6.819 / 6.869: Advances in Computer Vision

Image Retrieval:

Retrieval: Information, images, objects, large-scale

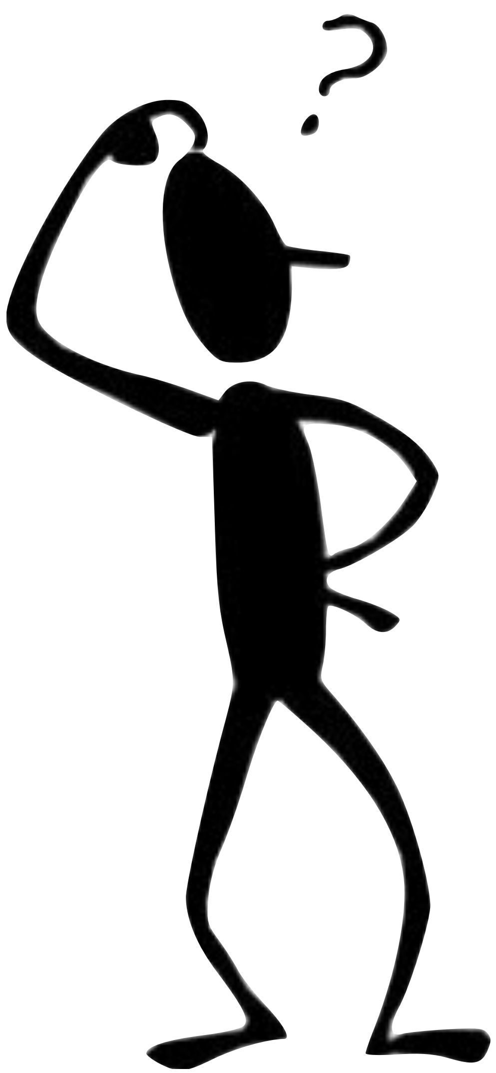
Website:
http://6.869.csail.mit.edu/fa15/

Instructor: Yusuf Aytar

Lecture TR 9:30AM – 11:00AM
(Room 34-101)

What is Image Retrieval ?

User



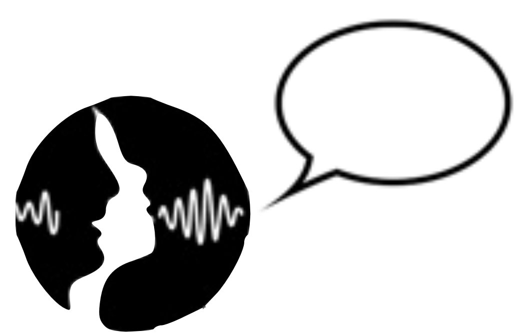
Query

Text

Fall in Boston

Image

Speech



Retrieval Results

Applications

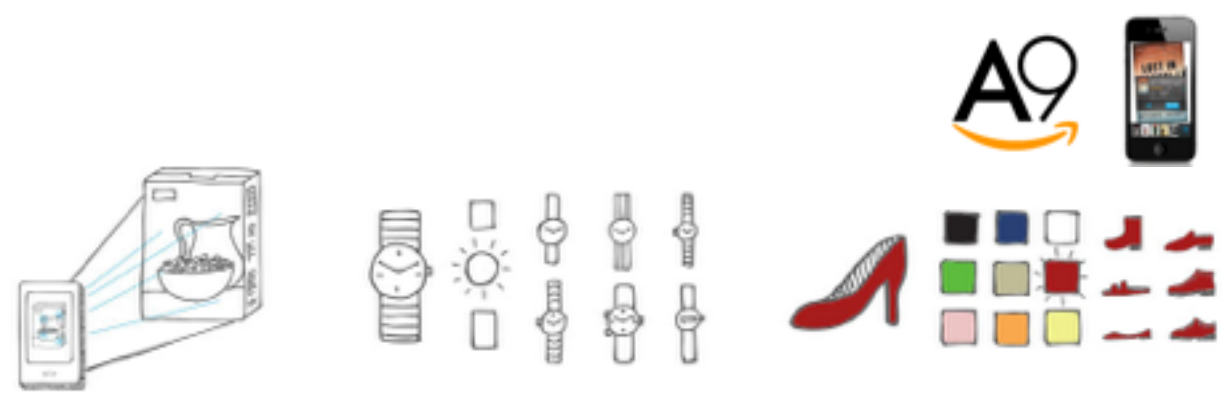
Art Retrieval

PLINKART

Medical Image Retrieval

Product Image Retrieval
(Reviews, other prices etc.)

Fashion Image Retrieval



Overview

Information Retrieval

Bag of Words, TF-IDF, Cosine Similarity, Inverted Index

Object Instance Retrieval

Bag of Visual Words, Video Google, Object Instance Retrieval

Fast Object Detection/Retrieval

Fast detection, Part representations, Generalization from exemplar

Large Scale Image Search

KD-trees, Locality Sensitive Hashing, Semantic Hashing, Compact Codes

Information Retrieval

Bag of Words (BOW)

The last duel
After quarrelling over a bank loan, two men took part in the last fatal duel staged on Scottish soil. BBC News's James Landale retraces the steps of his ancestor, who made that final challenge.

Bank	: 1
Loan	: 1
Water	: 0
Farmer	: 0

Doc-1

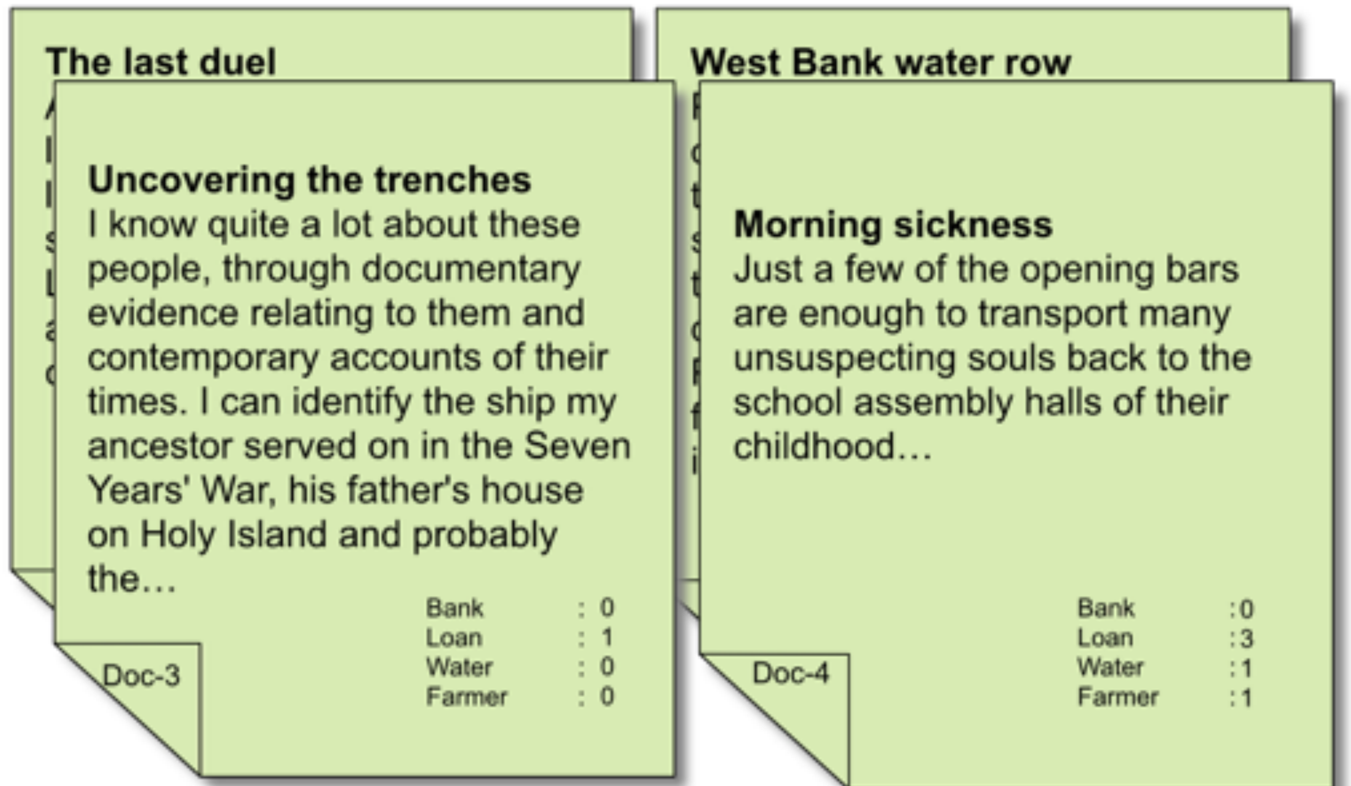
West Bank water row
Palestinians have accused Israel of diverting water away from their towns in order to keep Jewish settlements in the occupied territories fully supplied. Israel denies the charge saying Palestinian farmers are to blame for using illegal connections to irrigate their fields.

Bank	: 1
Loan	: 0
Water	: 2
Farmer	: 0

Doc-2

A widely used document representation method

Term Frequency (TF)



Lexicon \ Documents	Doc-1	Doc-2	Doc-3	Doc-4
Bank	1	1	0	0
Loan	1	0	0	3
Water	0	2	1	1
Farmer	0	0	0	1

Normalization

Lexicon \ Documents	Doc-1	Doc-2	Doc-3	Doc-4
Bank	0.5	0.33	0	0
Loan	0.5	0	0	0.6
Water	0	0.66	1	0.2
Farmer	0	0	0	0.2

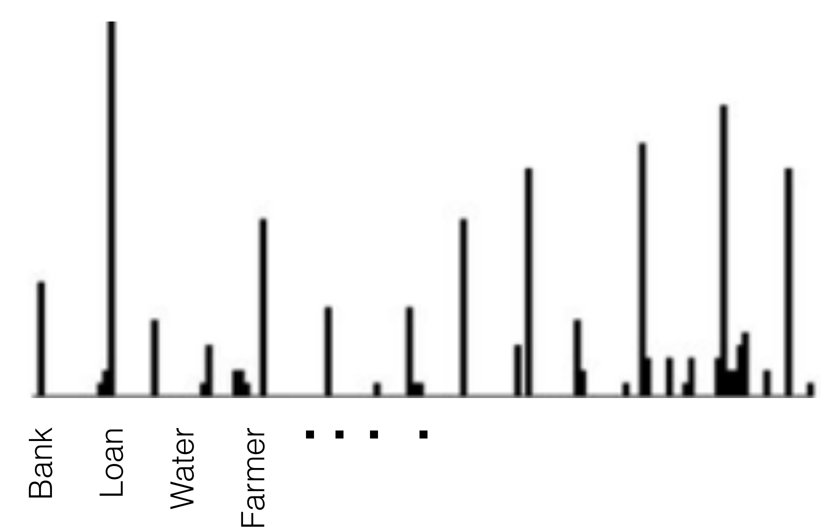
Inverse Document Frequency (IDF)

IDF of i^{th} word: $$idf_i = \log\left(\frac{n}{df_i}\right)$$

The last duel
After quarrelling over a bank loan, two men took part in the last fatal duel staged on Scottish soil. BBC News's James Landale retraces the steps of his ancestor, who made that final challenge.

