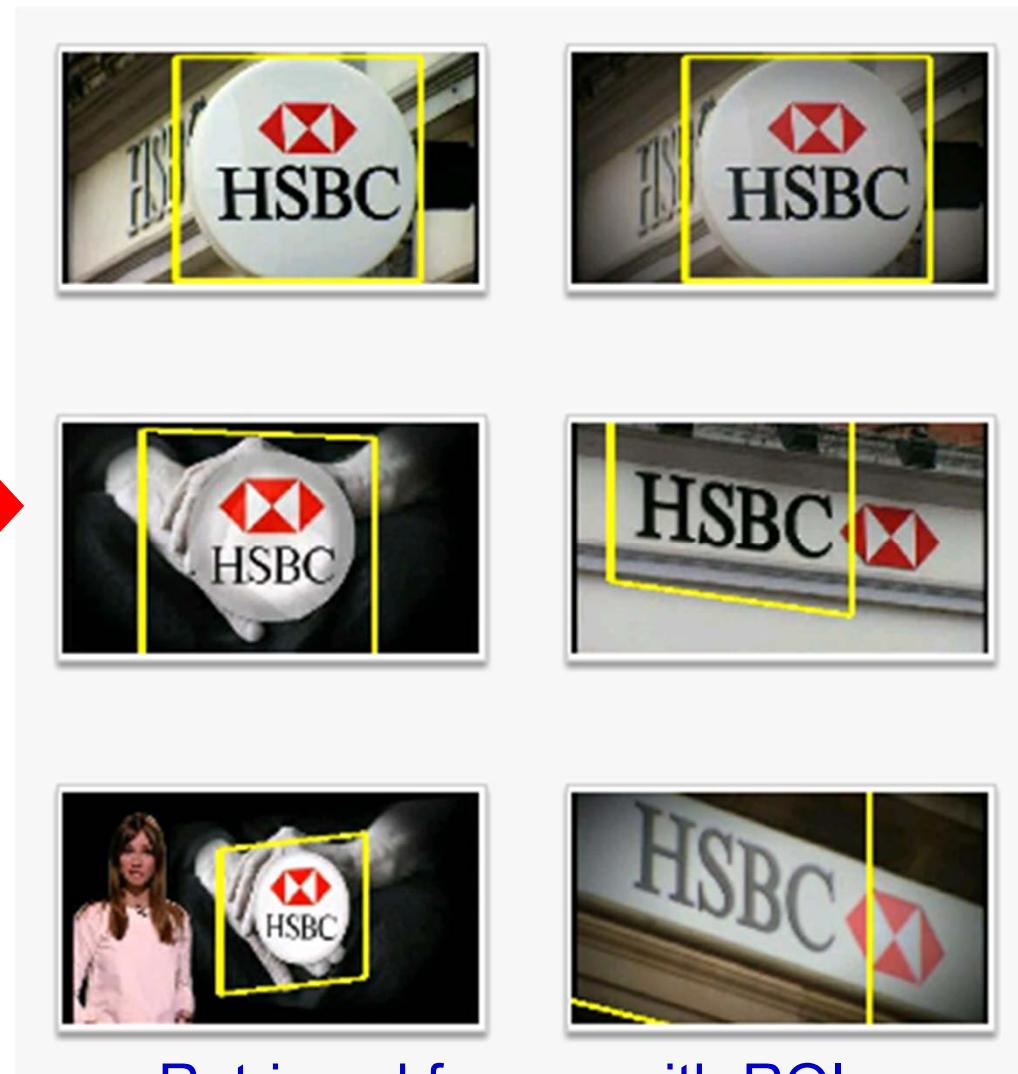
Query by example image: retrieve specific objects, unaffected by: scale, viewpoint, lighting, partial occlusion

Query image



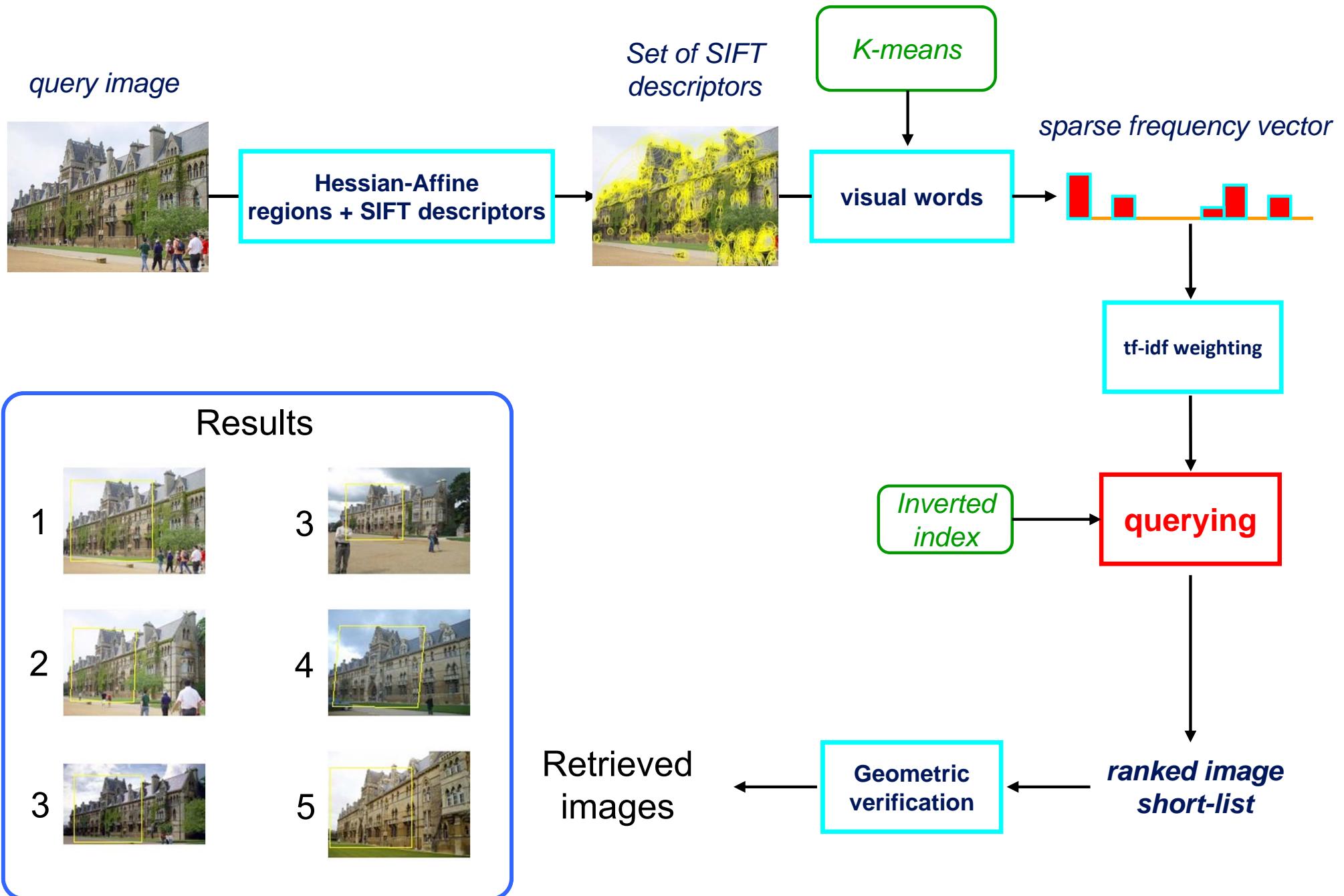
Download from web
(external)

(not category recognition)

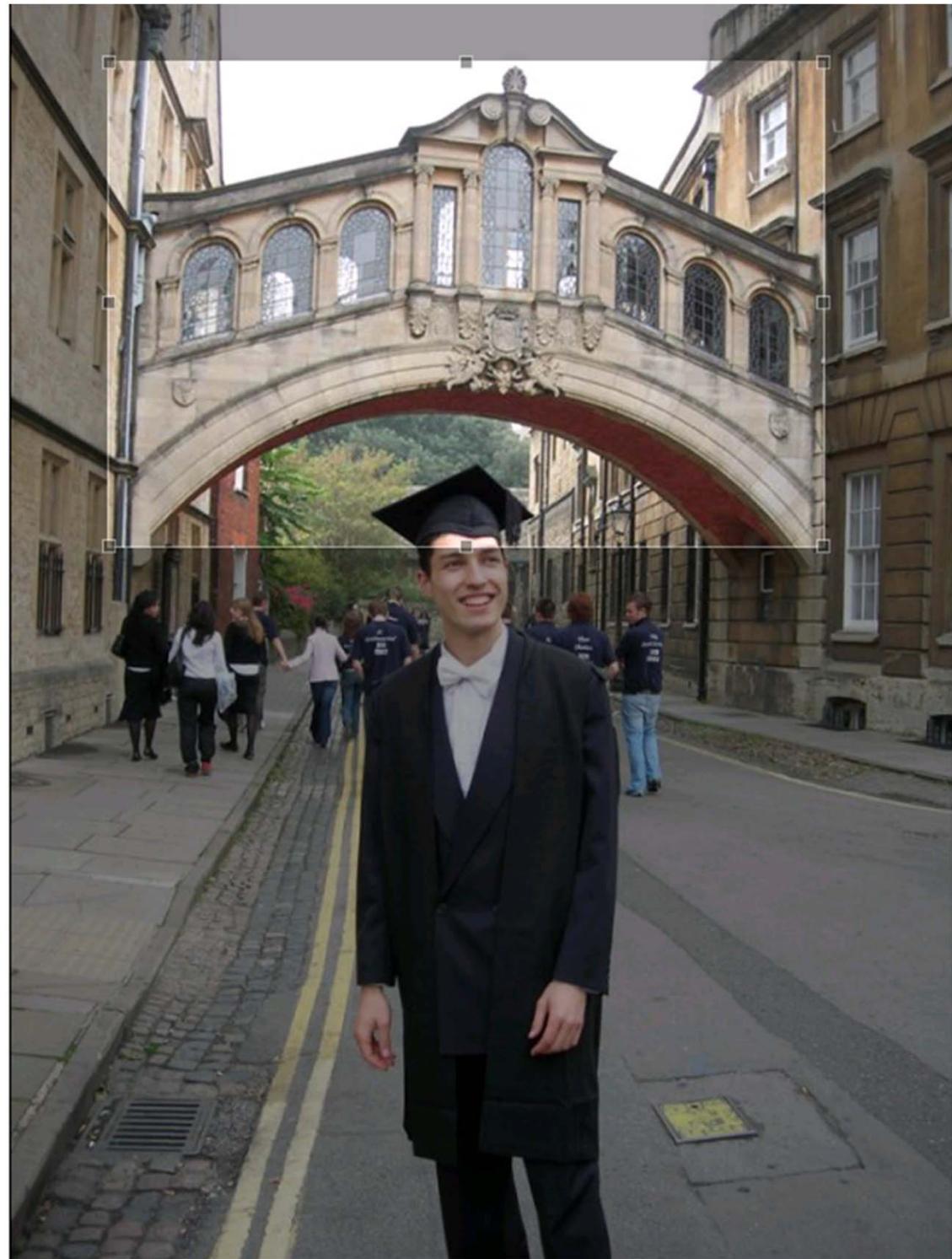


Retrieved frames with ROI

Visual engine: bag of visual words particular object retrieval



Example



Search

Search results 1 to 20 of 104844

1



ID: oxc1_hertford_000011

Score: 1816.000000

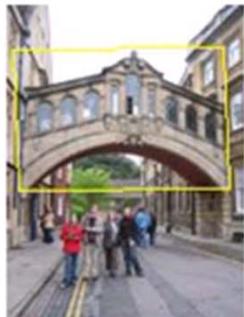
Putative: 2325

Inliers: 1816

Hypothesis: 1.000000 0.000000 0.000015 0.000000 1.000000 0.000031

[Detail](#)

2



ID: oxc1_all_souls_000075

Score: 352.000000

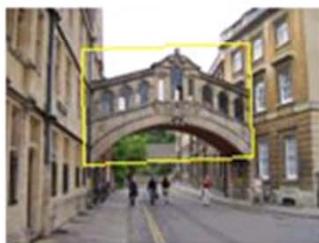
Putative: 645

Inliers: 352

Hypothesis: 1.162245 0.041211 -70.414459 -0.012913 1.146417 91.276093

[Detail](#)

