CS294-43: Visual Object and Activity Recognition

Prof. Trevor Darrell

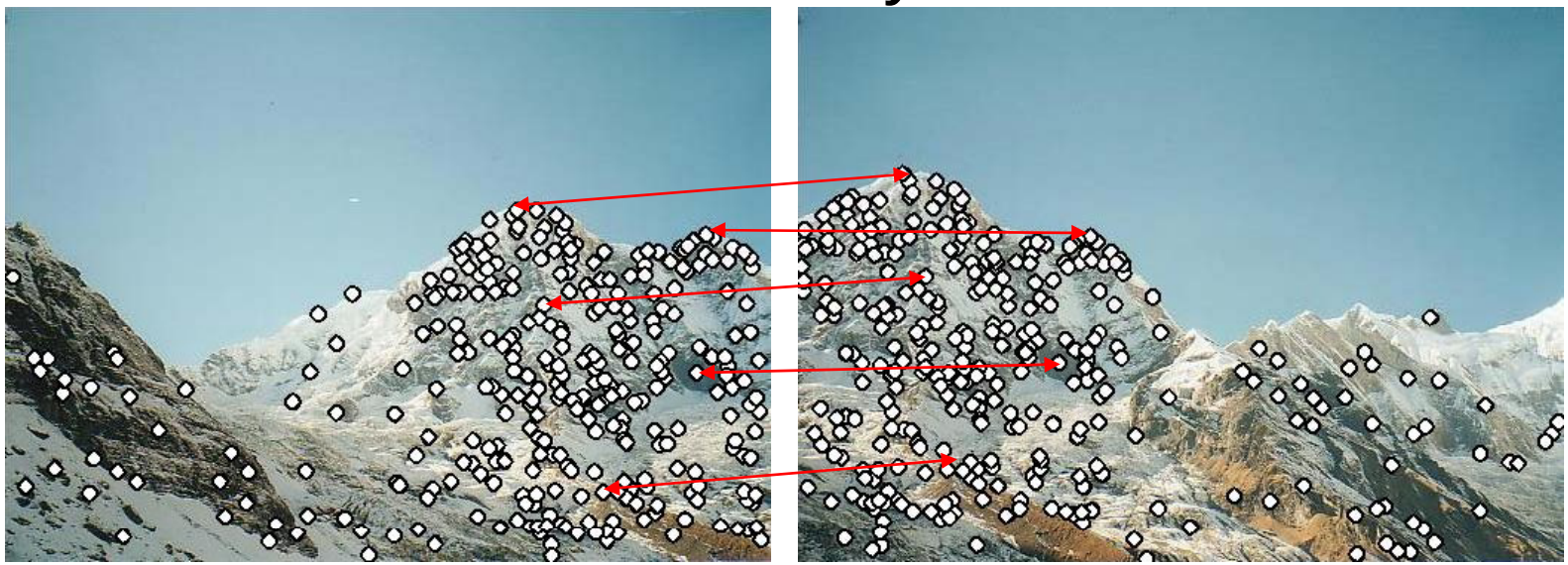
Jan 27th: Instance Recognition and Retrieval

Today

- SIFT
- Video Google
- Total Recall
- Photo Tourism

Correspondence

- Fundamental to many of the core vision problems
  - Recognition
  - Motion tracking
  - Multiview geometry
- Local features are the key



Images from: M. Brown and D. G. Lowe. Recognising Panoramas. In Proceedings of the the International Conference on Computer Vision (ICCV2003)

Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

Local Features: Detectors vs. Descriptors

Detected
Interest Points/Regions

Descriptors

<0 12 31 0 0 23 …>

<5 0 0 11 37 15 …>

<14 21 10 0 3 22 …>

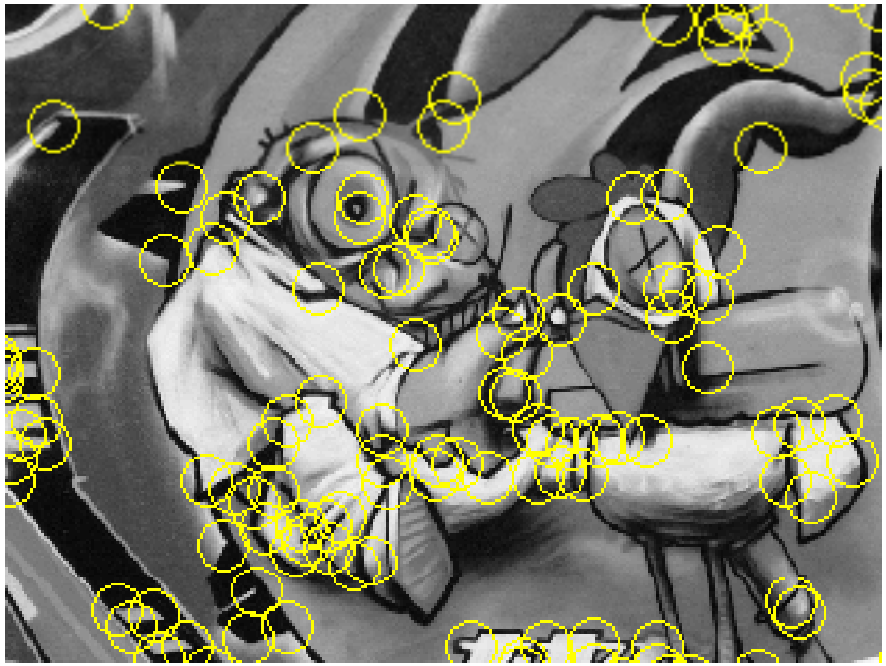
Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

Ideal Interest Points/Regions

- Lots of them
- Repeatable
- Representative orientation/scale
- Fast to extract and match

Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

Keypoint Localization

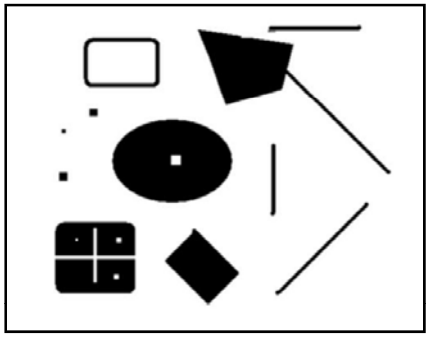


- **Goals:**
  - **Repeatable detection**
  - **Precise localization**
  - **Interesting content**

⇒ ***Look for two-dimensional signal changes***

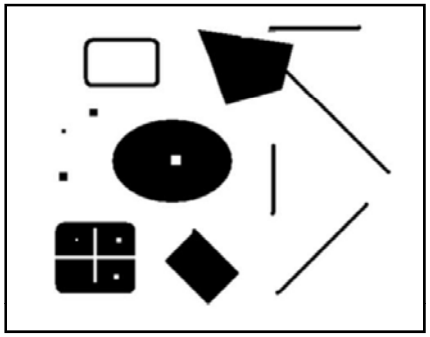
Harris Detector [Harris88]

Intuition: **Search for local neighborhoods where the image content has two main directions (eigenvectors).**

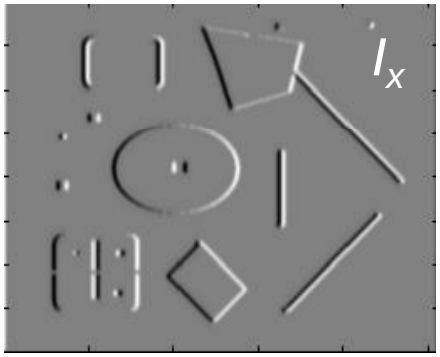


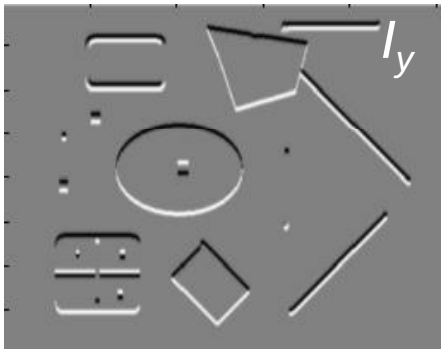
Harris Detector [Harris88]

Intuition: **Search for local neighborhoods where the image content has two main directions (eigenvectors).**



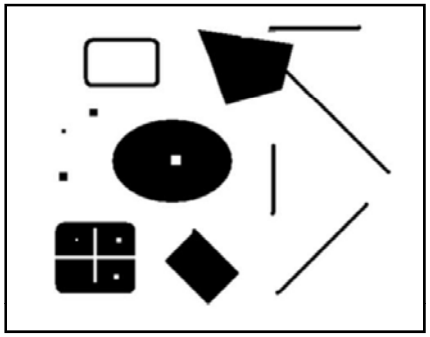
1. Image derivatives
   $g_x(\sigma_D)$, $g_y(\sigma_D)$,



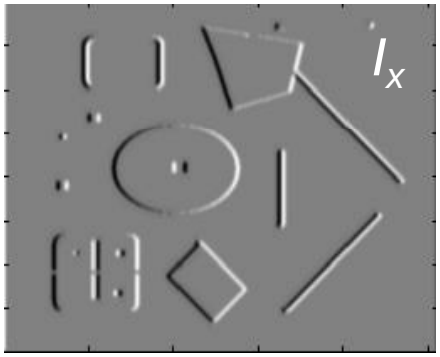


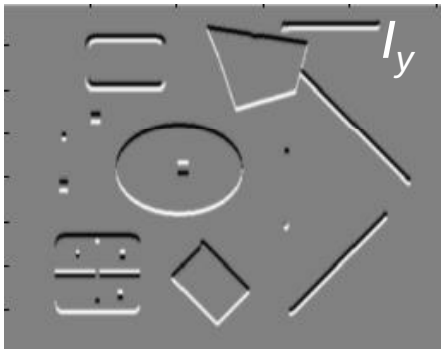
Harris Detector [Harris88]

Intuition: **Search for local neighborhoods where the image content has two main directions (eigenvectors).**

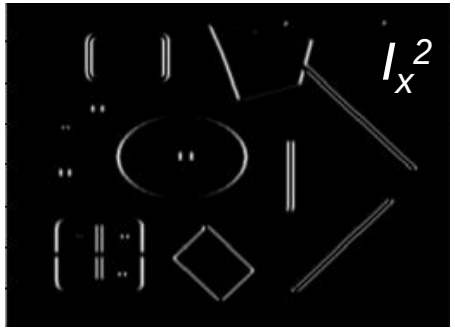


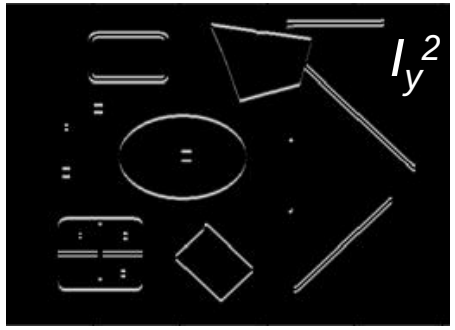
1. Image derivatives
   $g_x(\sigma_D)$, $g_y(\sigma_D)$,

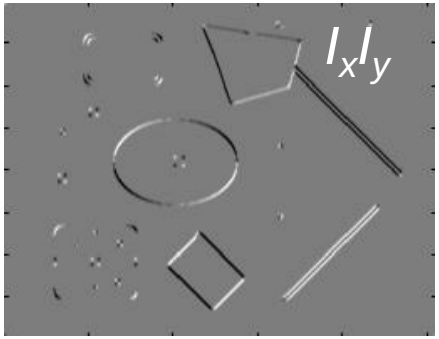




2. Square of derivatives





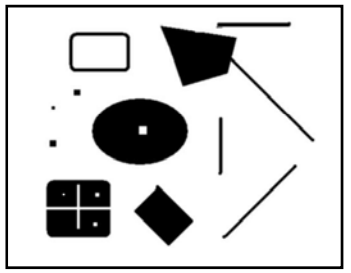


Slide credit K. Grauman, B. Leibe AAAI08 Short Course

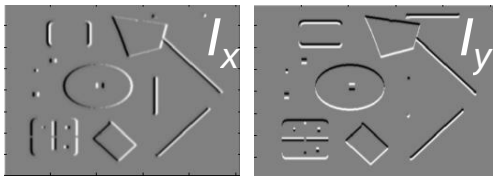
Harris Detector [Harris88]

Second moment matrix (autocorrelation matrix):

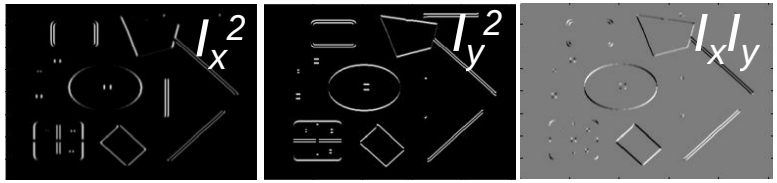
$$\mu(\sigma_I, \sigma_D) = g(\sigma_I) * \begin{bmatrix} I_x^2(\sigma_D) & I_x I_y(\sigma_D) \\ I_x I_y(\sigma_D) & I_y^2(\sigma_D) \end{bmatrix}$$



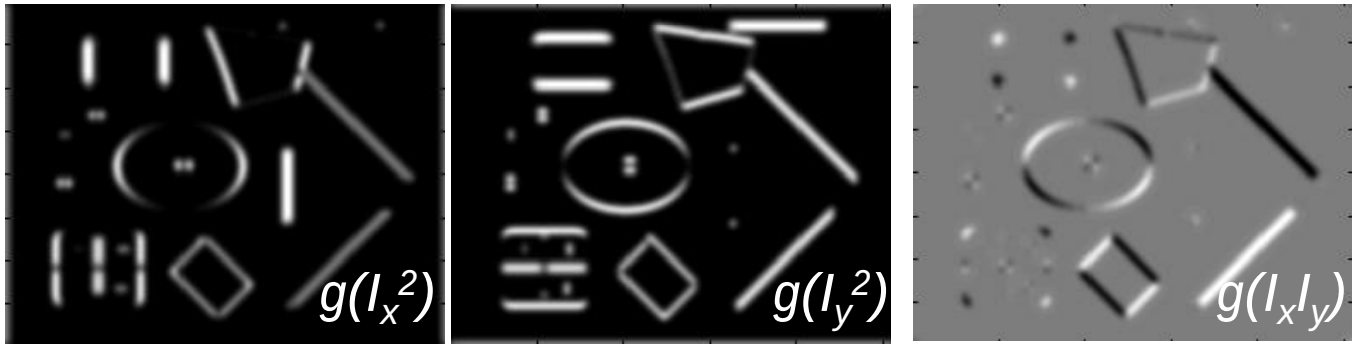
1. Image derivatives



2. Square of derivatives



3. Gaussian filter $g(\sigma_I)$

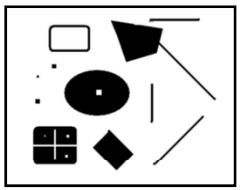


Slide credit K. Grauman, B. Leibe AAAI08 Short Course

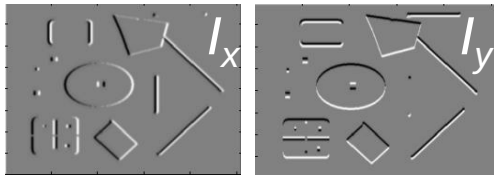
Harris Detector [Harris88]

Second moment matrix (autocorrelation matrix):

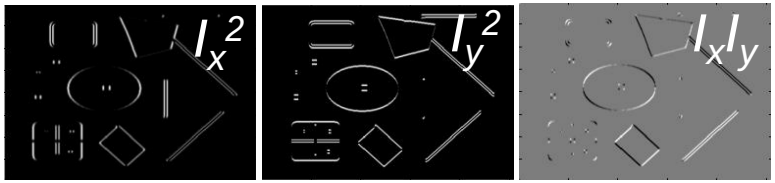
$$\mu(\sigma_I, \sigma_D) = g(\sigma_I) * \begin{bmatrix} I_x^2(\sigma_D) & I_x I_y(\sigma_D) \\ I_x I_y(\sigma_D) & I_y^2(\sigma_D) \end{bmatrix}$$



1. Image derivatives



2. Square of derivatives



3. Gaussian filter $g(\sigma_I)$

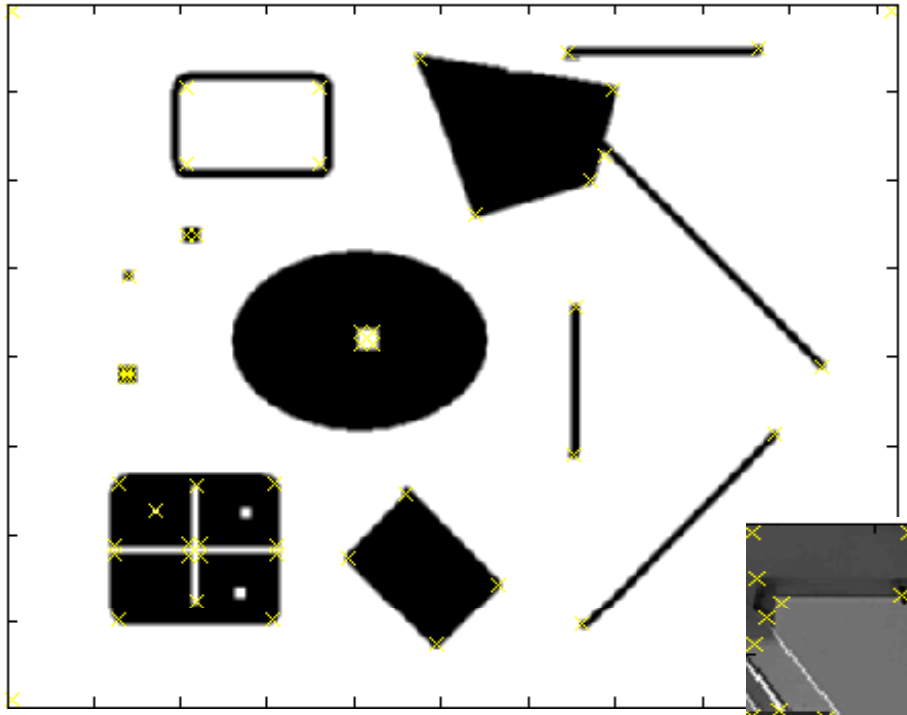
4. Cornerness function – both eigenvalues are strong

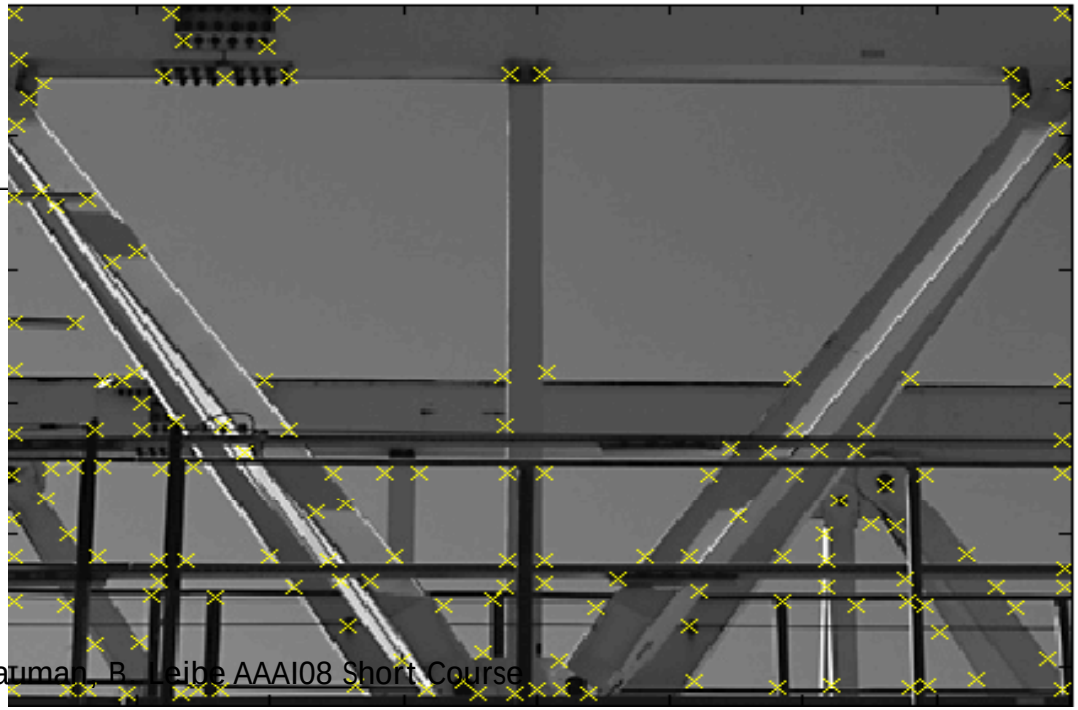
$$har = \det[\mu(\sigma_I, \sigma_D)] - \alpha[\text{trace}(\mu(\sigma_I, \sigma_D))] =$$

$$g(I_x^2)g(I_y^2) - [g(I_x I_y)]^2 - \alpha[g(I_x^2) + g(I_y^2)]^2$$

5. Non-maxima suppression

Harris Detector – Responses [Harris88]





Effect: **A very precise corner detector.**

Harris Detector – Responses [Harris88]

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Automatic Scale Selection

$$f(I_{i_1 \ldots i_m}(x,\sigma)) \quad = \quad f(I_{i_1 \ldots i_m}(x',\sigma'))$$

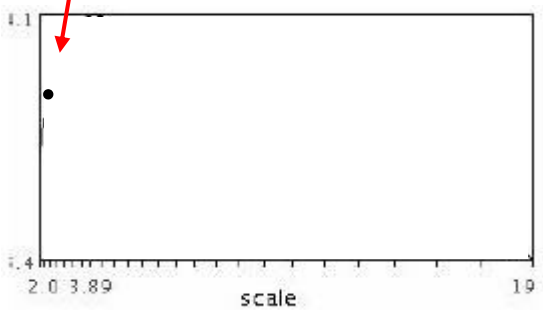
Same operator responses if the patch contains the same image up to scale factor

How to find corresponding patch sizes?

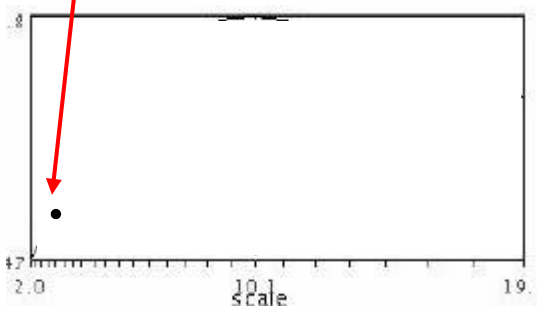
Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Automatic Scale Selection

- **Function responses for increasing scale (scale signature)**



$f(I_{i_1 \ldots i_m}(x,\sigma))$

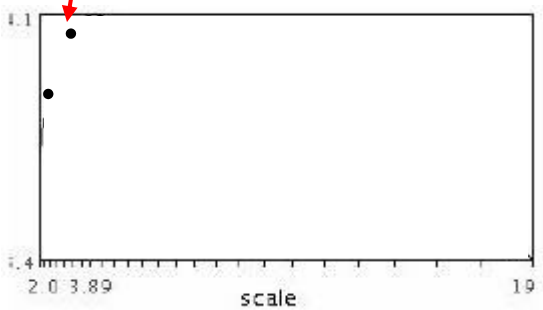


$f(I_{i_1 \ldots i_m}(x',\sigma))$

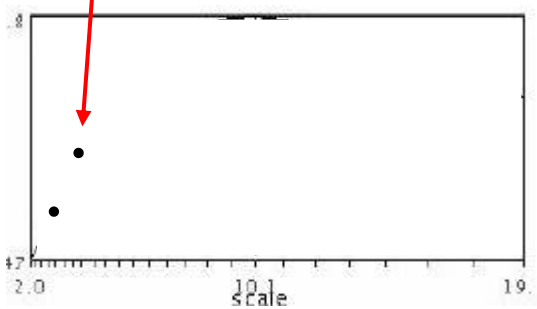
Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Automatic Scale Selection