Doc-1

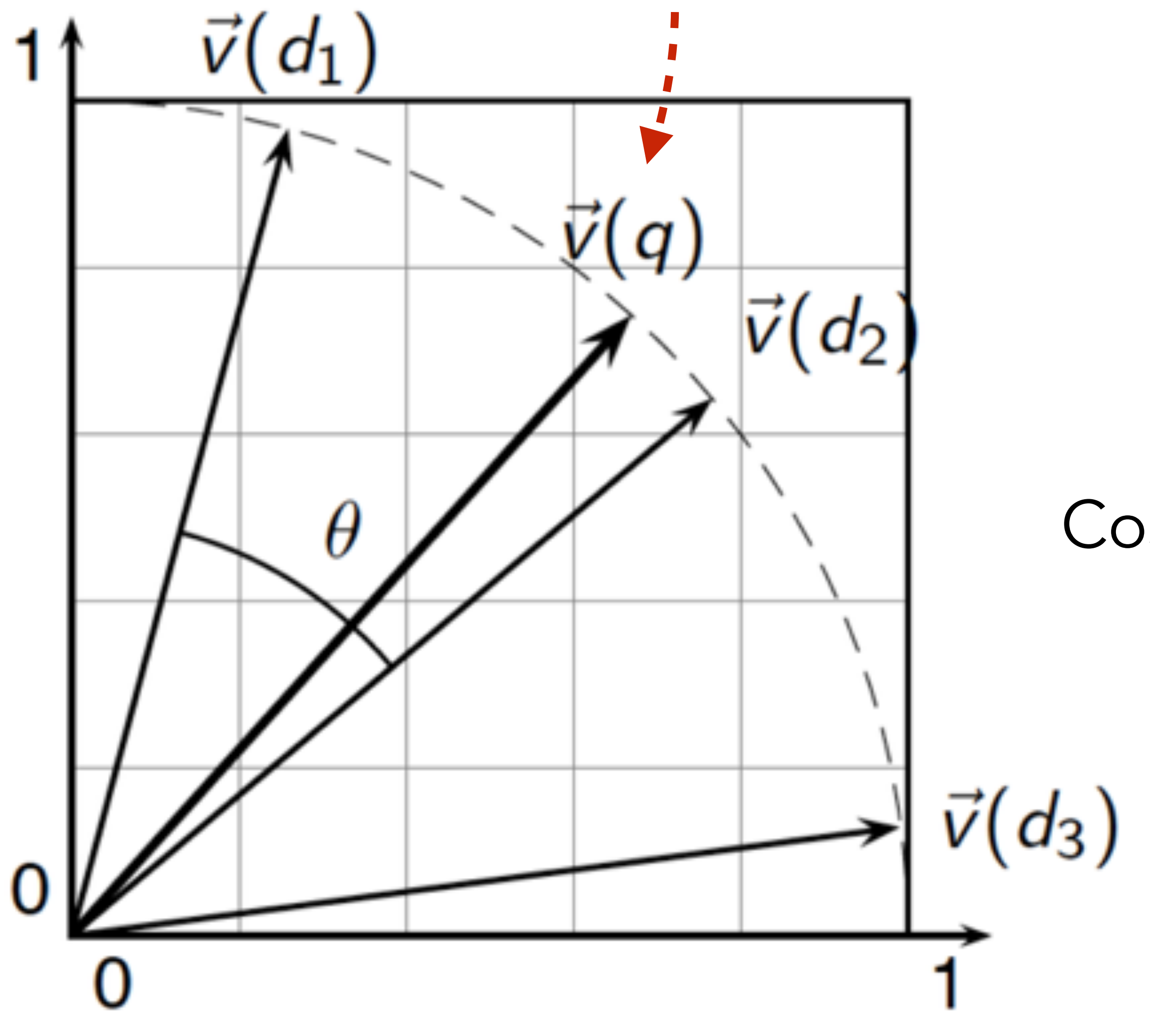
$\vec{v}(d_1) =$



$tf \times idf$

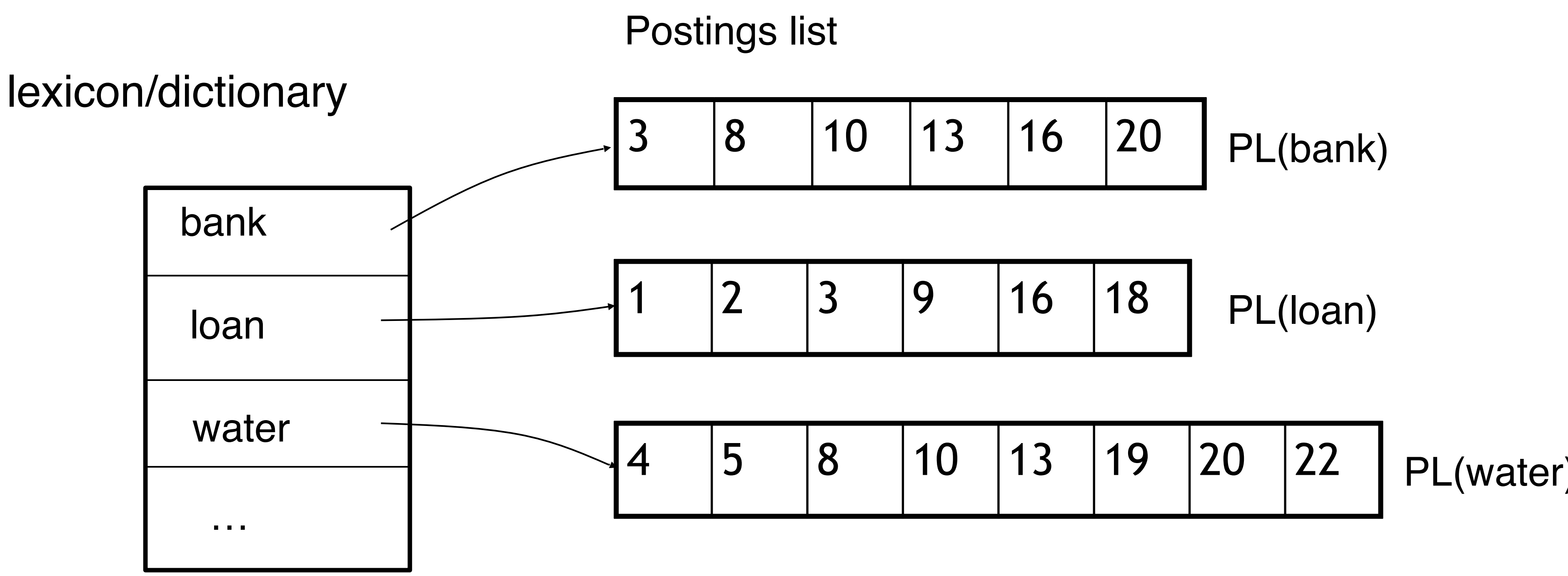
Cosine Similarity

Query: **fall in Boston**



Cosine Similarity Score = $\vec{v}(q)^T \vec{v}(d_1)$

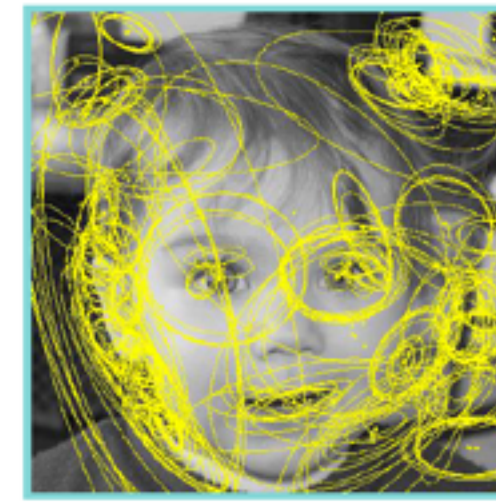
Inverted Index



Allows quick lookup of document ids with a particular word

Bag of Words &
Object Instance Retrieval

Feature Detectors



Roles of the detector:

- provide invariance to transformations
- **reduce the number of descriptors**

Popular detectors:

- Maximally Stable Extremal Regions (MSER)
  Matas, Chum, Urban, Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions", BMVC'02.
- Difference of Gaussians (DoG)
  Lowe, "Distinctive image features from scale-invariant keypoints", IJCV'04.
- Harris-Affine and Hessian-Affine
  Mikolajczyk, Schmid, "Scale and affine invariant interest point detectors", IJCV'04.

→ See also Mikolajczyk et al., "A comparison of affine region detectors", IJCV'05.

Dense descriptors are also possible

- Mainly for classification → let the classifier decide
  Leung, Malik, "Representing and recognizing the visual appearance of materials using 3D textons", IJCV'01.
- But also for image/scene/object retrieval
  Gordo, Rodriguez, Perronnin, Valveny, "Leveraging category-level labels for instance-level image retrieval", CVPR'12.

Slide Credit - F. Perronnin, LSVR tutorial at CVPR'13

Feature Descriptors

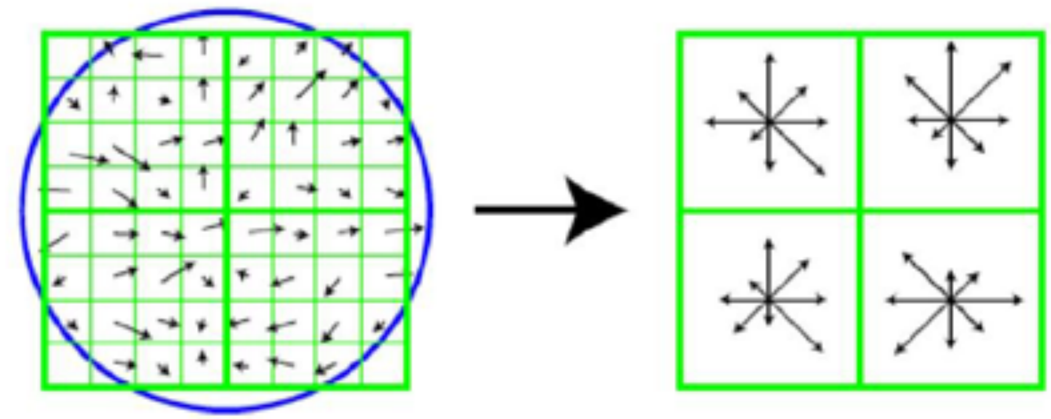
Description of a patch after orientation/scale/photometric normalization

Most widely-used patch descriptor: SIFT

Lowe, "Distinctive image features from scale-invariant keypoints", IJCV'04.

- 8 orientations of the gradient
- 4x4 spatial grid

$\rightarrow$ 128 dimensions



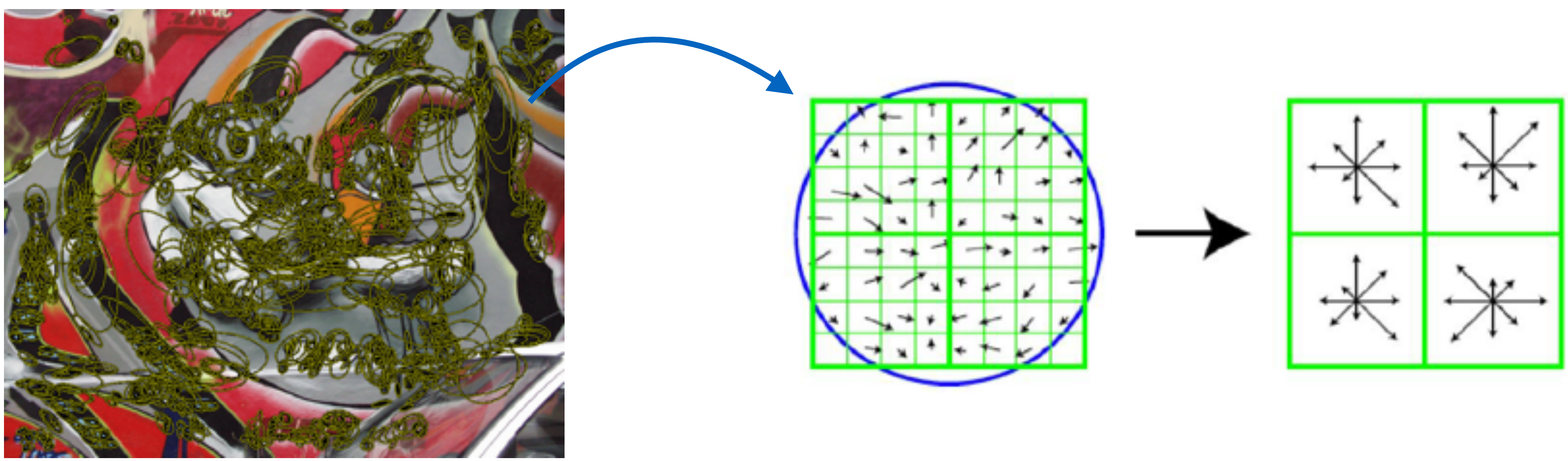
Many descriptors derive from SIFT:

- **More efficient: SURF**
  Bay, Tuytelaars, Van Gool, "SURF: speeded up robust features", ECCV'06.
- **More compact: CHOG, DAISY**
  Chandrasekhar et al, "Compressed histograns of gradients: a low-bit rate descriptor", IJCV'11.
  Tola, Lepetit, Fua, "DAISY: an efficient dense descriptor applied to wide baseline stereo", TPAMI'10.
- With color: color SIFT
  Van de Weijer, Schmid, "Coloring local feature extraction", ECCV'06.
  Burghouts and Geseborek, "Performance evaluation of local colour invariants", CVIU'09.

Slide Credit - F. Perronnin, LSVR tutorial at CVPR'13

Video Google

Feature Detectors / Descriptors

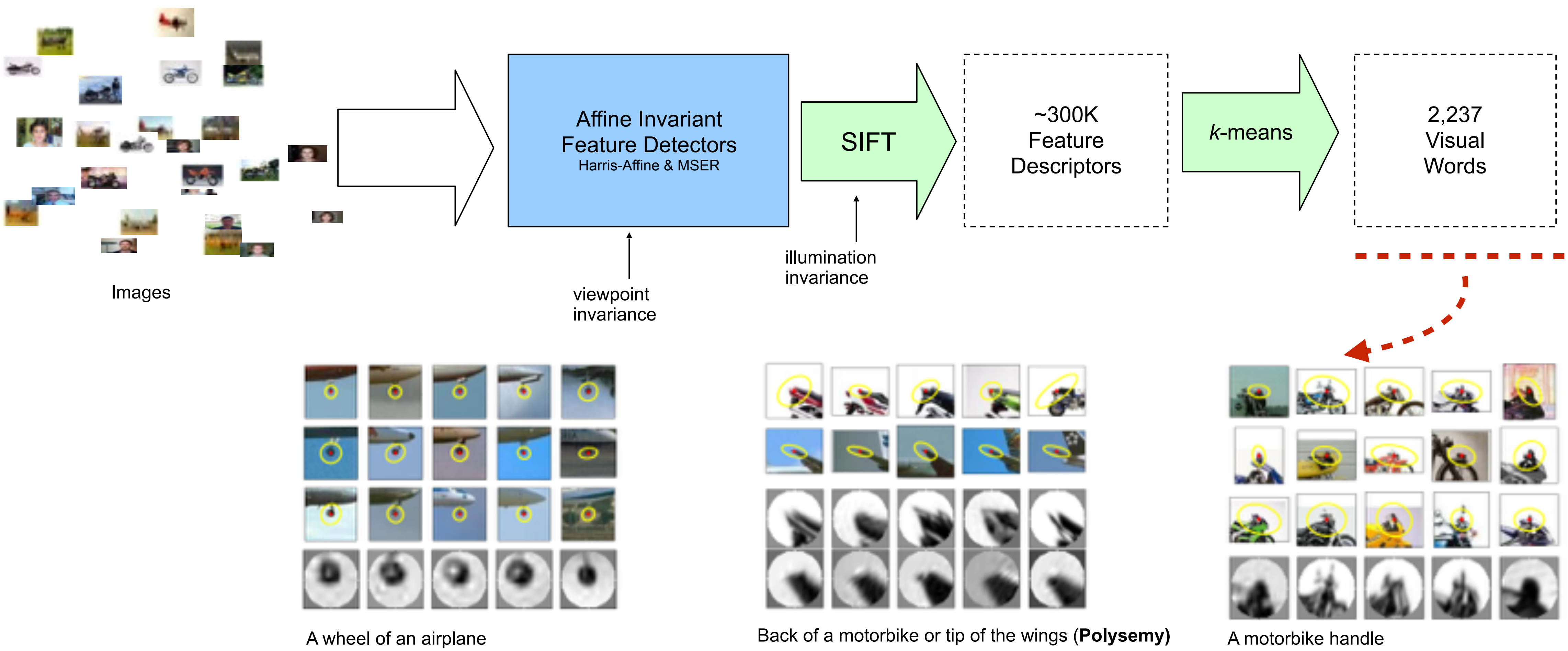


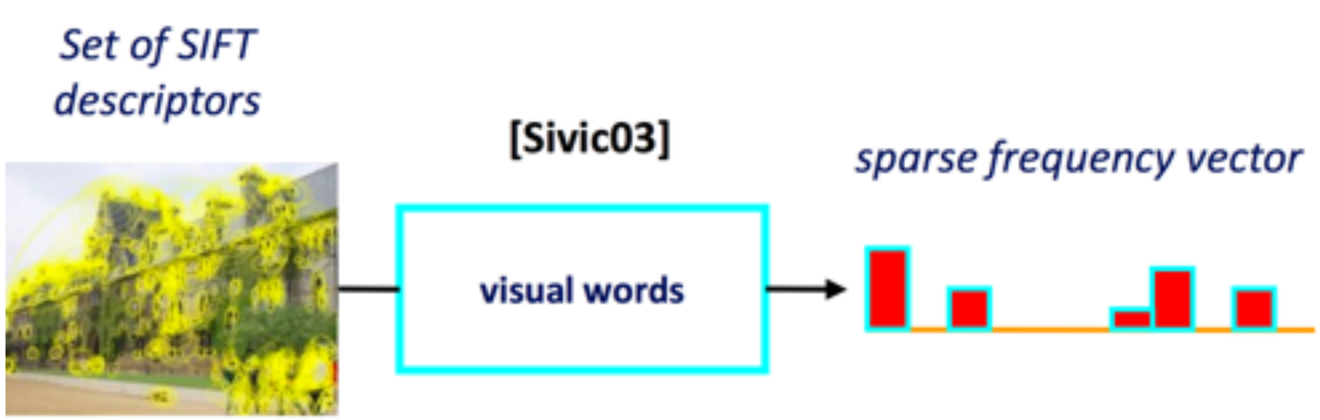
Harris-Affine & Hessian Affine as the **feature detectors**