3



ID: oxc1_hertford_000064

Score: 278.000000

Putative: 527

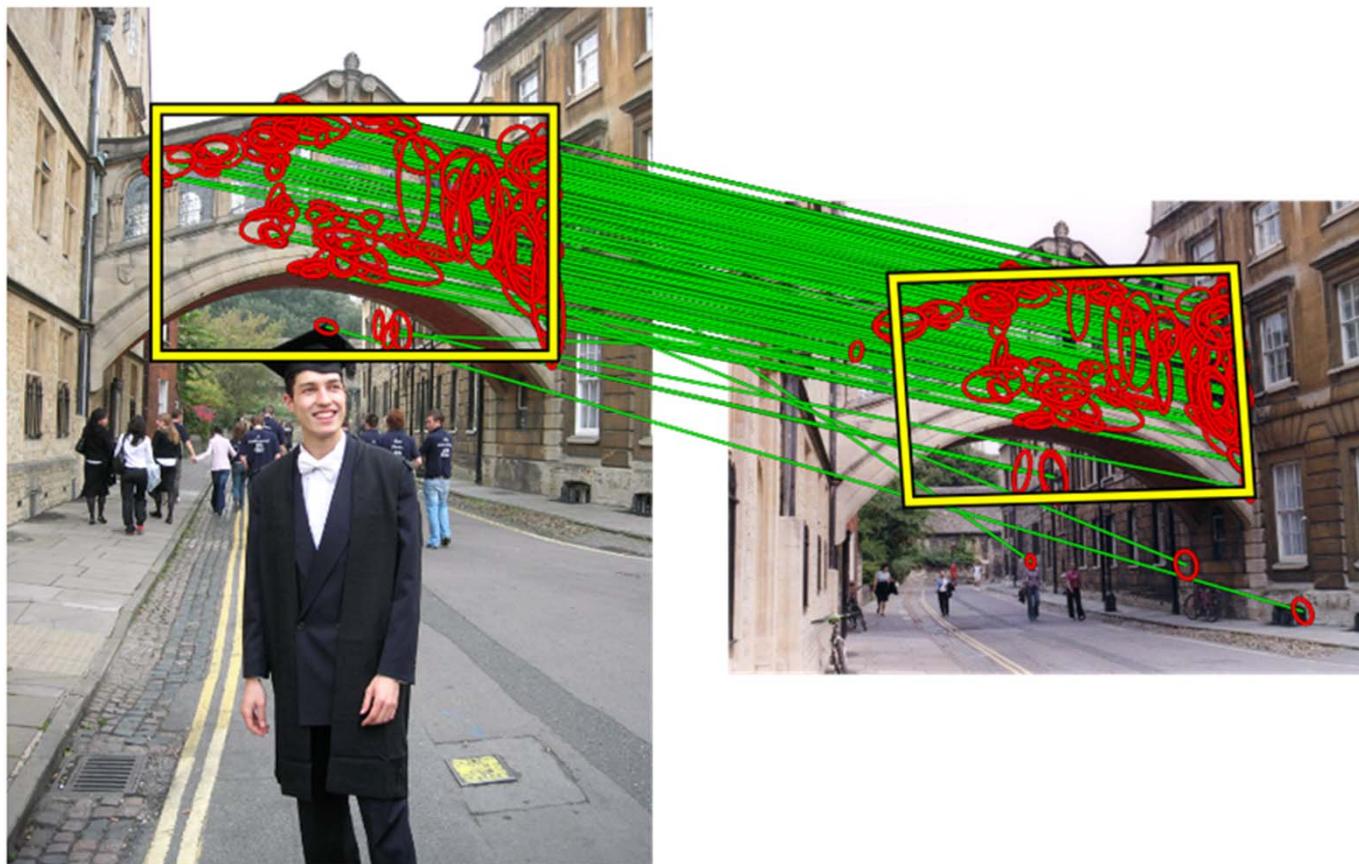
Inliers: 278

Hypothesis: 0.928686 0.026134 169.954620 -0.041703 0.937558 97.962112

[Detail](#)

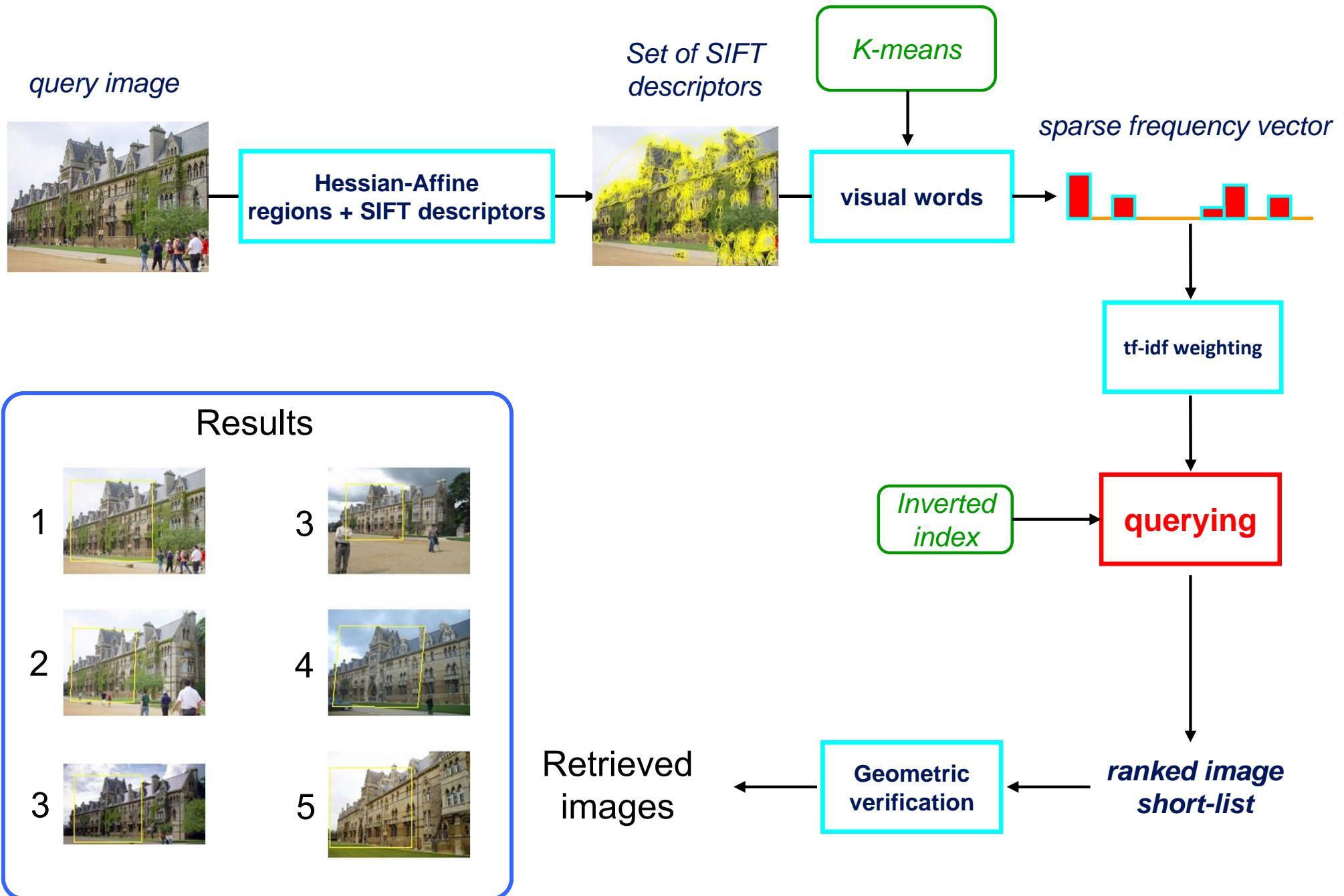
Spatial verification

Use the spatial distribution of the detections in the image to improve retrieval quality – re-rank short list by number of matches



SIFT matches consistent with an affine transformation

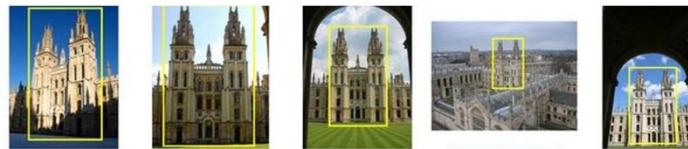
Visual engine: bag of visual words particular object retrieval



Oxford buildings dataset

- Landmarks plus queries used for evaluation

All Soul's



Ashmolean



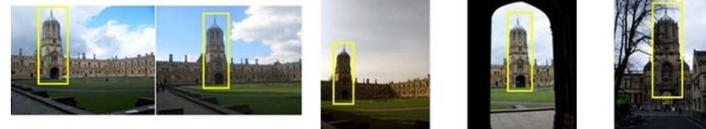
Balliol



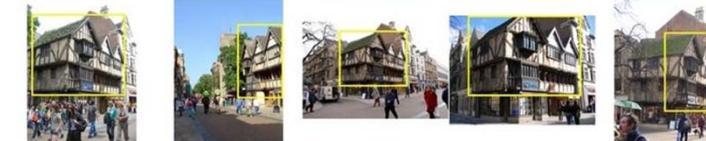
Bodleian



Thom Tower



Cornmarket



Bridge of Sighs



Keble



Magdalen



University Museum

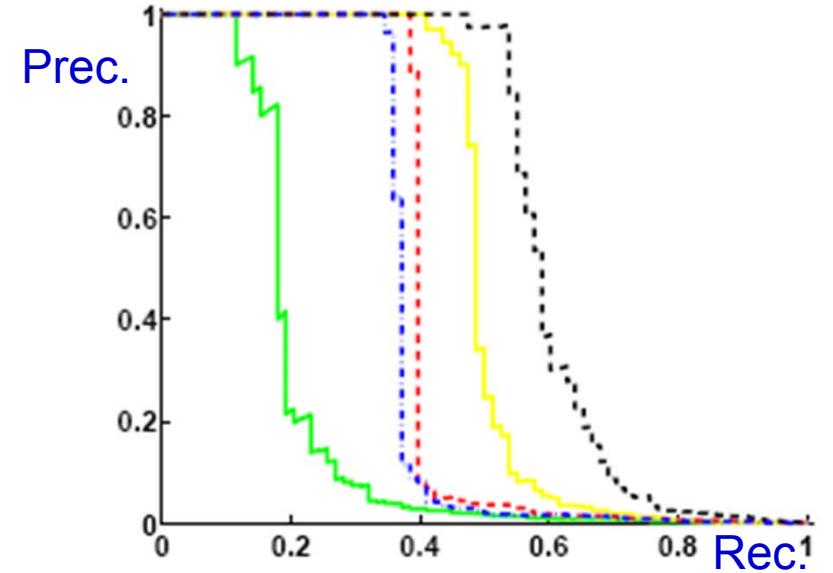


Radcliffe Camera



- Ground truth obtained for 11 landmarks over 5062 images
- Evaluate performance by Precision - Recall curves

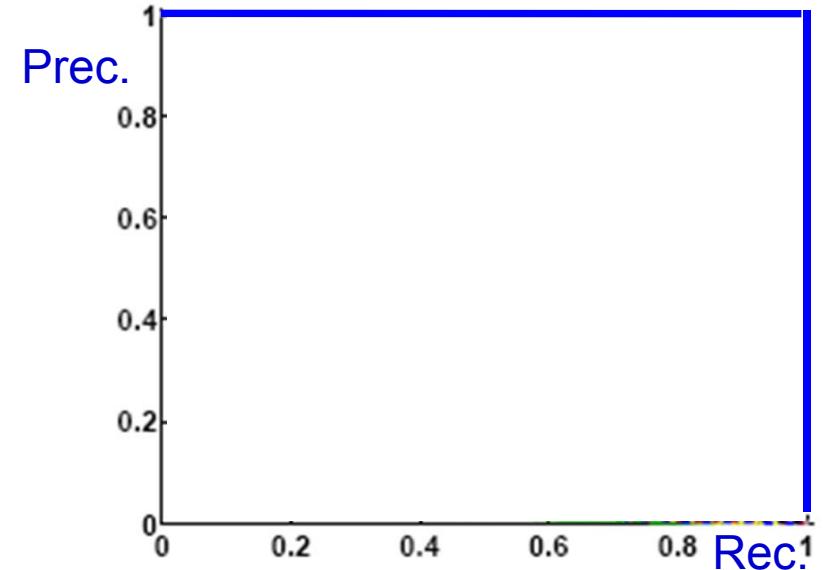
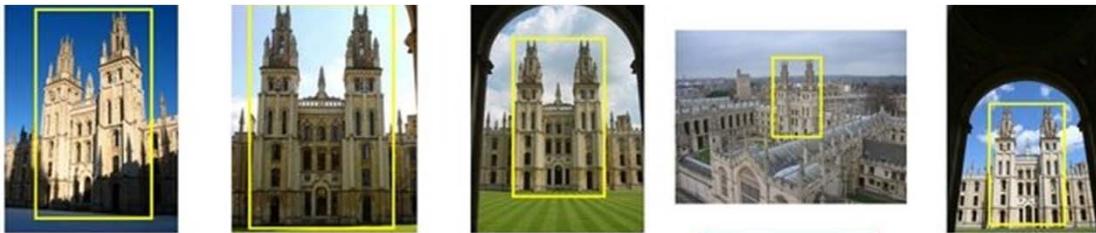
Query images



- high precision at low recall (like Google)
- variation in performance over query
- none retrieve all instances

Total Recall

Query images



Retrieve **all** occurrences of an object in the corpus

Improving SIFT

- Histogram measures such as Hellinger or χ^2 , outperform Euclidean distance when comparing histograms (e.g. image classification, object category detection, texture classification etc).
- And these can be implemented efficiently using approximate feature maps in the case of additive kernels
- SIFT is a histogram: can performance be boosted using a better distance measure?