- **Function responses for increasing scale (scale signature)**



$f(I_{i_1 \ldots i_m}(x,\sigma))$

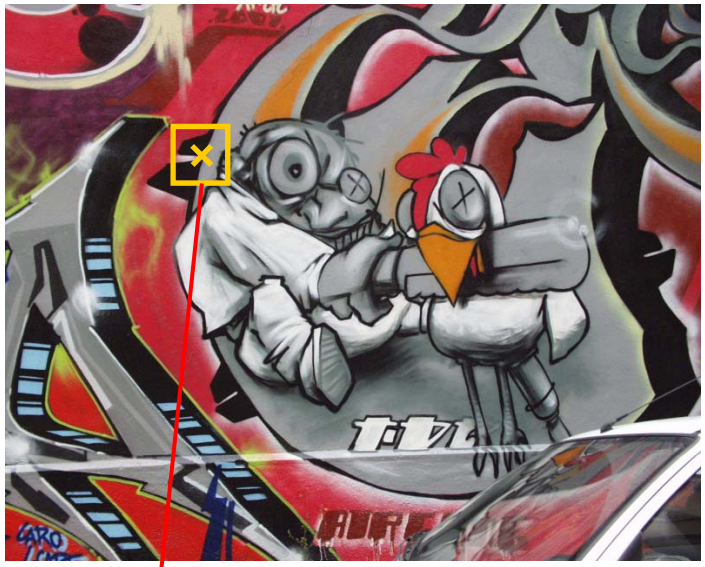


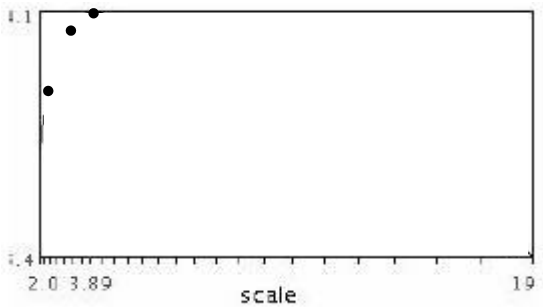
$f(I_{i_1 \ldots i_m}(x',\sigma))$

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

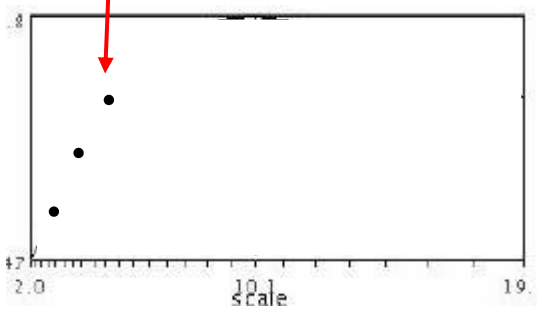
Automatic Scale Selection

- **Function responses for increasing scale (scale signature)**





$f(I_{i_1 \ldots i_m}(x,\sigma))$

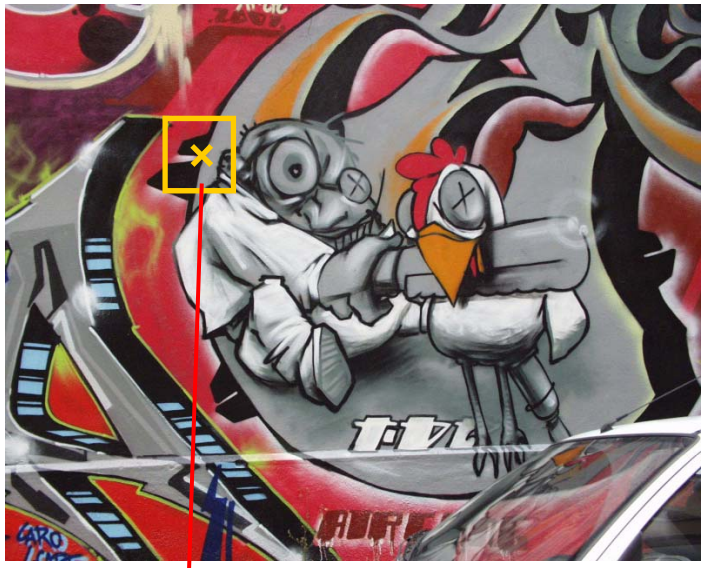


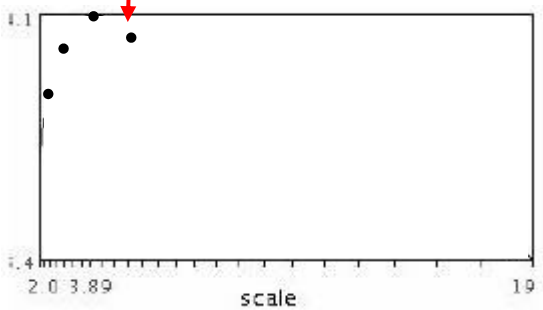
$f(I_{i_1 \ldots i_m}(x',\sigma))$

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

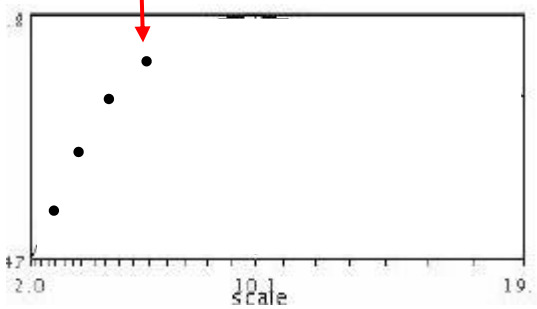
Automatic Scale Selection

- **Function responses for increasing scale (scale signature)**





$f(I_{i_1 \ldots i_m}(x,\sigma))$

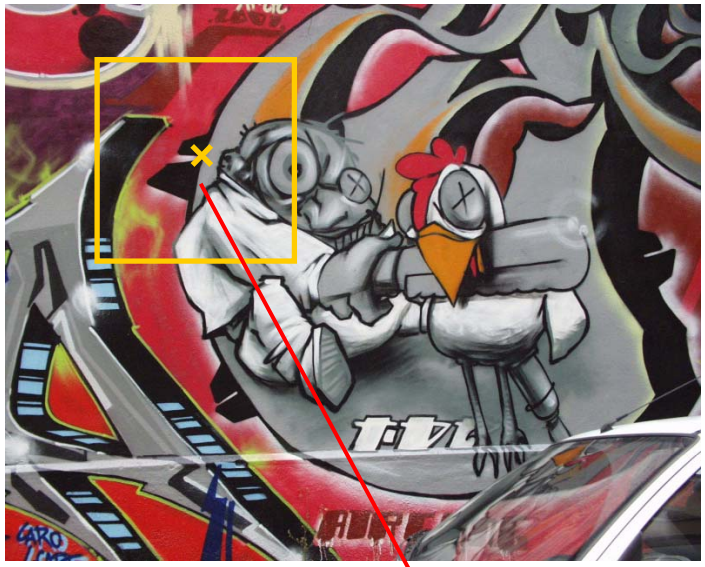


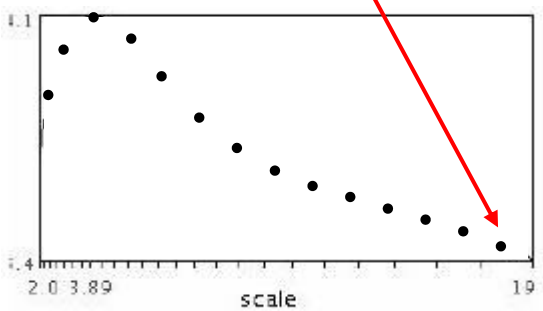
$f(I_{i_1 \ldots i_m}(x',\sigma))$

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

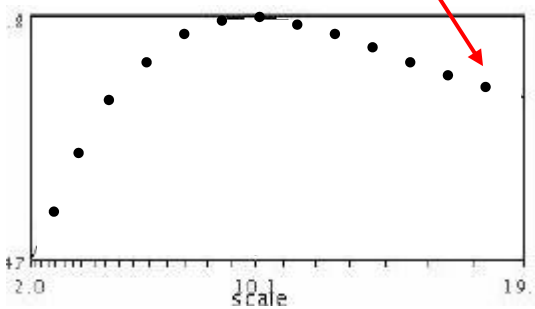
Automatic Scale Selection

- **Function responses for increasing scale (scale signature)**





$f(I_{i_1 \ldots i_m}(x,\sigma))$

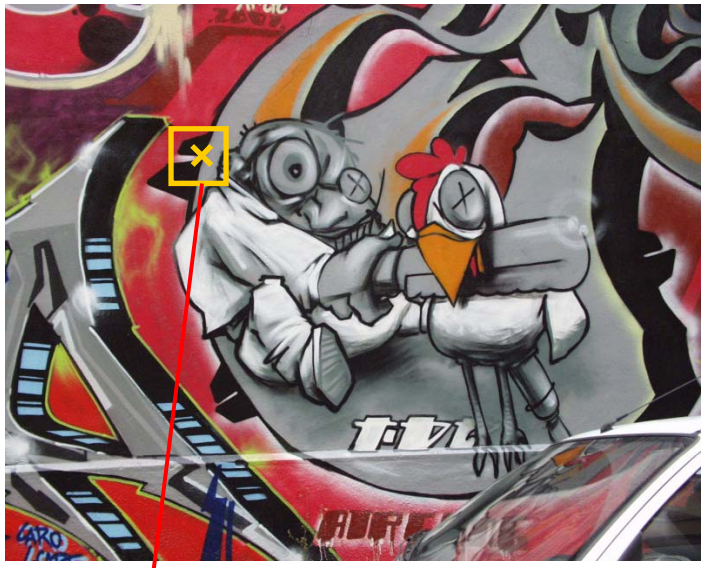


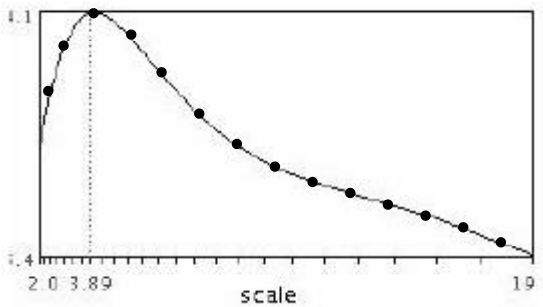
$f(I_{i_1 \ldots i_m}(x',\sigma))$

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

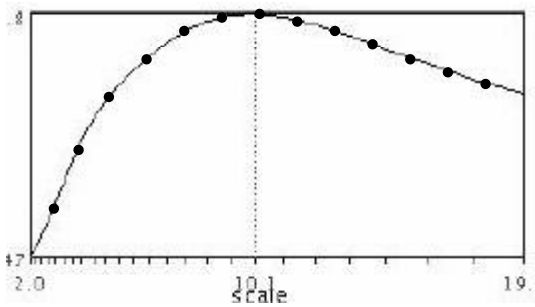
Automatic Scale Selection

- **Function responses for increasing scale (scale signature)**





$f(I_{i_1 \ldots i_m}(x,\sigma))$

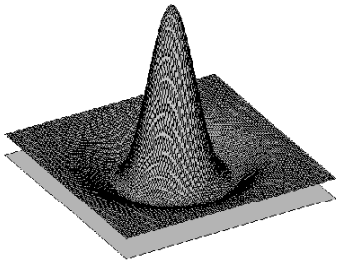


$f(I_{i_1 \ldots i_m}(x',\sigma'))$

Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Laplacian-of-Gaussian (LoG) scale detection

- **Laplacian also measures bandpass contrast…**
- **which ‘scale’ has most ‘contrast’?**



$$L_{xx}(\sigma) + L_{yy}(\sigma)$$

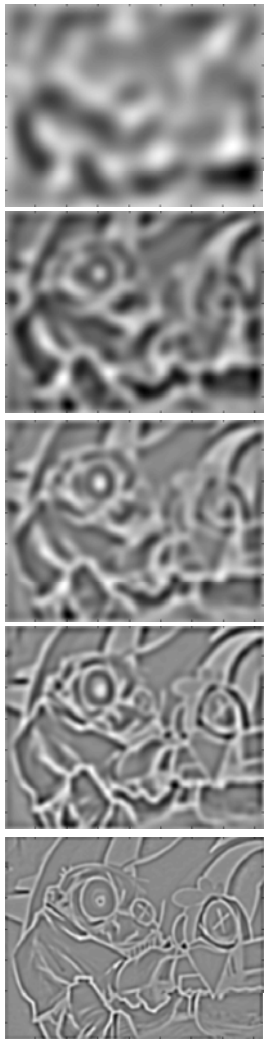
σ^5

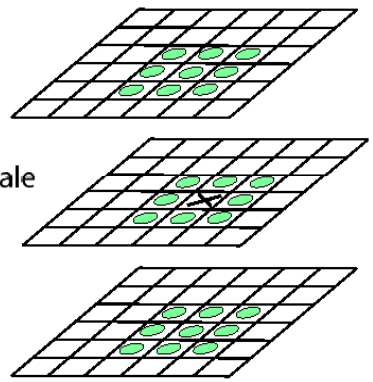
σ^4

σ^3

σ^2

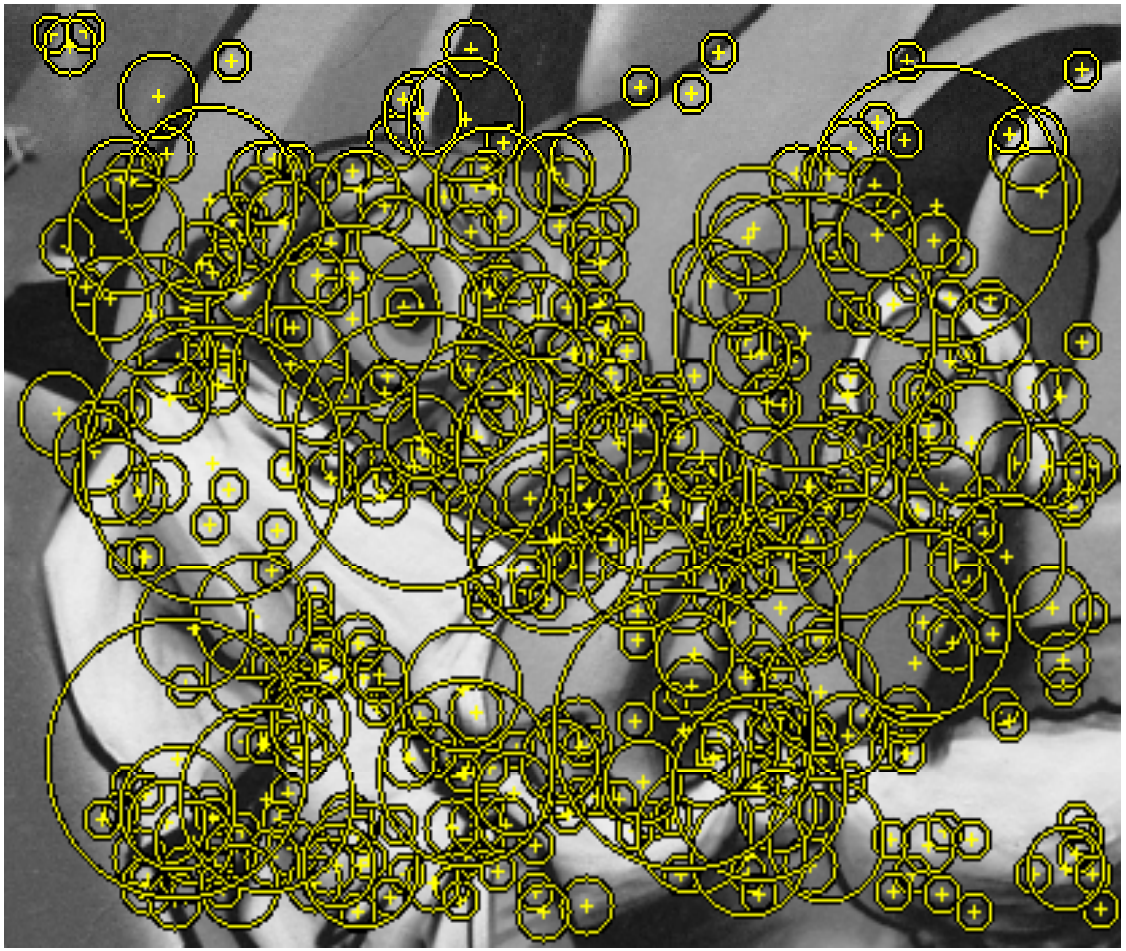
σ





⇒ **List of** *(x, y, s)*

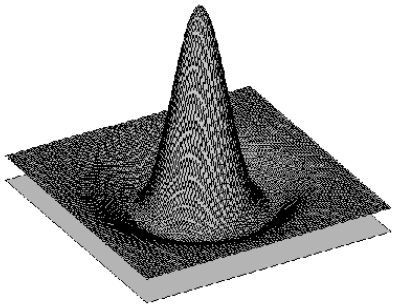
Results: Laplacian-of-Gaussian



Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Difference-of-Gaussian (DoG)

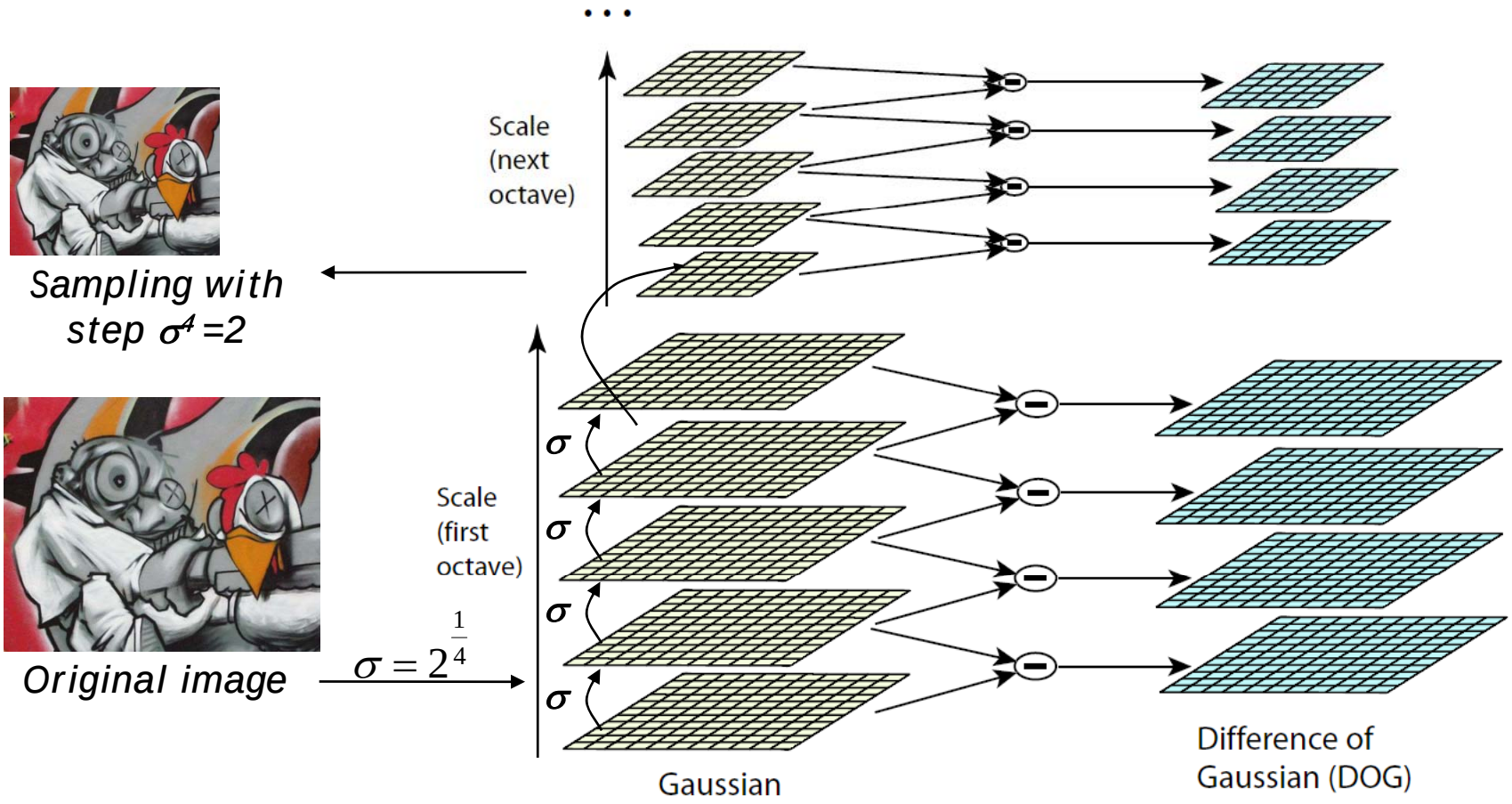
- **Difference of Gaussians as approximation of the Laplacian-of-Gaussian**



Slide credit K. Grauman, B. Leibe AAAI08 Short Course

DoG – Efficient Computation

- **Computation in Gaussian scale pyramid**



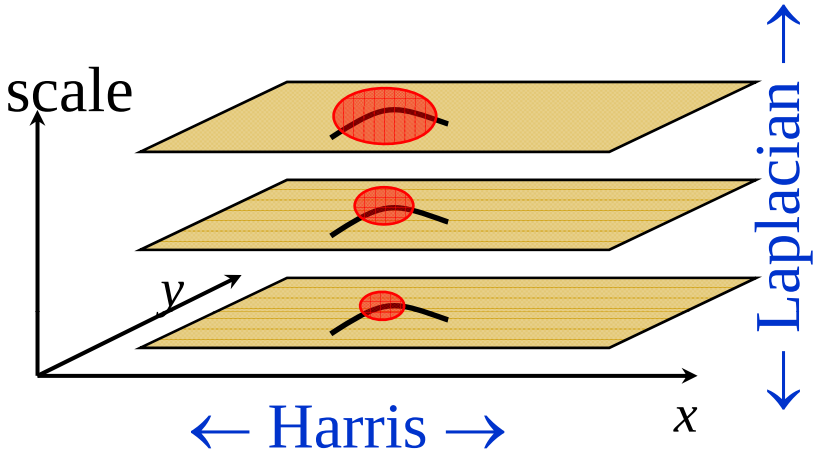
Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Results: Lowe's DoG

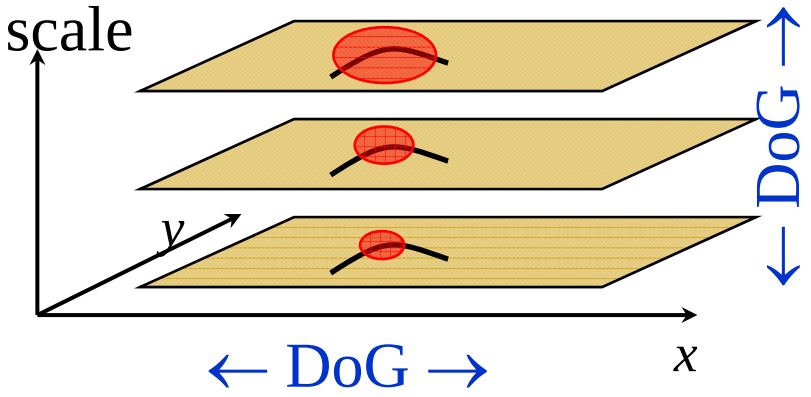
Slide credit K. Grauman, B. Leibe AAAI08 Short Course

Finding Keypoints – Scale, Location

- Harris-Laplacian[1]
  Find local maximum of:
  - Laplacian in scale
  - Harris corner detector in space (image coordinates)



- SIFT[2]
  Find local maximum of:
  - Difference of Gaussians in space and scale

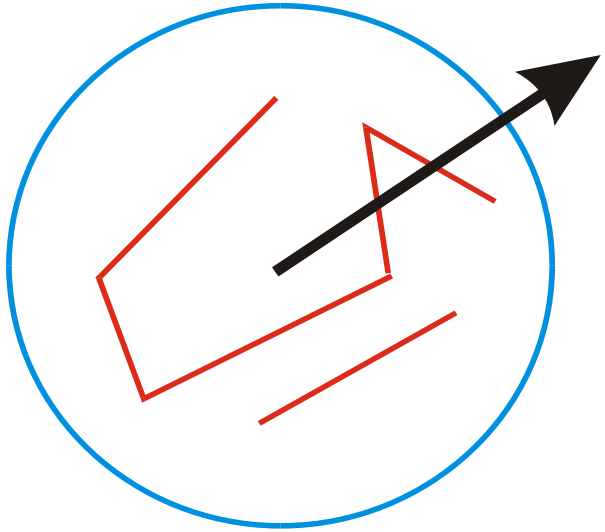


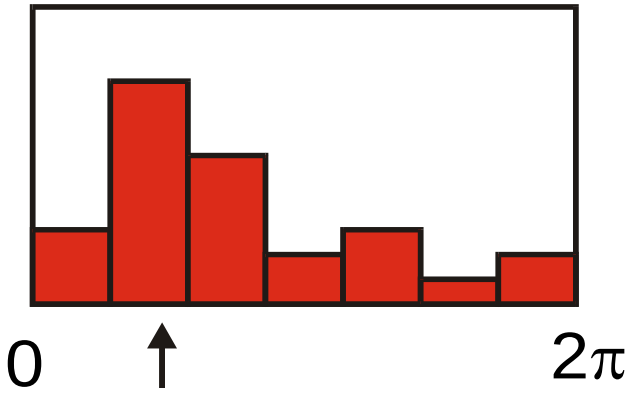
[1] K.Mikolajczyk, C.Schmid. "Indexing Based on Scale Invariant Interest Points". ICCV 2001
[2] D.Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". ICCV 1999

Slide credit David Lee / David Lowe

Finding Keypoints – Orientation

- Create histogram of local gradient directions computed at selected scale
- Assign canonical orientation at peak of smoothed histogram
- Each key specifies stable 2D coordinates (x, y, scale, orientation)

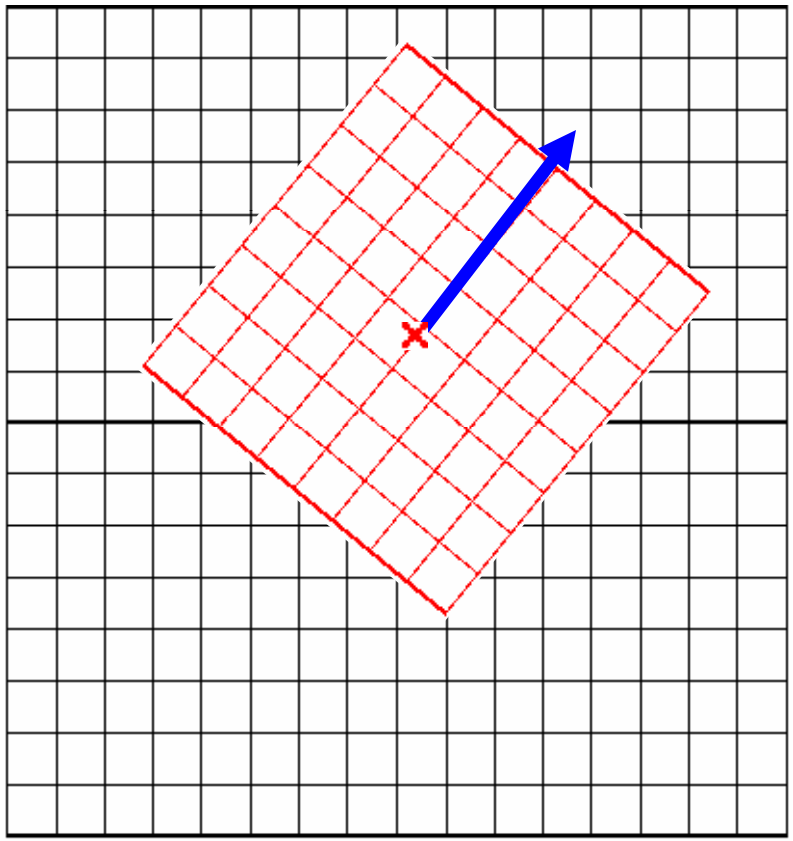




Slide credit David Lee / David Lowe

Finding Keypoints – Orientation

- Assign dominant orientation as the orientation of the keypoint



Slide credit David Lee / David Lowe

SIFT Descriptor

- 4x4 Gradient window
- Histogram of 4x4 samples per window in 8 directions
- Gaussian weighting around center(σ is 0.5 times that of the scale of a keypoint)
- 4x4x8 = 128 dimensional feature vector

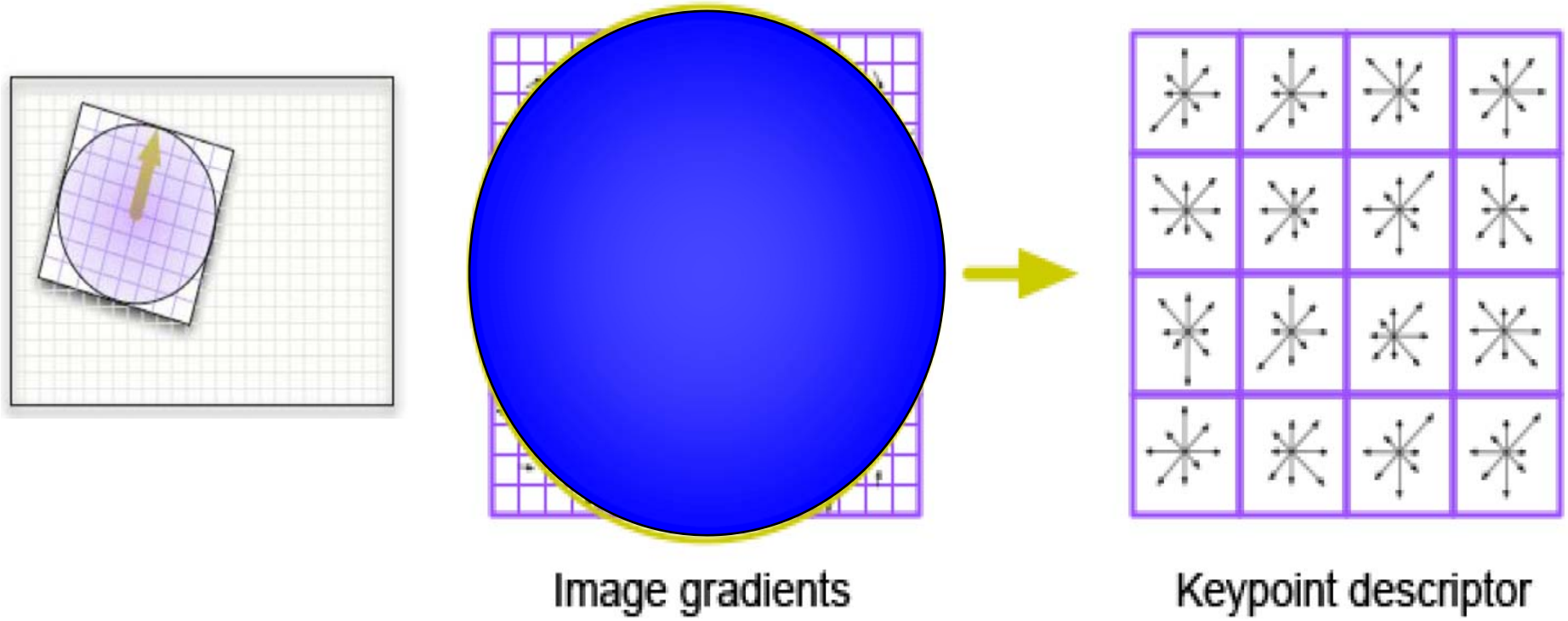


Image from: Jonas Hurrelmann

Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

SIFT Descriptor – Lighting changes

- Gains do not affect gradients
- Normalization to unit length removes contrast
- Saturation affects magnitudes much more than orientation
- Threshold gradient magnitudes to 0.2 and renormalize

Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

Performance

- Very robust
  - 80% Repeatability at:
    - 10% image noise
    - 45° viewing angle
    - 1k-100k keypoints in database
- Best descriptor in [Mikolajczyk & Schmid 2005]'s extensive survey
- 606+ citations on Google Scholar already for [2004] paper

Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

Typical Usage

- For set of database images:
  1. Compute SIFT features
  2. Save descriptors to database
- For query image:
  1. Compute SIFT features
  2. For each descriptor:
     - Find closest descriptors (L2 distance) in database
  3. Verify matches
     - Geometry
     - Hough transform

Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

Nearest-neighbor matching to feature database

- Hypotheses are generated by **approximate nearest neighbor** matching of each feature to vectors in the database
  - SIFT use best-bin-first (Beis & Lowe, 97) modification to k-d tree algorithm
  - Use heap data structure to identify bins in order by their distance from query point

- **Result:** Can give speedup by factor of 1000 while finding nearest neighbor (of interest) 95% of the time

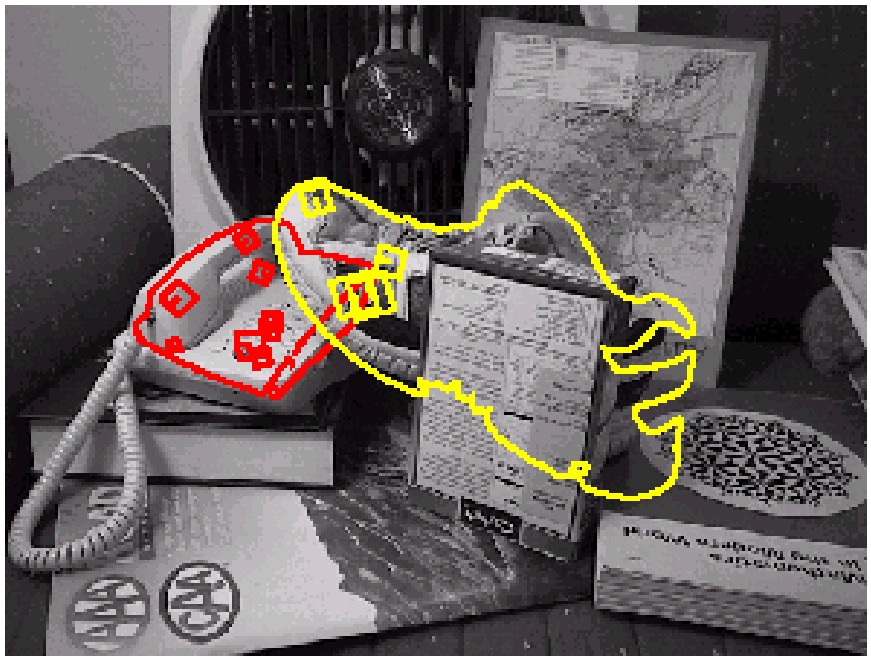
Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