SIFT as the **feature descriptor**

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

Video Google
Bag of Visual Words





Discovering objects and their location in images, Sivic et. al., ICCV 2005
Video Google: A Text Retrieval Approach to Object Matching in Videos, Josef Sivic and Andrew Zisserman, ICCV 2003
http://www.robots.ox.ac.uk/~vgg/publications/2012/Arandjelovic12/presentation.pdf

Video Google
Large scale object instance retrieval

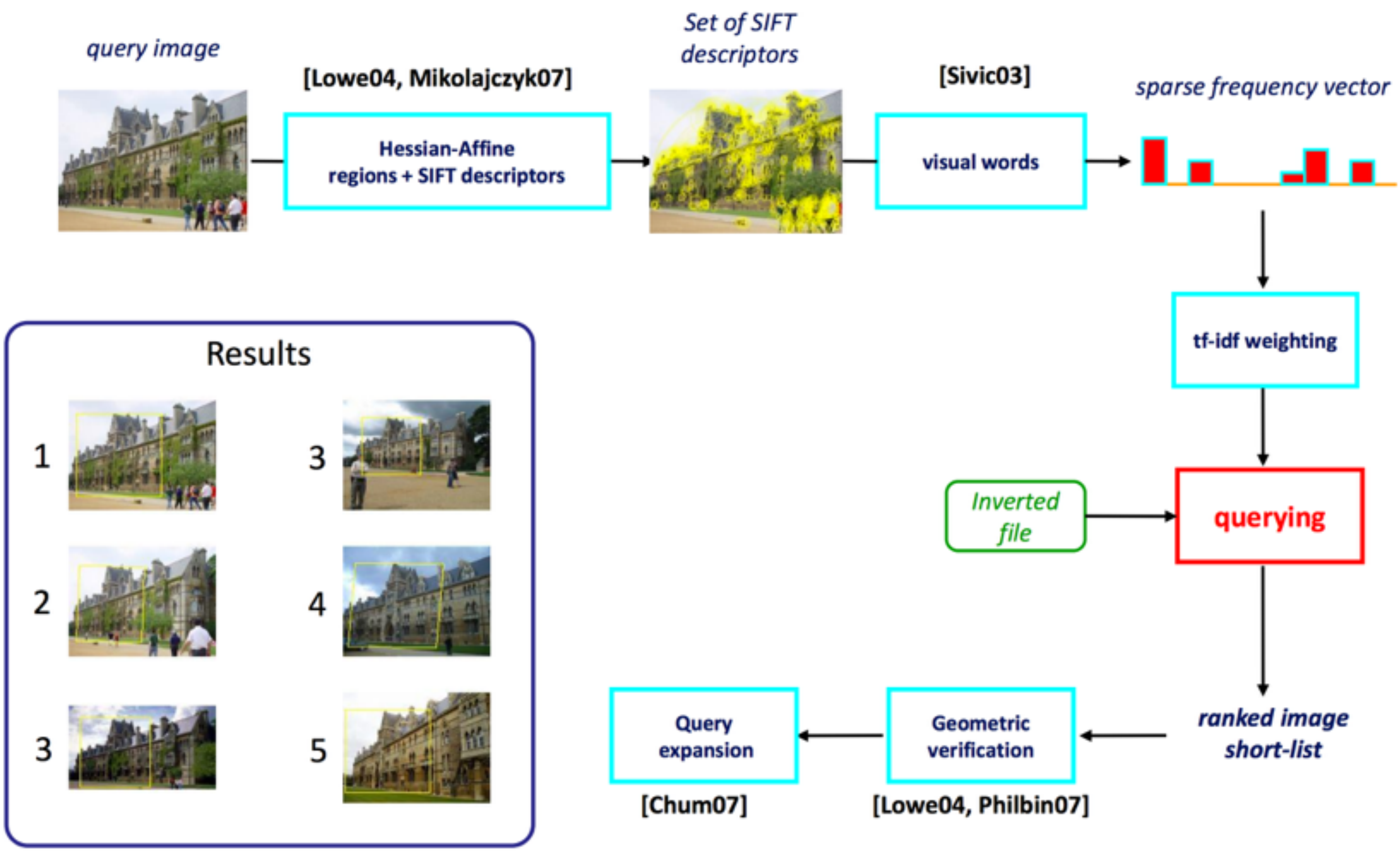
- Find all instances of the query object in a large scale dataset
- Do it instantly (< 1sec), and be robust to scale, viewpoint, lighting, partial occlusion

Video Google: A Text Retrieval Approach to Object Matching in Videos, Josef Sivic and Andrew Zisserman, ICCV 2003
http://www.robots.ox.ac.uk/~vgg/publications/2012/Arandjelovic12/presentation.pdf

Video Google

Particular object retrieval - Bag of visual words



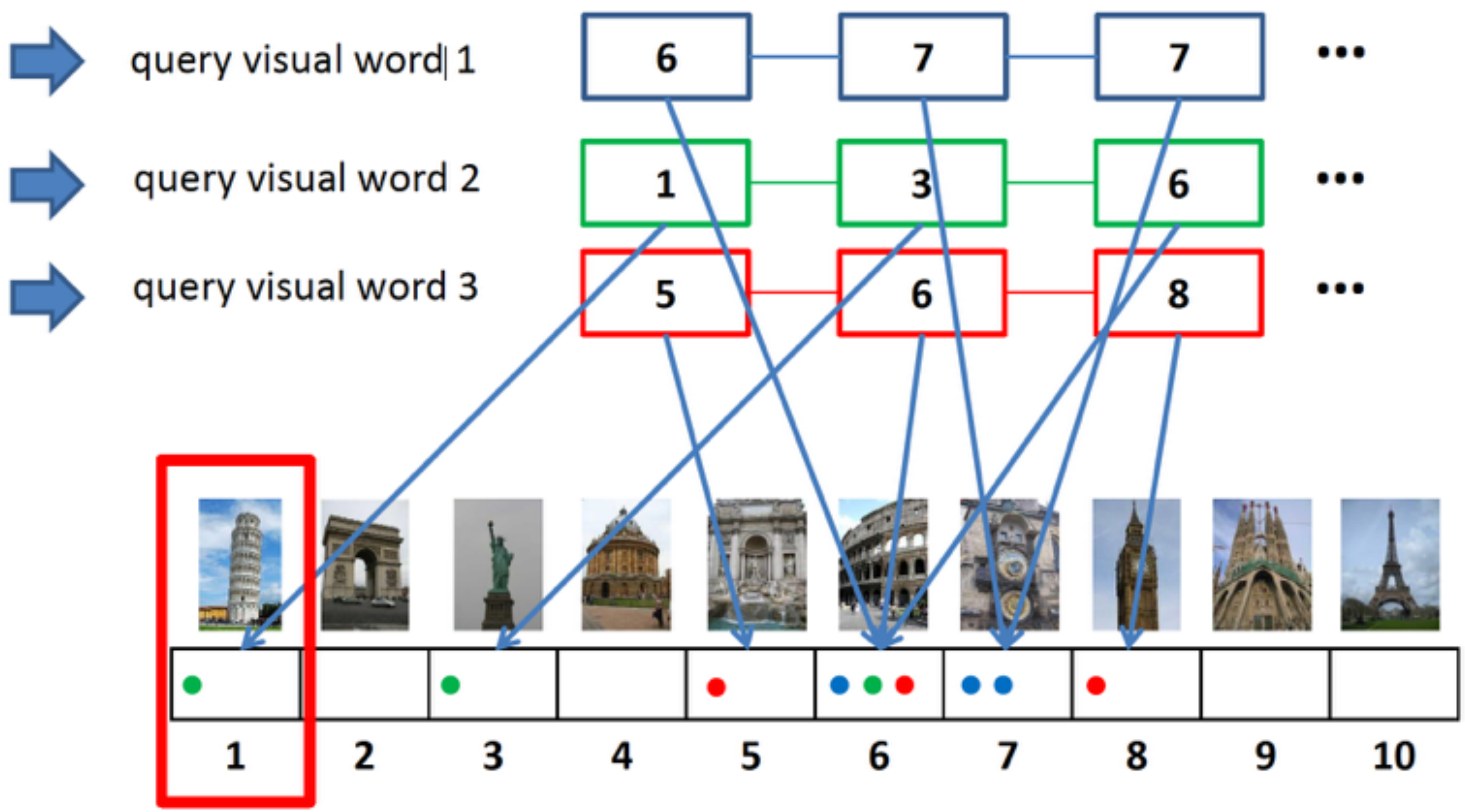
Object retrieval with large vocabularies and fast spatial matching, Philbin, Chum, Isard, Sivic, Zisserman, CVPR 2007

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

http://www.robots.ox.ac.uk/~vgg/publications/2012/Arandjelovic12/presentation.pdf

Video Google
BOW + Inverted File Indexing

$$\text{score} = \frac{\mathbf{q}^\top \mathbf{x}}{||\mathbf{x}||}$$



Slide Credit - Chum, LSVR tutorial at CVPR'13

Spatial Verification

1. Image Query

2. Initial Retrieval Set

3. Spatial Verification

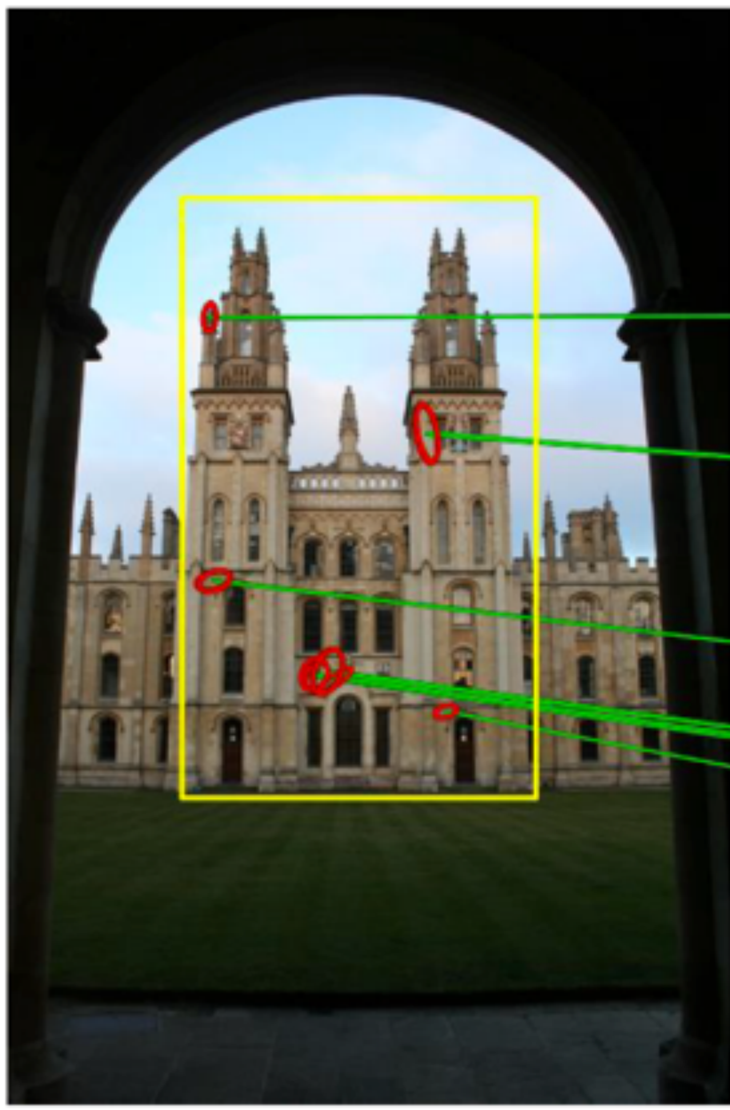
Object retrieval with large vocabularies and fast spatial matching, Philbin, Chum, Isard, Sivic, Zisserman, CVPR 2007

http://www.robots.ox.ac.uk/~vgg/publications/2012/Arandjelovic12/presentation.pdf