Hellinger distance

Hellinger kernel (Bhattacharyya's coefficient) for L1 normalized histograms x and y :

$$H(x, y) = \sum_{i=1}^n \sqrt{x_i y_i}$$

Distances and kernels x and y L2 normalized

$$\begin{aligned} d_E(x, y)^2 &= \|x - y\|_2^2 \\ &= \|x\|_2^2 + \|y\|_2^2 - 2x^\top y \\ &= 2 - 2 \sum_{i=1}^n x_i y_i \end{aligned}$$


kernel

Hellinger distance

Hellinger kernel (Bhattacharyya's coefficient) for L1 normalized histograms x and y :

$$H(x, y) = \sum_{i=1}^n \sqrt{x_i y_i}$$

Distances and kernels x and y L1 normalized

$$\begin{aligned} d_E(\sqrt{x}, \sqrt{y})^2 &= \|\sqrt{x} - \sqrt{y}\|_2^2 \\ &= \|\sqrt{x}\|_2^2 + \|\sqrt{y}\|_2^2 - 2\sqrt{x}^\top \sqrt{y} \\ &= 2 - 2 \sum_{i=1}^n \underbrace{\sqrt{x_i y_i}}_{\text{kernel}} \end{aligned}$$

Hellinger distance

Hellinger kernel (Bhattacharyya's coefficient) for L1 normalized histograms x and y :

$$H(x, y) = \sum_{i=1}^n \sqrt{x_i y_i}$$

Explicit feature map of x into x' :

- L1 normalize x
- element-wise square root x to give x'

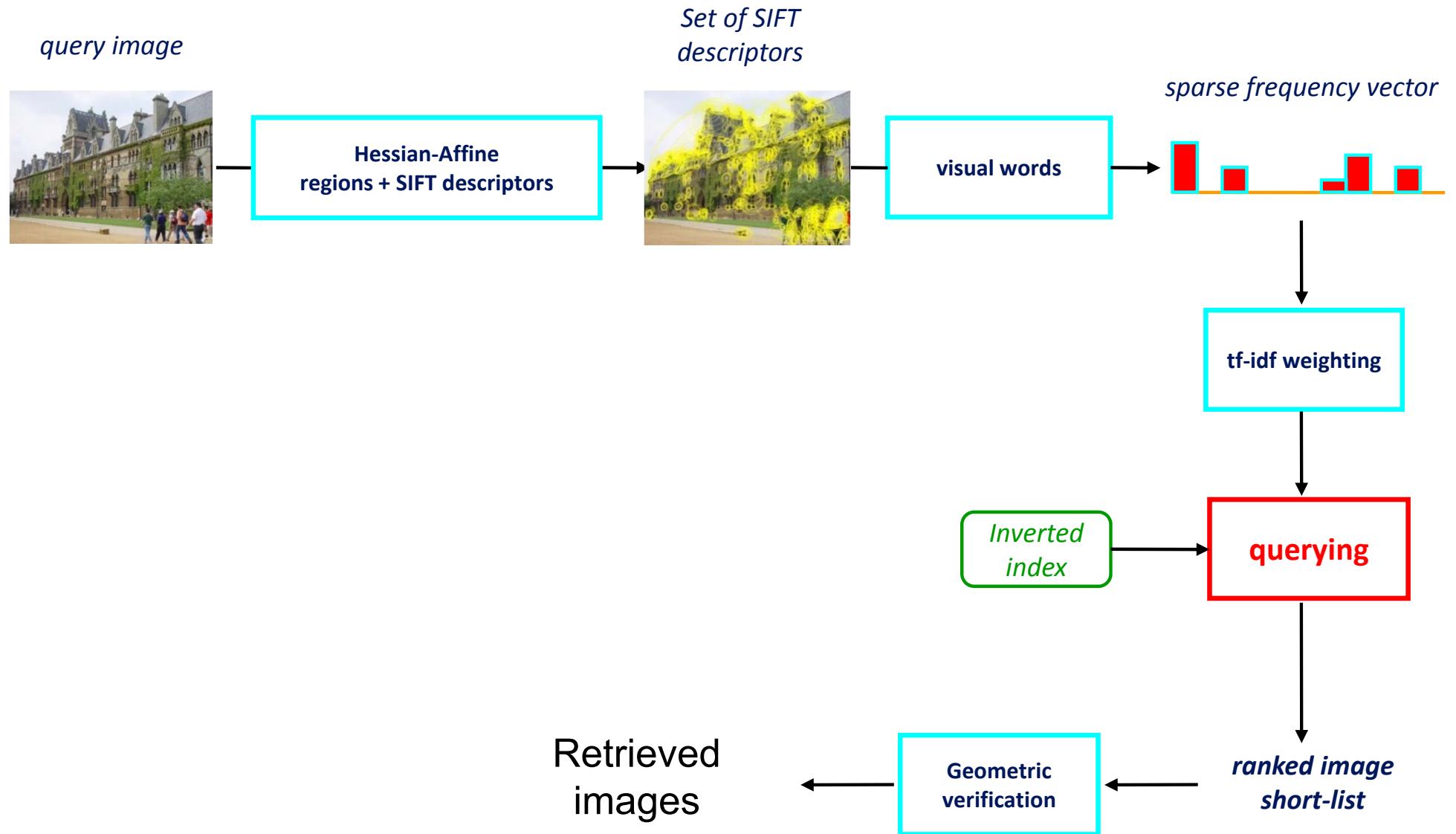
then x' is L2 normalized

Euclidean distance in the feature map space is equivalent to Hellinger distance in the original space, since:

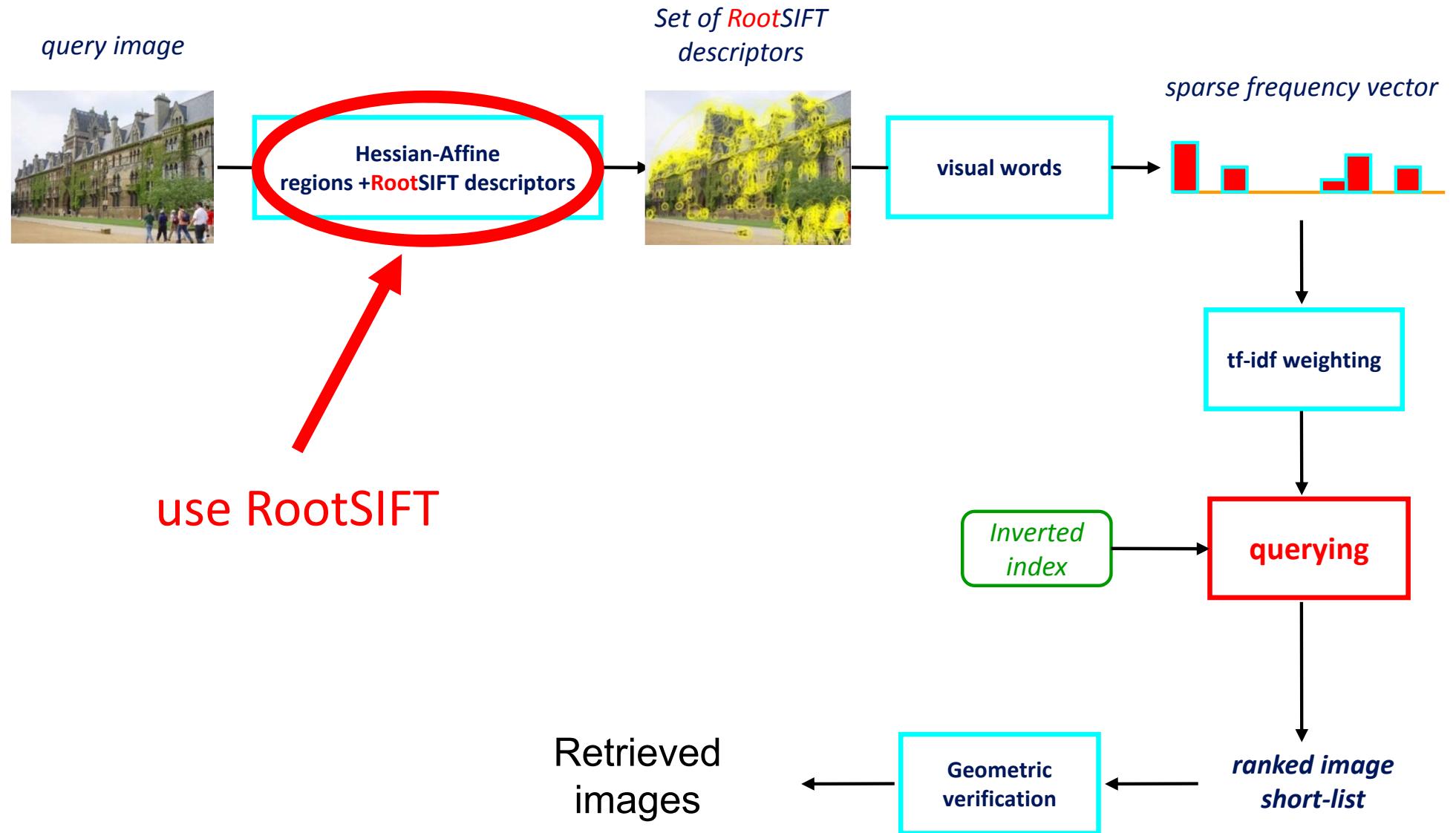
$$\mathbf{x}'^\top \mathbf{y}' = H(\mathbf{x}, \mathbf{y})$$

} RootSIFT

Bag of visual words particular object retrieval



Bag of visual words particular object retrieval



RootSIFT: mAP performance

Philbin *et al.* 2007: bag of visual words either with

- tf-idf ranking,
- or tf-idf ranking and spatial reranking

Evaluate on:

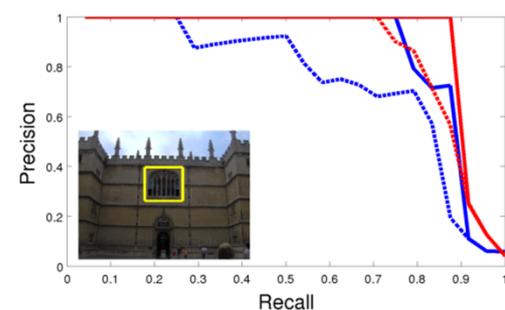
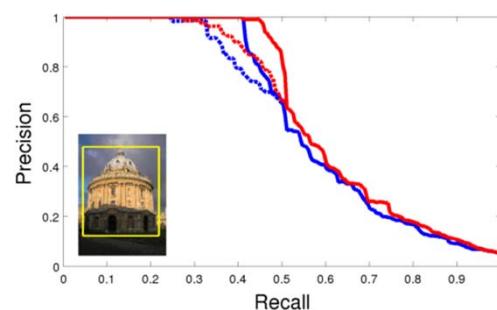
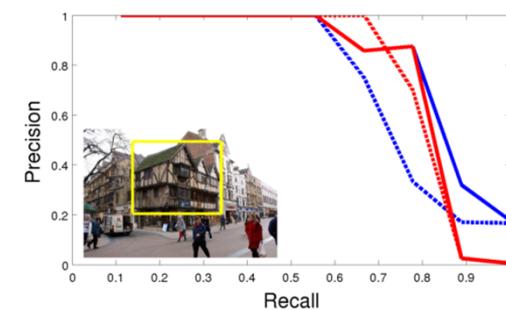
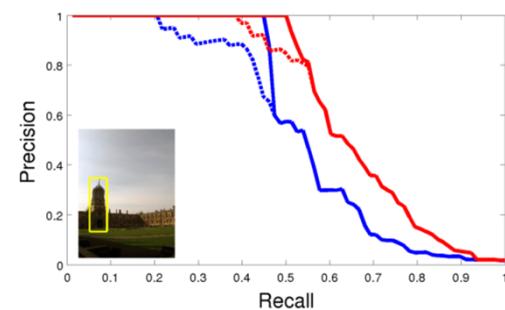
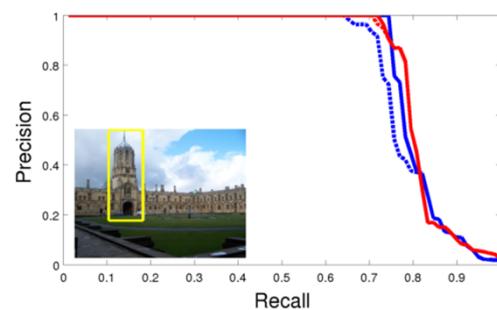
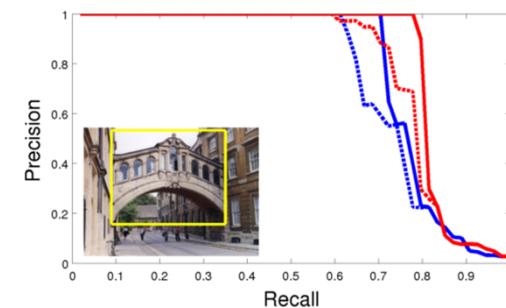
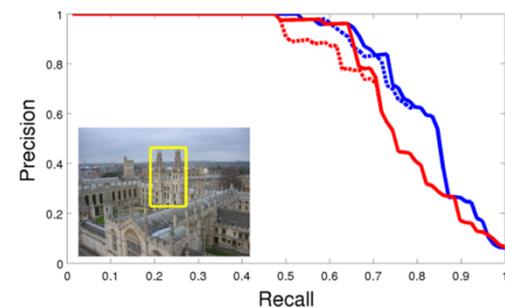
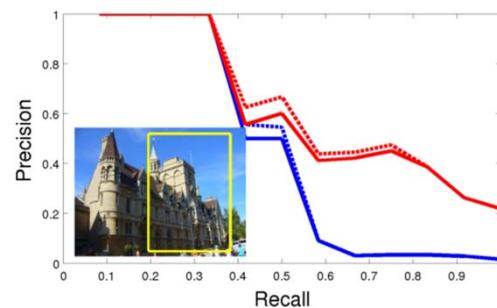
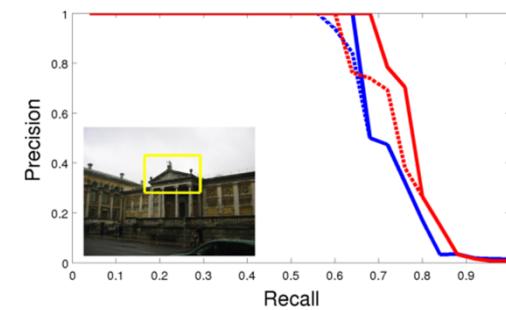
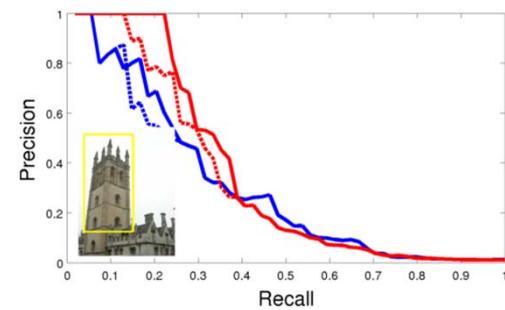
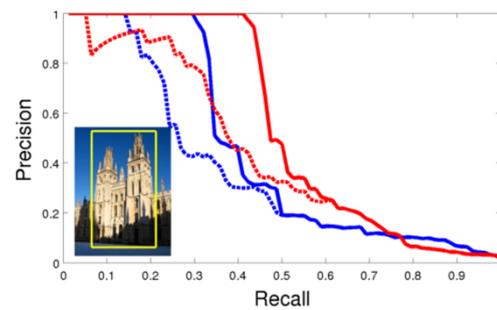
- Oxford 5k buildings,
- and on Oxford105k (5k buildings + 100k distractor images)

mean Average Precision (mAP)

Retrieval method	Oxford 5k	Oxford 105k
SIFT: tf-idf ranking	0.636	0.515
SIFT: tf-idf with spatial reranking	0.672	0.581
RootSIFT: tf-idf ranking	0.683	0.581
RootSIFT: tf-idf with spatial reranking	0.720	0.642



RootSIFT: results, Oxford 5k

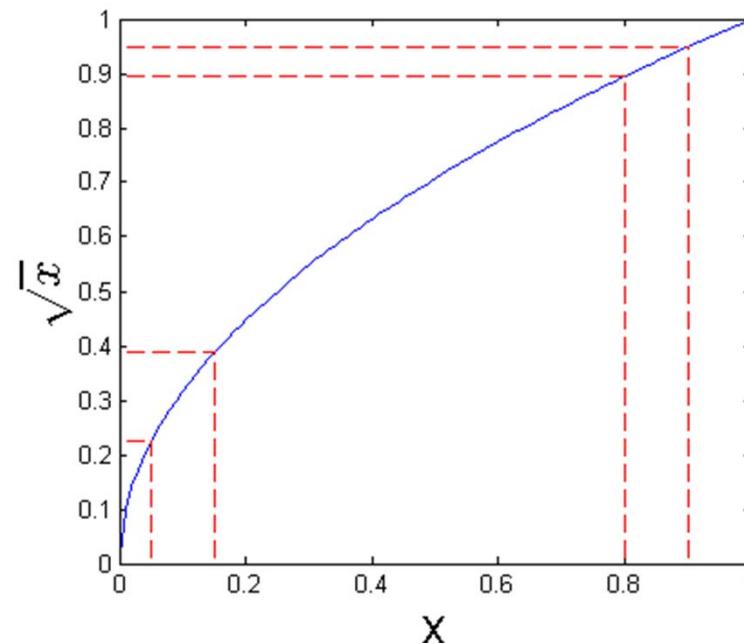


tfidf: dashed
spatial rerank: solid
RootSIFT: red
SIFT: blue

Why does it work better?

Intuition: Euclidean distance can be dominated by large bin values. Hellinger distance is more sensitive to smaller bin values

$$H(x, y) = \sum_{i=1}^n \sqrt{x_i y_i}$$



RootSIFT Advantages

- Extremely simple to implement and use
 - one line of Matlab code to convert SIFT to RootSIFT:
$$\text{rootsift} = \text{sqrt}(\text{sift} / \text{sum}(\text{sift}));$$
- Conversion from SIFT to RootSIFT can be done on-the-fly
- No need to re-compute stored SIFT descriptors for large image datasets
- Applications throughout computer vision
 - k-means, approximate nearest neighbours, soft-assignment to visual words, Fisher vector coding, PCA, descriptor learning, hashing methods, product quantization etc.

There is a magic bullet

Other significant improvements ...

Discriminative learning of descriptors, a better SIFT, e.g.

- Winder et al CVPR 09, Brown et al PAMI 2011, Philbin et al ECCV 10
- Convex learning of pooling regions and projection - Simonyan et al ECCV12

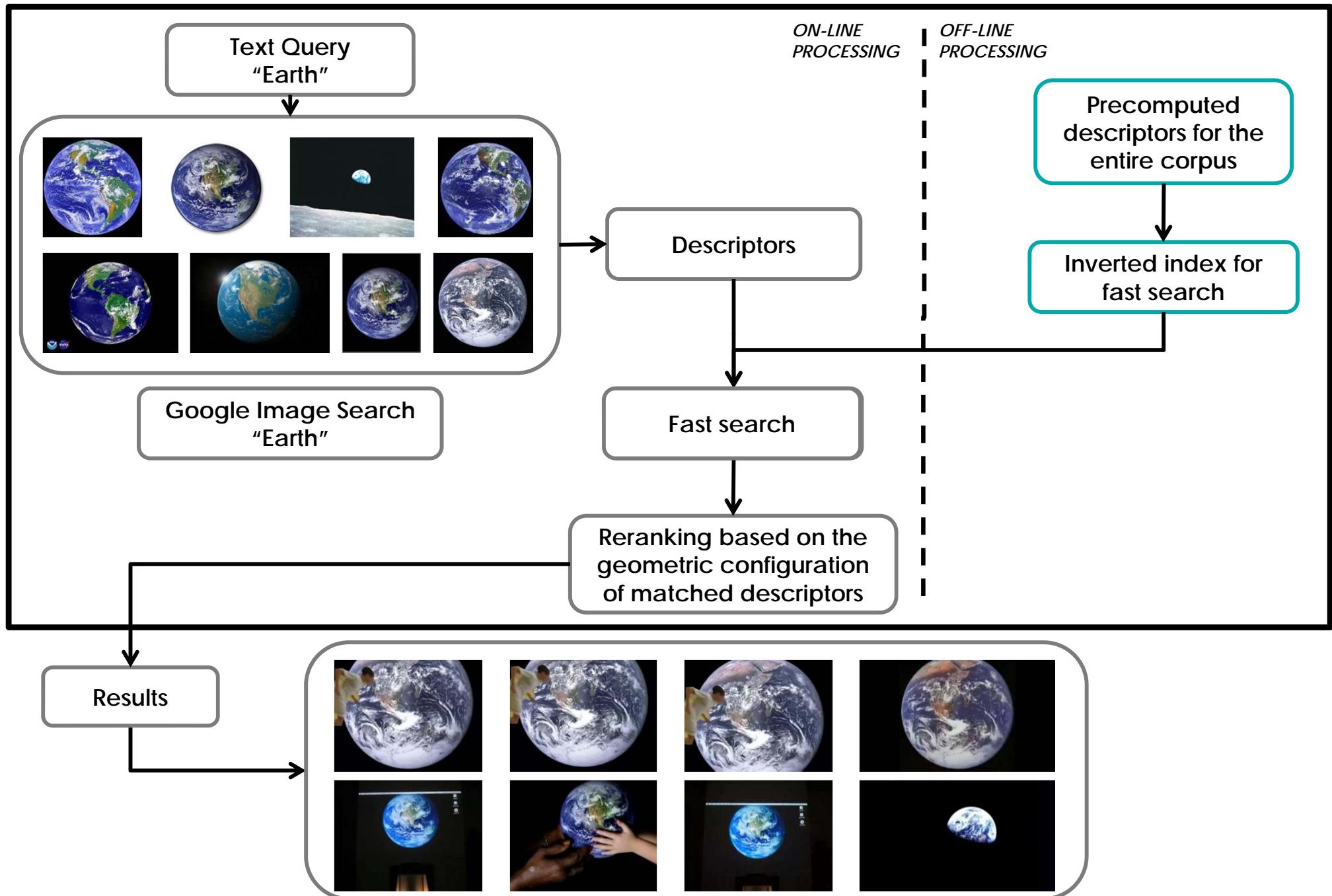
Closer representation of descriptor (reduce quantization errors), e.g.

- Philbin et al CVPR 08, Jegou et al ECCV 08, Mikulik et al ECCV 10
- Product Quantization on residuals – Jegou et al PAMI 2011

Query expansion, e.g.

- Chum et al ICCV 07, Turcot & Lowe ICCV 09 (workshop), Chum et al CVPR11
- Discriminative query expansion – Arandjelovic & Zisserman CVPR 2012

On-the-fly Instance Search



How are positive images used for instance search?

Compute a BOW feature vector x_i for each positive image

Possibilities:

- Average feature vectors x_i into q and query with q
- Query with each feature vector x_i in turn and combine ranked results



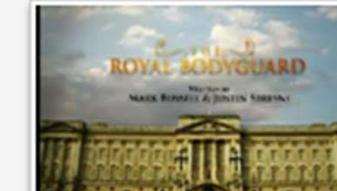
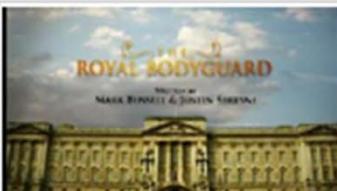
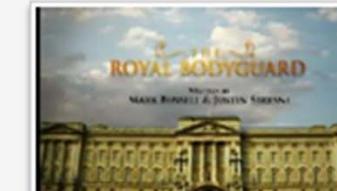
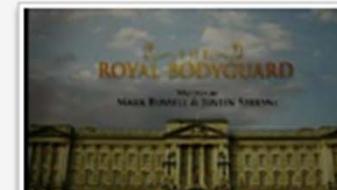
Video dataset: BBC TV

- 4372 broadcasts from BBC 1, 2, 3 & 4
- Programmes from late 2011 to early 2012 from prime time slot (7pm-12pm) over five months
- 3007 hours of video represented by 1 frame per second
- 11M seconds of data, 3M keyframes
- Frames are 480 x 270 pixels



Instance Search – Example ‘Buckingham Palace’