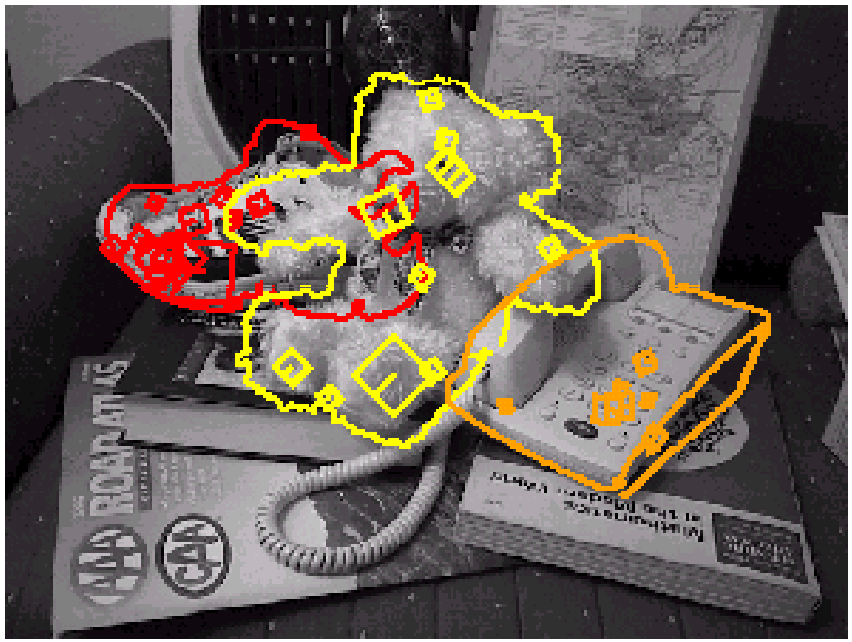
3D Object Recognition

- Only 3 keys are needed for recognition, so extra keys provide robustness

Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

Recognition under occlusion

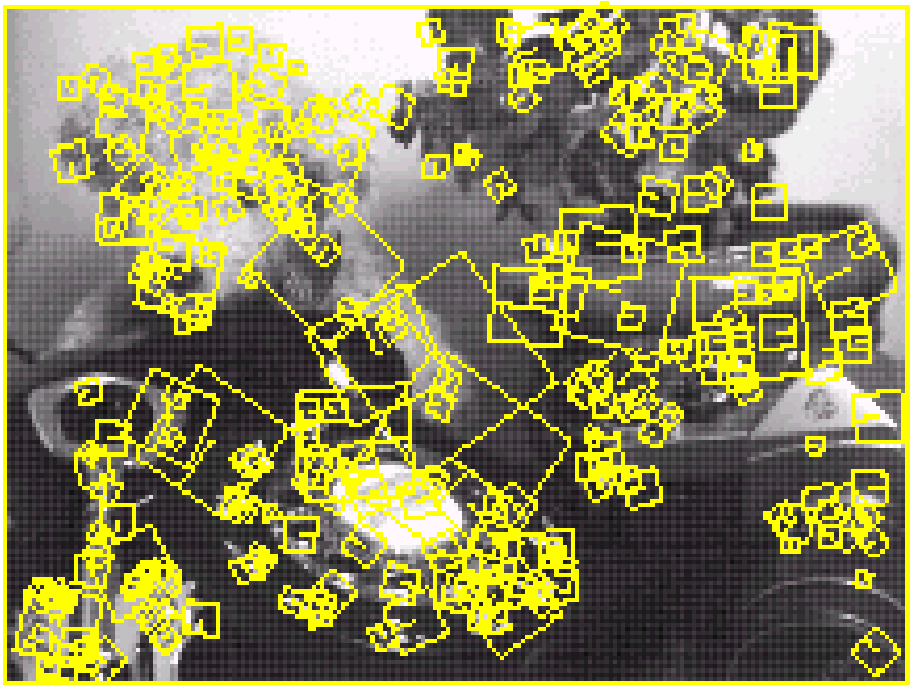




Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

Test of illumination Robustness

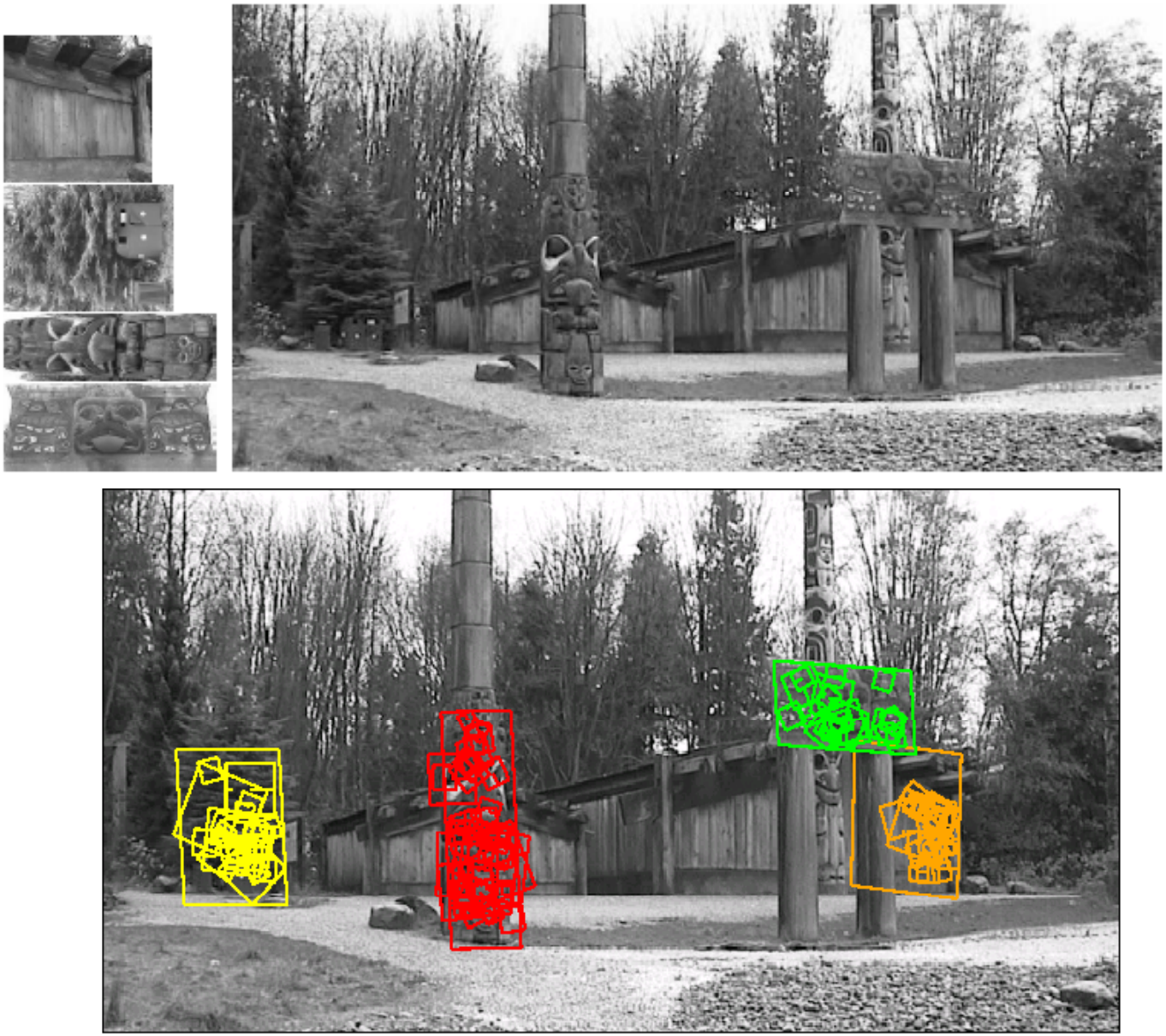
- Same **image** under differing illumination



273 keys verified in final match

Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

Location recognition



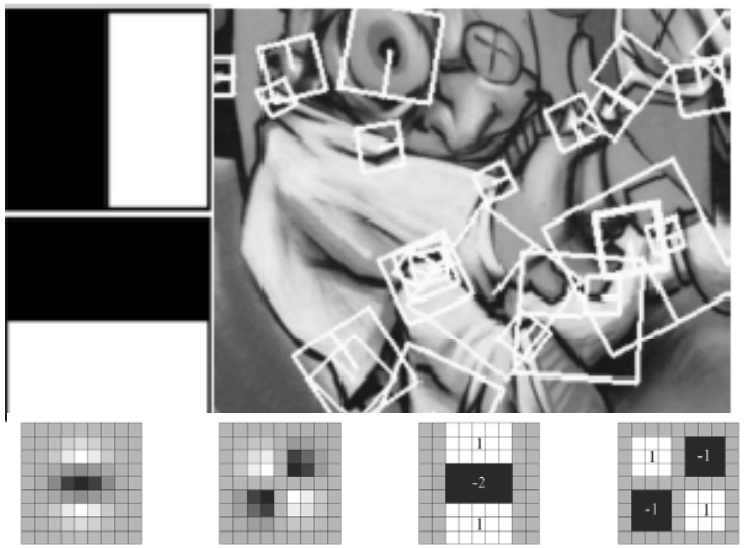
Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

Image Registration Results

[Brown & Lowe 2003]

Slide credit: O. Pele, S. Thrun, J. Košecká, N. Kumar

Local Descriptors: SURF



- **Fast approximation of SIFT idea**
  - **Efficient computation by 2D box filters & integral images**
    ⇒ 6 times faster than SIFT
  - **Equivalent quality for object identification**

- **GPU implementation available**
  - **Feature extraction @ 100Hz (detector + descriptor, 640×480 img)**
  - http://www.vision.ee.ethz.ch/~surf

[Bay, ECCV'06], [Cornelis, CVGPU'08]

Scaling up: particular object retrieval

Example I: Visual search in feature films

Visually defined query

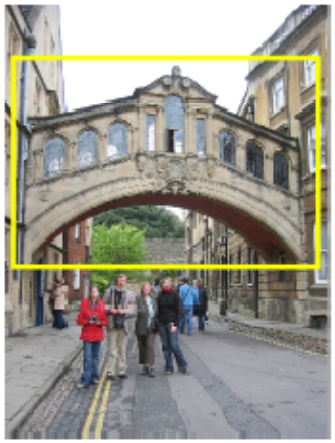
"Find this clock"

"Find this place"

"Groundhog Day" [Rammis, 1993]

Slide credit: J. Sivic

Example II: Search photos on the web for particular places



Find these landmarks

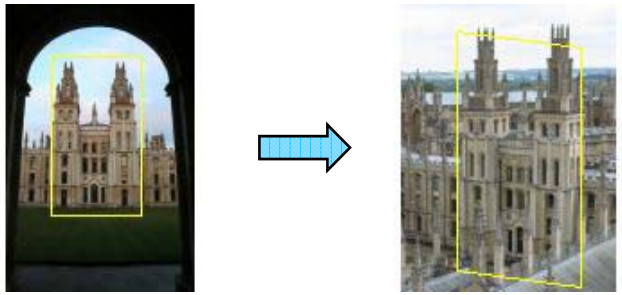
...in these images and 1M more

Slide credit: J. Sivic

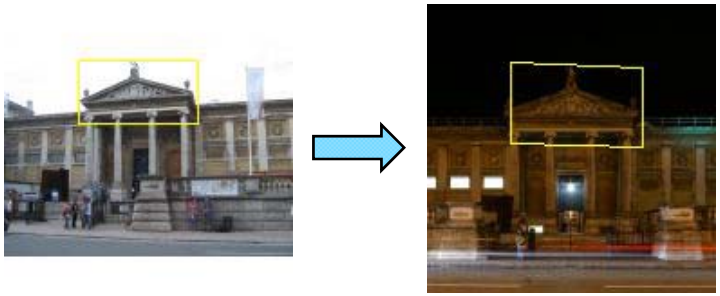
Why is it difficult?

Want to find the object despite possibly large changes in scale, viewpoint, lighting and partial occlusion

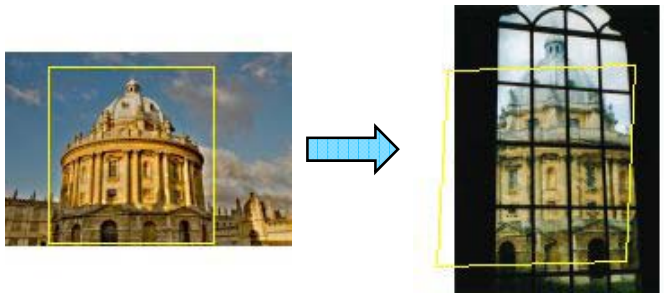
Scale



Viewpoint



Lighting



Occlusion

Slide credit: J. Sivic

The need for large-scale visual search

Flickr: has 2 billion photographs, more than 1 million added daily

Company collections

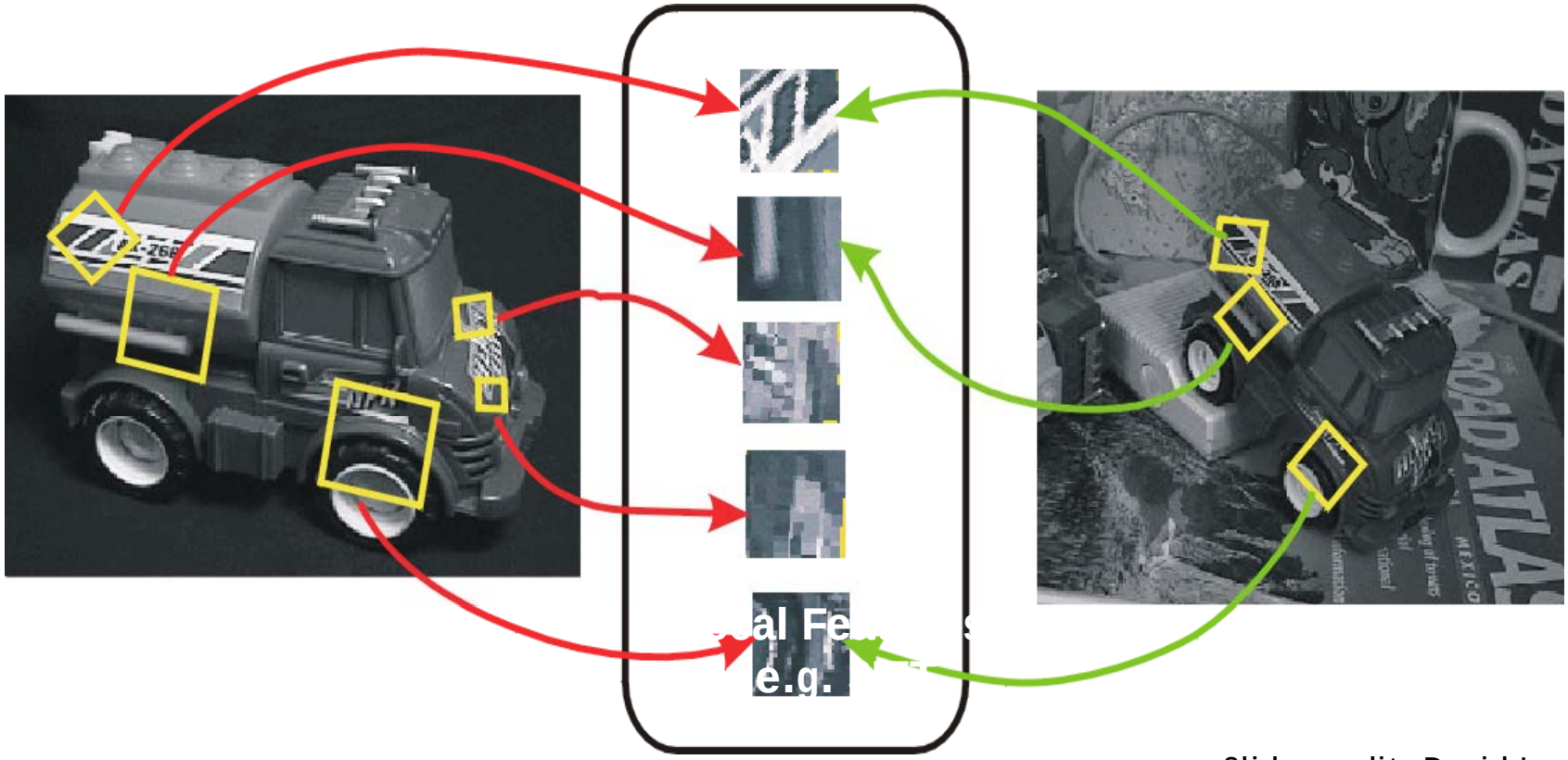
Personal collections: 10000s of digital camera photos and mpegs

Vast majority will have minimal, if any, textual annotation.
Yet text is the only common way of searching / accessing documents (e.g. Google / Live search)

Slide credit: J. Sivic

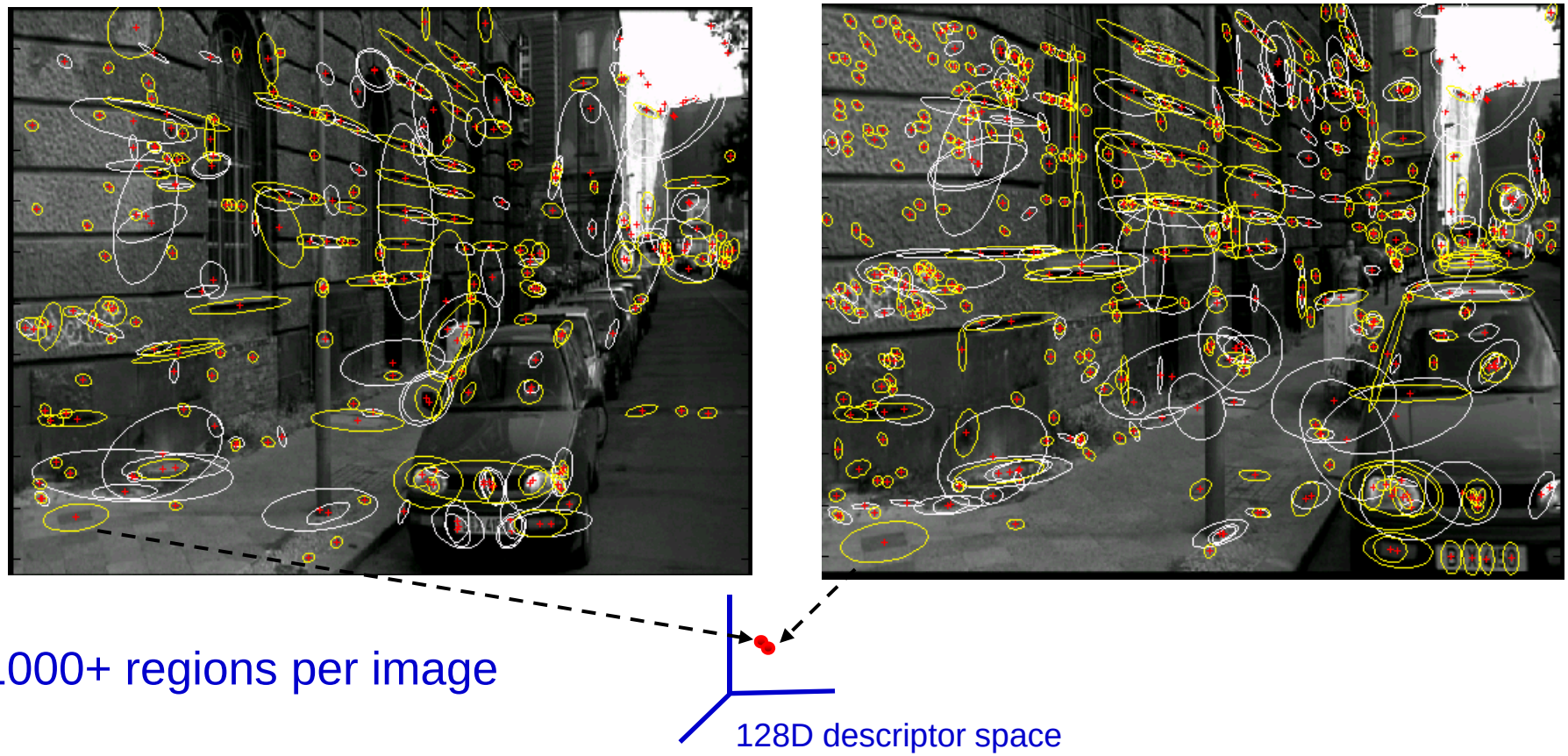
Recap: Image representation

- Image content is transformed into local features that are invariant to geometric and photometric transformations



Slide credit: David Lowe

Affine covariant regions (Feb 10th)

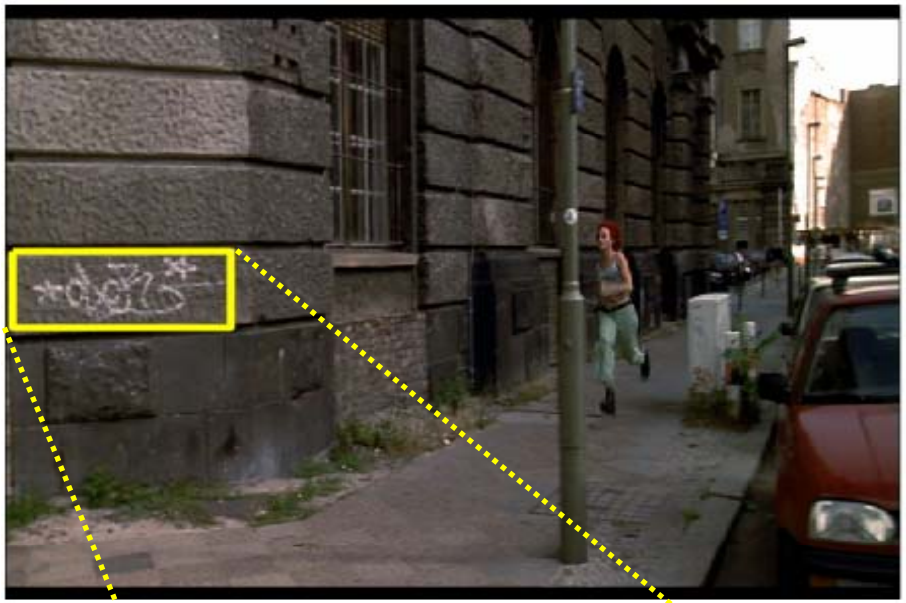


- a region's size and shape are not fixed, but
- automatically adapts to the image intensity to cover the same physical surface
- i.e. pre-image is the same surface region

Represent each region by the 128-dimensional SIFT descriptor vector

Slide credit: J. Sivic

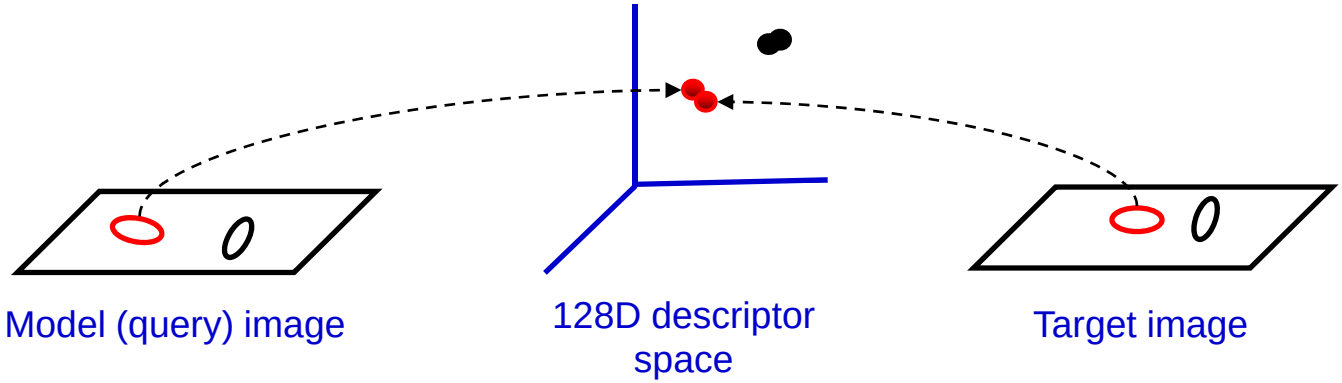
Example



Slide credit: J. Sivic

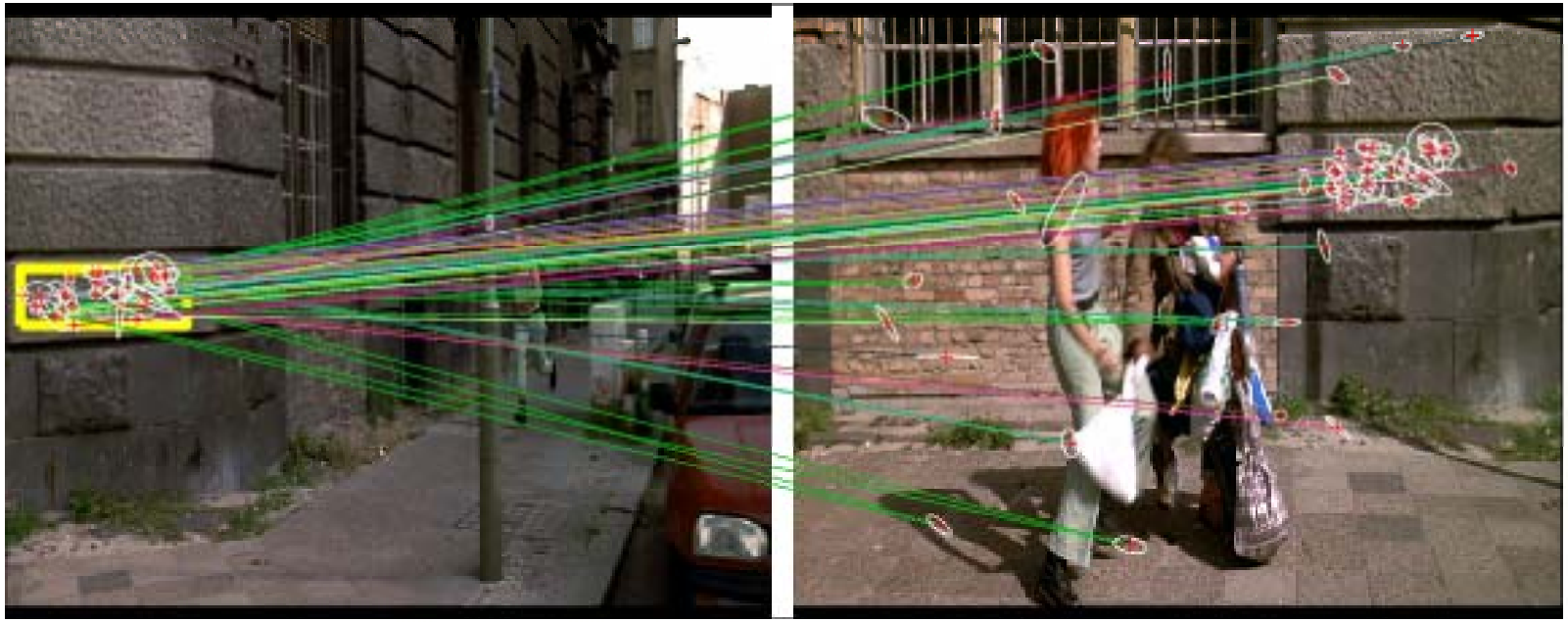
Object recognition

Establish correspondences between object model image and target image by nearest neighbour matching on SIFT vectors



Slide credit: J. Sivic

Problem with matching on descriptors alone



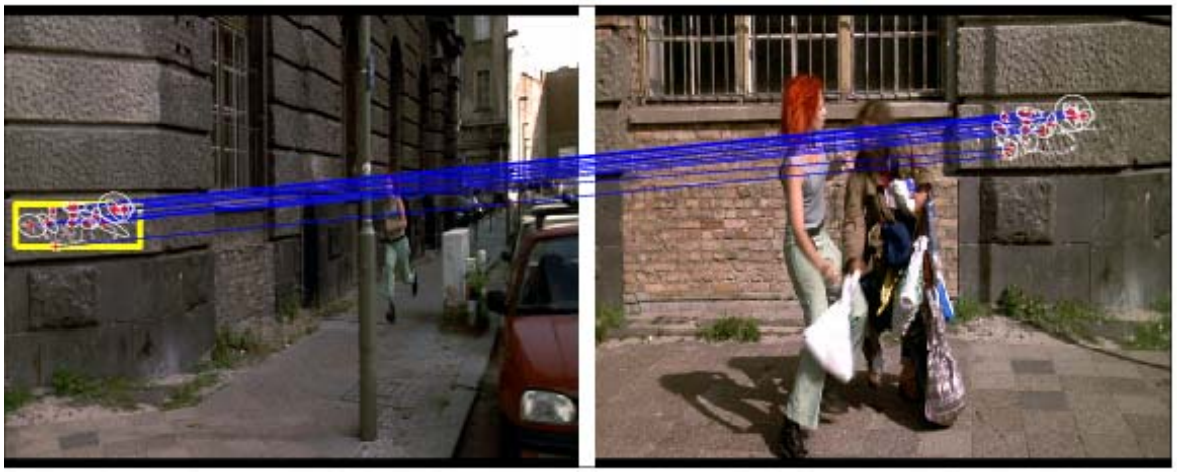
- too much individual invariance
- each region can affine deform independently (by different amounts)
- use semi-local and global spatial relations to verify matches, e.g.:
  - common affine transformation [Lowe ‘99] (strong requirement)
  - locally similar affine transformation [Ferrari ‘04]
  - spatial neighbours match spatial neighbours [Schmid ‘97]