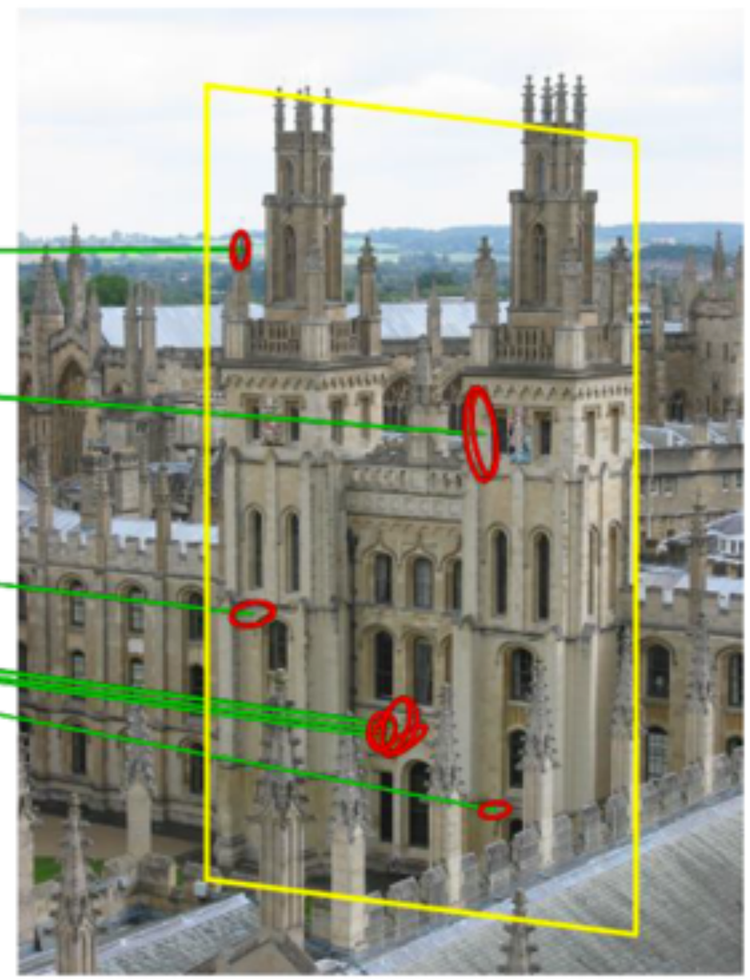
Query Expansion

Slide Credit - Chum, LSVR tutorial at CVPR'13

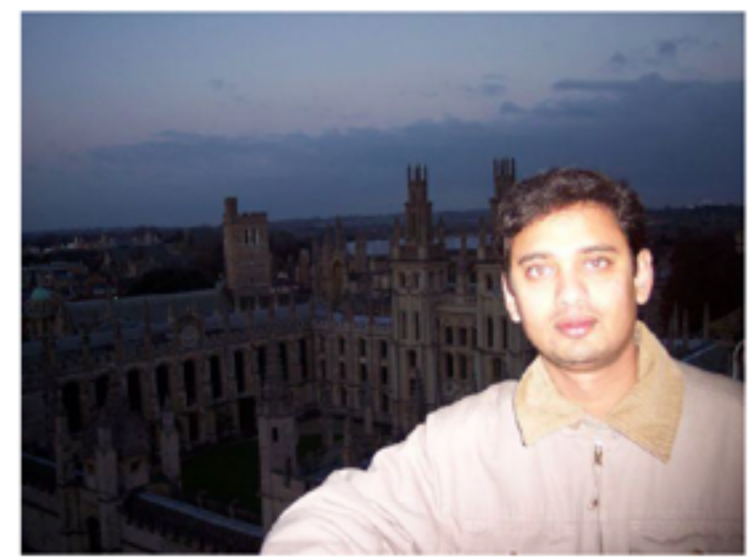
Query Expansion



Query Image



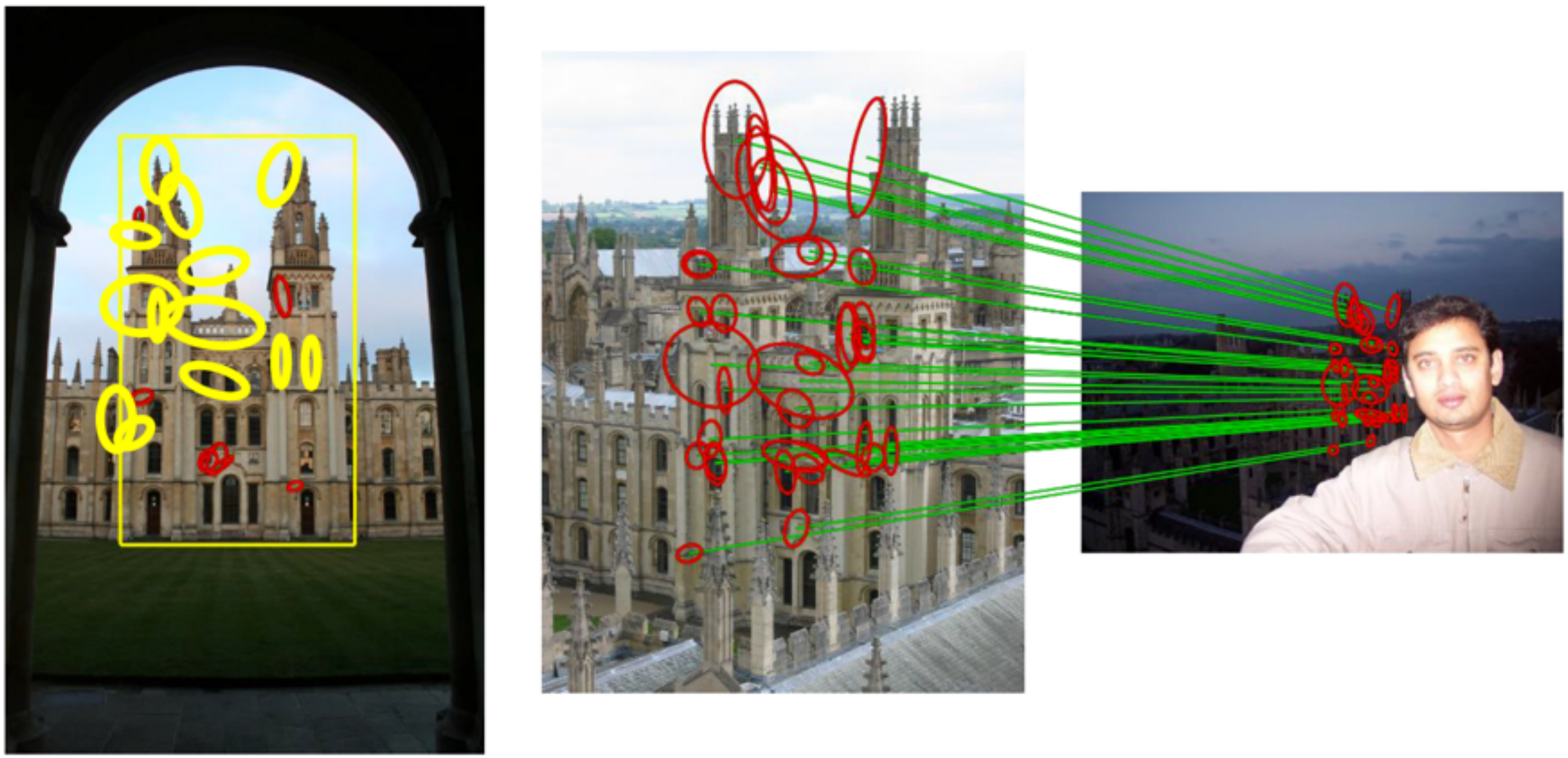
Retrieved image



Originally not retrieved

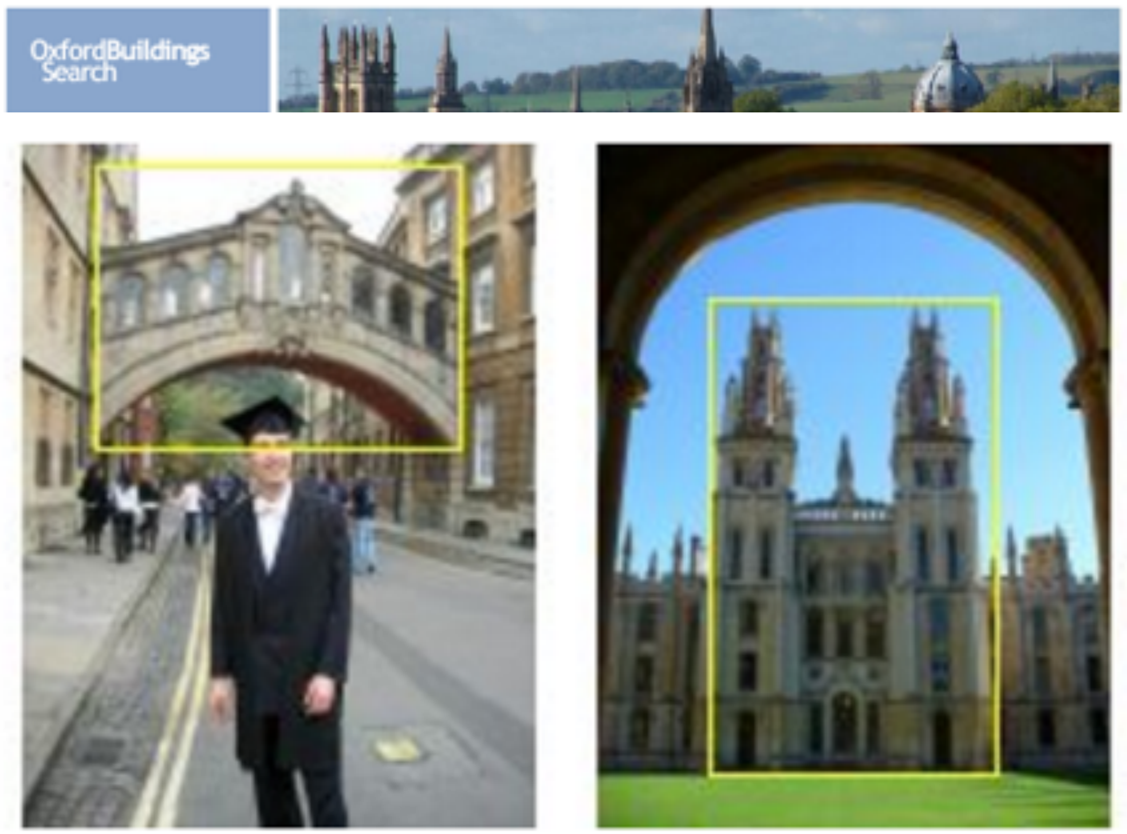
Slide Credit - Chum, LSVR tutorial at CVPR'13

Query Expansion



Slide Credit - Chum, LSVR tutorial at CVPR'13

Video Google - Object Instance Retrieval

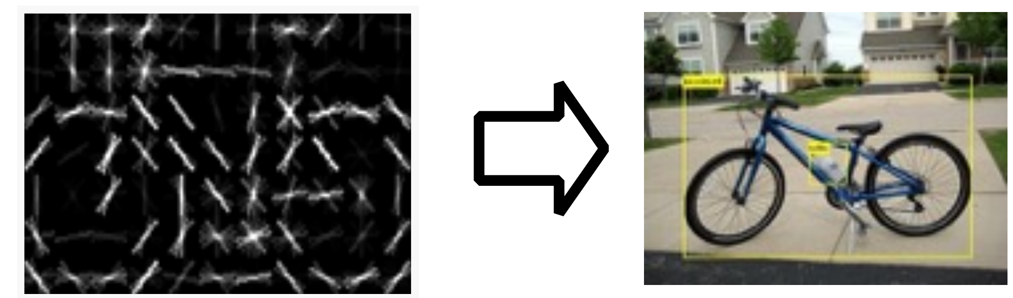


http://www.robots.ox.ac.uk/~vgg/demo/

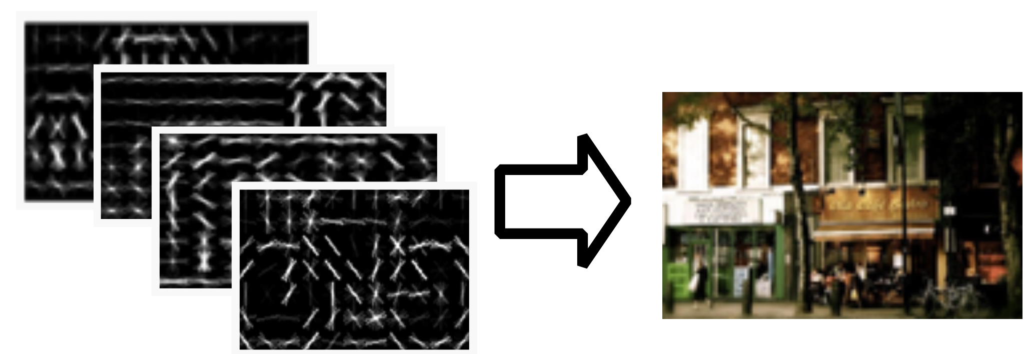
Immediate, scalable object category detection

Motivation: Object Detection

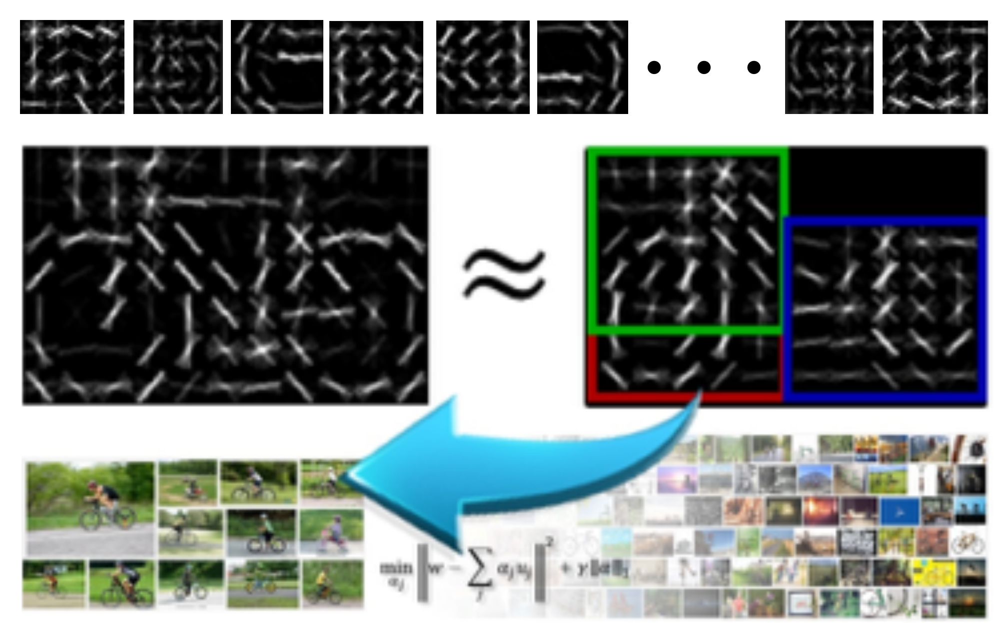
- Running ***a detector*** fast on a ***single image*** (Cascades, PQ, etc.) [Felzenszwalb-CVPR10,Vedaldi-CVPR12, Sadeghi-NIPS13].



- Running ***multiple detectors*** fast on a ***single image*** (Sparselets, etc.) [Song-ECCV12, Dean-CVPR13].



- Running ***a detector*** fast **(~1sec)** on a ***large-scale image dataset***, similar to Video Google [Sivic03] but for category detection.