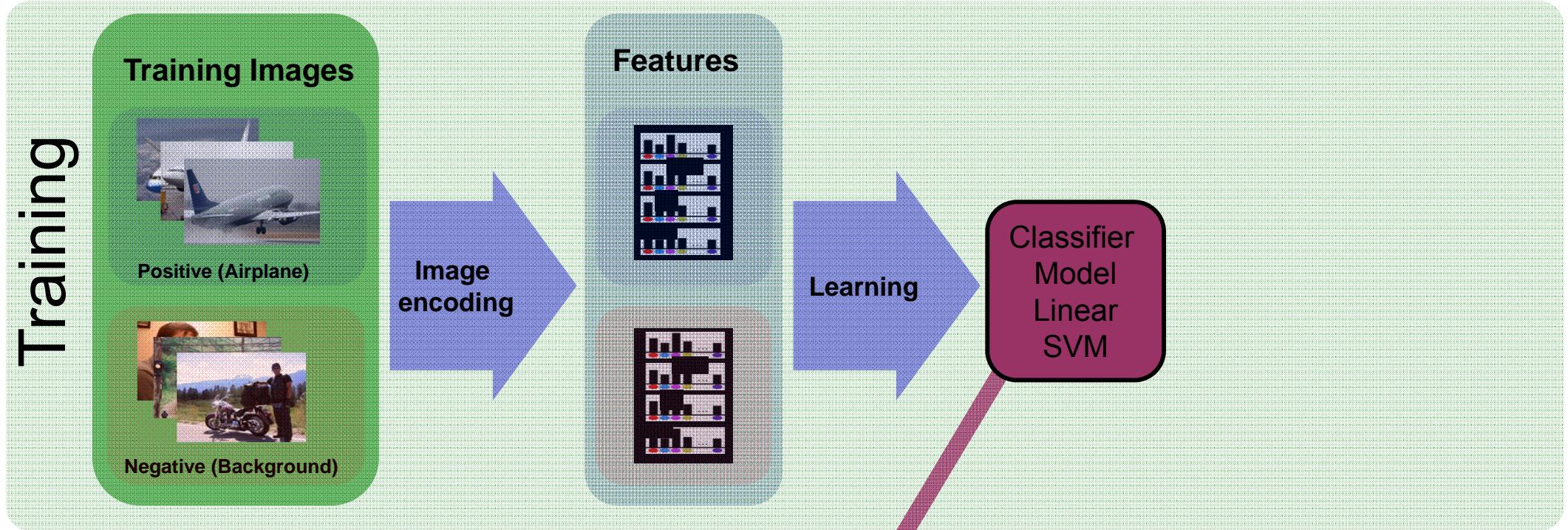
VISOR  buckingham palace BBCb |  Search  

 BBC News at Six	 The One Show	 The One Show	 The One Show	 World News Today
 The Diamond Queen	 The Queen's Palaces	 BBC London News	 The Royal Bodyguard	 The One Show
 The Royal Bodyguard	 The Diamond Queen	 The One Show	 The Royal Bodyguard	 The Royal Bodyguard
 BBC Weekend News	 The Royal Bodyguard	 The Royal Bodyguard	 Regimental Stories	 The Diamond Queen

2. On-the-fly Category Search

Image classification

Training



Testing

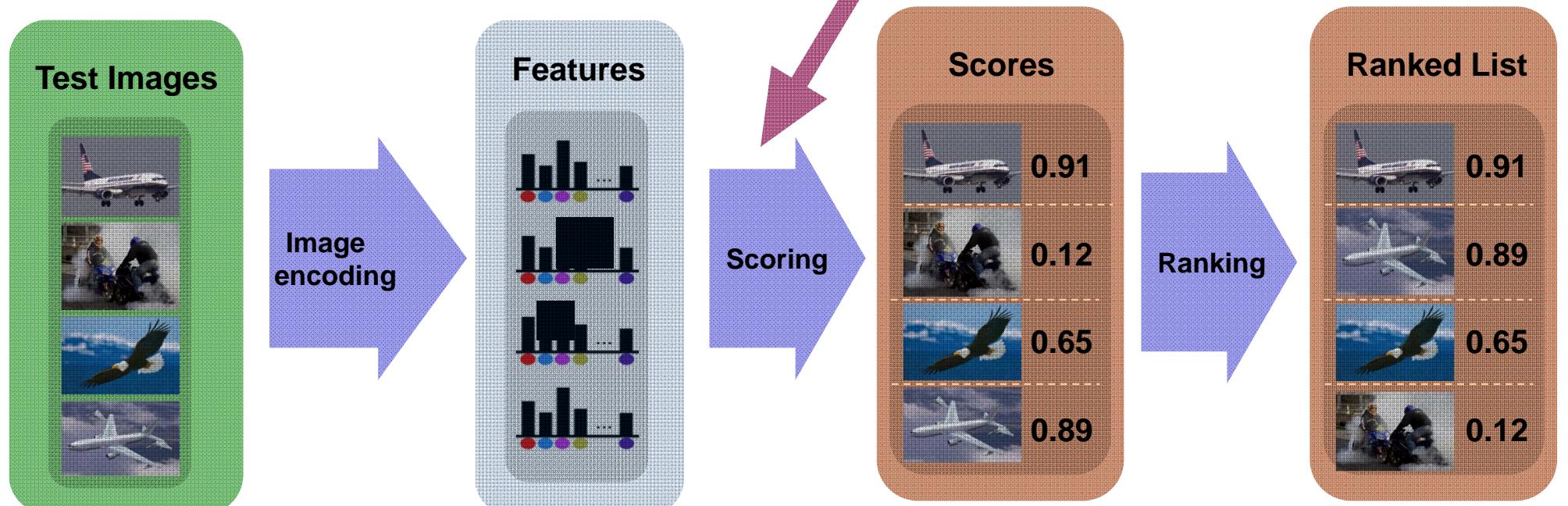


Image classification

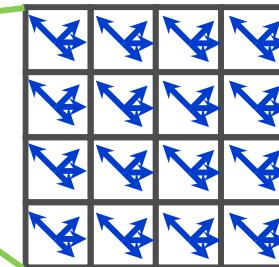
Classify an image by the objects/scenes it contains

- Review recent progress in encoding methods
- Choice of encoding method – trade off:
 - memory footprint
 - speed
 - performance

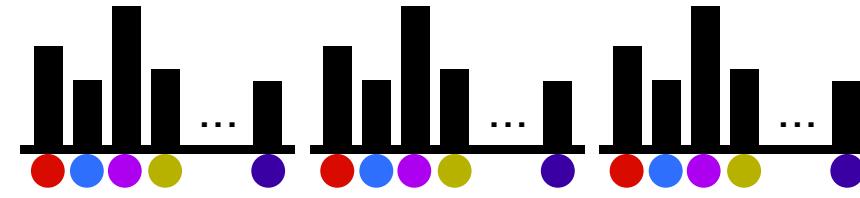
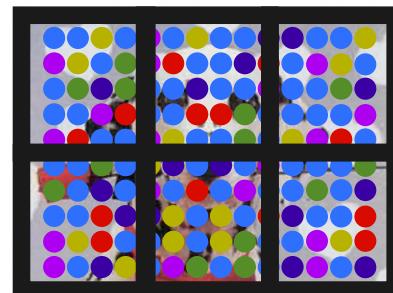
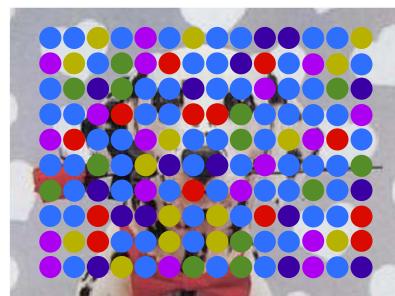
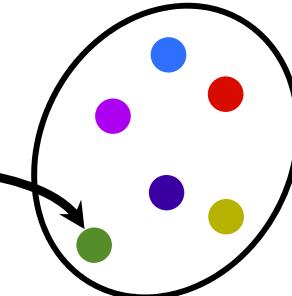
Image Encoding

Dense SIFT features

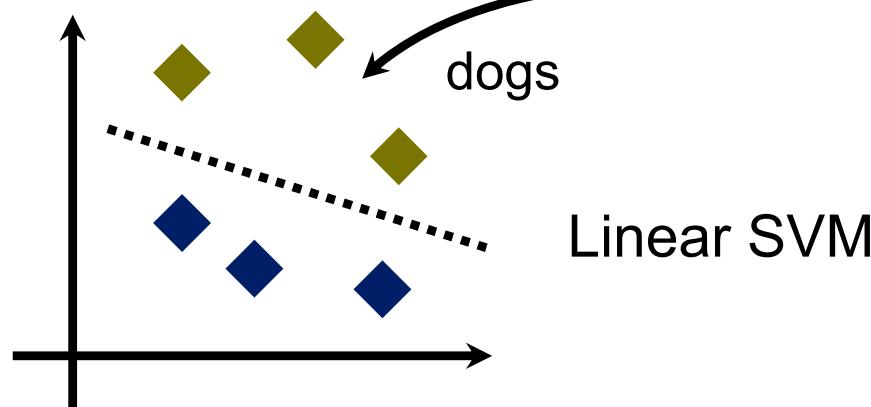
- Bag of Visual Words (BOW) pipeline



VQ



...



[Luong & Malik, 1999]

[Varma & Zisserman, 2003]

[Csurka et al, 2004]

[Vogel & Schiele, 2004]

[Jurie & Triggs, 2005]

[Lazebnik et al, 2006]

[Bosch et al, 2006]

Evolution of encodings ...

Soft and sparse assignments, e.g.

- Philbin et al CVPR 08, Gemert et al ECCV 08,
- Locality-constrained linear coding (LLC) – Wang et al CVPR 10

Representing SIFT distribution mean in voronoi cell, e.g.

- super-vector coding – Zhou et al ECCV 10
- VLAD – Jegou et al CVPR 10

Representing SIFT distribution mean and covariance in voronoi cell, e.g.

- Fisher vector – Perronnin et al CVPR 07 & 10, ECCV 10

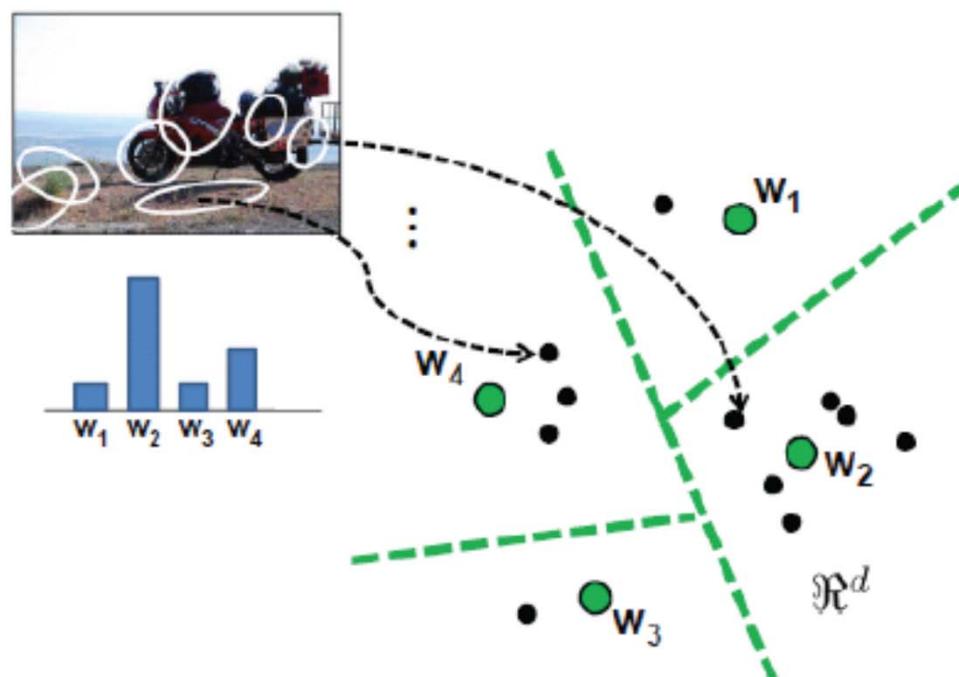
Improvements to normalization, PCA, whitening for VLAD/FV

- Chen et al 2011, Jegou & Chum ECCV 12
- All about VLAD – Arandjelovic & Zisserman CVPR 13