Slide credit: J. Sivic

Example

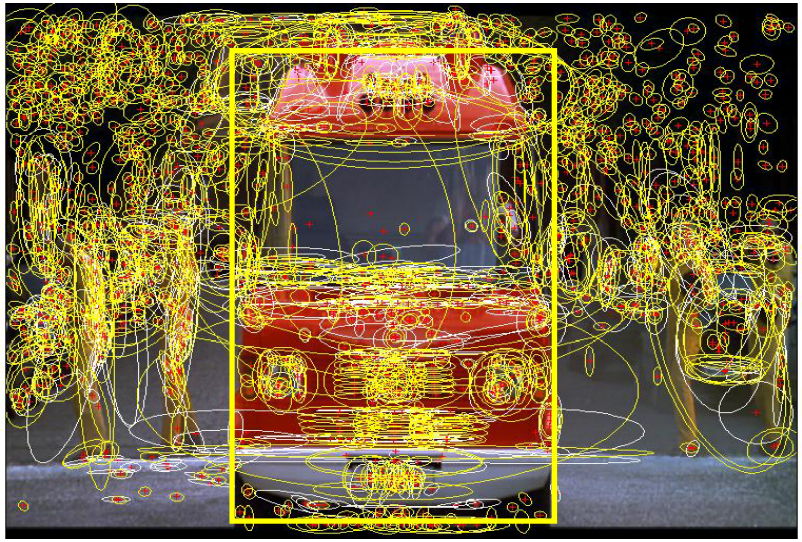
Initial matches

Spatial consistency required



Slide credit: J. Sivic

Example of object recognition



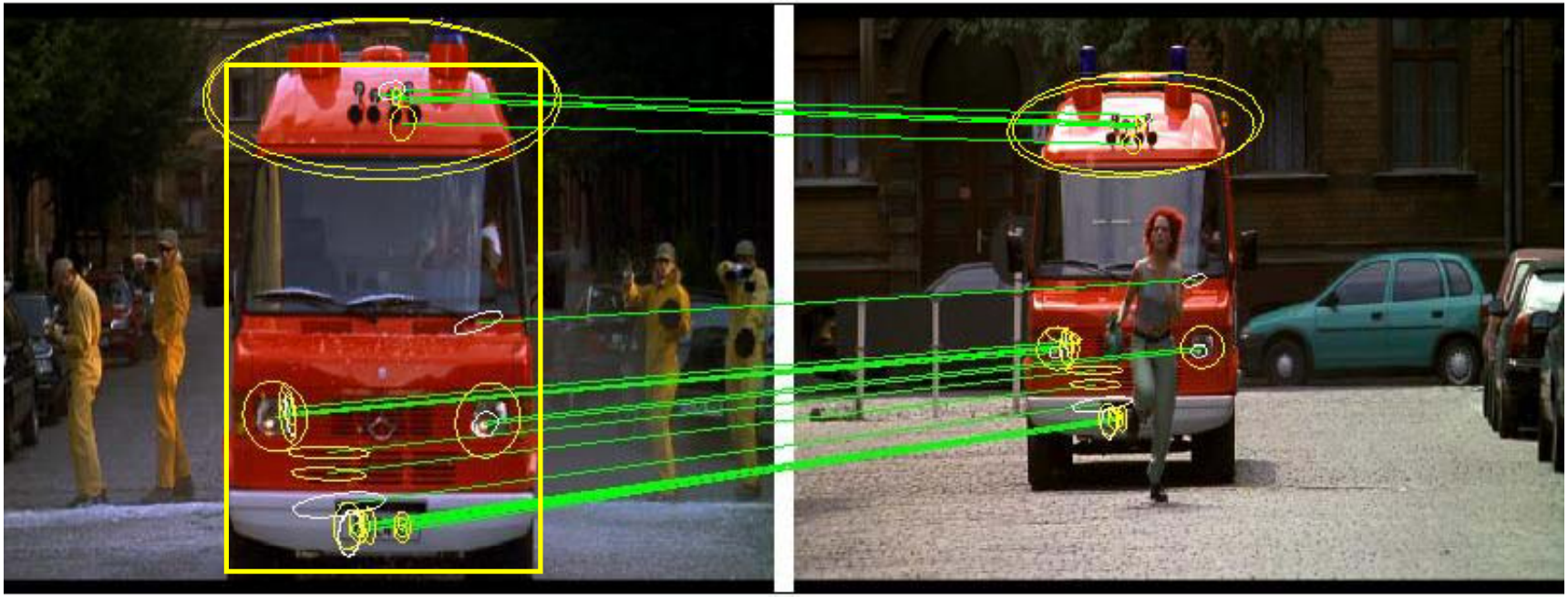
1000+ descriptors per frame

Shape adapted regions

Maximally stable regions

Slide credit: J. Sivic

Match regions between frames using SIFT descriptors and spatial consistency



Multiple regions overcome problem of partial occlusion

Shape adapted regions

Maximally stable regions

Slide credit: J. Sivic

Visual search using local regions

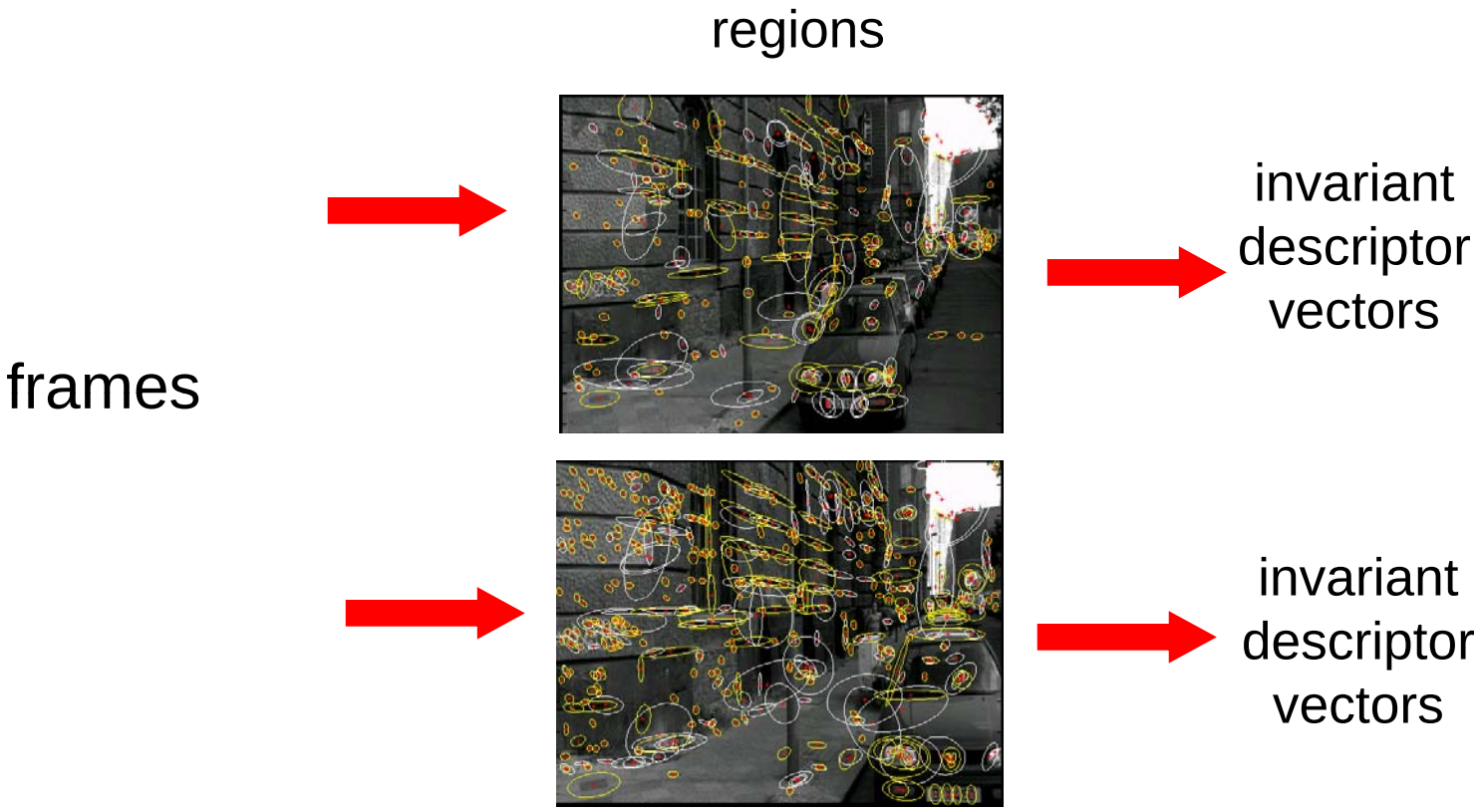
Schmid and Mohr ’97	– 1k images
Sivic and Zisserman’03	– 5k images
Nister and Stewenius’06	– 50k images (1M)
Philbin et al.’07	– 100k images
Chum et al.’07 + Jegou and Schmid’07	– 1M images
Chum et al.’08	– 5M images

Index 1 billion (10^9) images

– 200 servers each indexing 5M images?

Slide credit: J. Sivic

Outline of a retrieval strategy

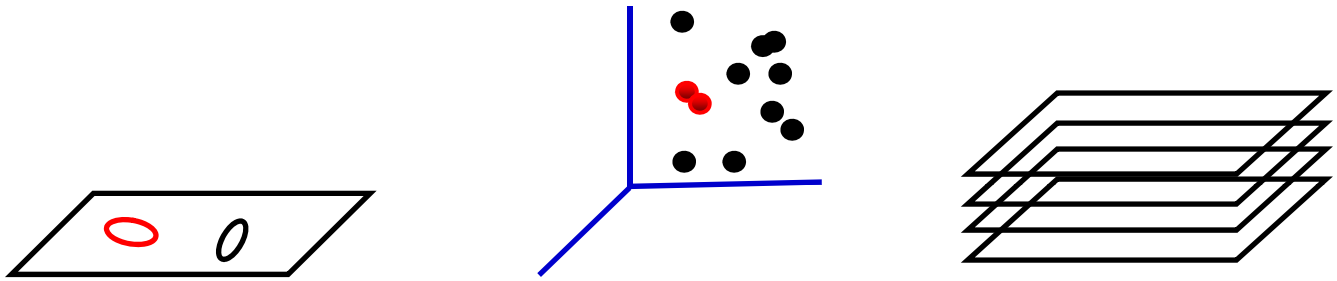


1. Compute affine covariant regions in each frame independently
2. “Label” each region by a vector of descriptors based on its intensity
3. Finding corresponding regions is transformed to finding nearest neighbour vectors
4. Rank retrieved frames by number of corresponding regions
5. Verify retrieved frame based on spatial consistency

Slide credit: J. Sivic

Visual retrieval / search

Establish correspondences between object model image and images in the database by **nearest neighbour matching** on SIFT vectors



Model image

128D descriptor space

Image database

Slide credit: J. Sivic

Indexing local features

With potentially thousands of features per image, and hundreds to millions of images to search, how to efficiently find those that are relevant to a new image?

- Low-dimensional descriptors : can use standard efficient data structures for nearest neighbor search
- High-dimensional descriptors: approximate nearest neighbor search methods more practical
- Inverted file indexing schemes

Slide credit: J. Sivic

Nearest-neighbor matching

Solve following problem for all feature vectors, $\mathbf{x_j}$, in the query image:

$$\forall j \; NN(j) = \arg\min_i ||\mathbf{x}_i - \mathbf{x}_j||$$

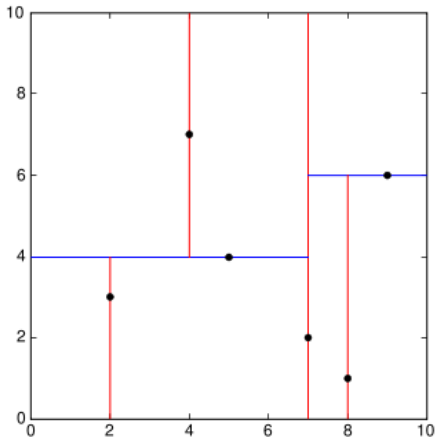
where $\mathbf{x_i}$ are features in database images.

Nearest-neighbour matching is the major computational bottleneck

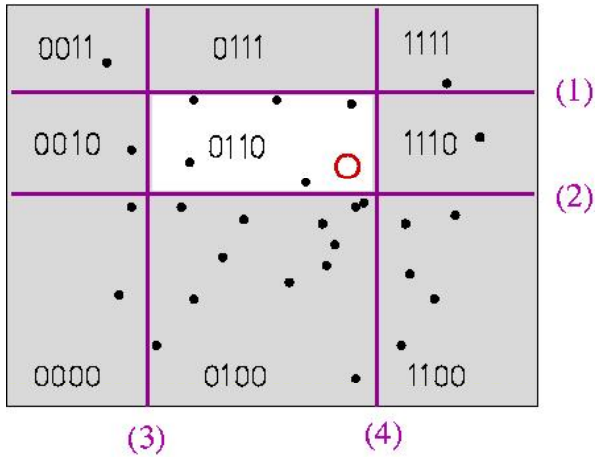
- Linear search performs *dn* operations for *n* features in the database and *d* dimensions
- No exact methods are faster than linear search for d>10
- Approximate methods can be much faster, but at the cost of missing some correct matches. Failure rate gets worse for large datasets.

Slide credit: J. Sivic

Indexing local features: approximate nearest neighbor search



Best-Bin First (BBF), a variant of k-d trees that uses priority queue to examine most promising branches first [Beis & Lowe, CVPR 1997]

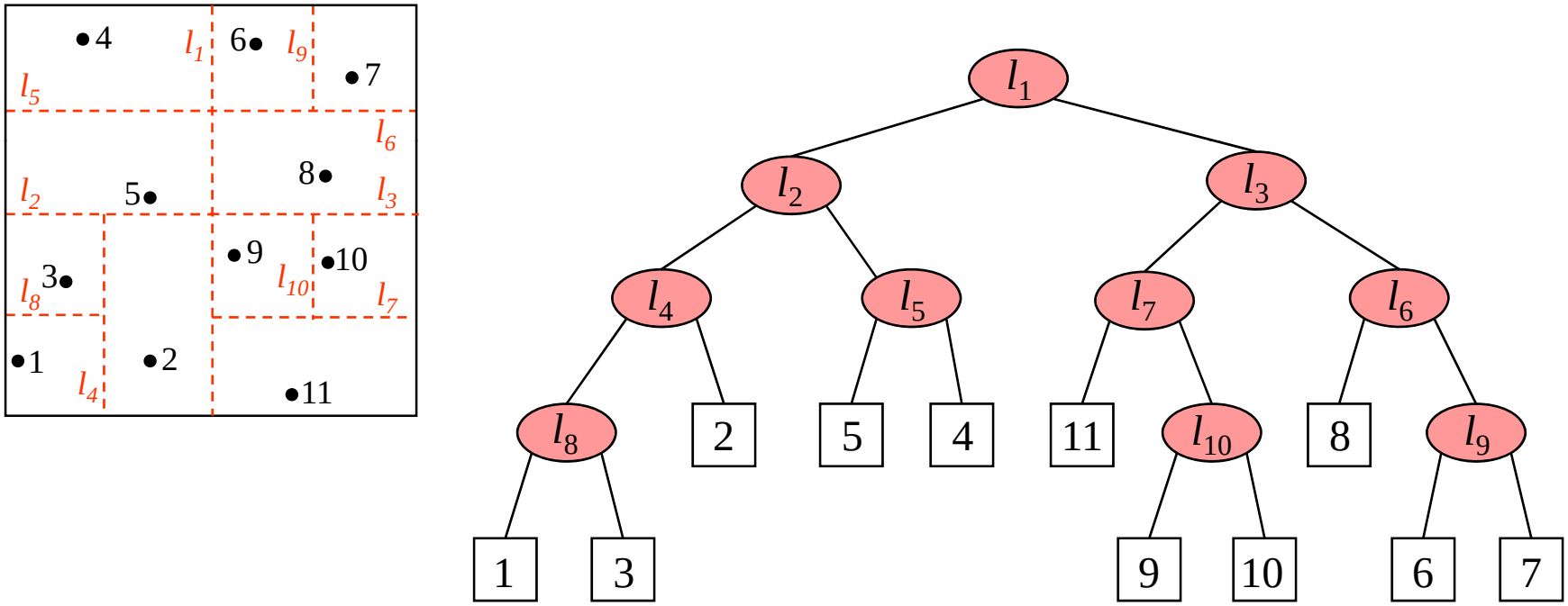


Locality-Sensitive Hashing (LSH), a randomized hashing technique using hash functions that map similar points to the same bin, with high probability [Indyk & Motwani, 1998]

Slide credit: J. Sivic

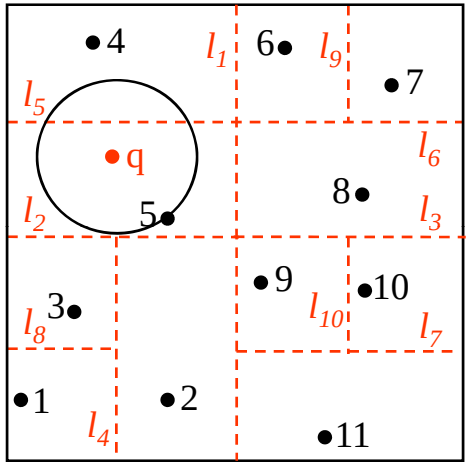
K-d tree construction

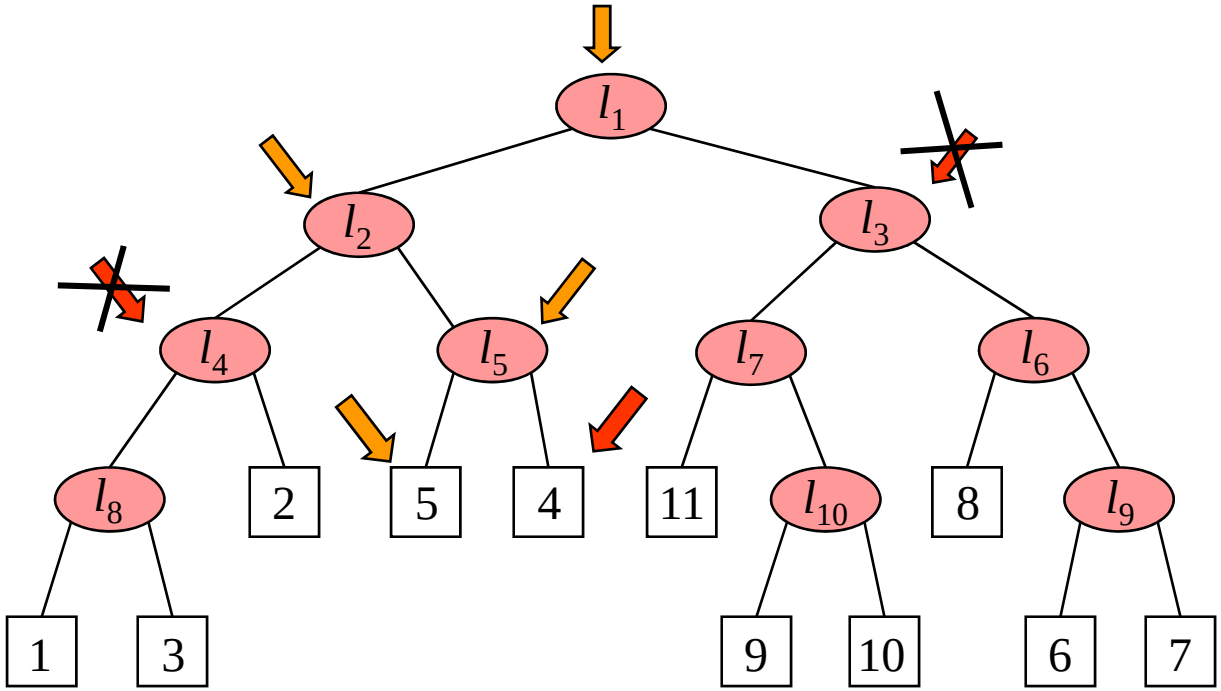
Simple 2D example



Slide credit: Anna Atramentov

K-d tree query





Slide credit: Anna Atramentov

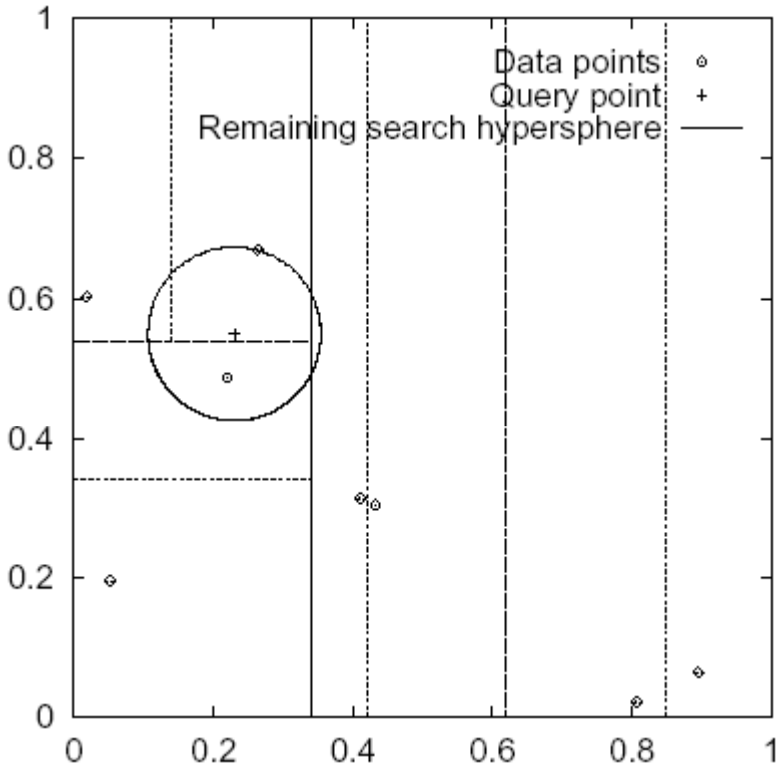
Approximate nearest neighbour K-d tree search

Key idea:

- Limit the number of neighbouring k-d tree bins to explore

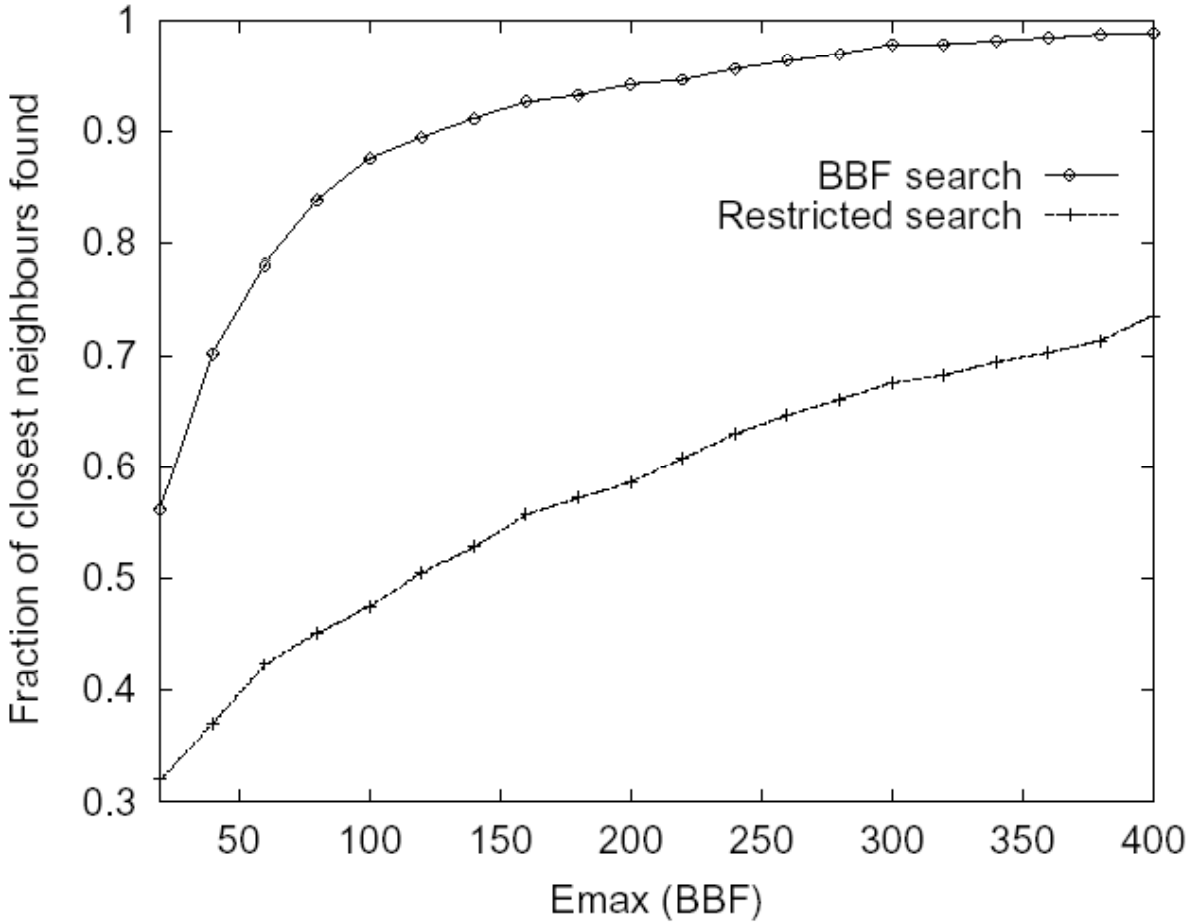
- Search k-d tree bins in order of distance from query

- Requires use of a priority queue



Slide credit: J. Sivic

Fraction of nearest neighbors found



100,000 uniform points in 12 dimensions.

Results:

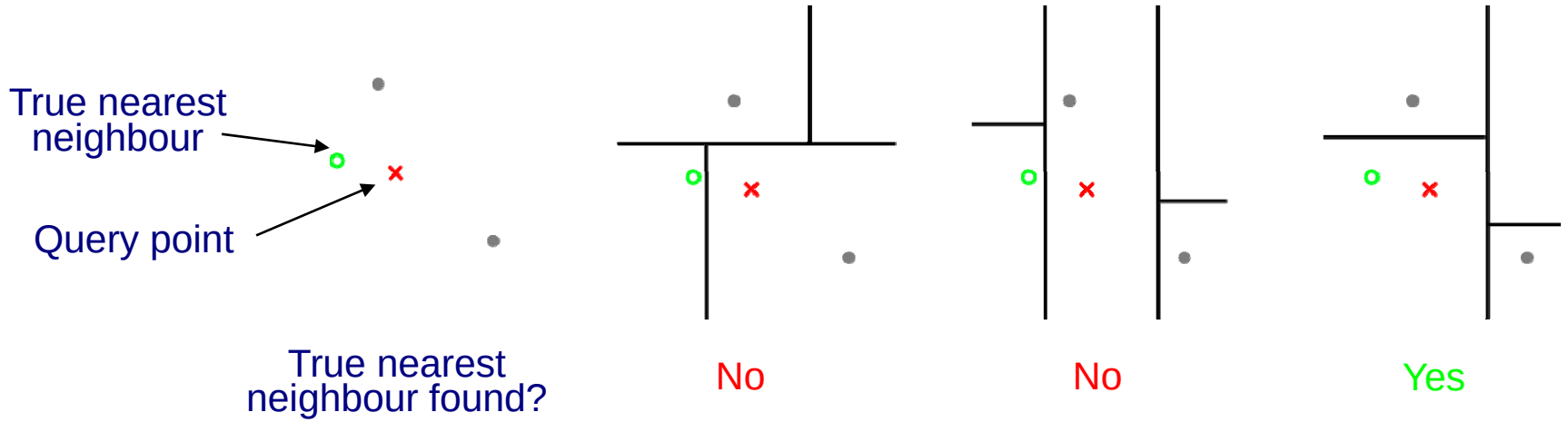
Speedup by several orders of magnitude over linear search

Slide credit: J. Sivic

Approximate nearest neighbour
K-d tree search

- How to choose the dimension to split and the splitting point?