Large Scale Object Instance Retrieval

query

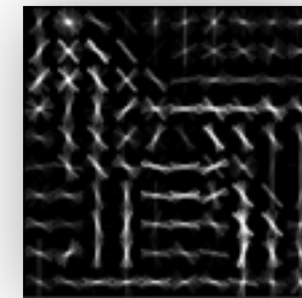
- Retrieve instantly (< 1sec)
- Robust to: scale, viewpoint, lighting, partial occlusion

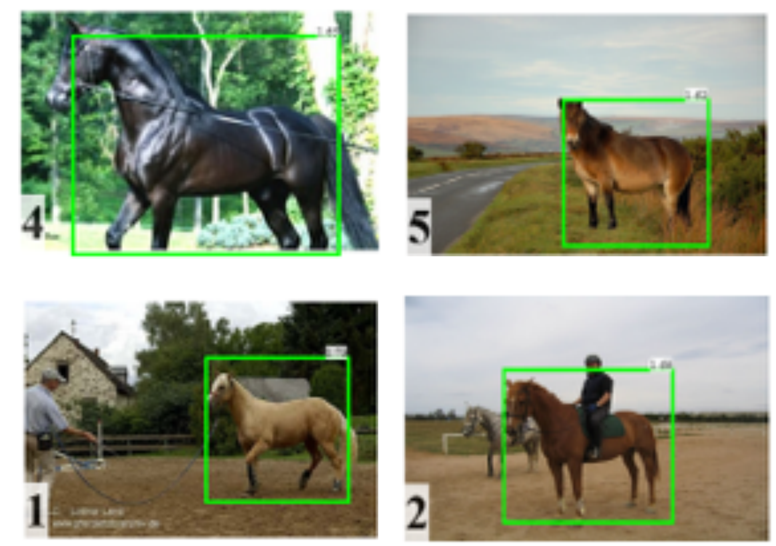
versus

retrieval results

Large Scale Object Category Detection

query





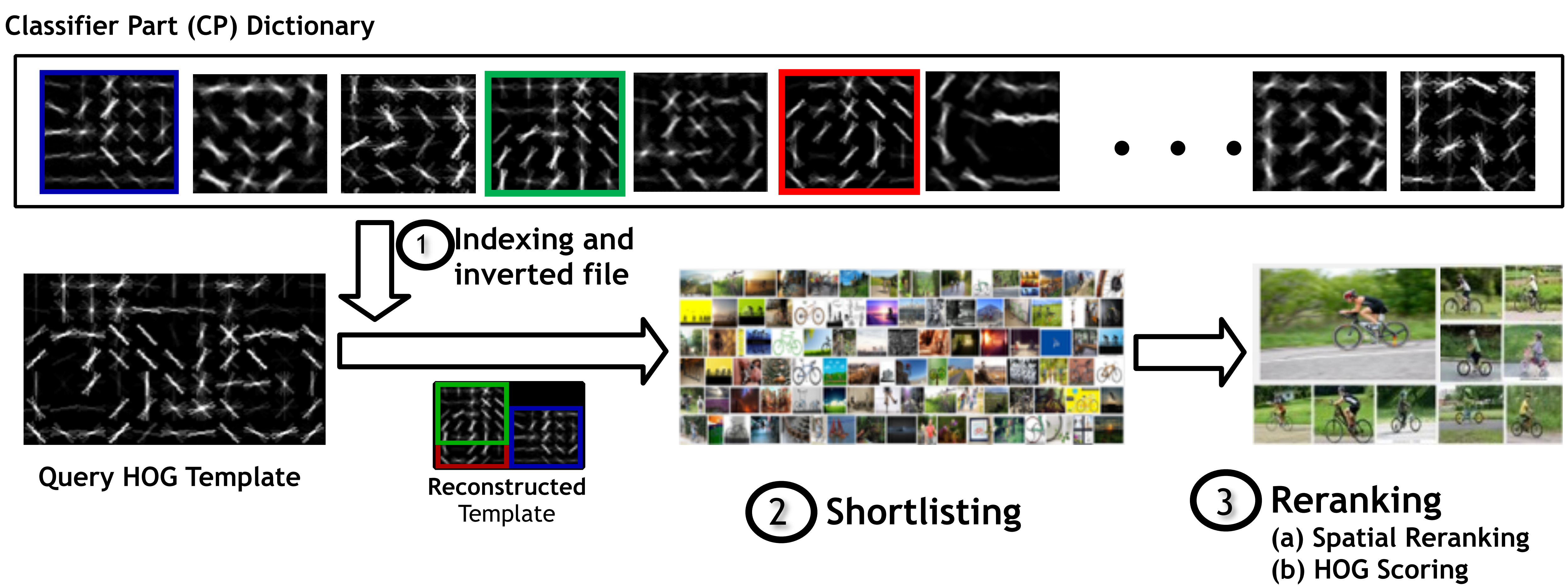
- Retrieve instantly (~ 1sec)
- Robust to: scale, viewpoint, lighting, partial occlusion and

Intra-class variance

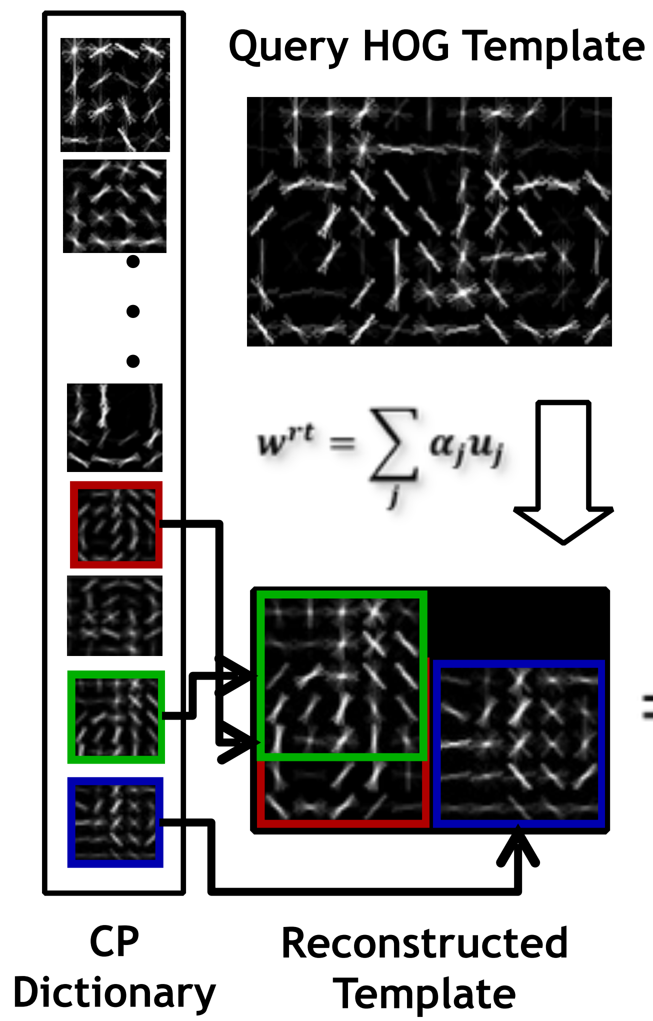
[Arandjelovic-CVPR12] -

Overview

Uses the three stages of Video Google revamped for object category detection

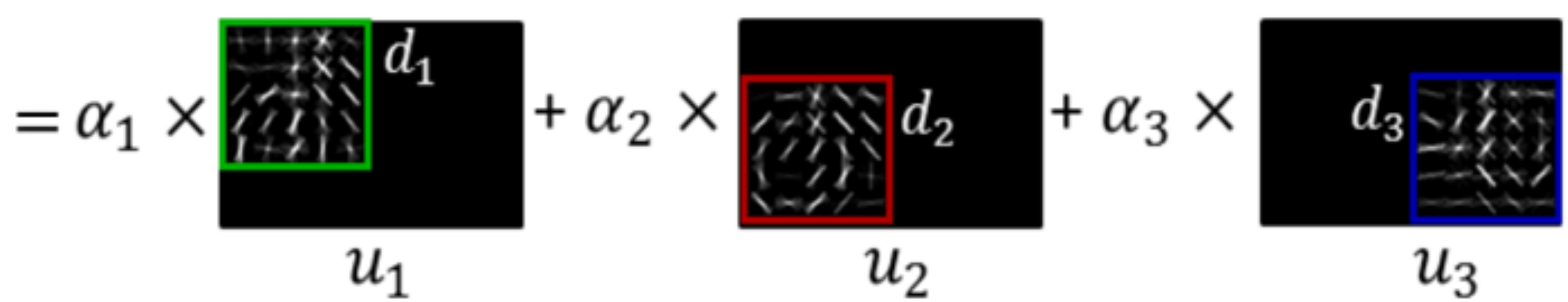


Query Representation



Query (HOG template) is first represented as a sparse combination of CPs.

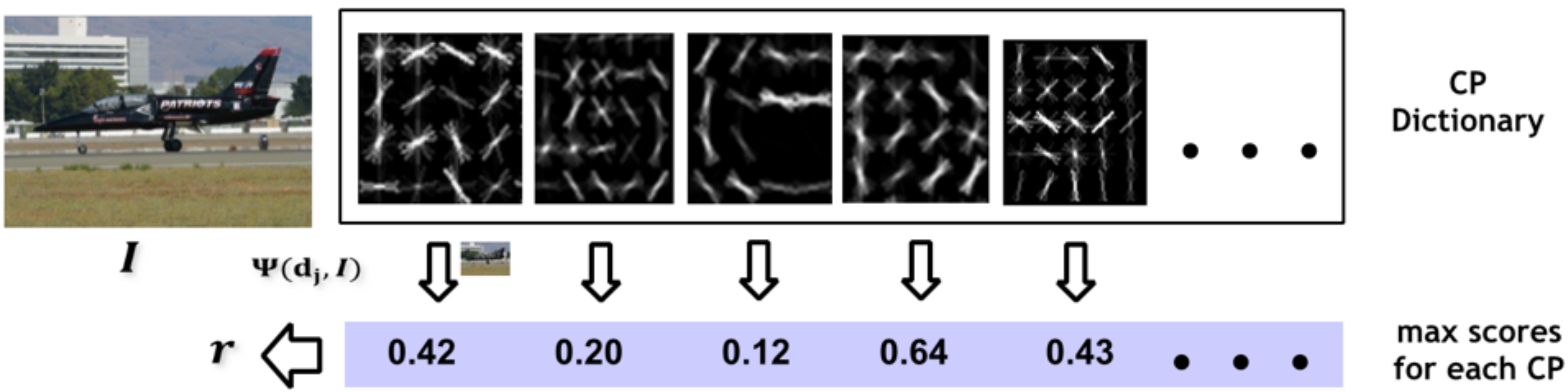
$$\min_{\alpha} \left\| w - \sum_j \alpha_j u_j \right\|^2 + \gamma \|\alpha\|_1 \qquad st: \alpha_j \geq 0$$



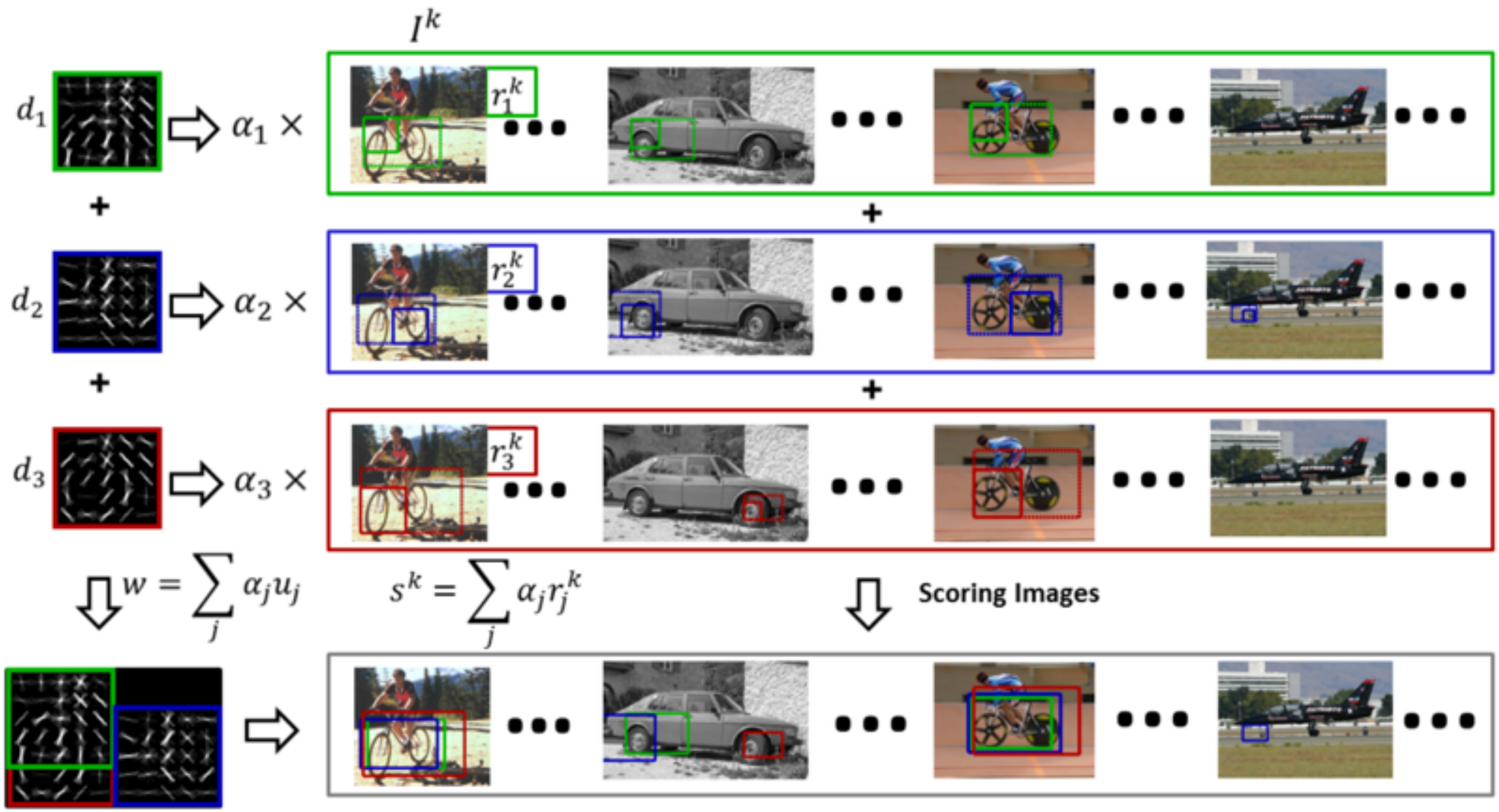
α and the spatial layouts of CPs define the reconstructed template w^{rt}

Image Representation

Representation: Image $\boldsymbol{I}$ is represented with vector $\boldsymbol{r} = [\boldsymbol{\Psi}(\boldsymbol{d_1}, \boldsymbol{I})\ \ \boldsymbol{\Psi}(\boldsymbol{d_2}, \boldsymbol{I}), \ldots, \boldsymbol{\Psi}(\boldsymbol{d_M}, \boldsymbol{I})]$, where the $\boldsymbol{i^{th}}$ component is the maximum response of the CP $\boldsymbol{d_i}$ sliding over $\boldsymbol{I}$.