Comparison & code: “The devil is in the details”, Chatfield et al, BMVC11

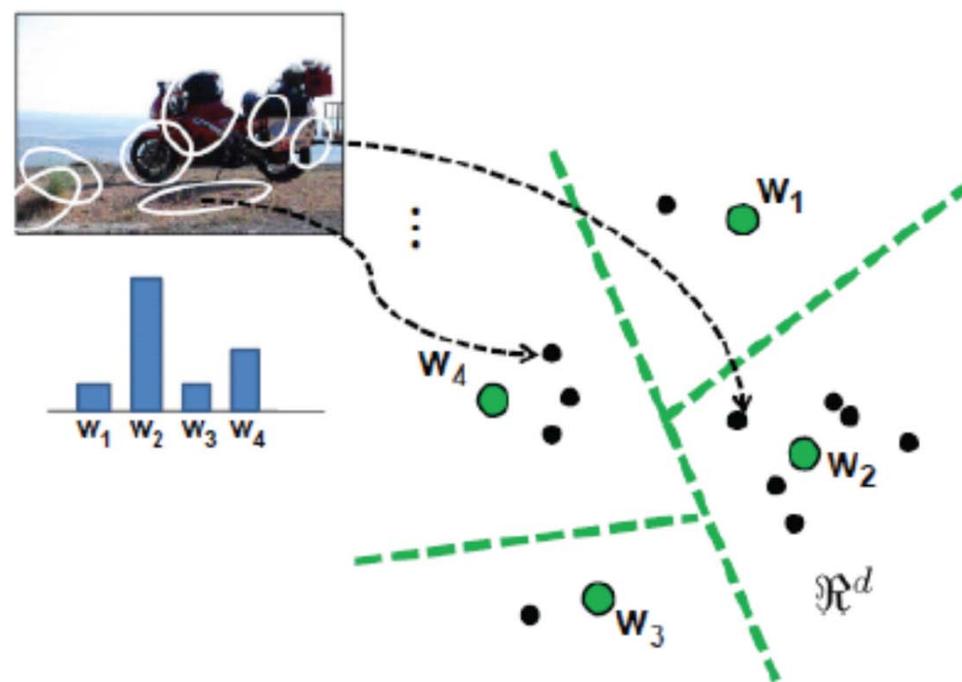
Encoding the Descriptor Distribution

- BOW only **counts** the number of SIFT descriptors assigned to each Voronoi cell



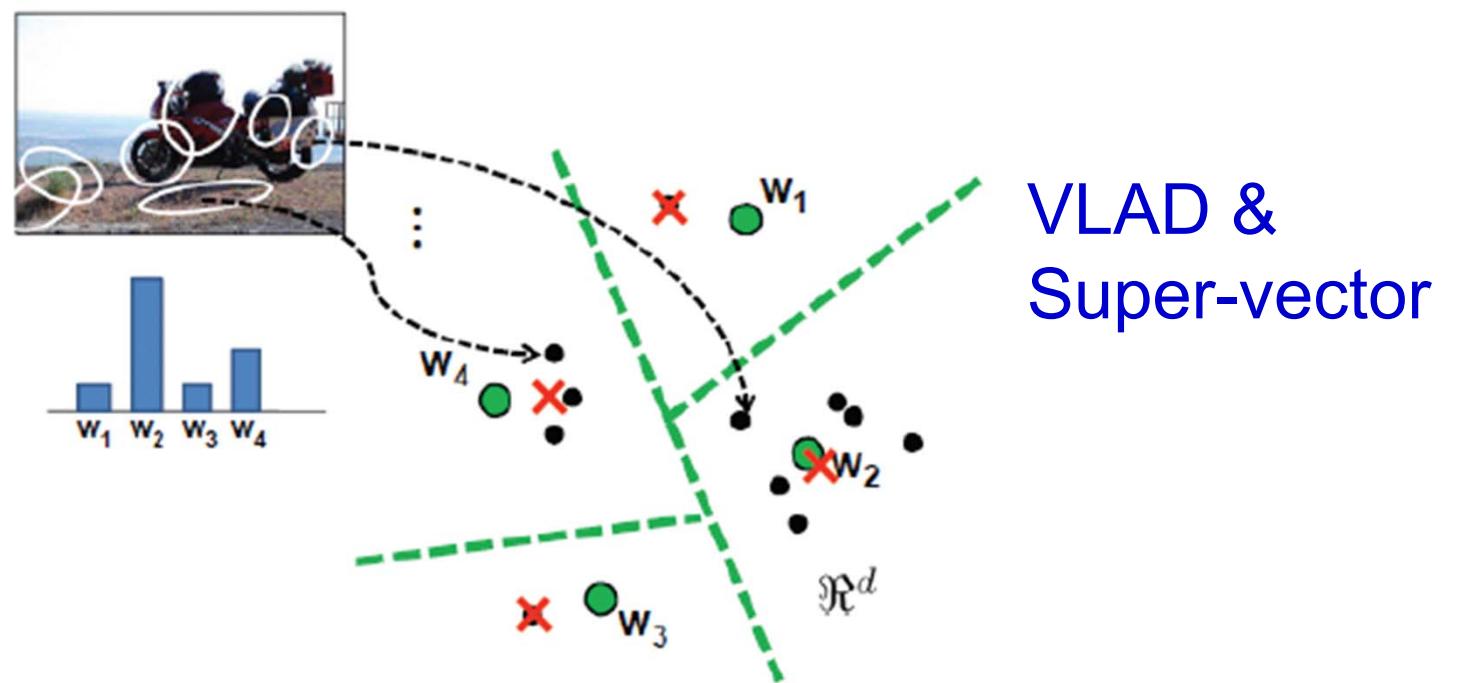
Encoding the Descriptor Distribution

- BOW only **counts** the number of SIFT descriptors assigned to each Voronoi cell
- Why not include **other statistics**? For instance



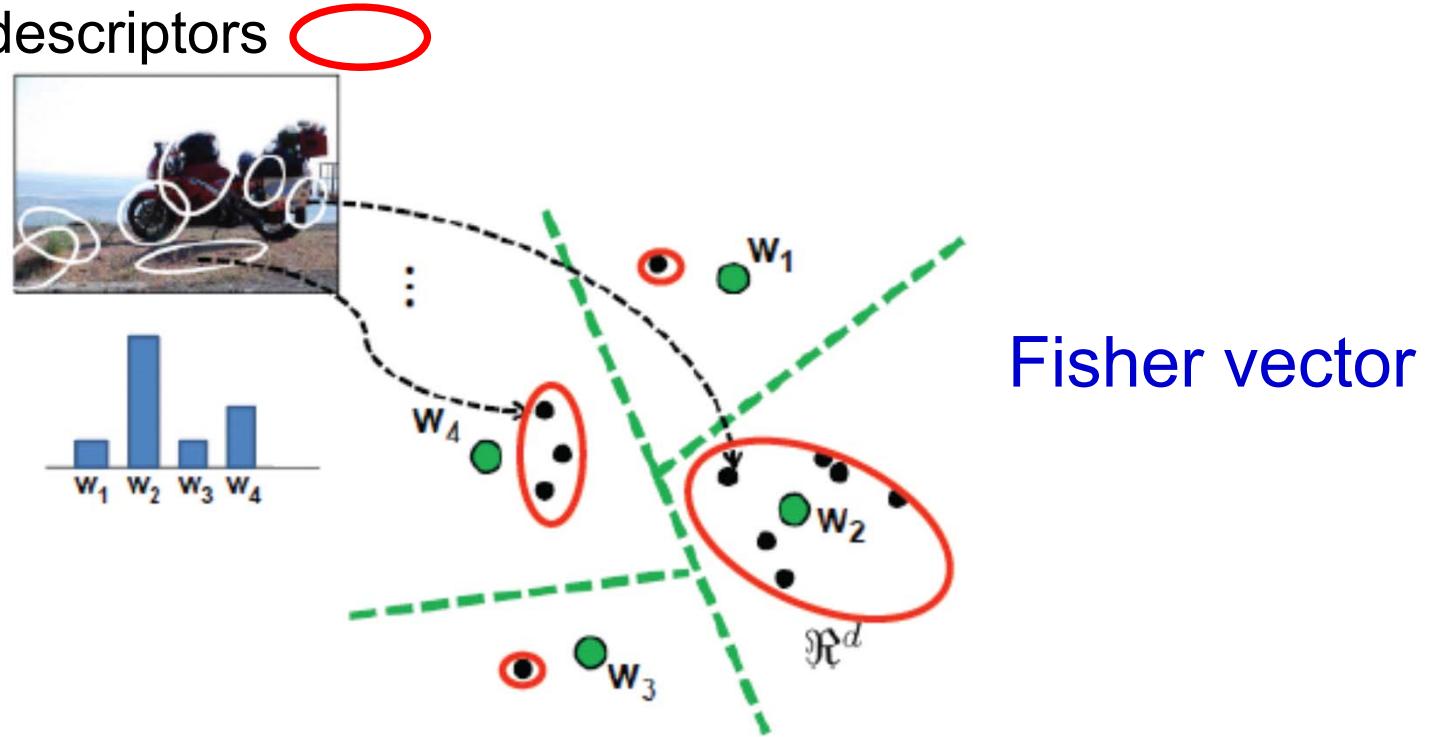
Encoding the Descriptor Distribution

- BOW only **counts** the number of SIFT descriptors assigned to each Voronoi cell
- Why not include **other statistics**? For instance
 - mean of descriptors \times



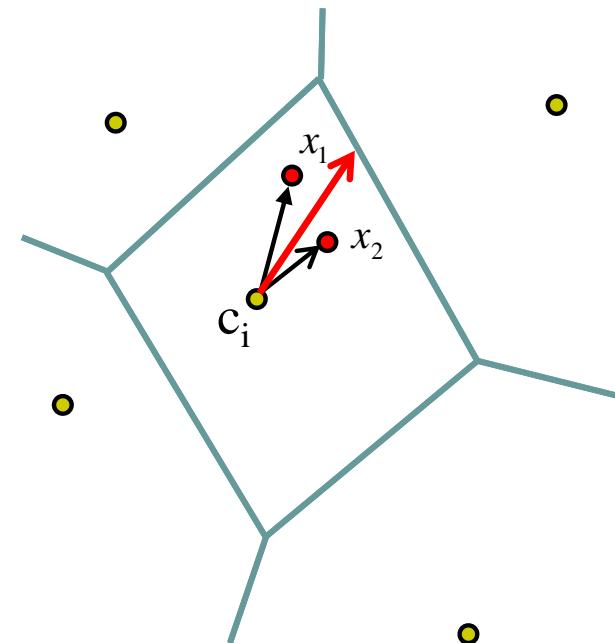
Encoding the Descriptor Distribution

- BOW only **counts** the number of SIFT descriptors assigned to each Voronoi cell
- Why not include **other statistics**? For instance
 - mean of descriptors
 - (co)variance of descriptors