- Multiple randomized trees increase the chances of finding nearby points

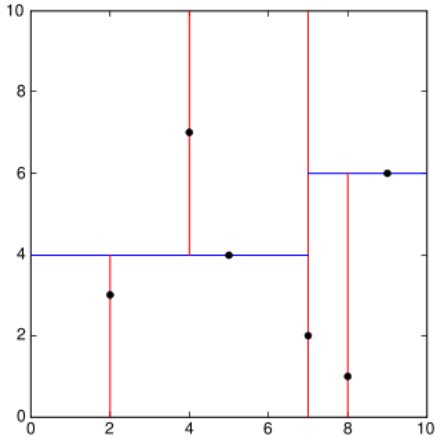


Finding (approximate) nearest neighbours in **O(logN)** time

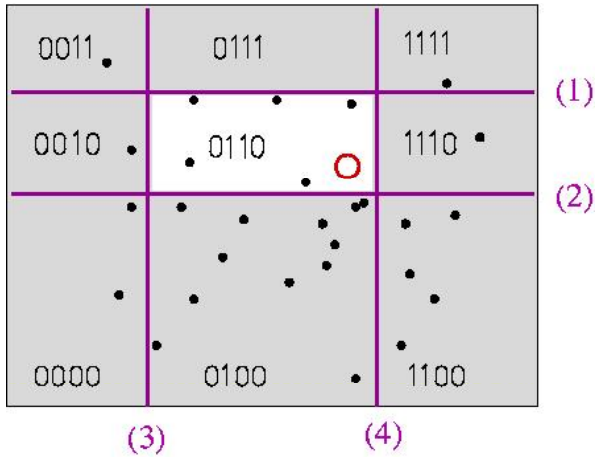
N … number of data points

Slide credit: J. Sivic

Indexing local features: approximate nearest neighbor search



Best-Bin First (BBF), a variant of k-d trees that uses priority queue to examine most promising branches first [Beis & Lowe, CVPR 1997]

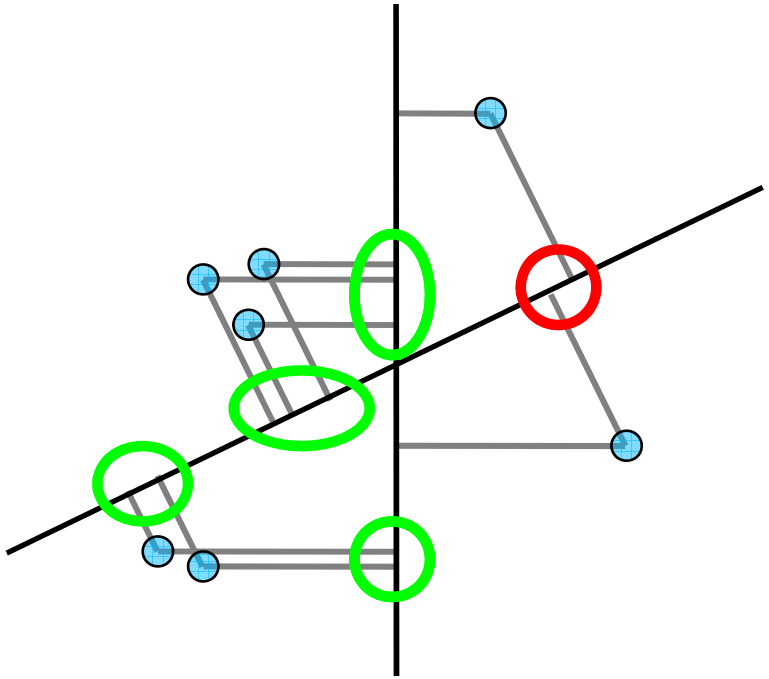


Locality-Sensitive Hashing (LSH), a randomized hashing technique using hash functions that map similar points to the same bin, with high probability [Indyk & Motwani, 1998]

Slide credit: J. Sivic

Locality Sensitive Hashing (LSH)

- Choose a random projection
- Project points
- Points close in the original space remain close under the projection
- Unfortunately, converse not true



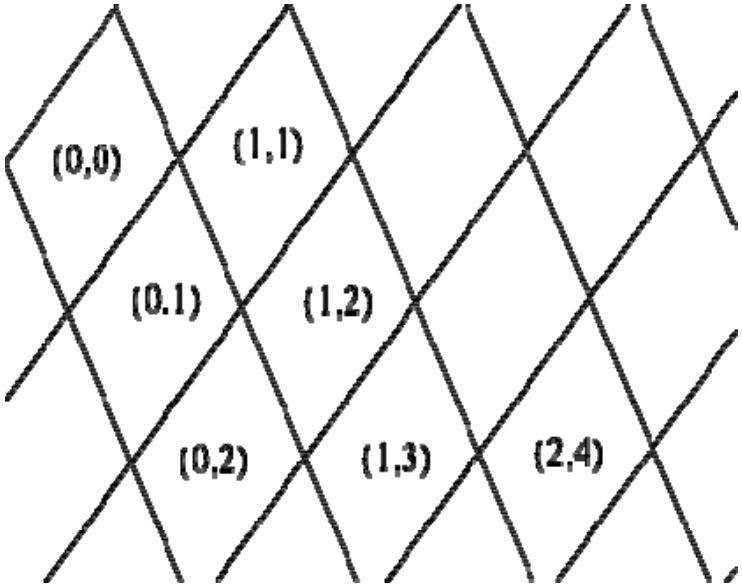
- Answer: use multiple quantized projections which define a high-dimensional “grid”

Slide credit: J. Sivic

Locality Sensitive Hashing (LSH)

- Cell contents can be efficiently indexed using a hash table

- Repeat to avoid quantization errors near the cell boundaries

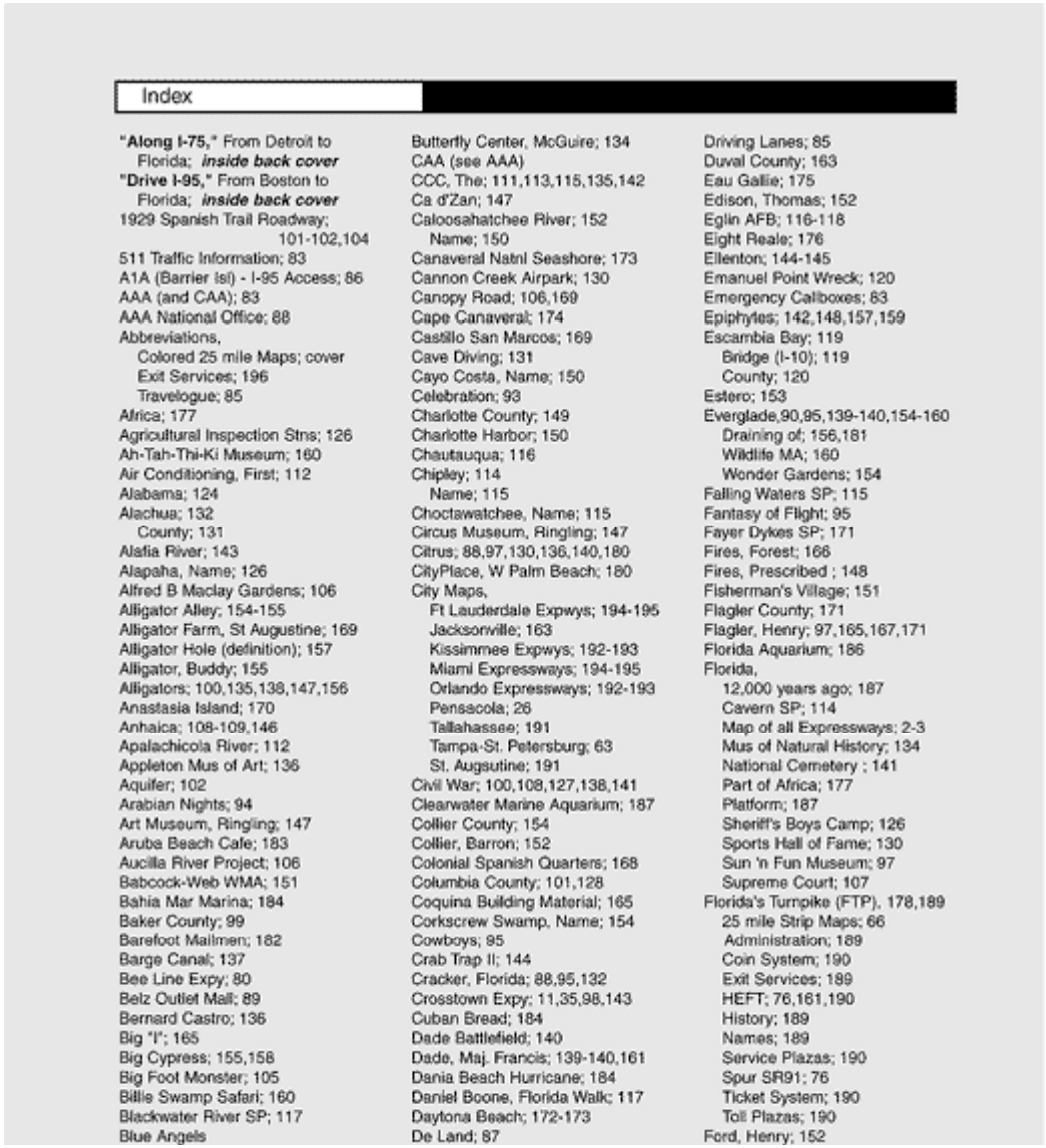


- Point that shares at least one cell = potential candidate

- Compute distance to all candidates

Slide credit: J. Sivic

Indexing local features: inverted file index



Index

For text documents, an efficient way to find all *pages* on which a *word* occurs is to use an index…

We want to find all *images* in which a *feature* occurs.

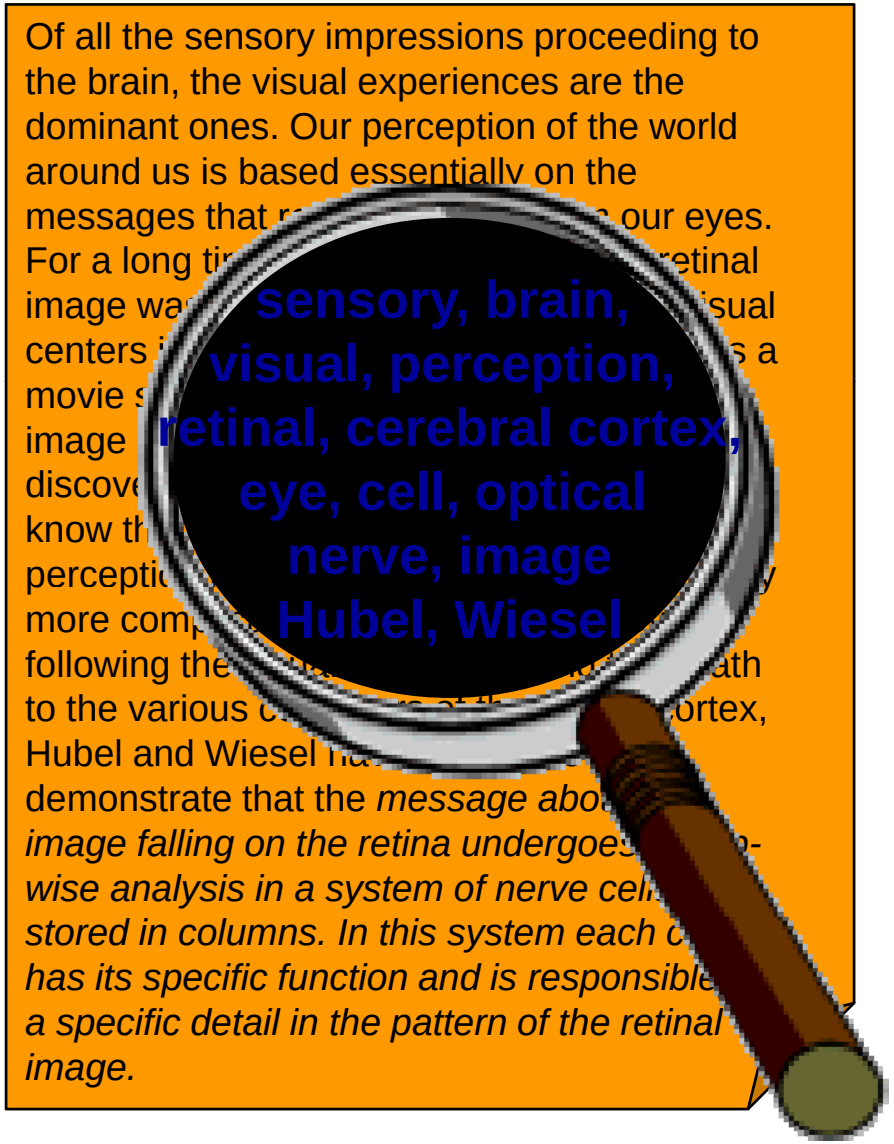
To use this idea, we'll need to map our features to "visual words".

Slide credit: J. Sivic

Object → Bag of ‘words’

Slide credit L. Fei-Fei

Analogy to documents



Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that r… our eyes. For a long ti… retinal image wa… isual centers i… s a movie s… image … discove… know th… perceptio… y more com… following the … ath to the various c… ortex, Hubel and Wiesel ha… demonstrate that the *message abo… image falling on the retina undergoes … wise analysis in a system of nerve cell… stored in columns. In this system each c… has its specific function and is responsible… a specific detail in the pattern of the retinal image.*

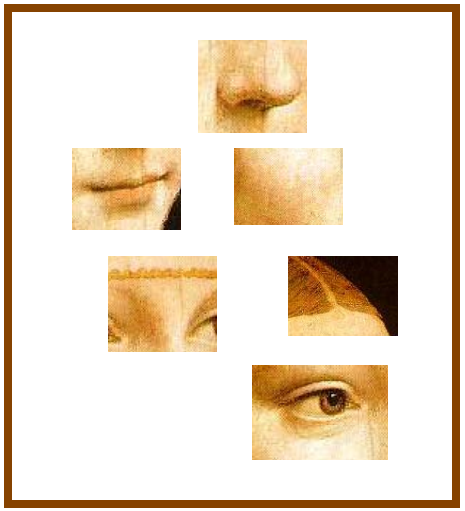
China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be created by a predicted 30%… $750bn, compared wi… $660bn. T… annoy th… China's… deliber… agrees … yuan is … governo… also need… demand so … country. China … e yuan against the dol… nd permitted it to trade within a narrow… but the US wants the yuan to be allowed… de freely. However, Beijing has made it cl… t it will take its time and tread carefully be… allowing the yuan to rise further in value.

Slide credit L. Fei-Fei

A clarification: definition of “BoW”

Looser definition

- Independent features



Slide credit L. Fei-Fei

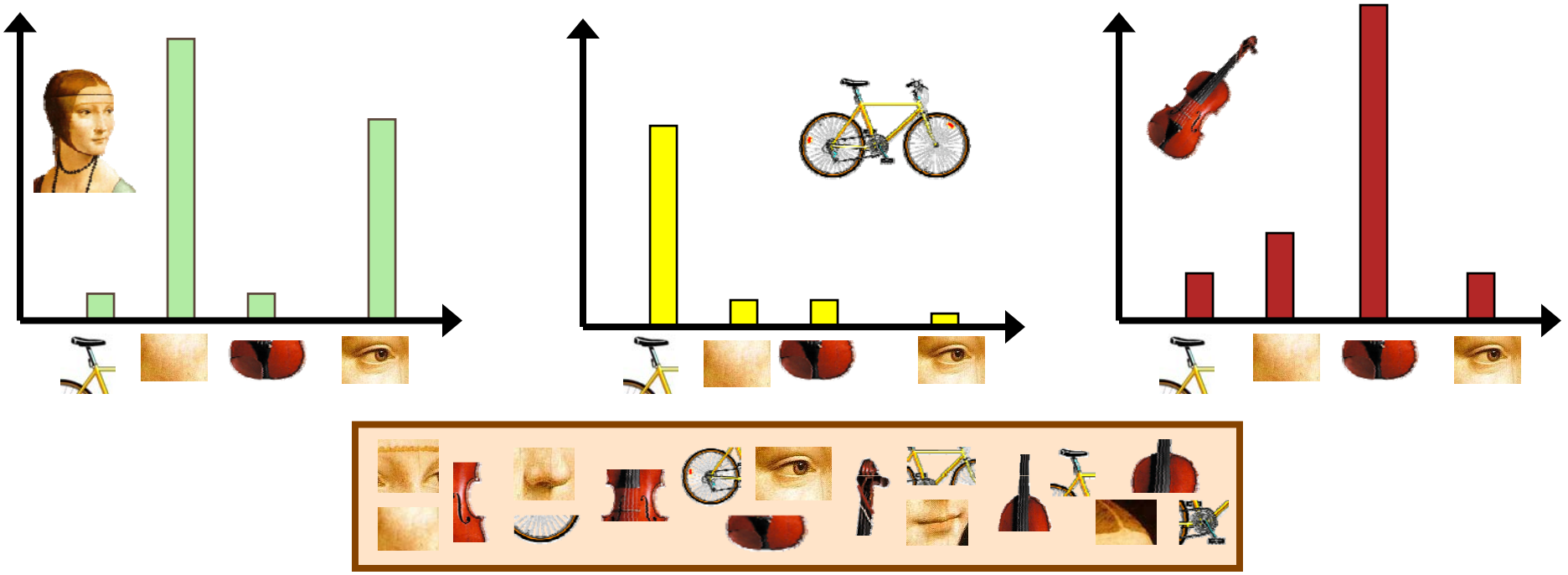
A clarification: definition of “BoW”

Looser definition

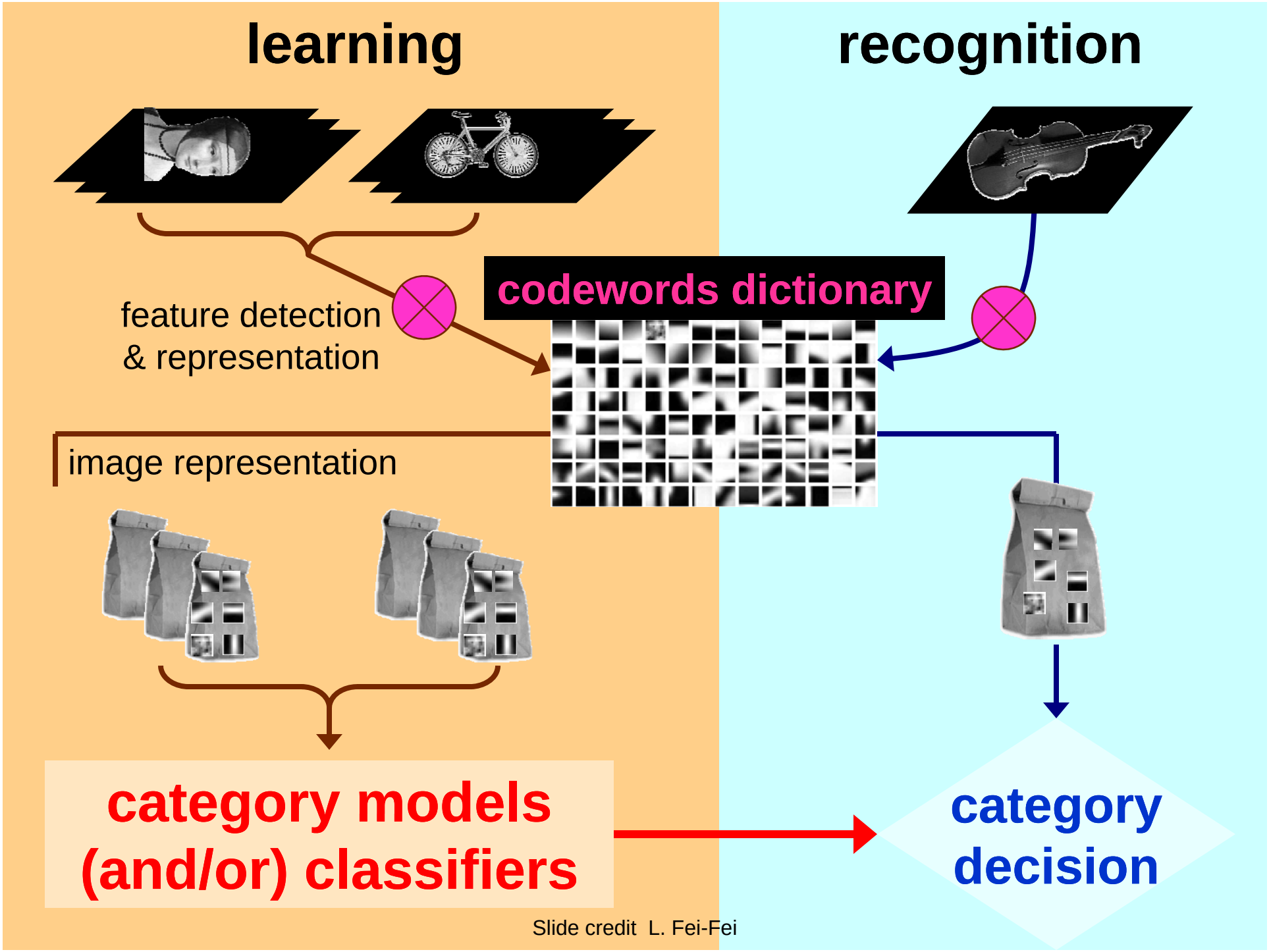
- Independent features

Stricter definition

- Independent features
- histogram representation

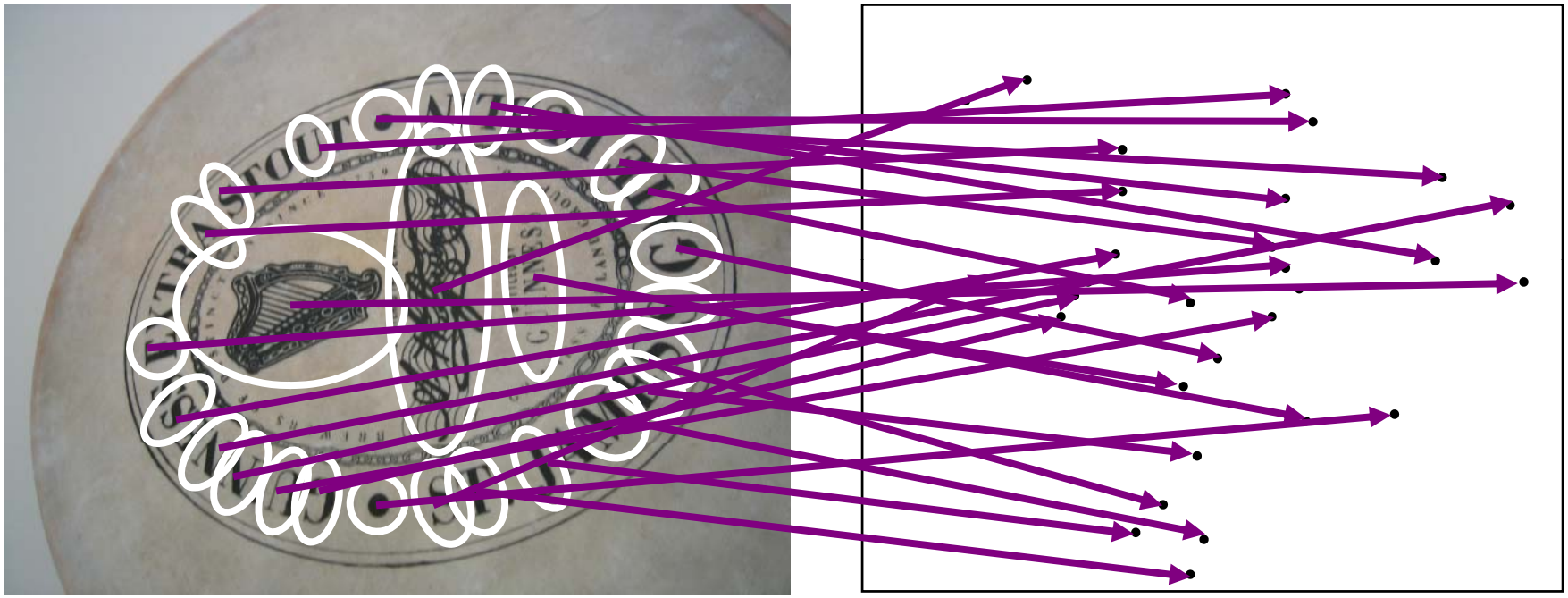


Slide credit L. Fei-Fei


learning
recognition
feature detection
& representation
codewords dictionary
image representation
category models
(and/or) classifiers
category
decision
Slide credit L. Fei-Fei

Visual words: main idea

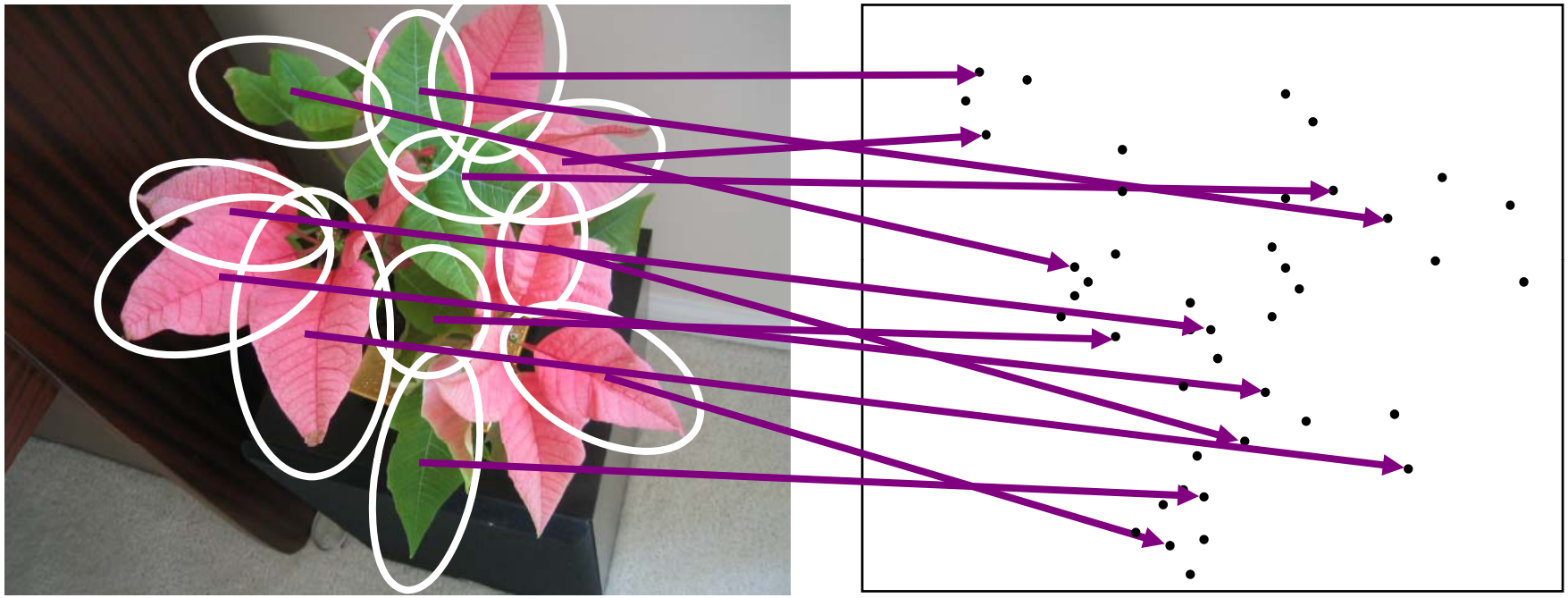
Extract some local features from a number of images …