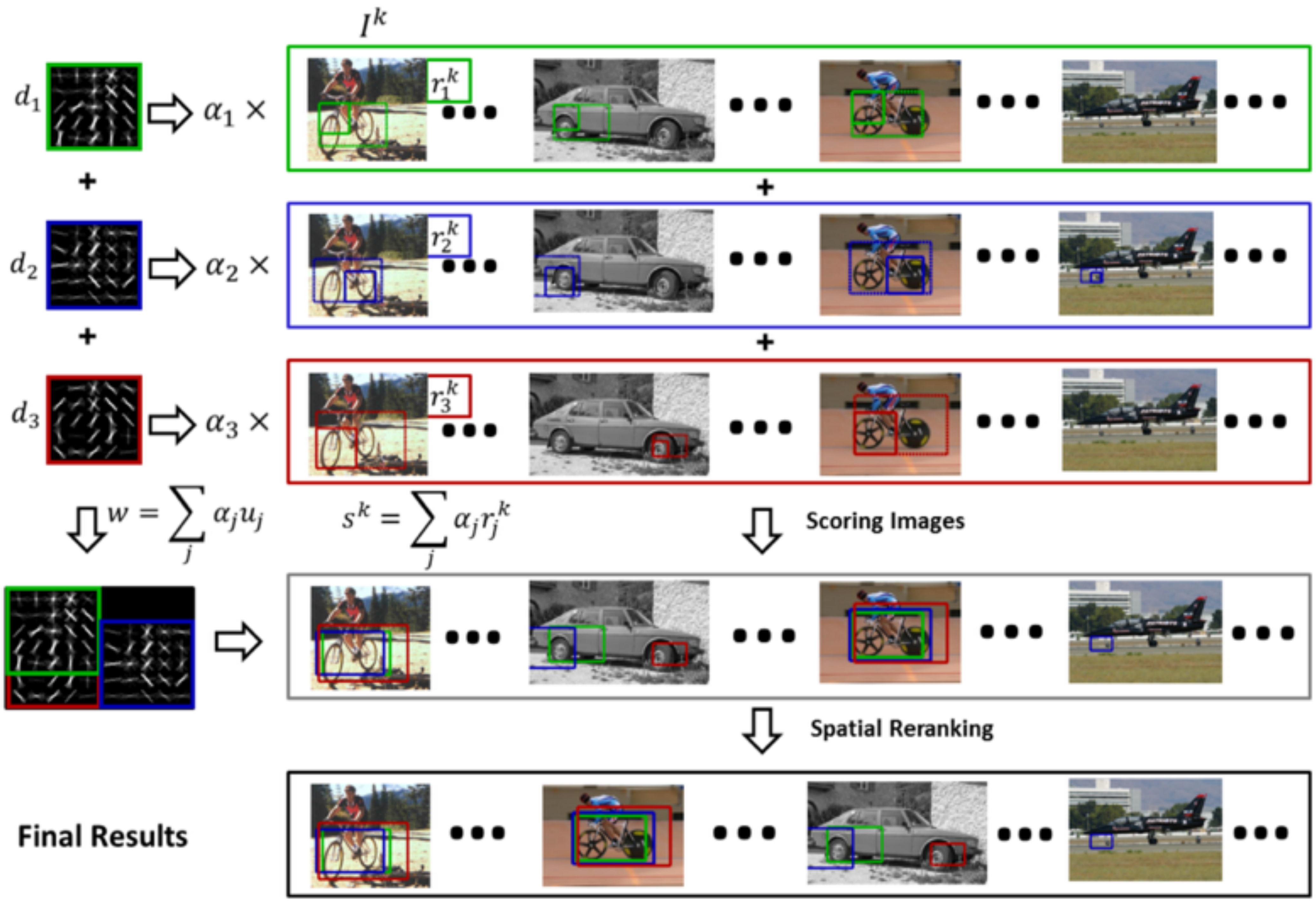
Shortlisting



Shortlisting: For the given template $\boldsymbol{w}^{rt}$, images are shortlisted using the score $\boldsymbol{\alpha}^T\boldsymbol{r}$ which is an upper bound on the maximum score of $\boldsymbol{w}^{rt}$ obtained by sliding over $\boldsymbol{I}$.

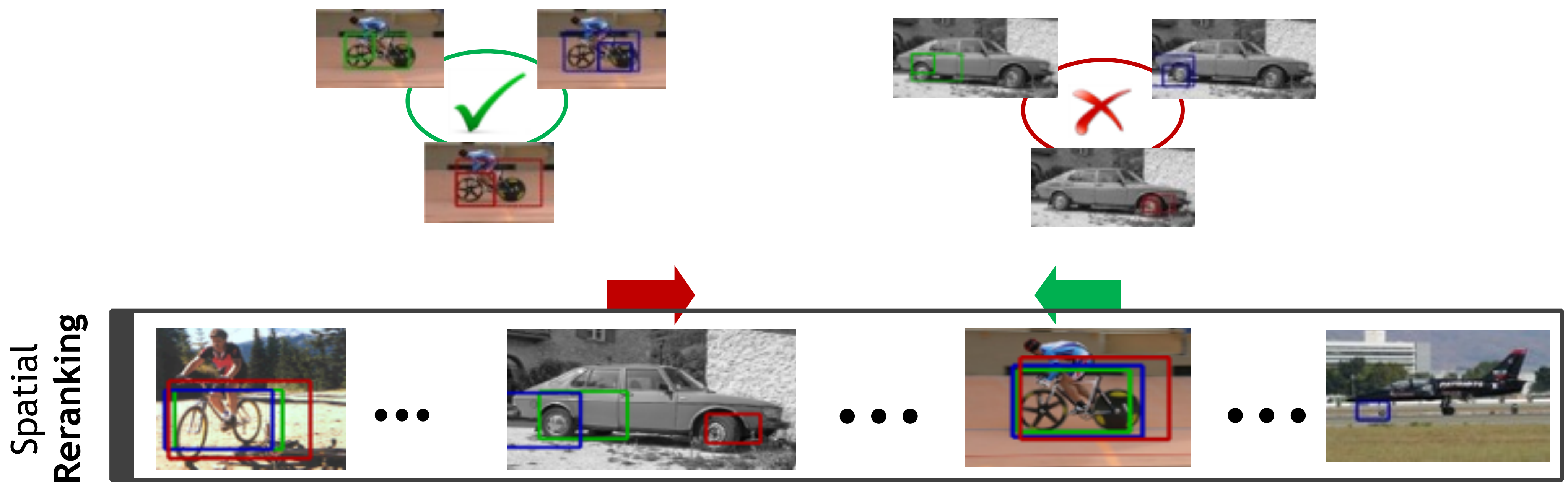
$$\boldsymbol{\alpha}^T\boldsymbol{r} = \sum_j \alpha_j \Psi(\boldsymbol{d}_j, \boldsymbol{I}) \geq \Psi(\boldsymbol{w}^{rt}, \boldsymbol{I})$$

Spatial Reranking



Reranking (Spatial + Original Template)

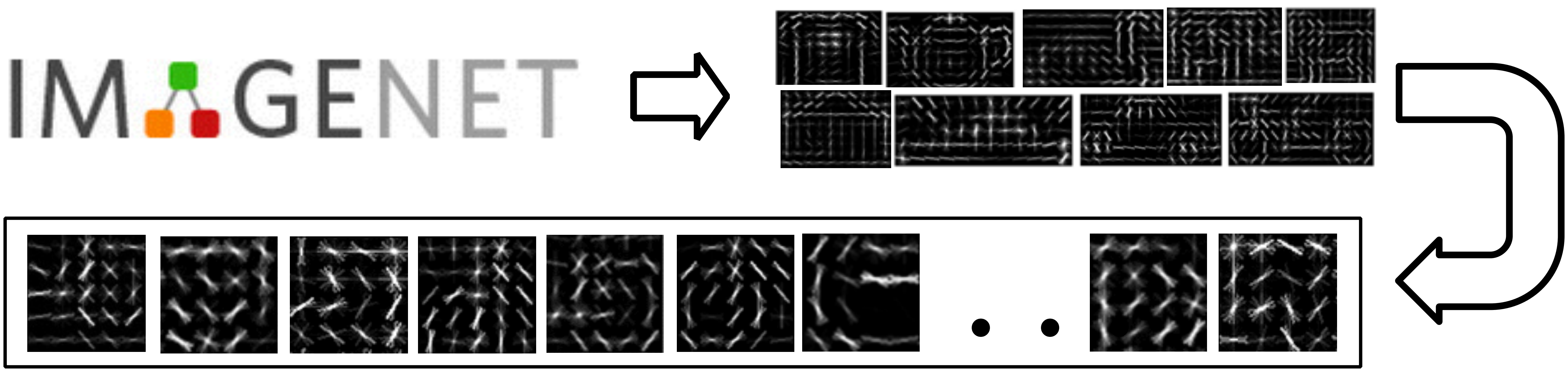
Shortlisted images are reranked via fast Hough-like voting of bounding box candidates suggested by each CP.



Retrieved bounding box candidates are re-scored using the original HOG template with fast and memory efficient PQ compression.

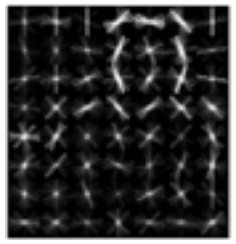
Dictionary & Dataset

10K dictionaries of sizes 3x3 – 7x7 HOG cells are extracted from DPMs trained from 1000 **ImageNet** categories.

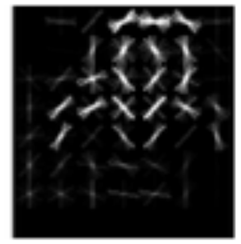


Tests are performed on **PASCAL VOC07** test set (5K images) and validation sets (100K images) of **ImageNet** 2011 and 2012 challenges.