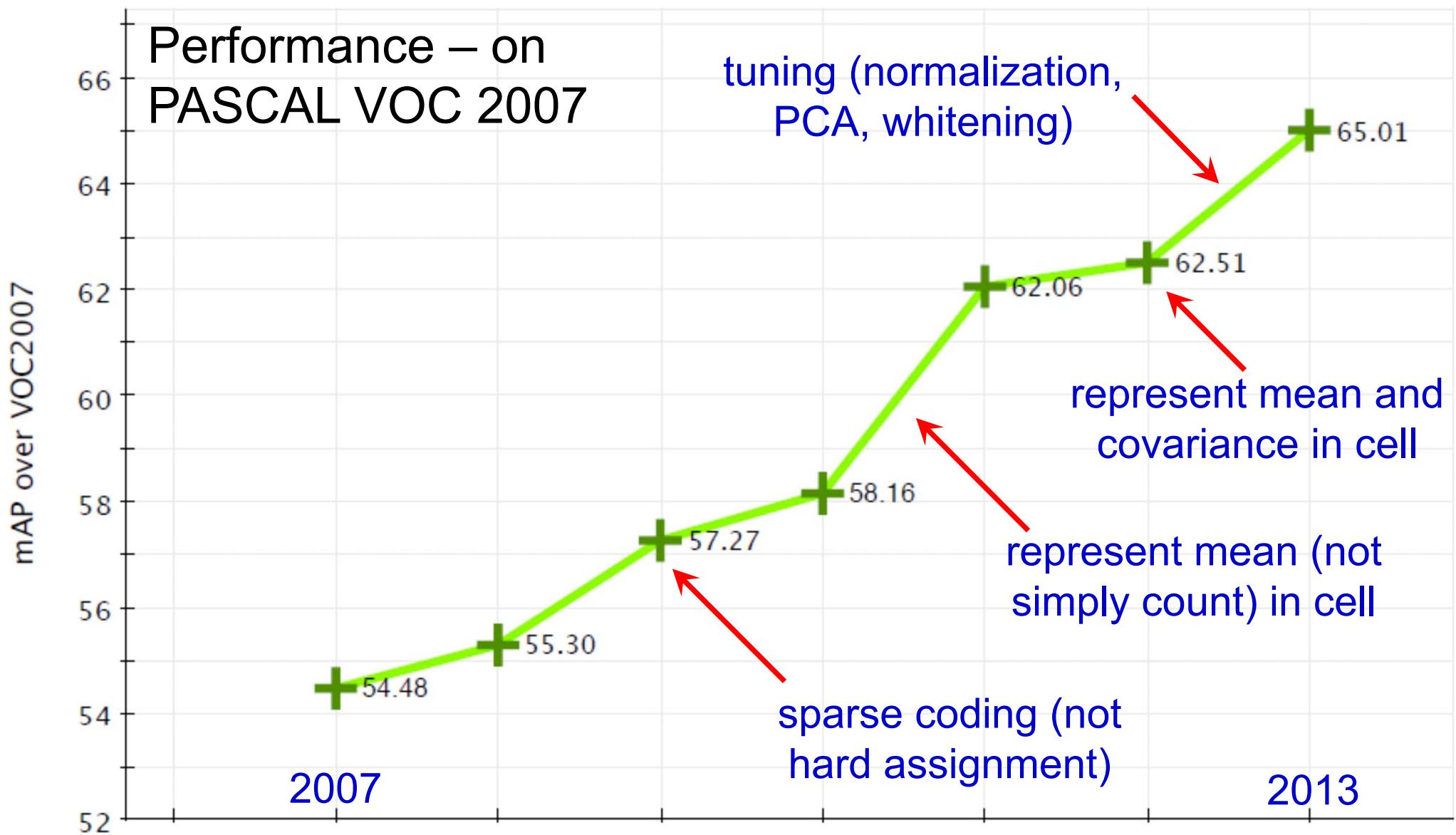


VLAD – Encoding

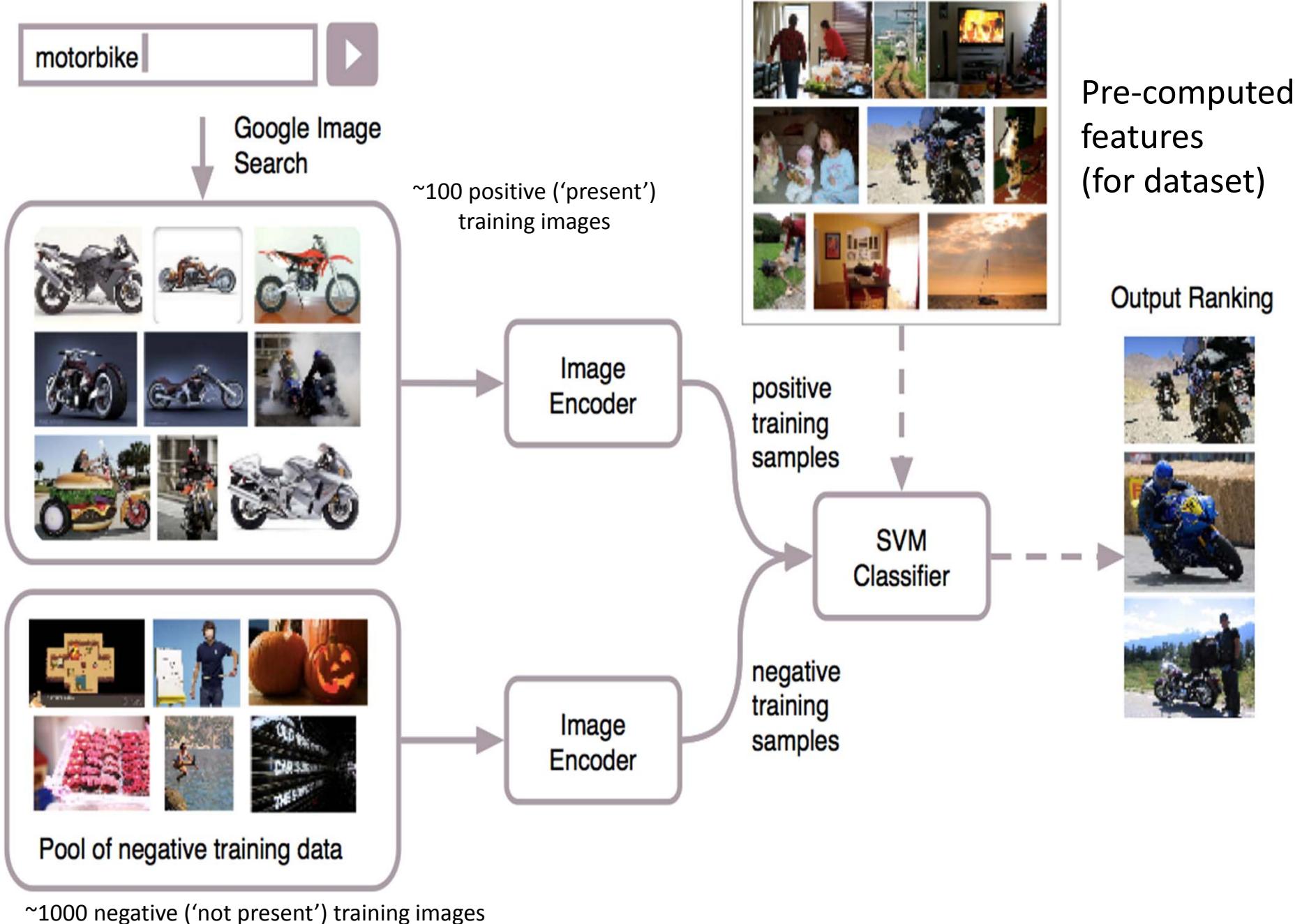
- **VLAD : vector of locally aggregated descriptors**
- Learn a vector quantizer (k -means): $c_1, \dots, c_i, \dots, c_k$, with c_i centroid of dim. d
- For a given image
 - ▶ assign each SIFT descriptor to closest center c_i
 - ▶ accumulate (sum) descriptors per cell
$$v_i := v_i + (x_j - c_i)$$
measure residual of vectors within a cell
- VLAD of dimension $D = k \times d$
(k typically between 16 and 512, $d = 128$ or less)
- The vector is square-root + L2-normalized



Evolution of Encoding Methods



On-the-fly Visual Category Search



Video dataset: BBC TV

- 4372 broadcasts from BBC 1, 2, 3 & 4
- Programmes from late 2011 to early 2012 from prime time slot (7pm-12pm) over five months
- 3007 hours of video represented by 1 frame per second
- 11M seconds of data, 3M keyframes
- Frames are 480 x 270 pixels



Visual Category Search – Examples ‘Car’

VISOR

car

BBCb

Search

Best of Top Gear
Series 16, Episode 1

Sun, 18 March 2012
BBC Two
10 occurrences

The screenshot shows a visual search interface named VISOR. At the top, there's a search bar with the word "car" and a dropdown menu set to "BBCb". Below the search bar is a timeline with several video clips. The first clip is titled "Best of Top Gear Series 16, Episode 1" and is dated "Sun, 18 March 2012" from "BBC Two", with "10 occurrences". It contains five thumbnail images of an orange sports car. The second clip is titled "Top Gear USA Series 2, Episode 2" and is dated "Fri, 13 January 2012" from "BBC Three", with "18 occurrences". It contains five thumbnail images of various cars, including a blue pickup truck and a red sedan. The third clip is titled "Top Gear Series 18, Episode 2" and is dated "Sun, 05 February 2012" from "BBC Two", with "19 occurrences". It contains five thumbnail images of luxury sports cars. Each clip has a timestamp below it: "2m0s" for the first, "3m2s" for the second, and "14m48s" for the third. A vertical ellipsis icon is visible next to each clip.

2m0s

Top Gear USA
Series 2, Episode 2

Fri, 13 January 2012
BBC Three
18 occurrences

3m2s

14m48s

Top Gear
Series 18, Episode 2

Sun, 05 February 2012
BBC Two
19 occurrences

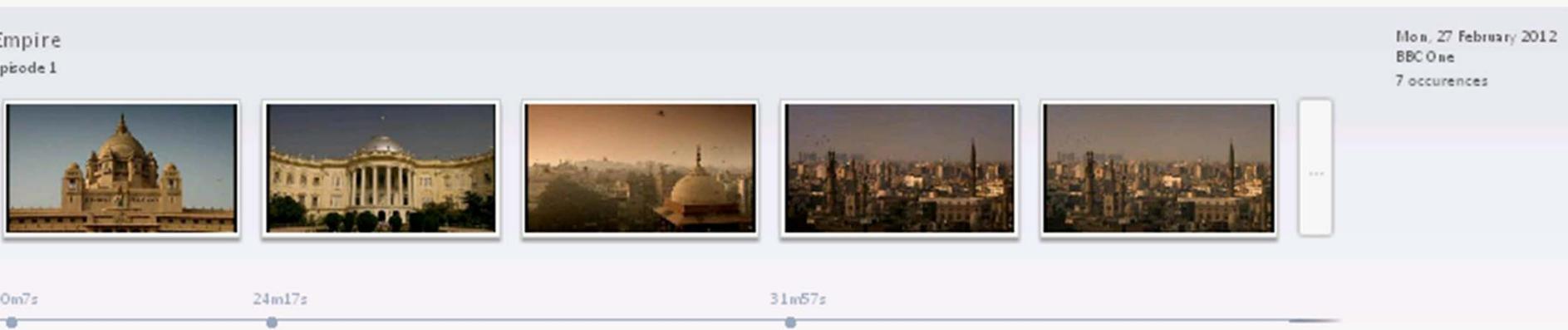
Visual Category Search – Examples ‘Cityscape’

VISOR  cityscape + BBCb Search Images processed in 40.292 · Model trained in 1.492 · F1 score 0.555

Search results page 1 of 1 (1,500 results)

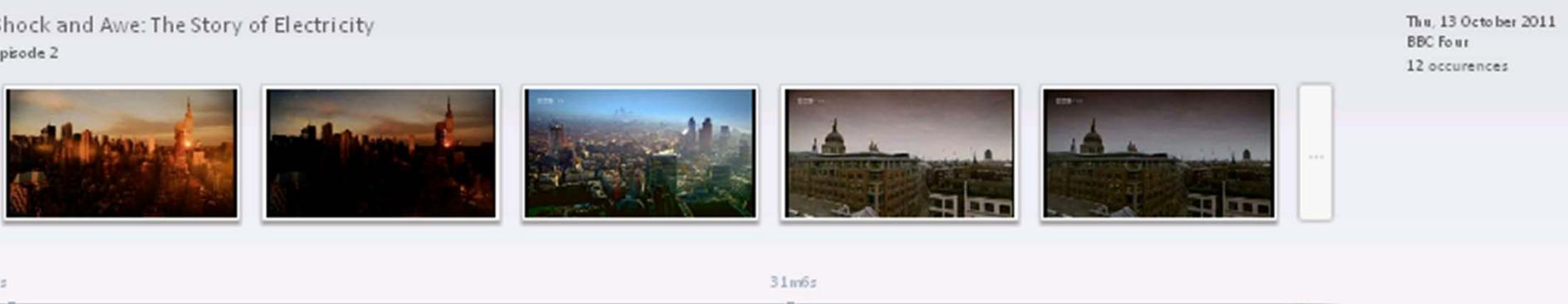
Empire
Episode 1

Mon, 27 February 2012
BBC One
7 occurrences


10m7s 24m17s 31m57s

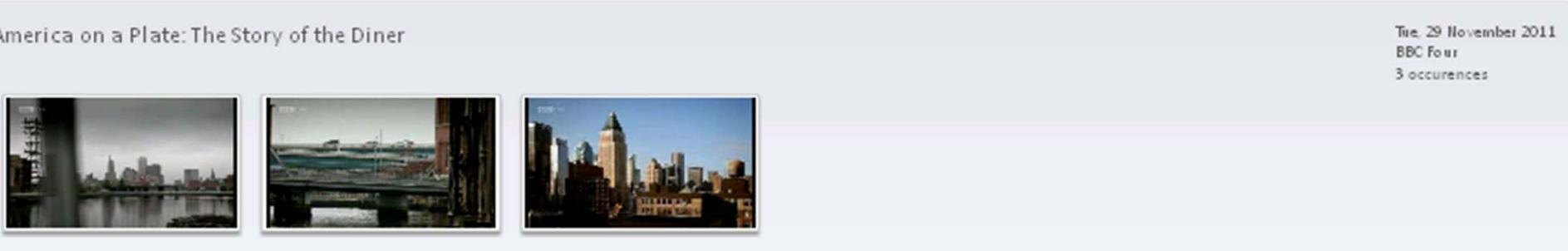
Shock and Awe: The Story of Electricity
Episode 2

Thu, 13 October 2011
BBC Four
12 occurrences


3s 31m6s

America on a Plate: The Story of the Diner

Tue, 29 November 2011
BBC Four
3 occurrences



VLAD Data Stats

3 Million key frames

Total size of original descriptors: $328k \times 4 \times 3M = 3936 \text{ GB}$

Dimensionality reduction 328k \rightarrow 8k using PCA
(mAP 62.06 \rightarrow 60.30)