e.g., SIFT descriptor space: each point is 128-dimensional

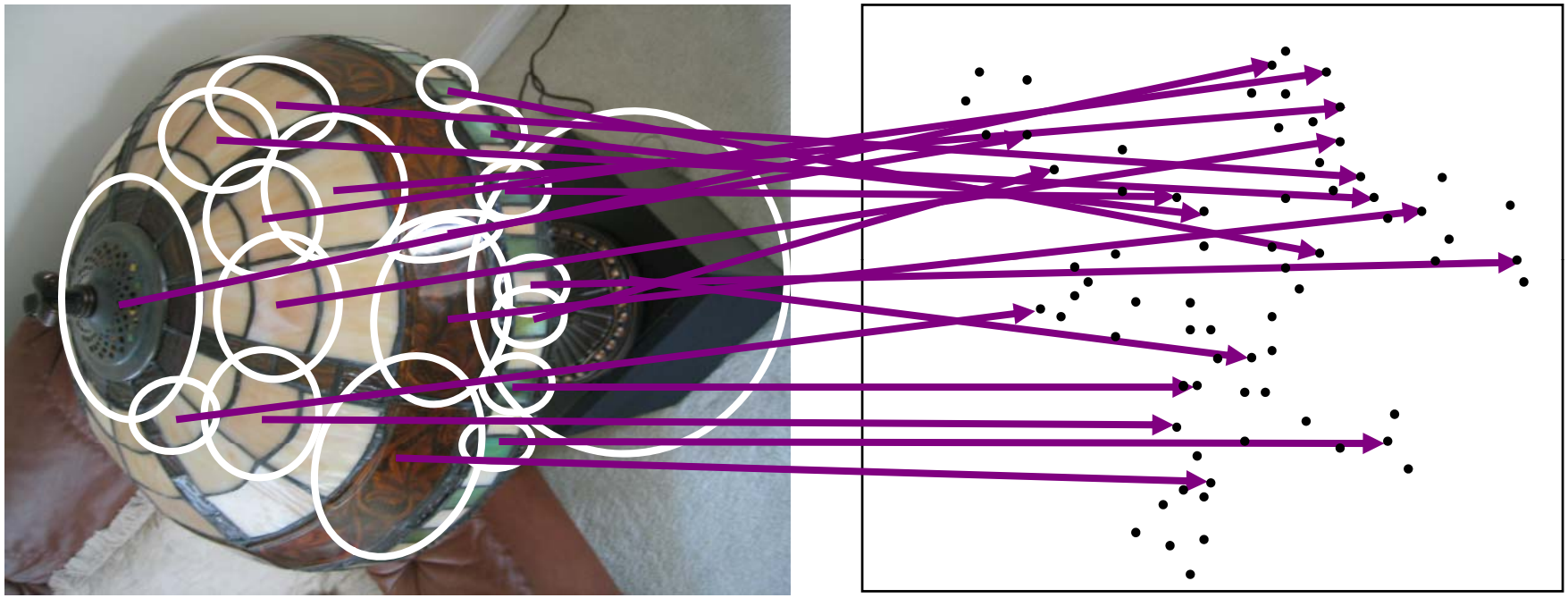
Slide credit: D. Nister

Visual words: main idea



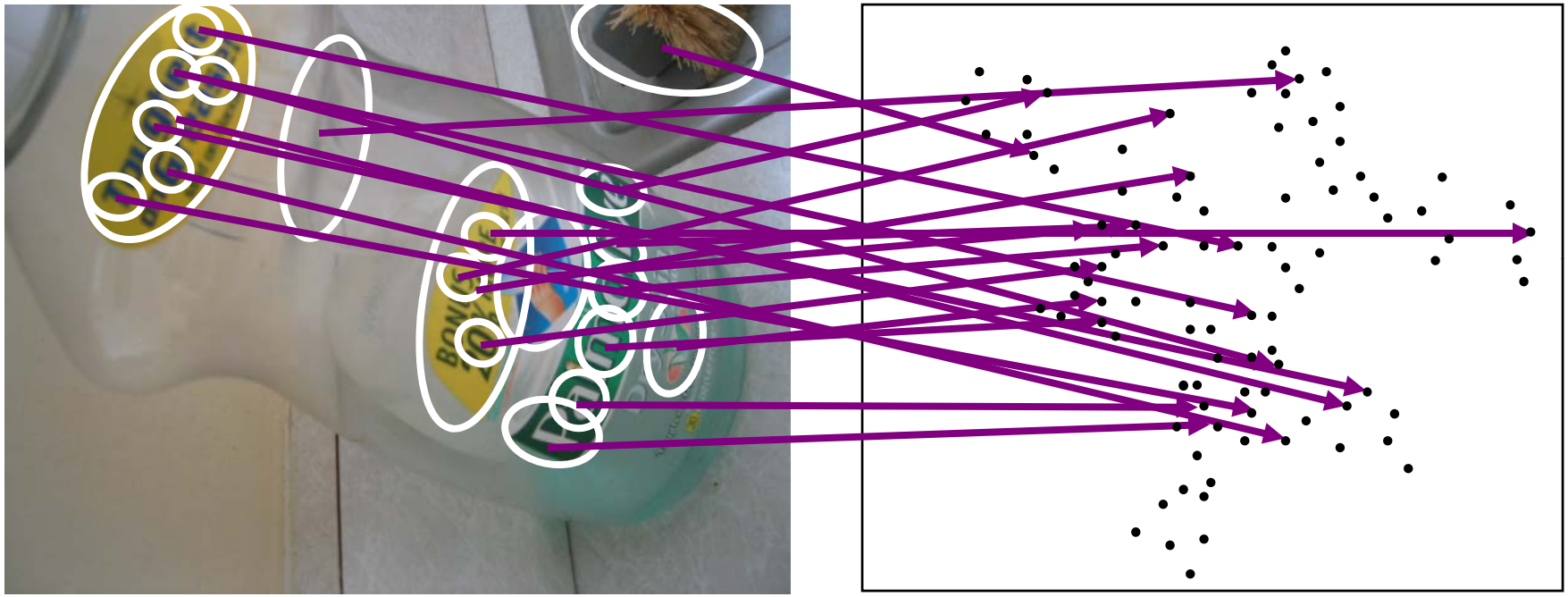
Slide credit: D. Nister

Visual words: main idea

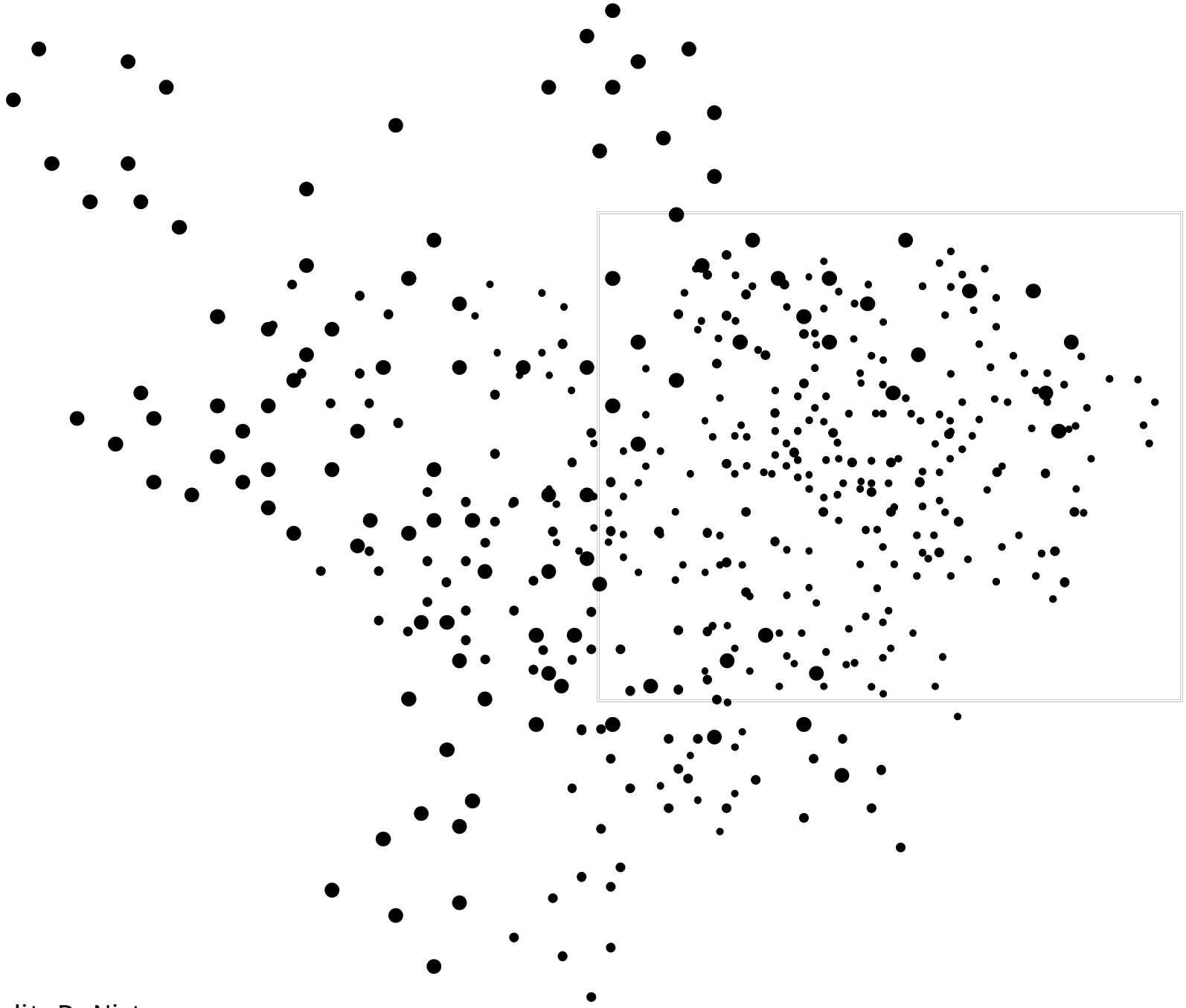


Slide credit: D. Nister

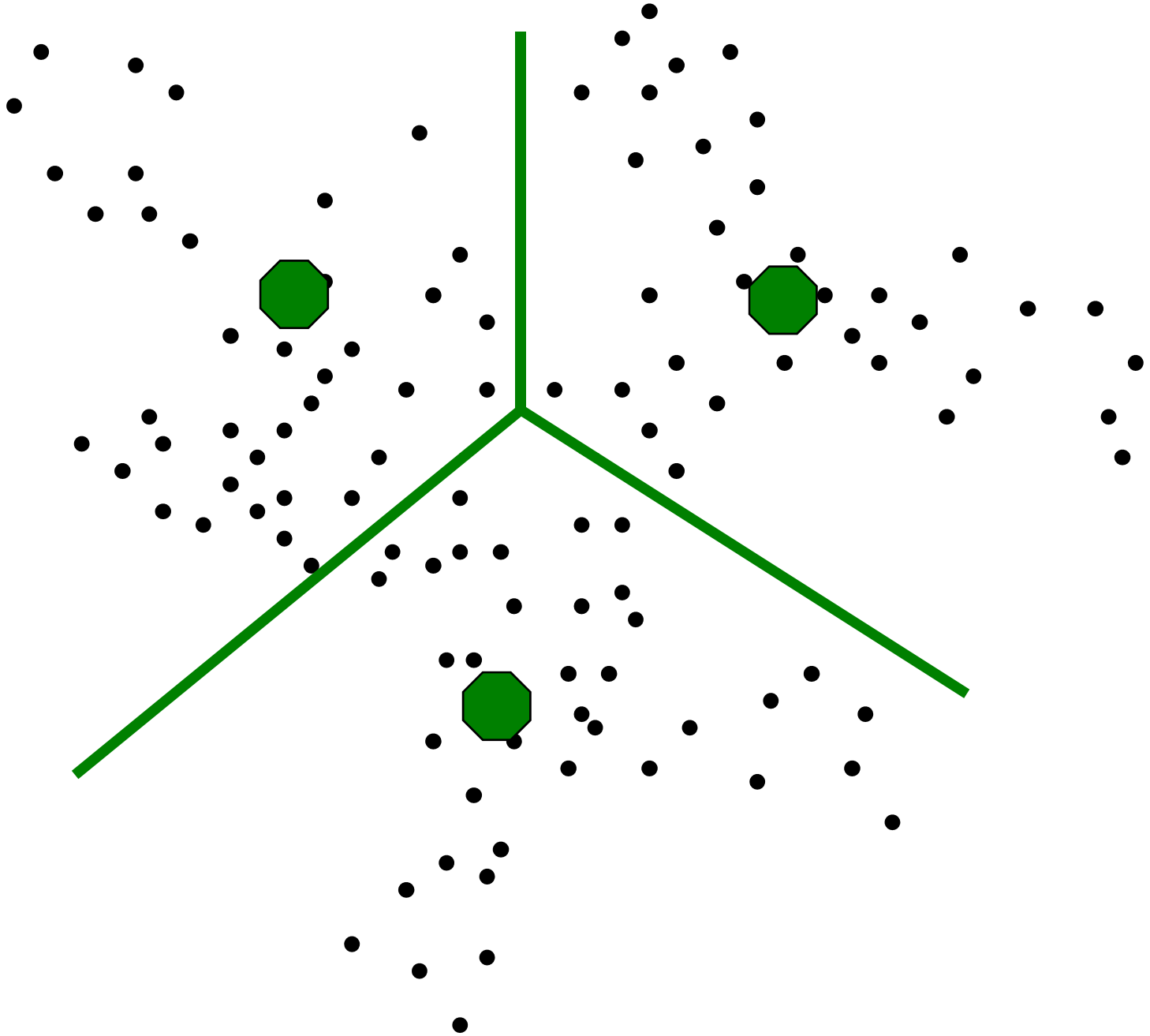
Visual words: main idea



Slide credit: D. Nister



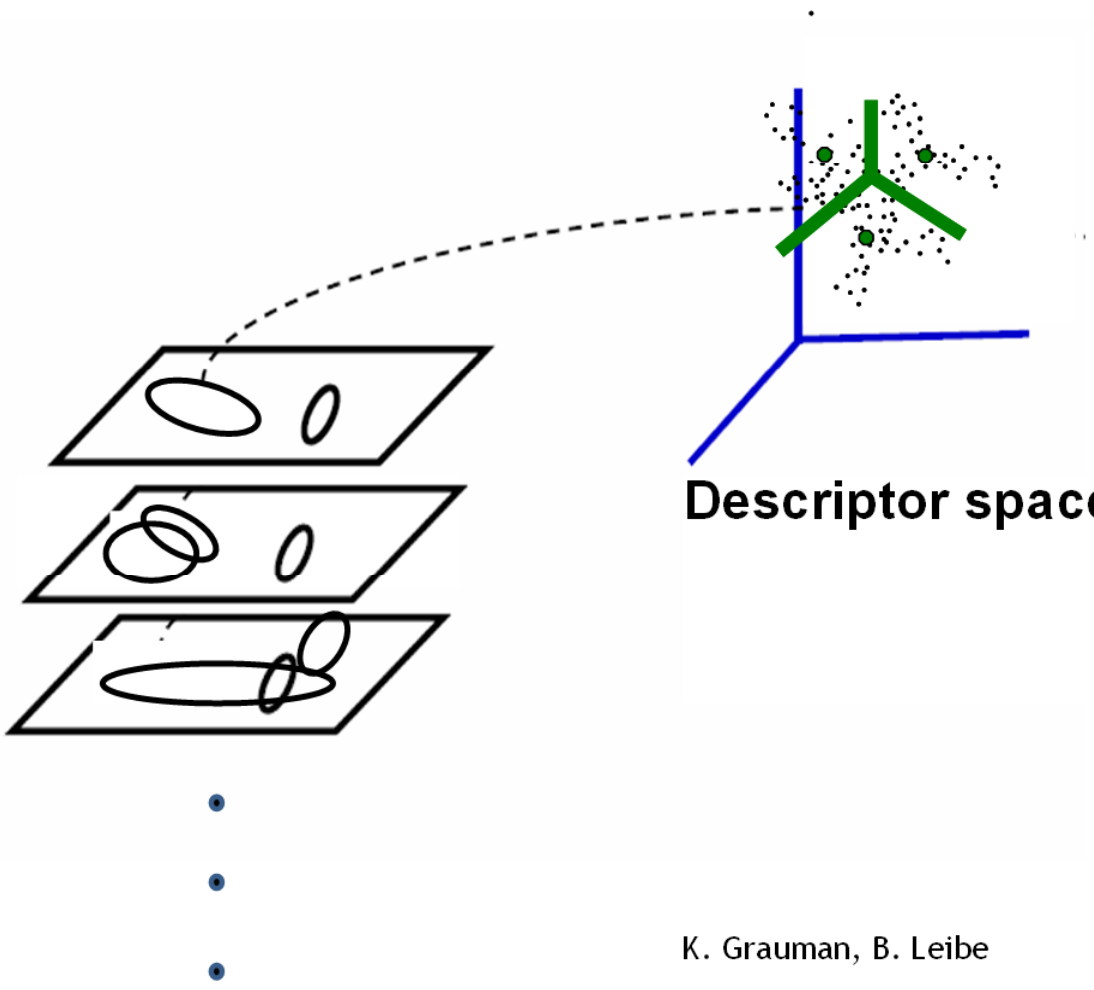
Slide credit: D. Nister



Slide credit: D. Nister

Visual words: main idea

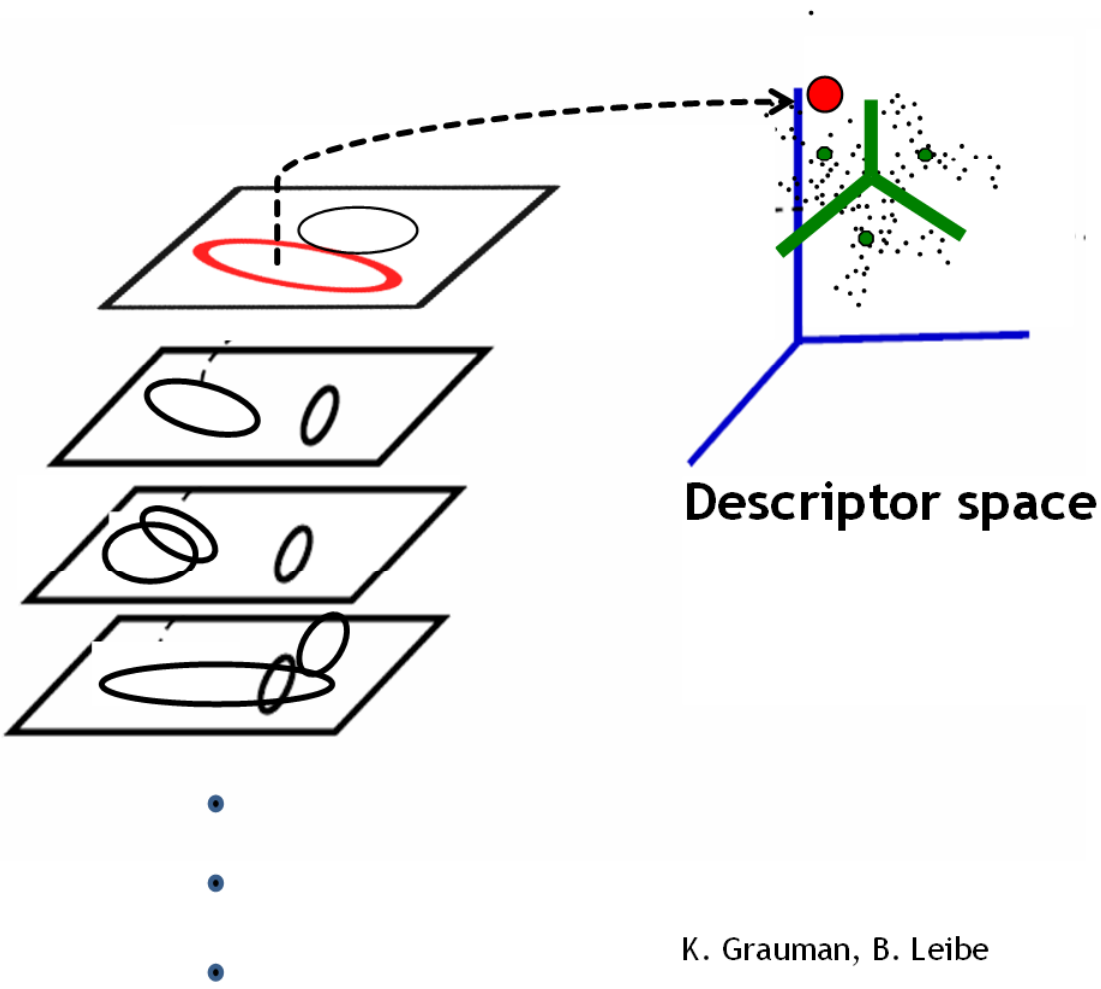
Map high-dimensional descriptors to tokens/words by quantizing the feature space



- Quantize via clustering, let cluster centers be the prototype "words"

Visual words: main idea

Map high-dimensional descriptors to tokens/words by quantizing the feature space

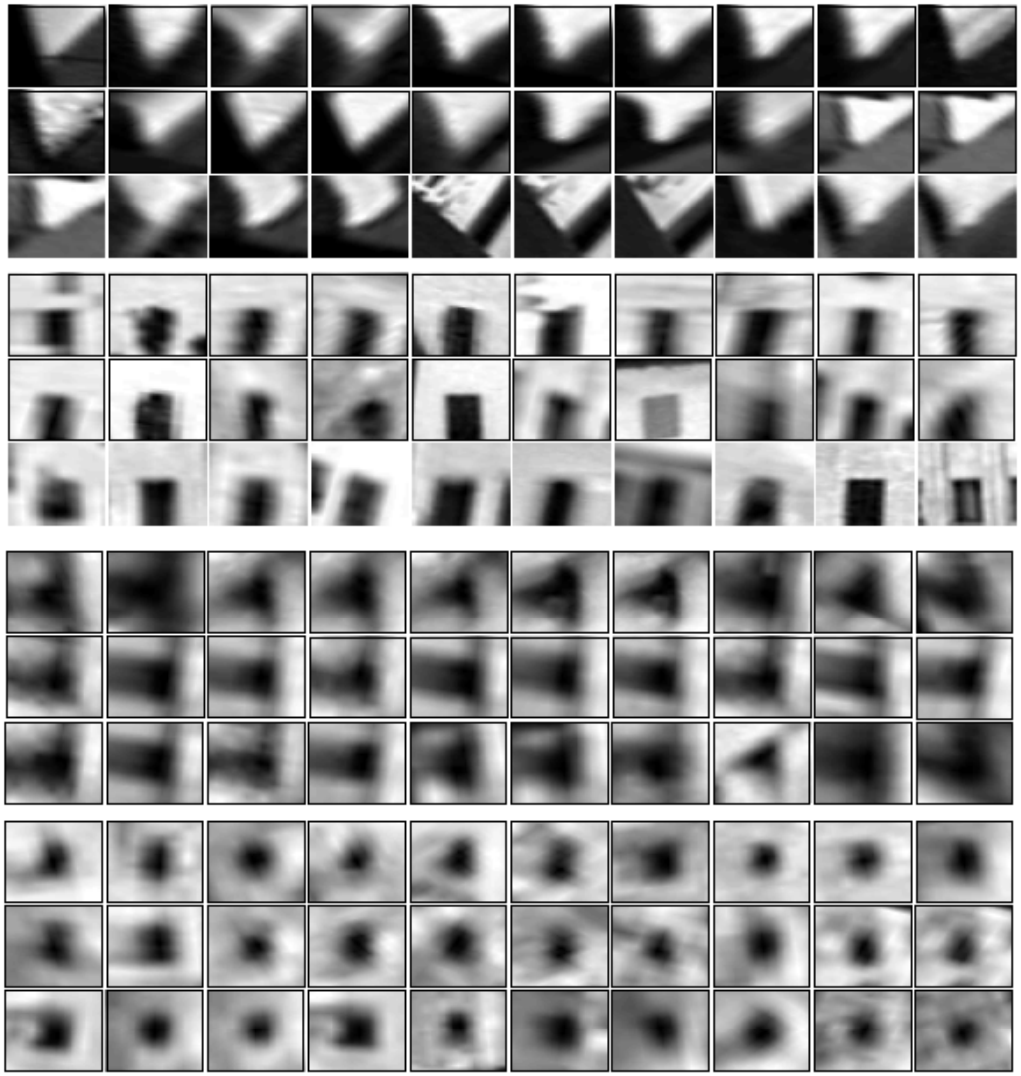


- Determine which word to assign to each new image region by finding the closest cluster center.

K. Grauman, B. Leibe

Visual words

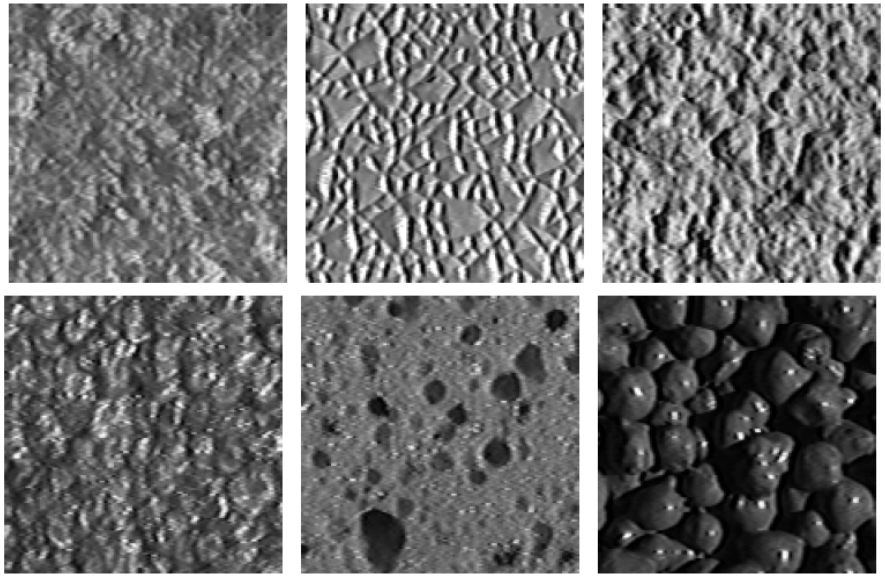
Example: each group of patches belongs to the same visual word



Slide credit: J. Sivic

Visual words

- First explored for texture and material representations
- *Texton* = cluster center of filter responses over collection of images
- Describe textures and materials based on distribution of prototypical texture elements.



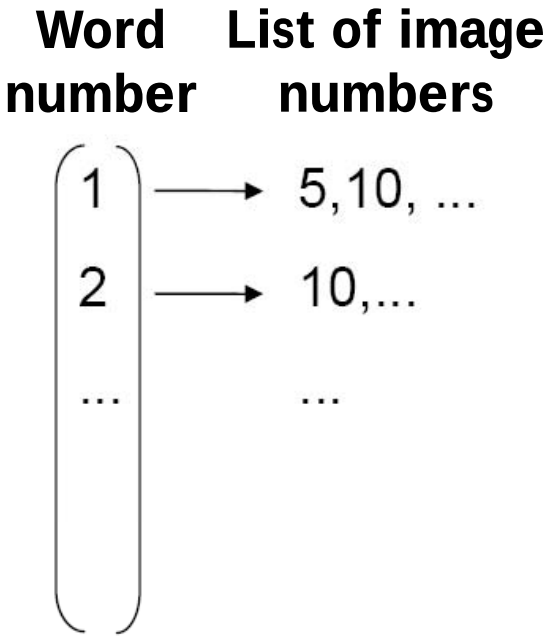
Leung & Malik 1999; Varma & Zisserman, 2002; Lazebnik, Schmid & Ponce, 2003;

Slide credit: J. Sivic

Inverted file index for images comprised of visual words

frame #5

frame #10



- Score each image by the number of common visual words (tentative correspondences)
- But: does not take into account spatial layout of regions

Slide credit: J. Sivic

Clustering / quantization methods

• k-means (typical choice), agglomerative clustering, mean-shift,…

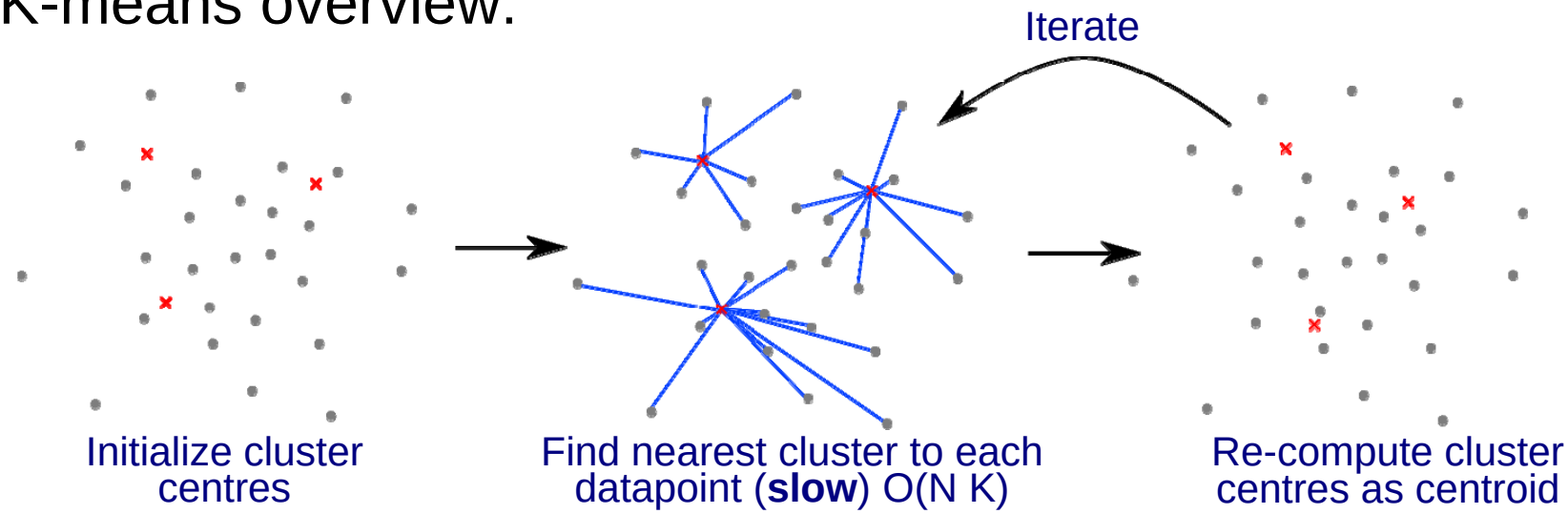
• Hierarchical clustering: allows faster insertion / word assignment while still allowing large vocabularies

- Vocabulary tree [Nister & Stewenius, CVPR 2006]

Slide credit: J. Sivic

Quantization using K-means

- K-means overview:



- K-means provably locally minimizes the sum of squared errors (SSE) between a cluster centre and its points

- But: The quantizer depends on the initialization.

- The nearest neighbour search is the bottleneck

Slide credit: J. Sivic

Approximate K-means

- Use the approximate nearest neighbour search (randomized forest of kd-trees) to determine the closest cluster centre for each data point.

- Original K-means complexity = O(N K)
- Approximate K-means complexity = O(N log K)

- Can be scaled to very large K.