Detection Results

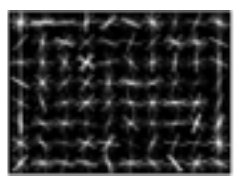


Person Template

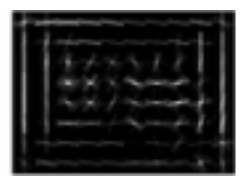


Reconstructed Person Template

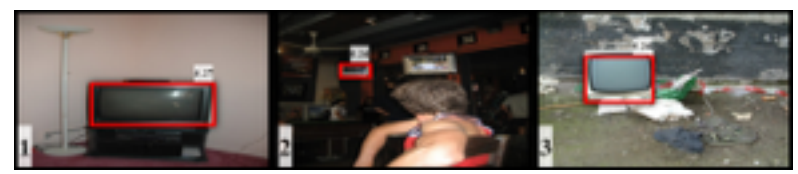
Top 3 retrievals



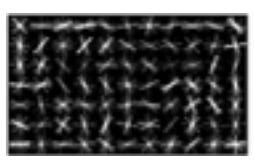
TV/Monitor Template



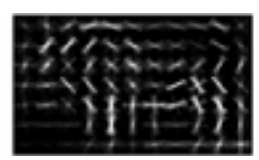
Reconstructed TV/Monitor Template



Top 3 retrievals

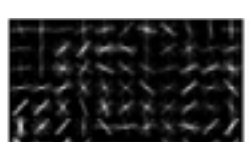


Cow Template



Reconstructed Cow Template

Top 3 retrievals



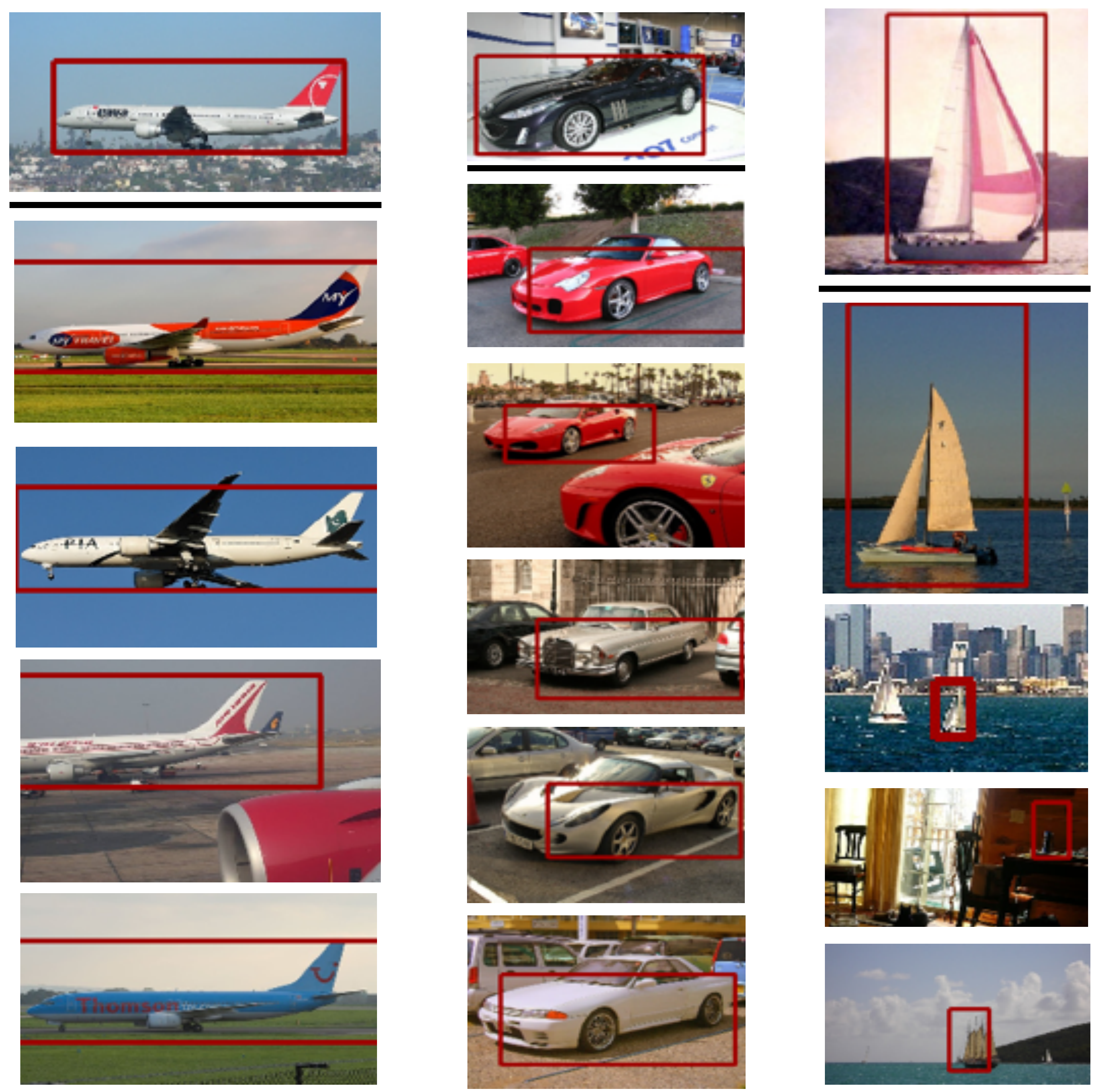
Motorbike Template

Reconstructed Motorbike Template

Top 3 retrievals

Immediate, scalable object category detection, Y. Aytar, A. Zisserman, CVPR 2014

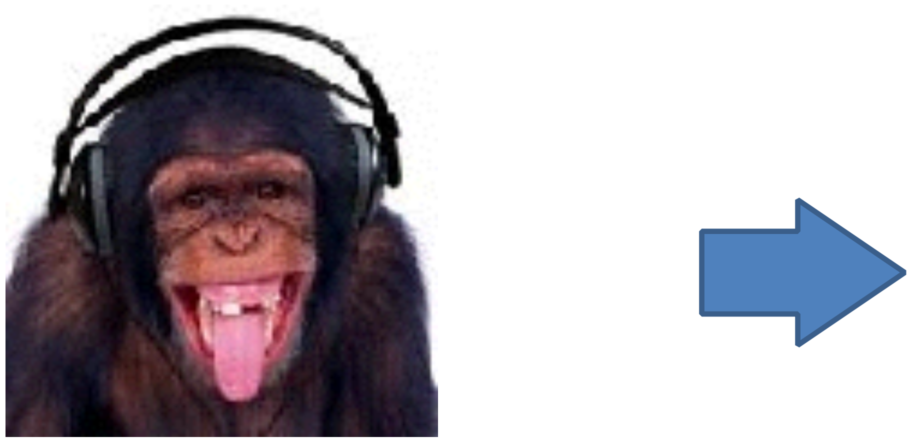
Exemplar SVM Results



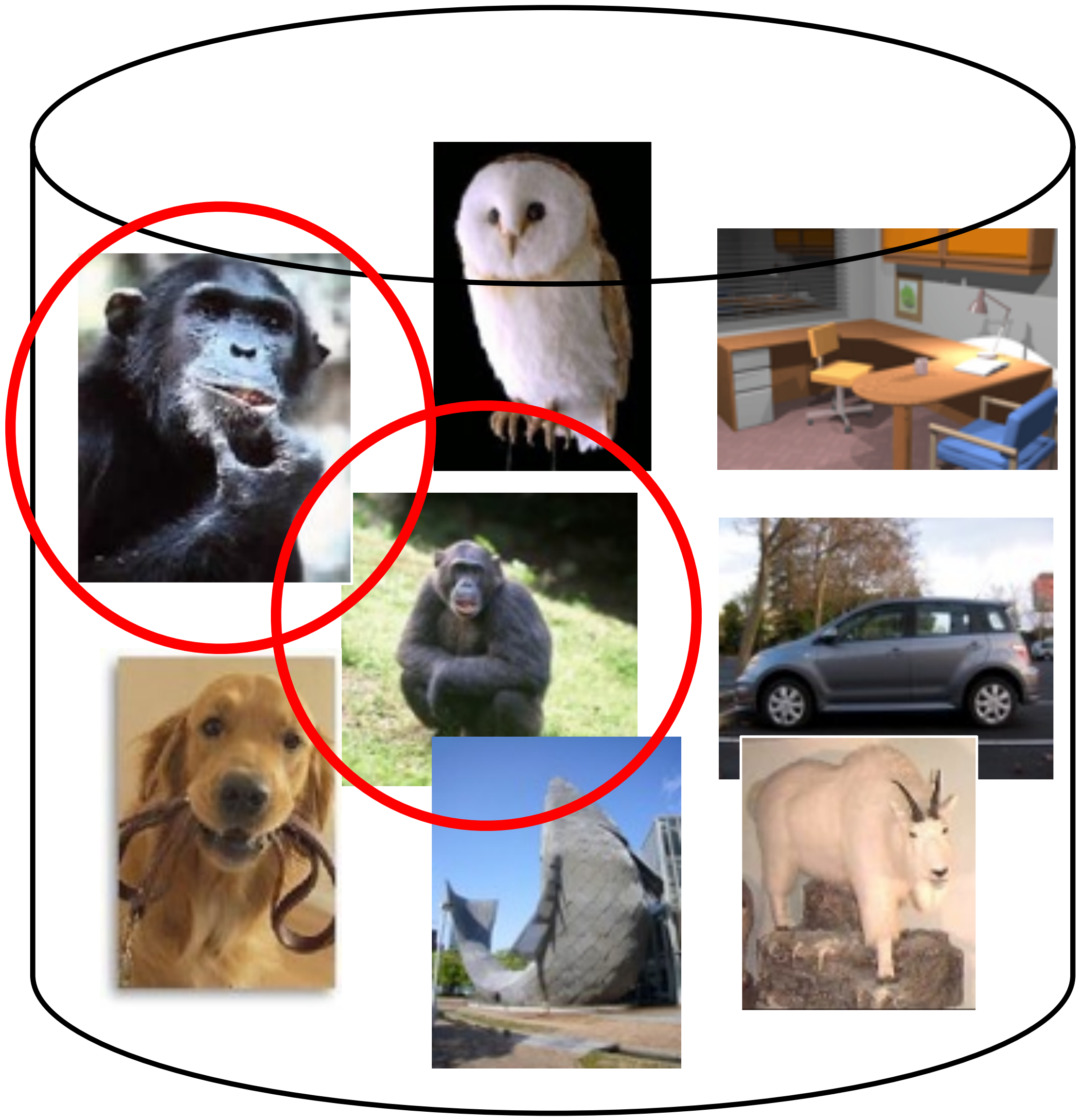
Large Scale Image Search

Large Scale Image Search

- Find similar images in a large database



Fast & Accurate



Slide Credit - Kristen Grauman et al

Large Scale Image Search

Internet contains billions of images

Search the internet

The Challenge:

Need way of measuring similarity between images
(distance metric learning)

Needs to scale to Internet (How?)

Requirements for image search

- Search must be both **fast, accurate** and **scalable to large data set**

- Fast
  - Kd-trees: tree data structure to improve search speed
  - Locality Sensitive Hashing: hash tables to improve search speed
  - Small code: binary small code (010101101)
- Scalable
  - Require very little memory, enabling their use on standard hardware or even on handheld devices
- Accurate
  - Learned distance metric

Categorization of existing large scale image search algorithms

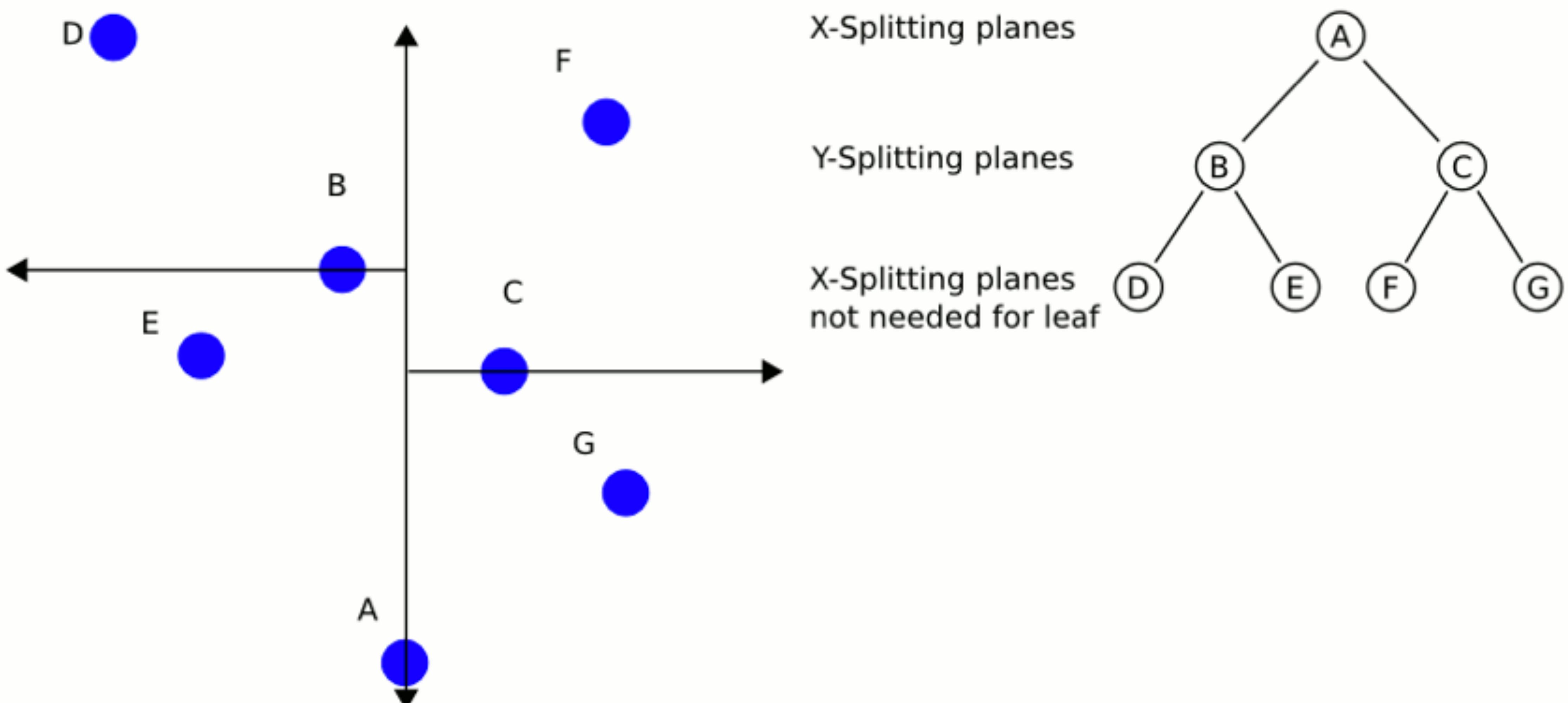
- Tree Based Structure
  - Spatial partitions (i.e. kd-tree) and recursive hyper plane decomposition provide an efficient means to search low-dimensional vector data exactly.

- Hashing
  - Locality-sensitive hashing offers **sub-linear** time search by hashing highly similar examples together.

- Binary Small Code
  - Compact binary code, with a few hundred bits per image

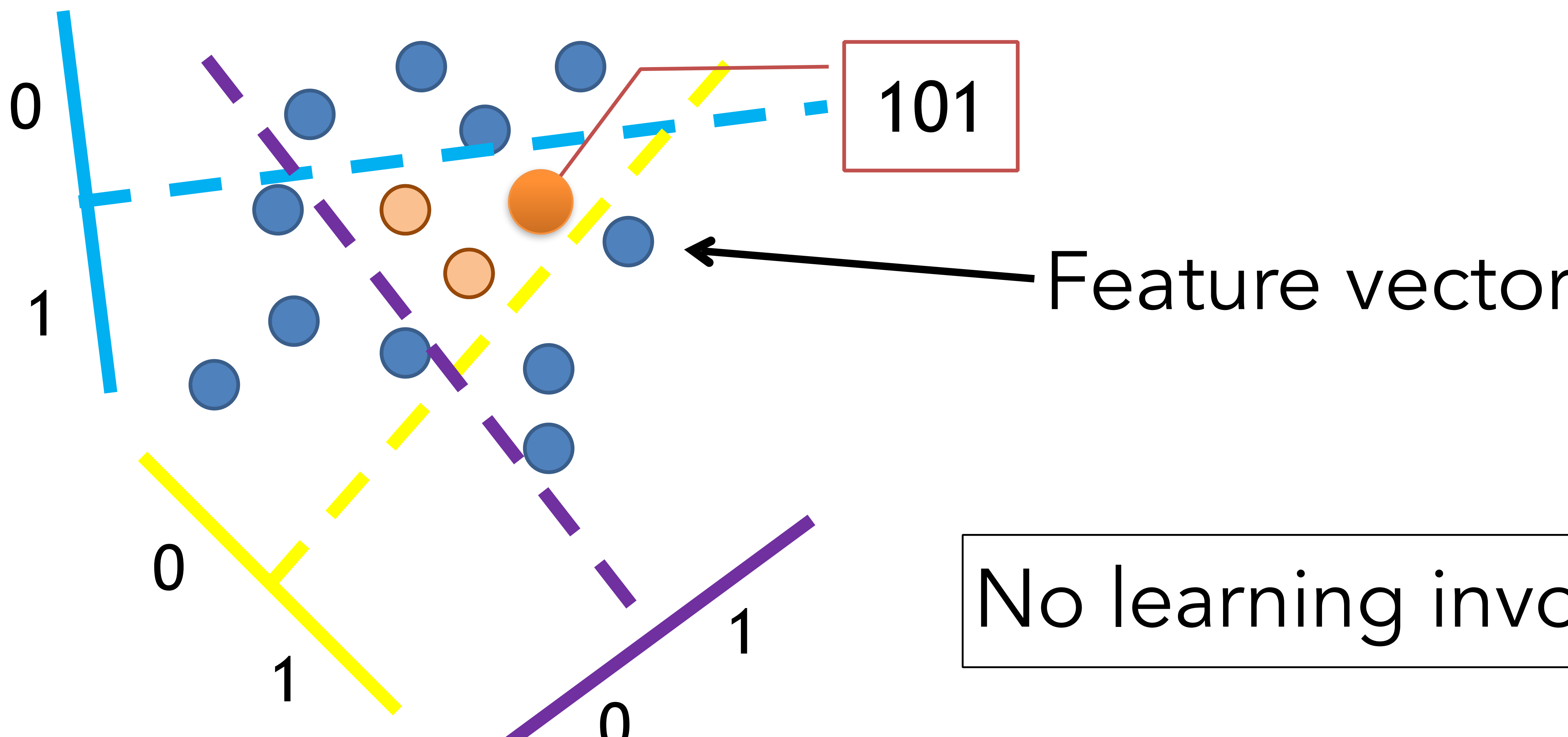
Tree Based Structure

- Kd-tree
  - The kd-tree is a binary tree in which every node is a k-dimensional point