- Memory footprint: $8k \times 4 \times 3M = 96 \text{ GB}$

Product Quantization: $8k \times 4 \rightarrow 2k$

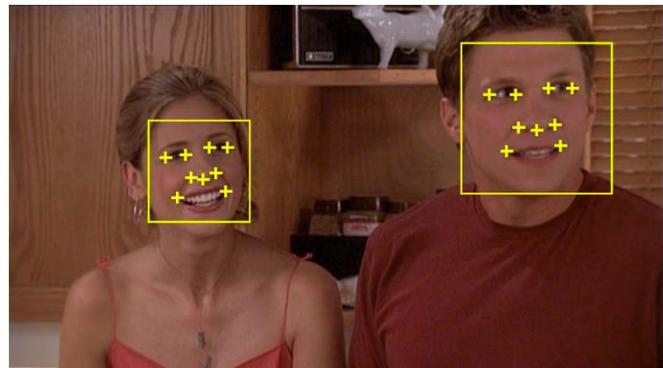
- Memory footprint: $2k \times 3M = 6 \text{ GB}$

Product Quantization for vector compression,
Jegou *et al.*, PAMI 2011

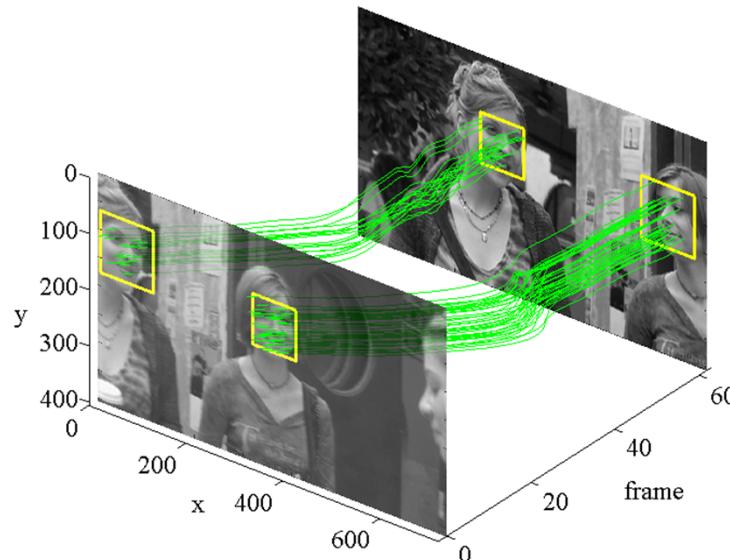
3. On-the-fly Face Search

Feature vectors for face (tracks)

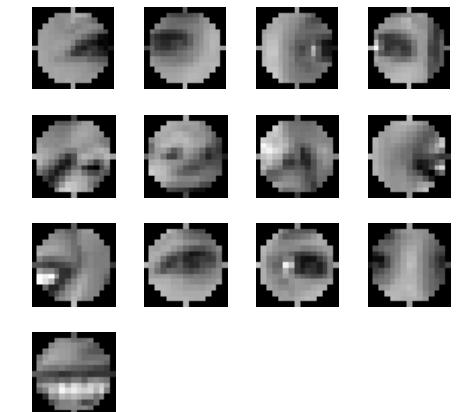
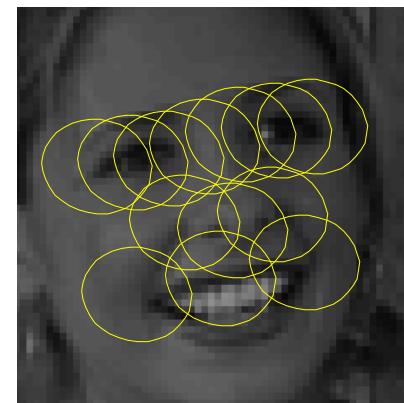
Face detection and facial landmark detection



Faces clustered into tracks

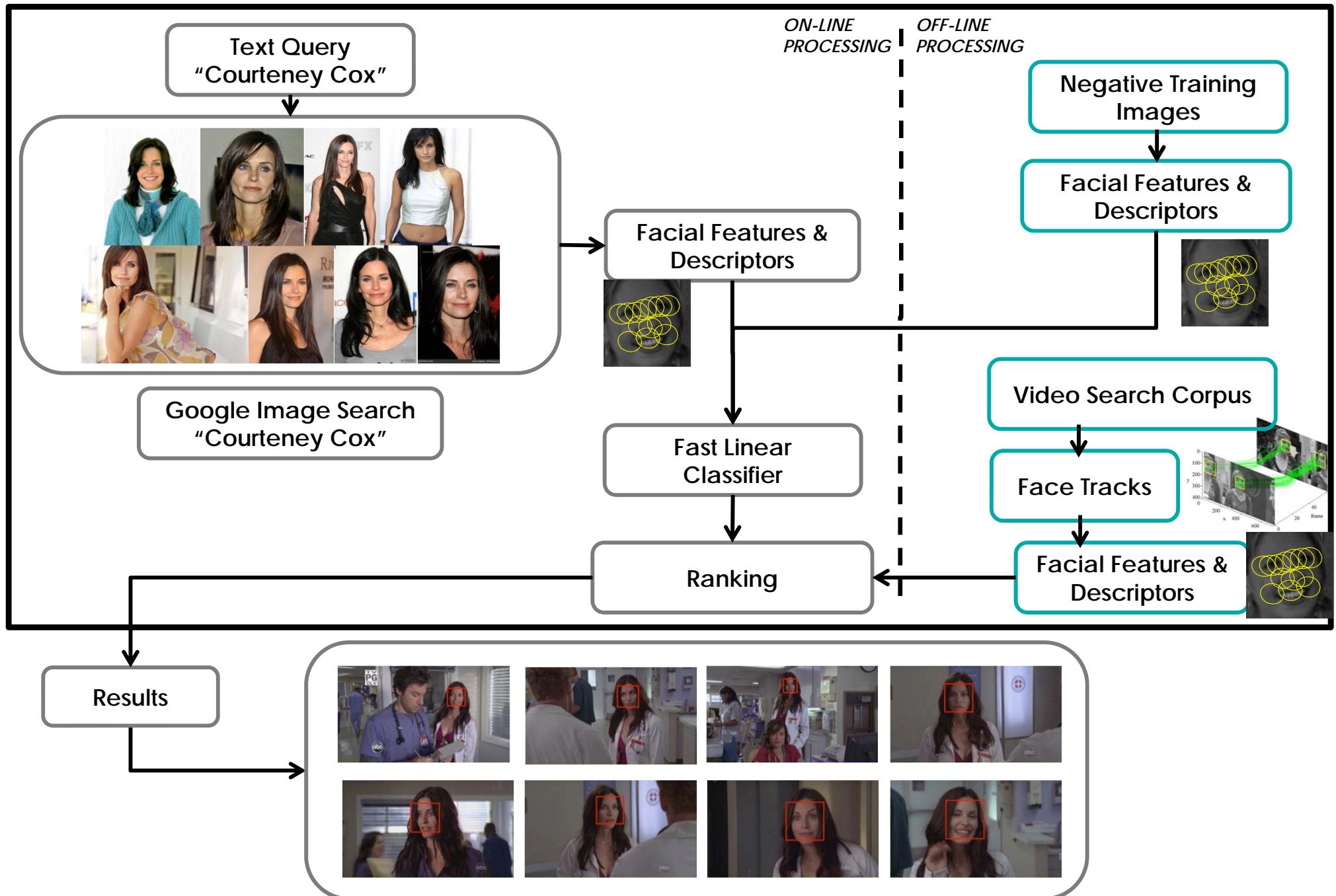


Feature region descriptors



Feature Vector

On-the-fly Person Retrieval



Face Search – Examples ‘Queen Elizabeth’

VISOR Search results page 1 of 250 (5,000 results)

queen elizabeth BBCb Search

The search results for 'Queen Elizabeth' on VISOR show a variety of images, including:

- Top row:
 - The World Against... (A portrait of Queen Elizabeth II on a £50 note)
 - BBC News at Ten (Queen Elizabeth II on a television screen)
 - BBC News at Ten (A £10 note featuring Queen Elizabeth II)
 - BBC News at Ten (Queen Elizabeth II sitting in a crowd)
 - The One Show (Queen Elizabeth II in a red hat)
- Second row:
 - World News Today (Queen Elizabeth II in a blue dress)
 - BBC News at Six (A £10 note featuring Queen Elizabeth II)
 - The Diamond Queen (Queen Elizabeth II in a white hat)
 - BBC Weekend News (A £10 note featuring Queen Elizabeth II)
 - Pramface (A woman holding a drawing of Queen Elizabeth II)
- Third row:
 - The Diamond Queen (Queen Elizabeth II in an orange dress)
 - BBC News at Six (Queen Elizabeth II in a green dress)
 - Dragons' Den: How... (Queen Elizabeth II smiling)
 - The Diamond Queen (Queen Elizabeth II in a red dress)
 - BBC News at Six (A £10 note featuring Queen Elizabeth II)
- Bottom row:
 - BBC News at Ten (A £10 note featuring Queen Elizabeth II)
 - Imagine (Queen Elizabeth II wearing a white fur-trimmed coat)
 - The Diamond Queen (Queen Elizabeth II smiling with others)
 - Imagine (Queen Elizabeth II in a dark coat)
 - Possession (A woman in a historical setting)

Video dataset: BBC TV

- 4372 broadcasts from BBC 1, 2, 3 & 4
- Programmes from late 2011 to early 2012 from prime time slot (7pm-12pm) over five months
- 3007 hours of video represented by 1 frame per second
- 11M seconds of data, 3M keyframes
- Frames are 480 x 270 pixels



Face Data Stats

- 3007 hours of video, 3 M shots
- 0.68 M shots have faces
- 0.8 M face tracks
- Total size of original descriptors: $4k \times 4 \times 0.8M = 12.8 \text{ GB}$
- Memory footprint (after PQ): $1k \times 0.8M = 0.8 \text{ GB}$
- NB no need for PCA dimensionality reduction here

Facial attributes – FaceTracer project

Examples:

- gender: male, female
- age: baby, child, youth, middle age, senior
- race: white, black, asian
- smiling, mustache, eye-wear, hair colour



Method

- **person independent** training set with attribute
- facial feature representation
- discriminative training of classifier for attribute

N. Kumar, P. N. Belhumeur and S. K. Nayar,

FaceTracer: A Search Engine for Large Collections of Images with Faces, *ECCV 2010*

Face Search – Examples ‘Moustache’

VISOR

moustache

BBCb

Search

The search results for 'moustache' in the VISOR interface show a grid of 20 thumbnail images, each with a caption below it. The thumbnails are arranged in four rows of five. The BBC logo is visible in the top right corner of the search bar.

Thumbnail 1	Thumbnail 2	Thumbnail 3	Thumbnail 4	Thumbnail 5
The League Cup Show	The Celebrity Appre...	Twenty Twelve	Twenty Twelve	When Rock Goes Acous...
9/11: Conspiracy...	Newsnight	BBC London News	Match of the Day 2	Later Live... with...
Later... with Jools...	Holy Flying Circus	Cruel Sea: The Penle...	The Queen's Palaces	American Football
The Celebrity Appre...	9/11: Conspiracy...	TV Greats: Our Favou...	MasterChef	MasterChef

Datasets

Description	BBC 1, 2, 3 & 4 prime time 5 months	BBC 1, 2, 3 & 4, Parliament & News 24 4 years
# broadcasts	4,372	56,078
video / hrs	3,007	39,289
# frames (1 per second)	11 M	141 M
# key frames (1 per shot)	3 M	34.6 M

Face Search on 40 k hrs – Example ‘Obama’

Search results page 8 of 250 (5,000 results)



BBC News



BBC News



This World



Newsnight



BBC News at Six



BBC News



BBC Weekend News



BBC News at Six



BBC News at Ten



World News Today



BBC News



BBC News at Five...



BBC News at Six



BBC News at Six



Newsnight



BBC News



BBC News at Six



President Obama at...



BBC News at Five...



BBC News at Ten

How can performance be improved?

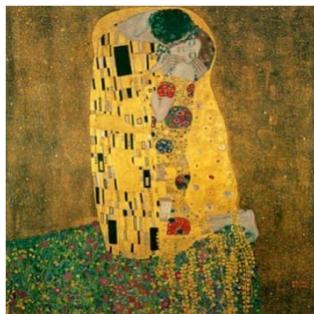
- Better face descriptor encoding
- See paper by Karen Simonyan *et al.* “Fisher Vector Faces in the Wild”, BMVC 2013

The Vision

All visual material (images, video) should be searchable for anything

- people, object categories, scene categories, particular objects, human actions and interactions, activities ...

and retrieved with high precision and high recall



On-the-fly papers

R. Arandjelović, A. Zisserman

Multiple queries for large scale specific object retrieval

British Machine Vision Conference, 2012

K. Chatfield, A. Zisserman

VISOR: Towards On-the-Fly Large-Scale Object Category Retrieval

Asian Conference on Computer Vision, 2012

O. M. Parkhi, A. Vedaldi, A. Zisserman

On-the-fly Specific Person Retrieval

International Workshop on Image Analysis for Multimedia Interactive Services,
2012