Slide credit: J. Sivic

Clustering / quantization methods

• k-means (typical choice), agglomerative clustering, mean-shift,…

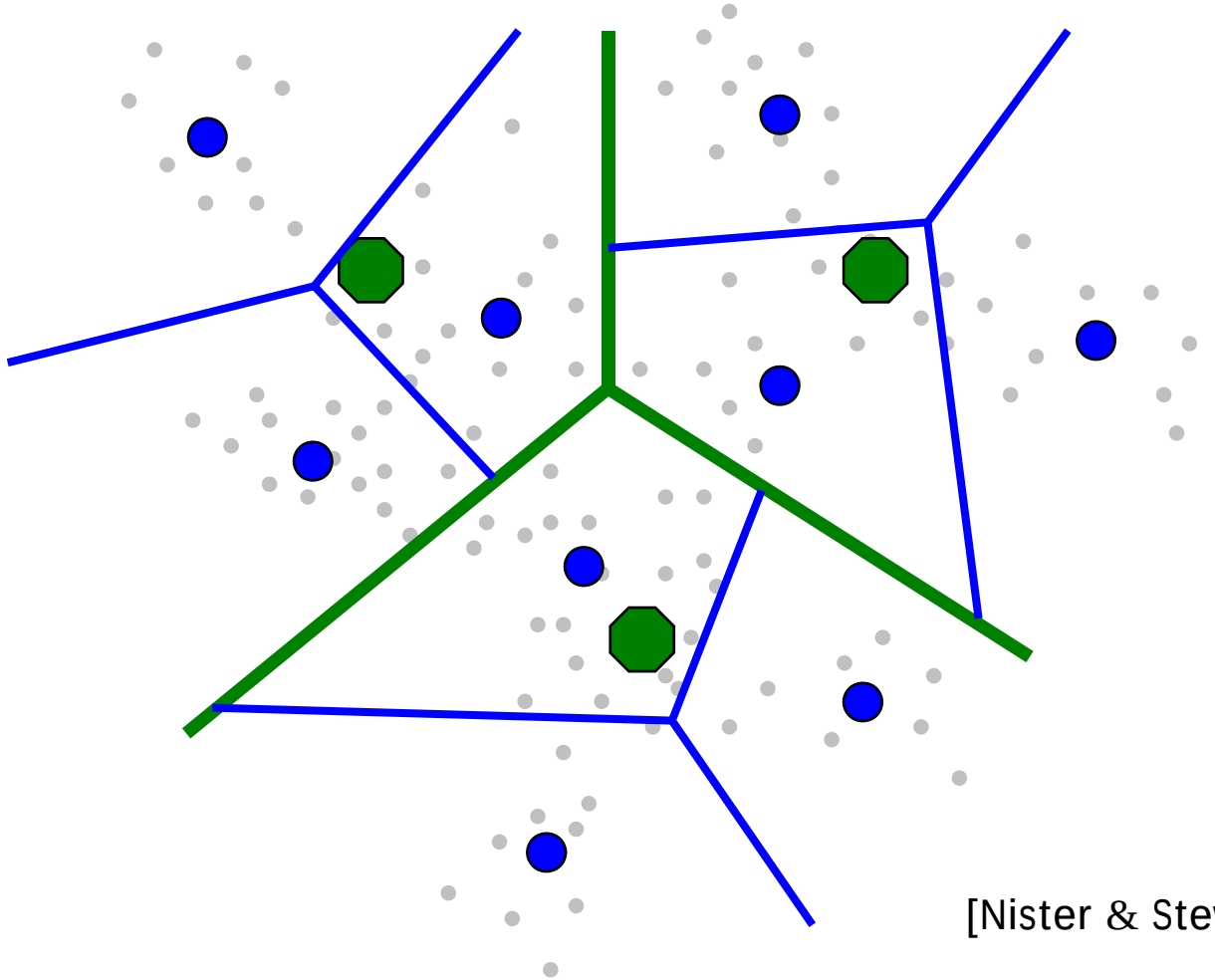
• Hierarchical clustering: allows faster insertion / word assignment while still allowing large vocabularies

- **Vocabulary tree [Nister & Stewenius, CVPR 2006]**

Slide credit: J. Sivic

Example: Recognition with Vocabulary Tree

Tree construction:

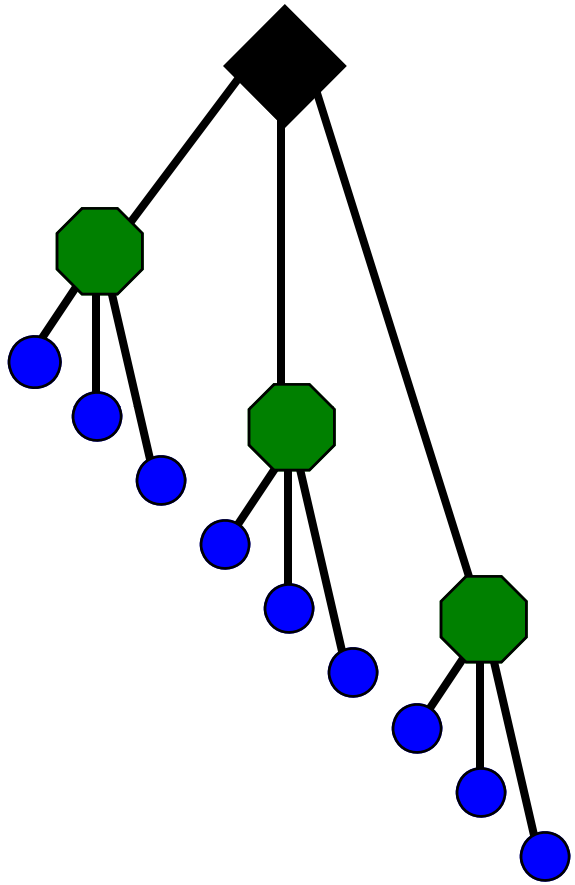


[Nister & Stewenius, CVPR'06]

Slide credit: David Nister

Vocabulary Tree

Training: Filling the tree

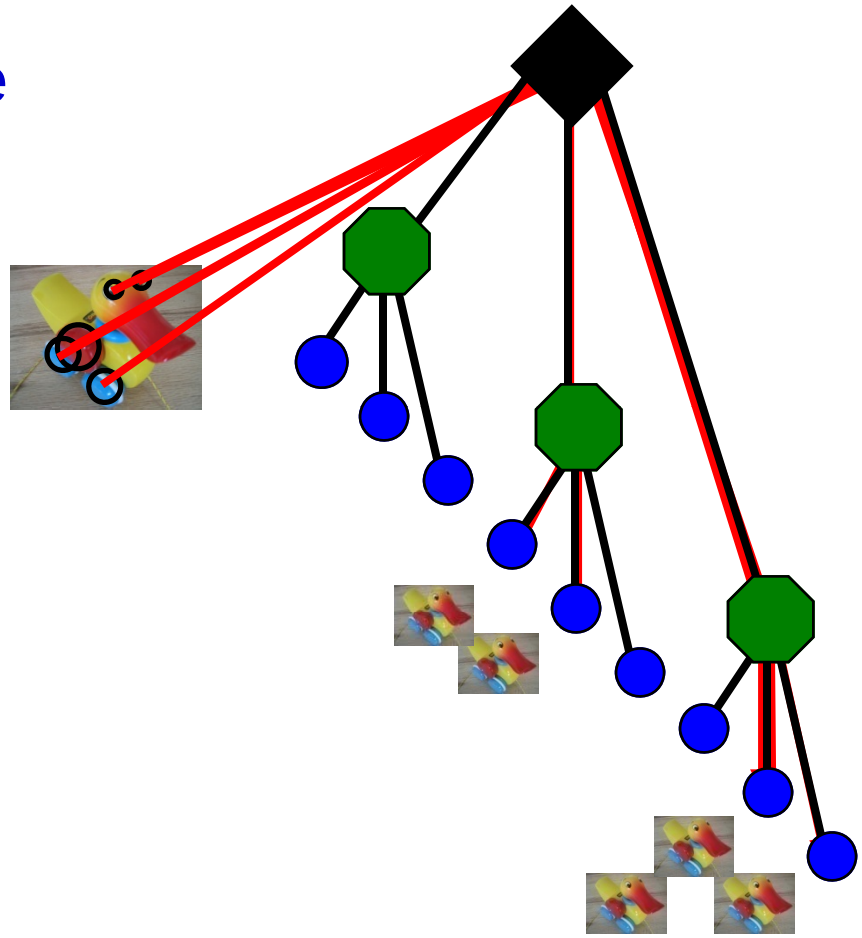


[Nister & Stewenius, CVPR'06]

Slide credit: David Nister

Vocabulary Tree

Training: Filling the tree

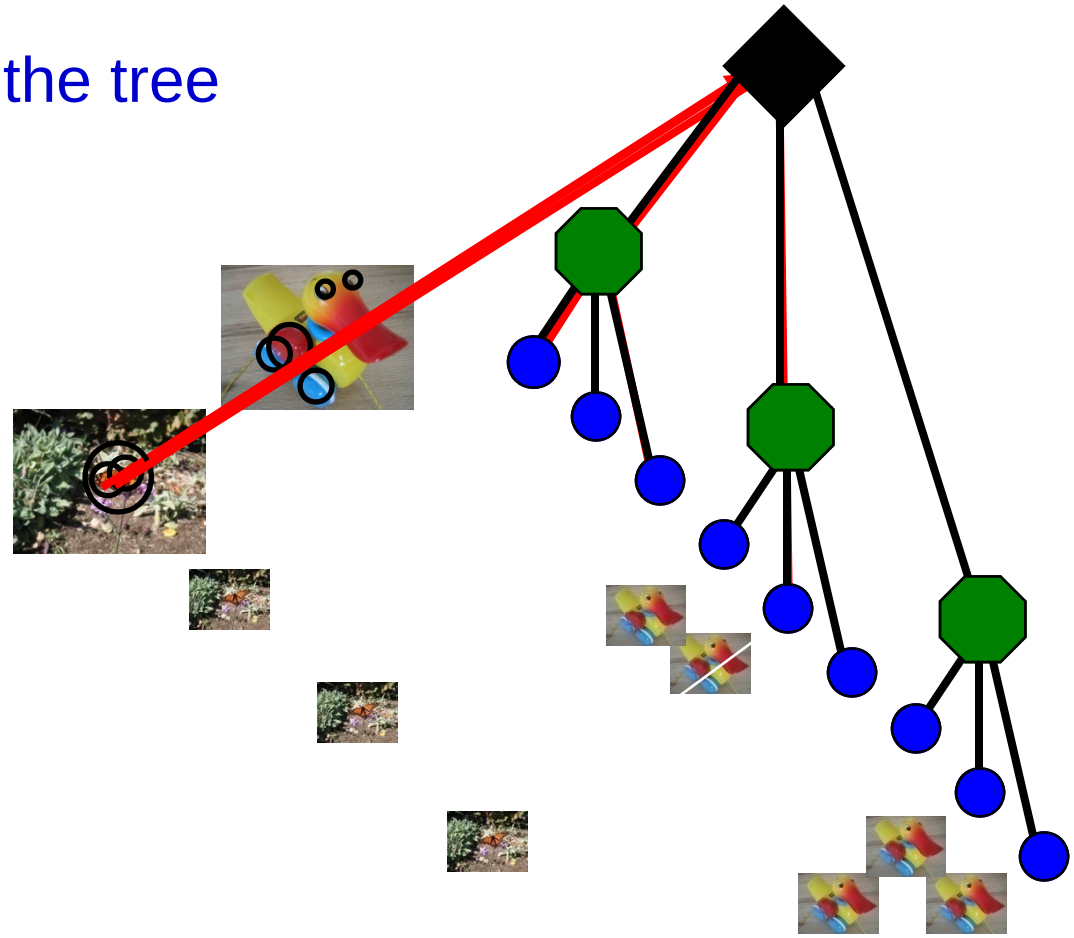


[Nister & Stewenius, CVPR'06]

Slide credit: David Nister

Vocabulary Tree

Training: Filling the tree

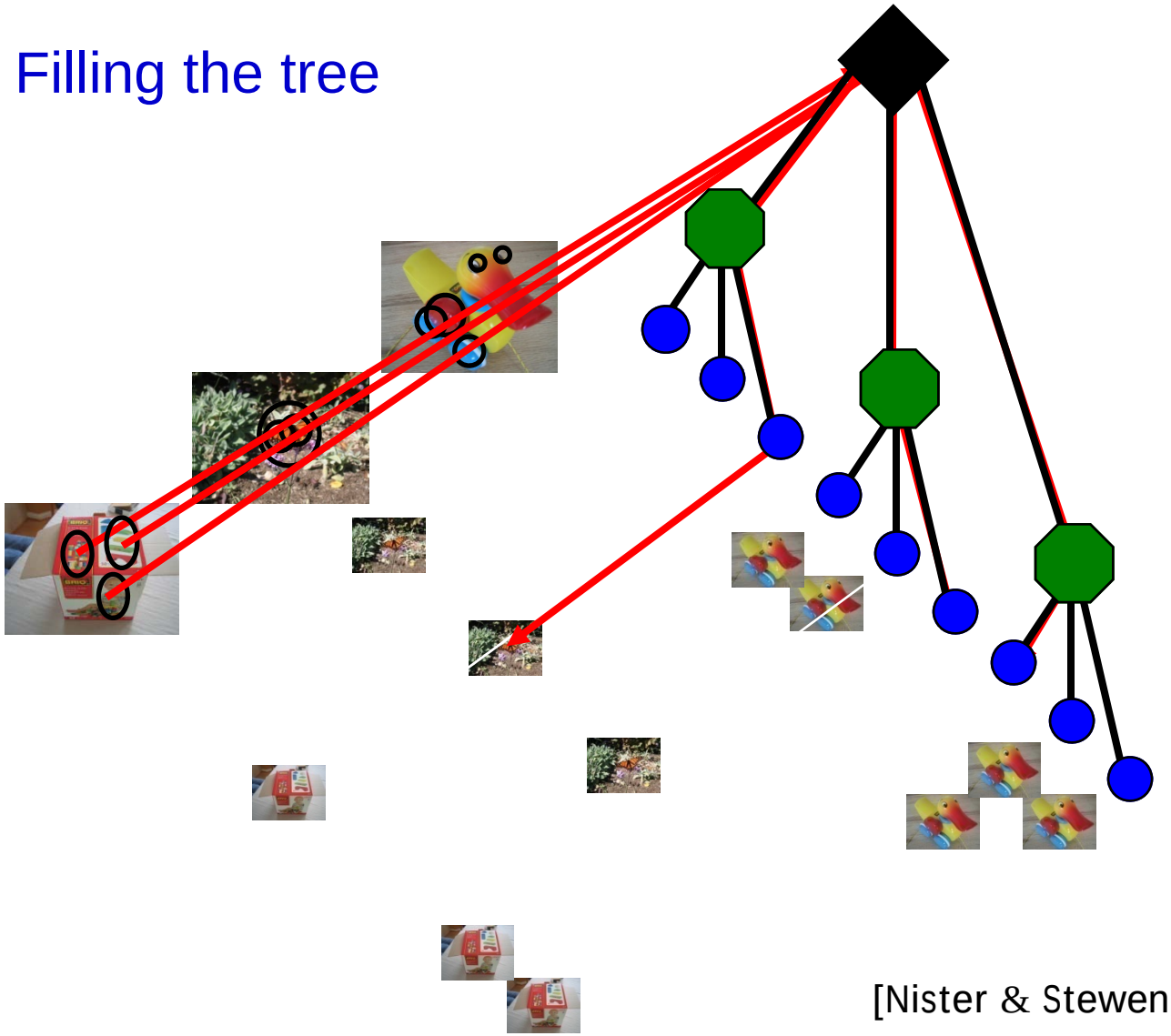


[Nister & Stewenius, CVPR'06]

Slide credit: David Nister

Vocabulary Tree

Training: Filling the tree

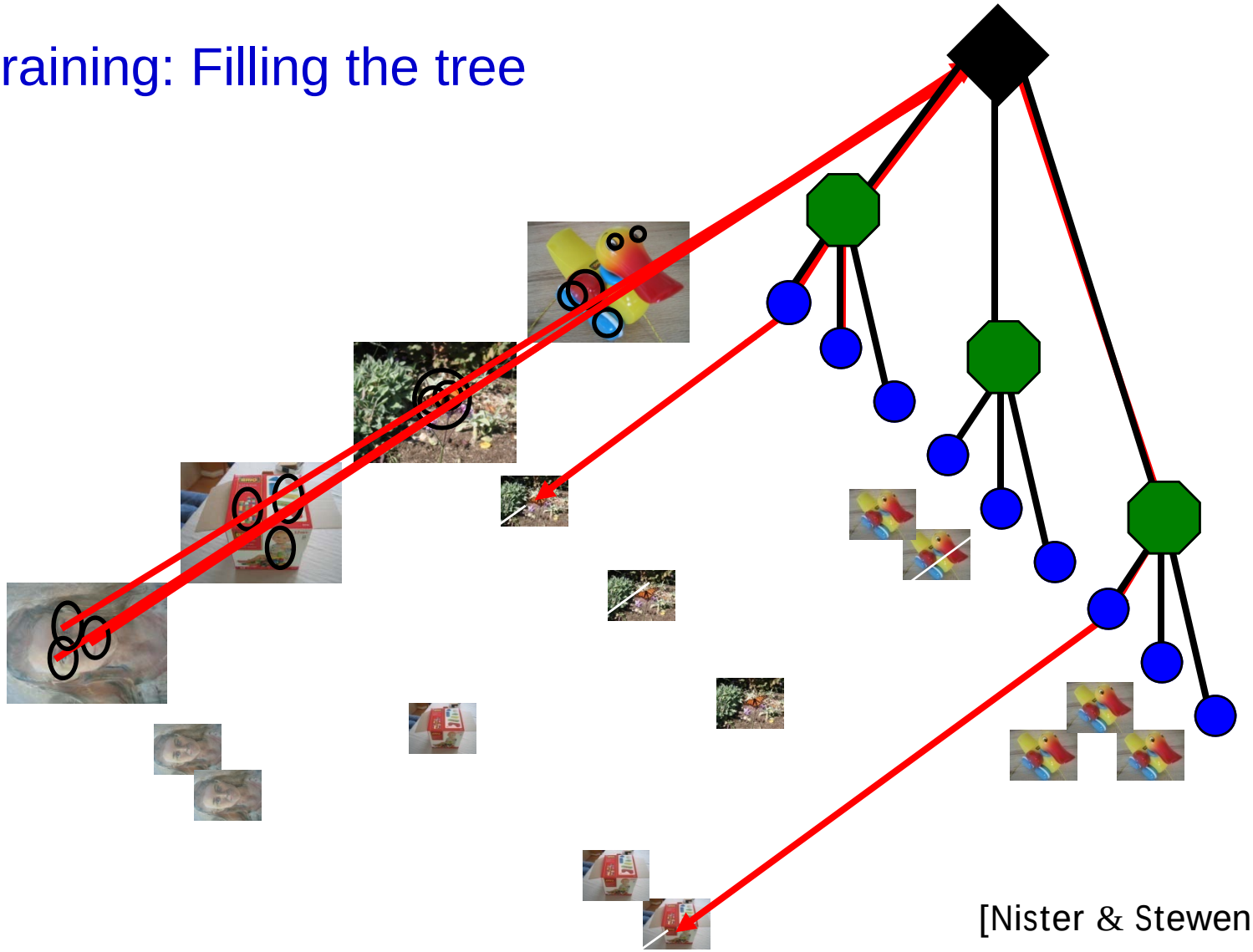


[Nister & Stewenius, CVPR'06]

Slide credit: David Nister

Vocabulary Tree

Training: Filling the tree

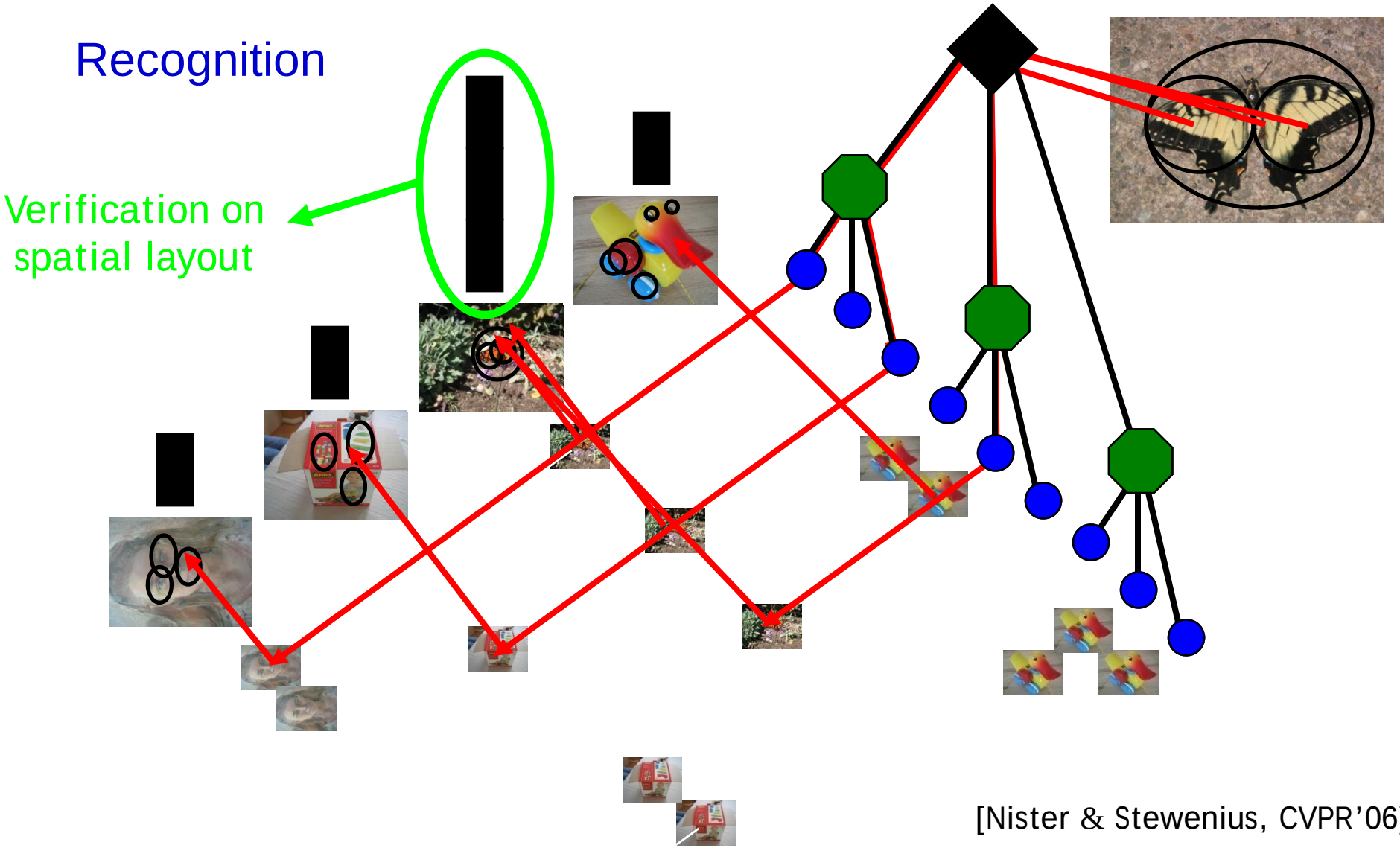


[Nister & Stewenius, CVPR'06]

Slide credit: David Nister

Vocabulary Tree

Recognition



[Nister & Stewenius, CVPR'06]

Slide credit: David Nister

Vocabulary Tree: Performance

Evaluated on large databases

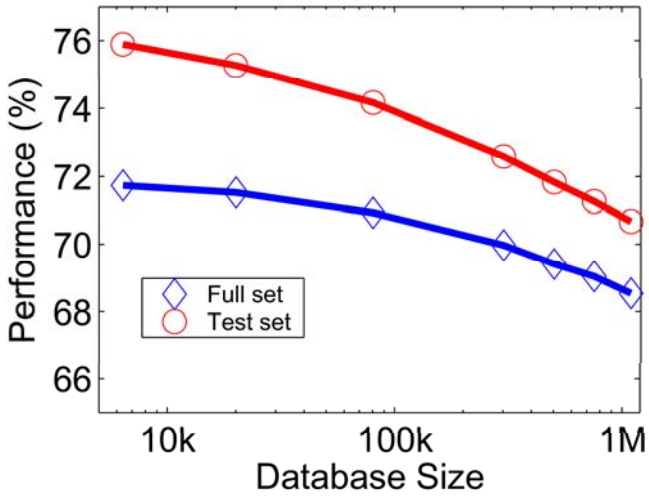
- Indexing with up to 1M images

Online recognition for database of 50,000 CD covers

- Retrieval in ~1s

Find experimentally that large vocabularies can be beneficial for recognition

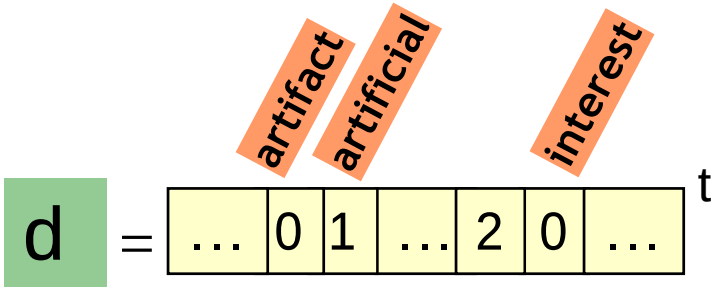
[Nister & Stewenius, CVPR'06]



Slide credit: J. Sivic

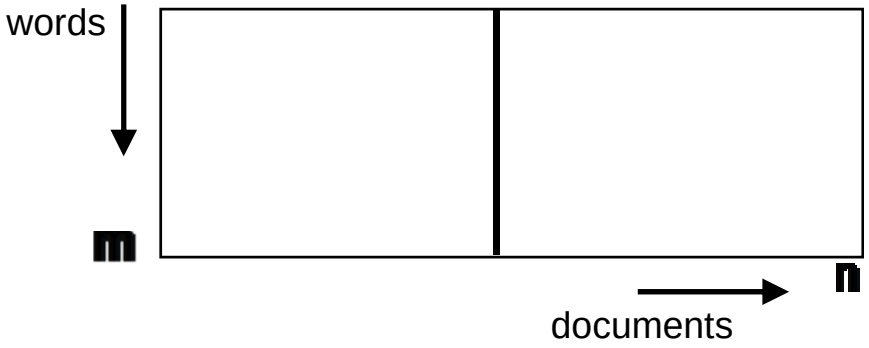
‘Bag of words’ model in text

Document = histogram of word frequencies (‘bag of words’ model)



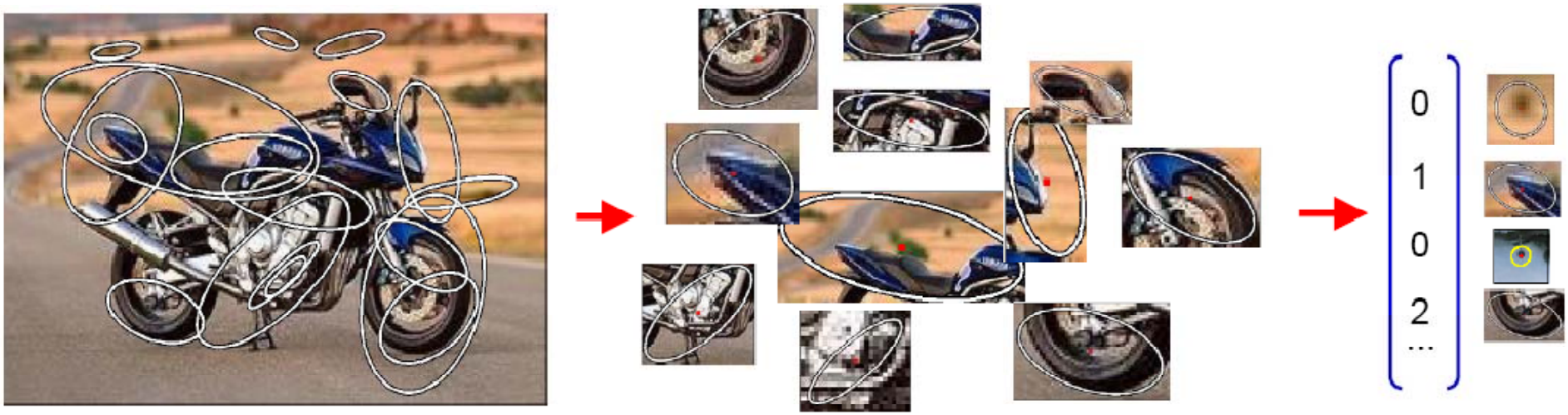
Hofmann 2001

Term-document matrix



Slide credit: J. Sivic

“Bag of visual words”



Slide credit: J. Sivic

Beyond Bag of Words

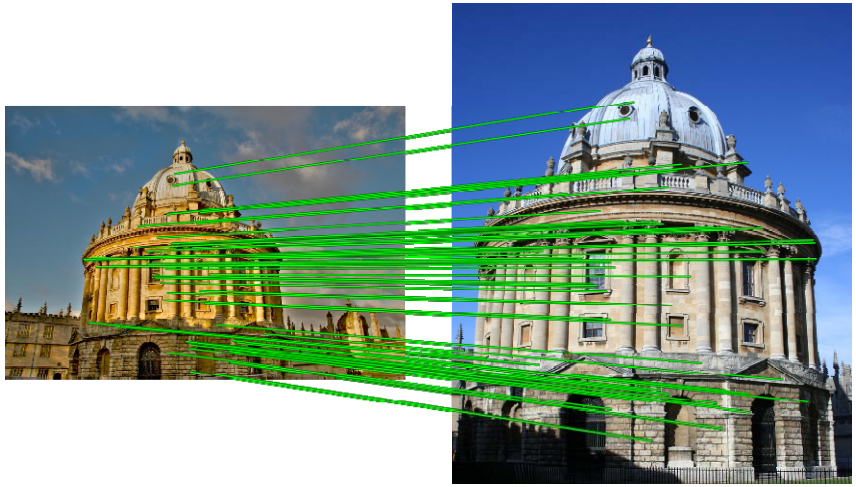
- Use the **position** and **shape** of the underlying features to improve retrieval quality

- Both images have many matches – which is correct?