Locality Sensitive Hashing

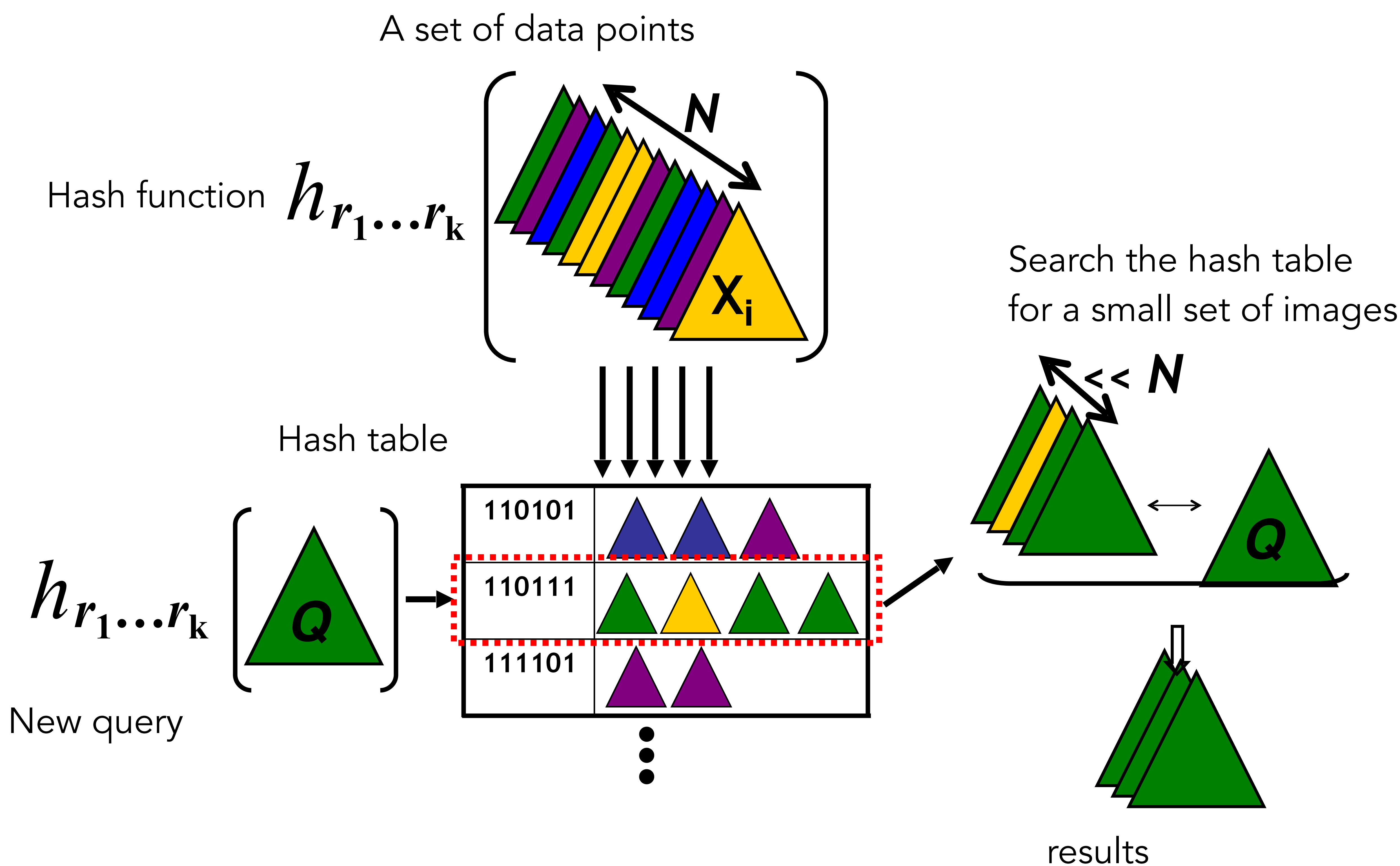
- Take random projections of data $r^T x$
- Quantize each projection with few bits



No learning involved

Slide Credit - Fergus et al

How to search from hash table?



Slide Credit - Kristen Grauman et al.

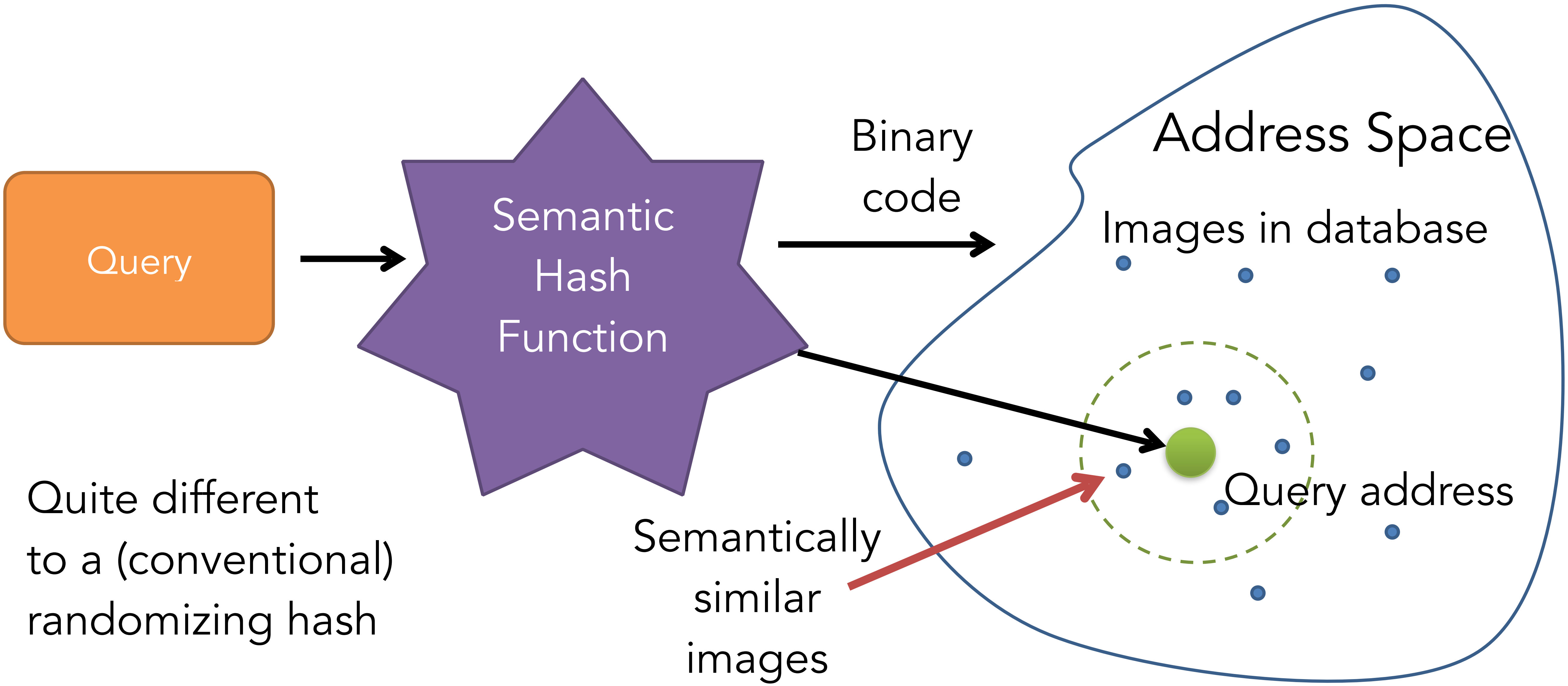
Binary codes for images

- Want images with similar content to have similar binary codes

- Use Hamming distance between codes
  - Number of bit flips
  - E.g.: Ham_Dist(10001010,10001110)=1
    Ham_Dist(10001010,11101110)=3

- Semantic Hashing [Salakhutdinov & Hinton, 2007]
  - Text documents

Semantic Hashing



Quite different
to a (conventional)
randomizing hash

– Find neighbors by exploring Hamming ball around query address
– Lookup time depends on radius of ball, NOT on # data points

Small Codes and Large Databases for Recognition, A. Torralba, R. Fergus, and Y. Weiss, CVPR 2008
Semantic Hashing, Ruslan Salakhutdinov and Geoffrey Hinton, International Journal of Approximate Reasoning, 2009

Slide Credit - Fergus et al

Compact Binary Codes

- Google has few billion images (10^9)
- PC has ~10 Gbytes (10^{11} bits)
- Codes must fit in memory (disk too slow)

Budget of **10^2 bits/image**

- 1 Megapixel image is 10^7 bits
- 32x32 color image is 10^4 bits

Semantic hash function must also reduce dimensionality

Slide Credit - Fergus et al

RBM architecture

- Network of binary stochastic units
- Hinton & Salakhutdinov, Science 2006

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i \in \text{visible}} b_i v_i - \sum_{j \in \text{hidden}} b_j h_j - \sum_{i,j} v_i h_i w_{ij}$$

Parameters: Weights w Biases b

Convenient conditional distributions:

$$p(h_j = 1|\mathbf{v}) = \sigma(b_j + \sum_i w_{ij} v_i)$$

$$p(v_i = 1|\mathbf{h}) = \sigma(b_i + \sum_j w_{ij} h_j)$$

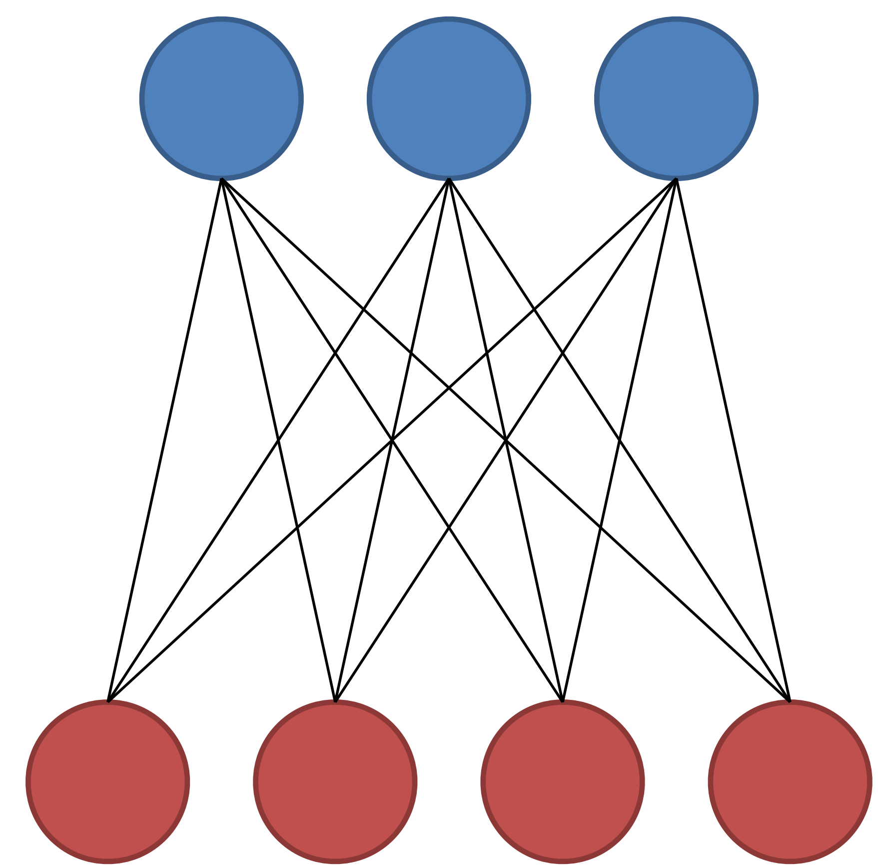
$\sigma(x) = 1/(1+e^{-x})$, the logistic function

Learn weights and biases using Contrastive Divergence

Hidden units: h

Symmetric weights w

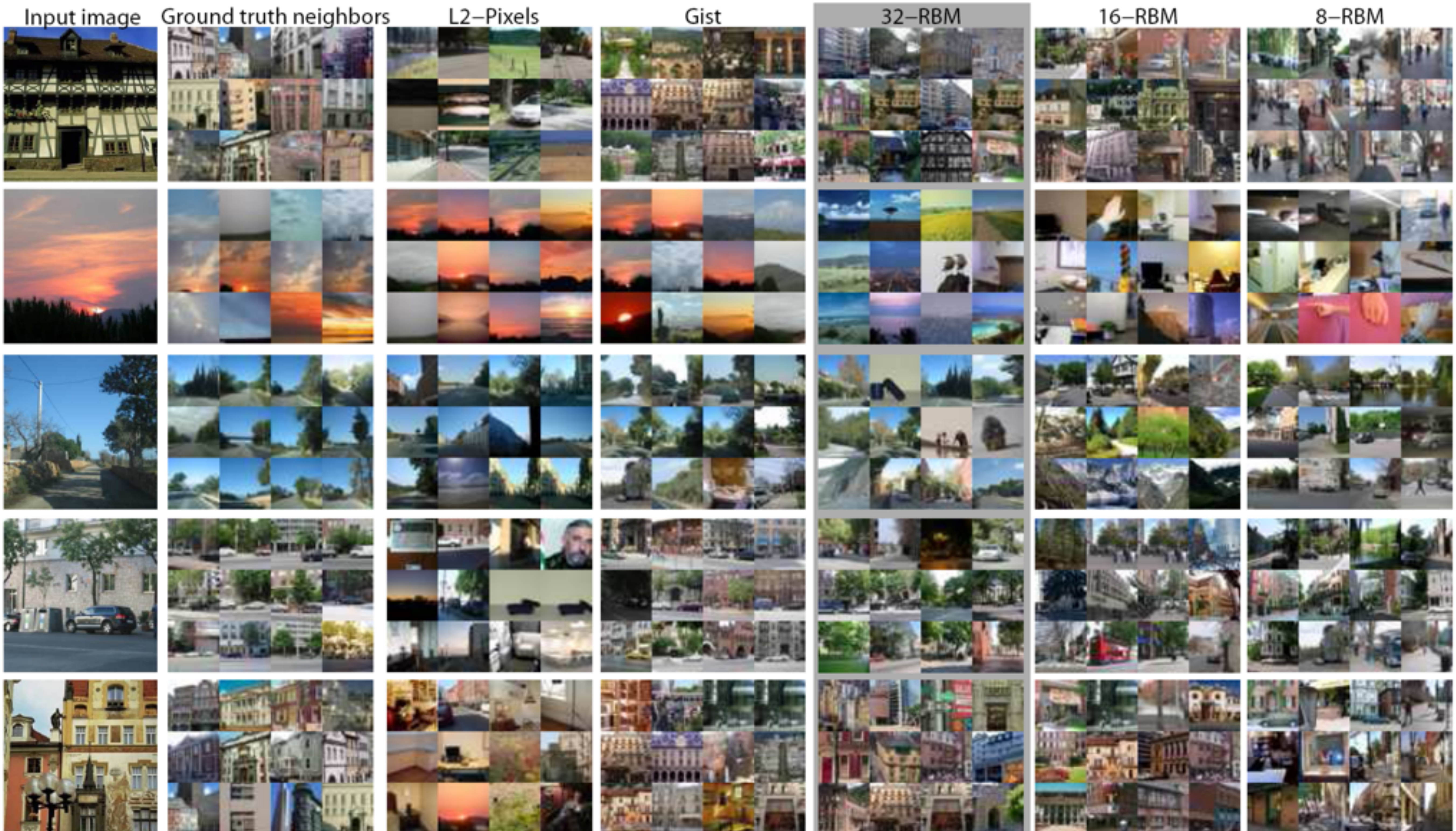
Visible units: v



Slide Credit - Fergus et al

Examples of LabelMe retrieval

12 closest neighbors under different distance metrics



Slide Credit - Fergus et al