Slide credit: J. Sivic

Beyond Bag of Words

- We can measure **spatial consistency** between the query and each result to improve retrieval quality



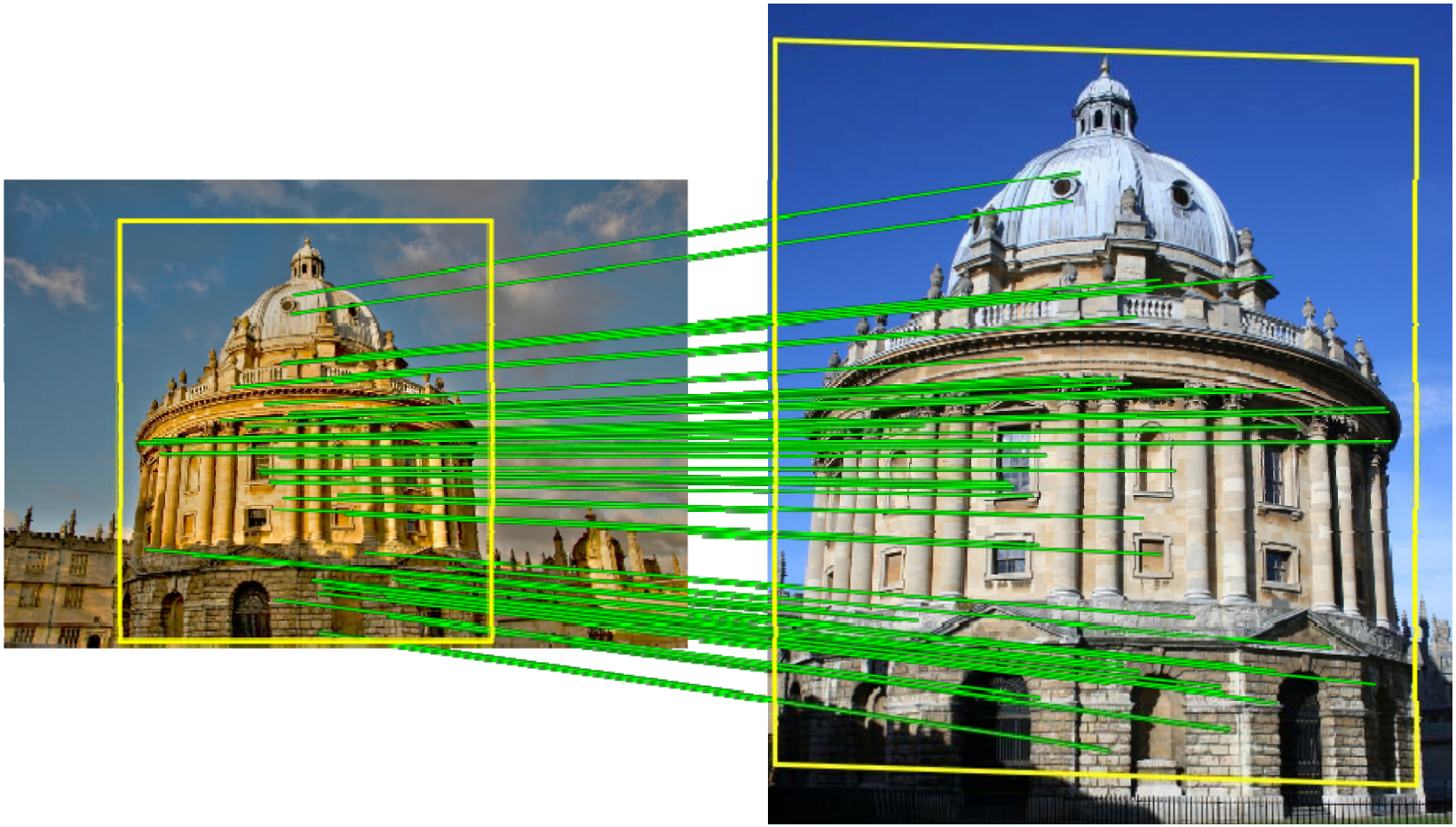
Many spatially consistent matches – **correct result**

Few spatially consistent matches – **incorrect result**

Slide credit: J. Sivic

Beyond Bag of Words

- Extra bonus – gives **localization** of the object



Slide credit: J. Sivic

Considered transformation

2D affine geometric transformation

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} \text{H} \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

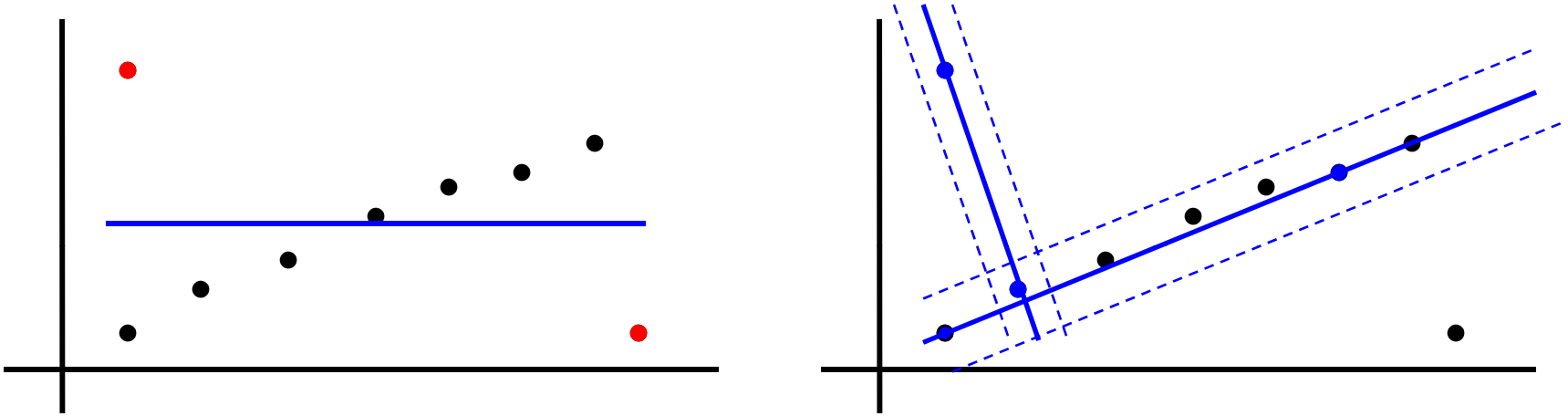
where H is a 2x2 non-singular matrix

Approximation to a planar homography (projective transformation)

Slide credit: J. Sivic

Review: Robust line estimation - RANSAC

Fit a line to 2D data containing outliers



There are two problems

1. a line fit which minimizes perpendicular distance
2. a classification into inliers (valid points) and outliers

Solution: use robust statistical estimation algorithm RANSAC (RANdom Sample Consensus) [Fishler & Bolles, 1981]

Slide credit: J. Sivic

RANSAC robust line estimation

Repeat

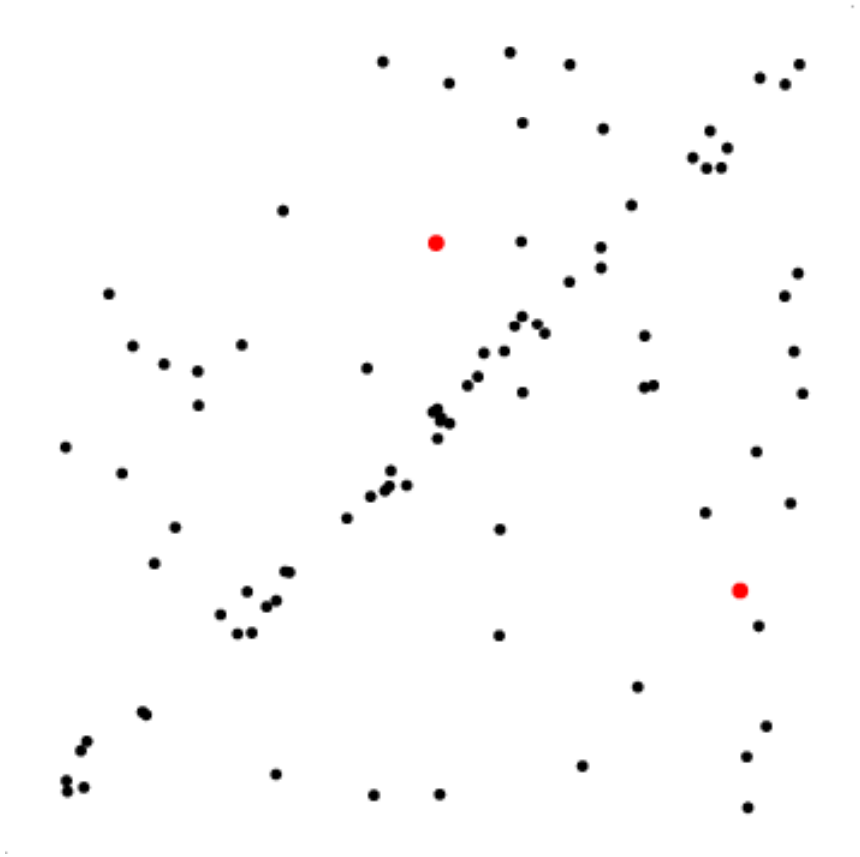
1. Select random sample of 2 points
2. Compute the line through these points
3. Measure support (number of points within threshold distance of the line)

Choose the line with the largest number of inliers

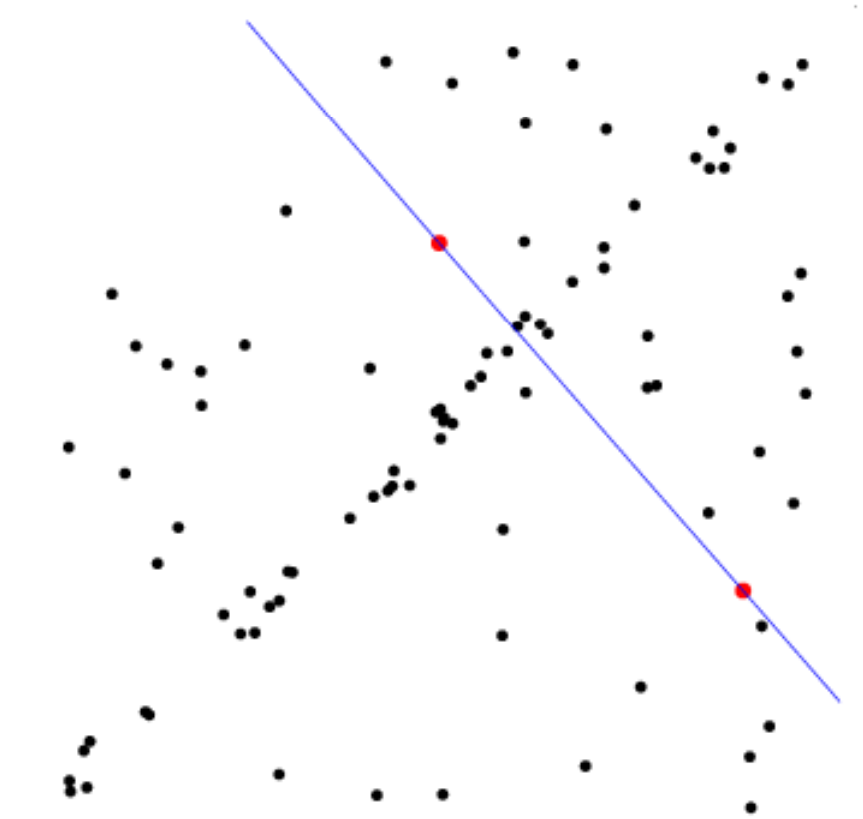
- Compute least squares fit of line to inliers (regression)

Slide credit: J. Sivic

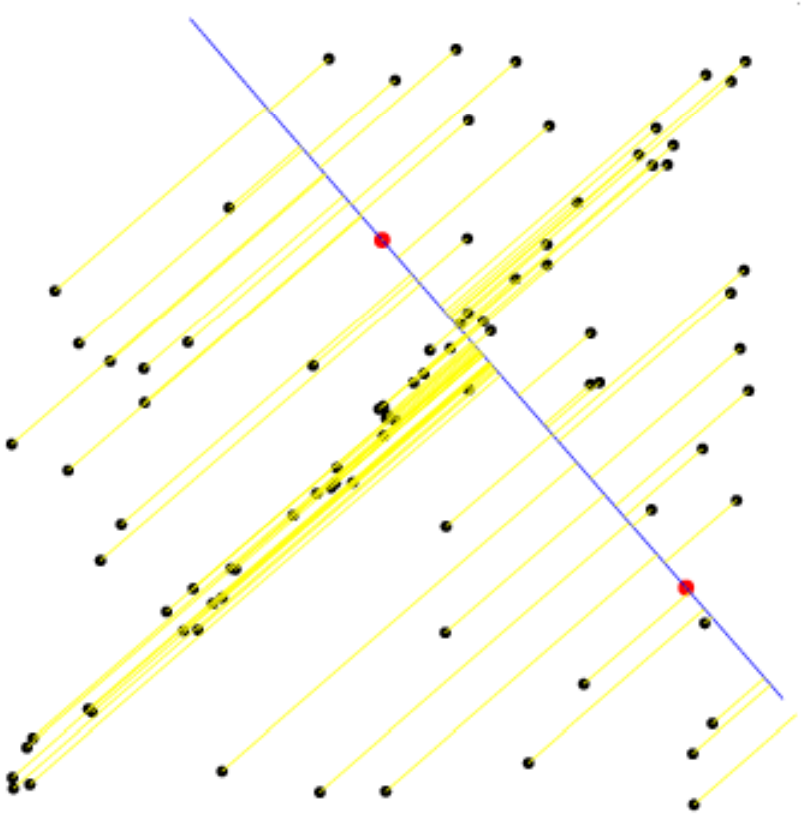
Slide credit: J. Sivic



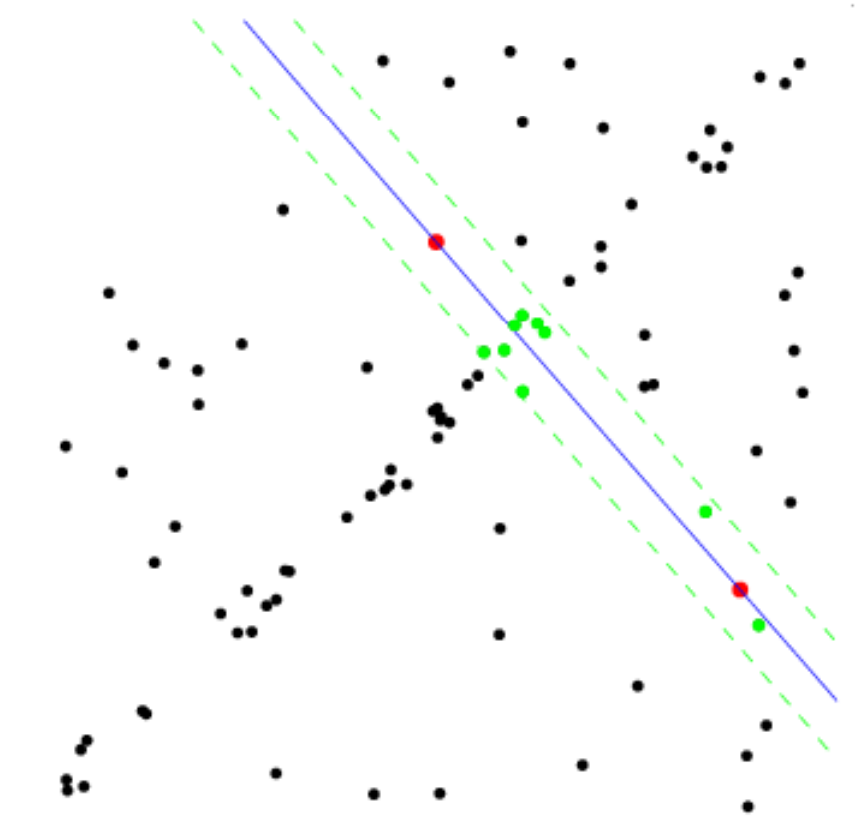
Slide credit: J. Sivic



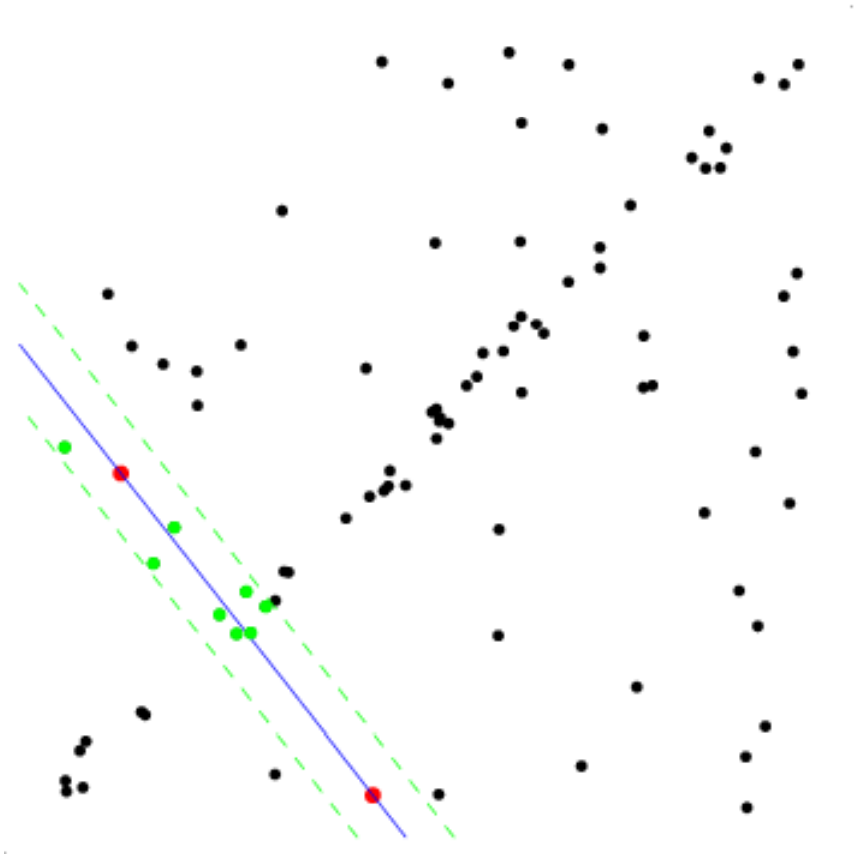
Slide credit: J. Sivic



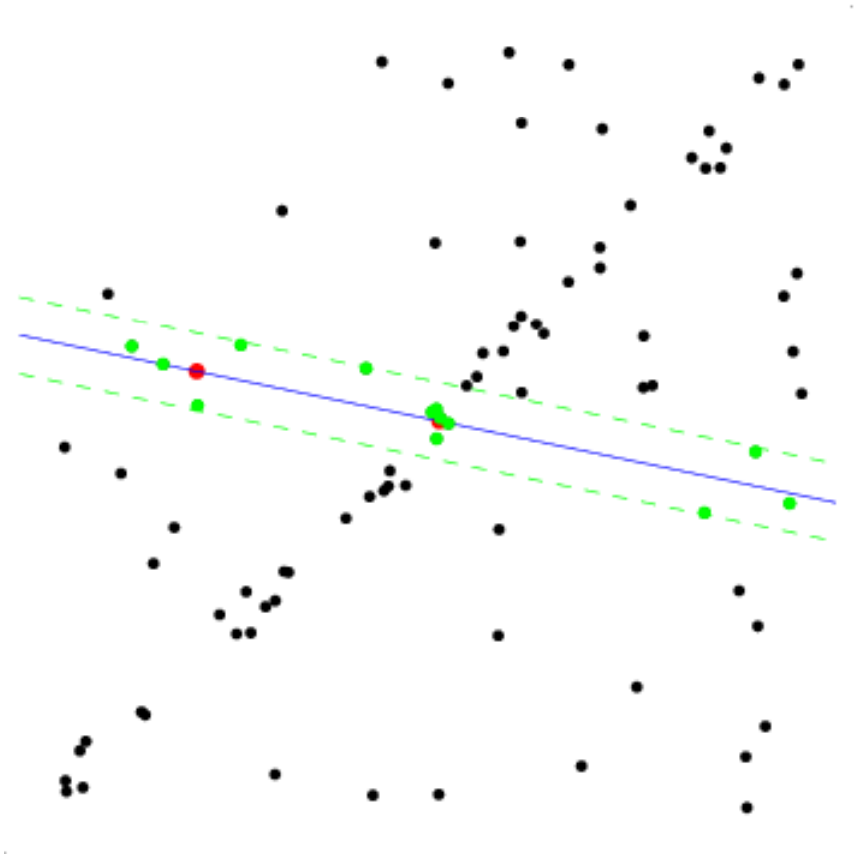
Slide credit: J. Sivic



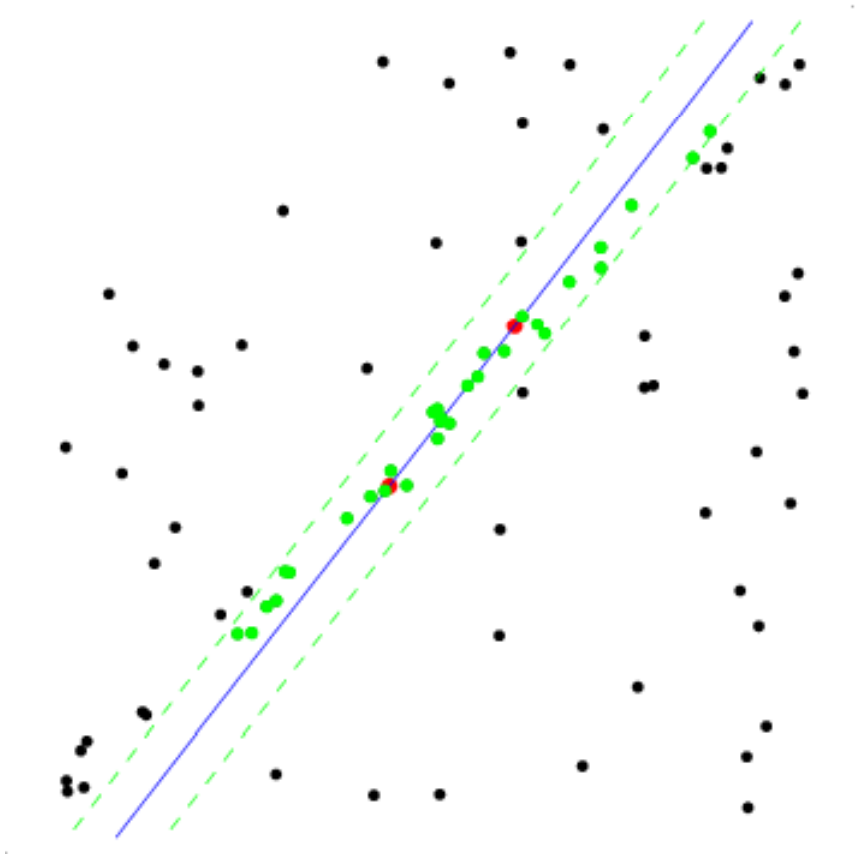
Slide credit: J. Sivic



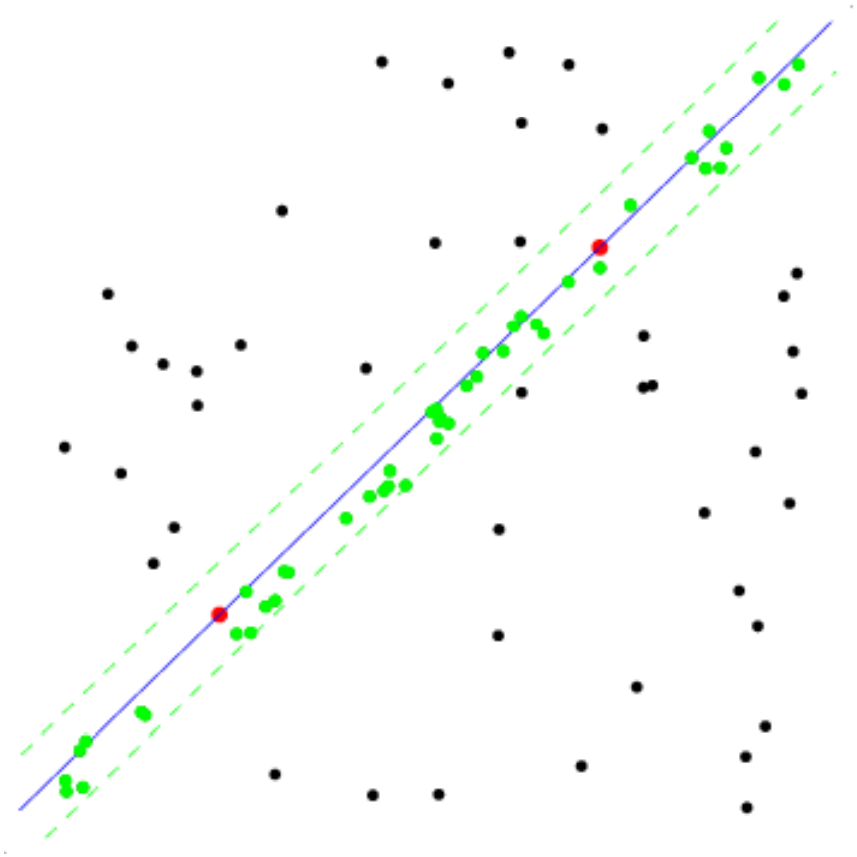
Slide credit: J. Sivic



Slide credit: J. Sivic



Slide credit: J. Sivic



Slide credit: J. Sivic

How many samples?

Number of samples *N*

- Choose *N* so that, with probability *p*, at least one random sample is free from outliers
- e.g.:
  - > *p*=0.99
  - > outlier ratio: *e*

Probability a randomly picked point is an inlier

$$\left(1-\underbrace{\left(1-e\right)^{s}}\right)^{N}=1-p$$

Probability of all points in a sample (of size s) are inliers

Source: M. Pollefeys

How many samples?

Number of samples *N*

- Choose *N* so that, with probability *p*, at least one random sample is free from outliers
- e.g.:
  - > *p*=0.99
  - > outlier ratio: *e*

Probability that all N samples (of size s) are corrupted (contain an outlier)

$$\left(1-\left(1-e\right)^{s}\right)^{N}=1-p$$

Probability of at least one point in a sample (of size s) is an outlier

$$N=\log\left(1-p\right)/\log\left(1-\left(1-e\right)^{s}\right)$$

	proportion of outliers *e*						
s	5%	10%	20%	30%	40%	50%	90%
1	2	2	3	4	5	6	43
2	2	3	5	7	11	17	458
3	3	4	7	11	19	35	4603
4	3	5	9	17	34	72	4.6e4
5	4	6	12	26	57	146	4.6e5
6	4	7	16	37	97	293	4.6e6
7	4	8	20	54	163	588	4.6e7
8	5	9	26	78	272	1177	4.6e8

Source: M. Pollefeys

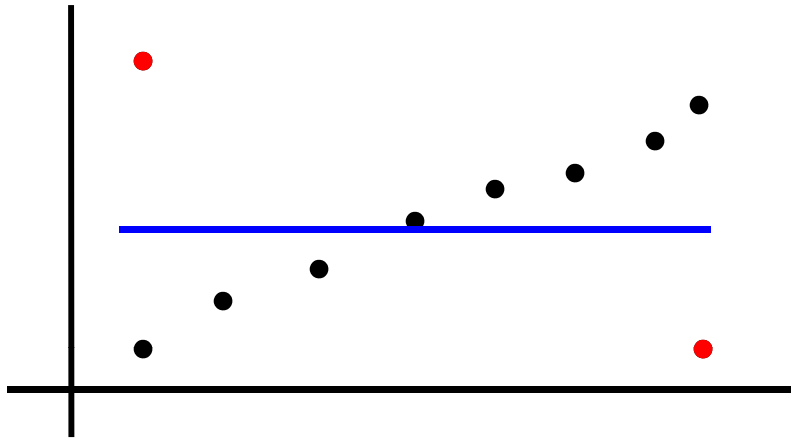
Example: line fitting

p = 0.99

s = ?

e = ?

N = ?



Source: M. Pollefeys

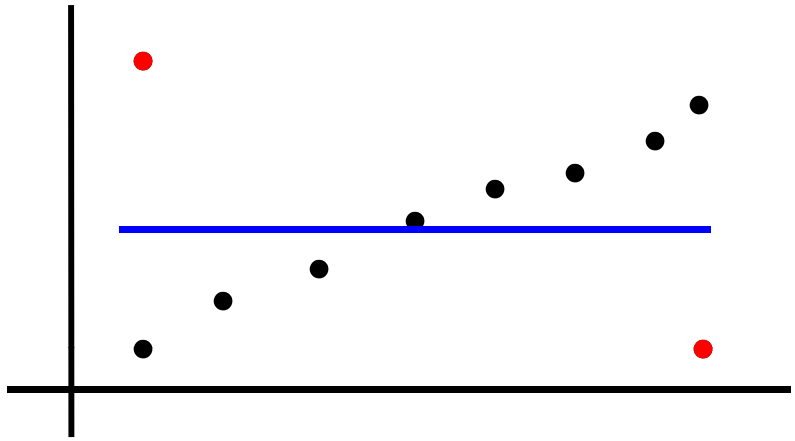
Example: line fitting

p = 0.99

s = 2

e = 2/10 = 0.2

N = 5



Compare with exhaustively trying all point pairs:

$$\binom{10}{2} = 10*9 / 2 = 45$$

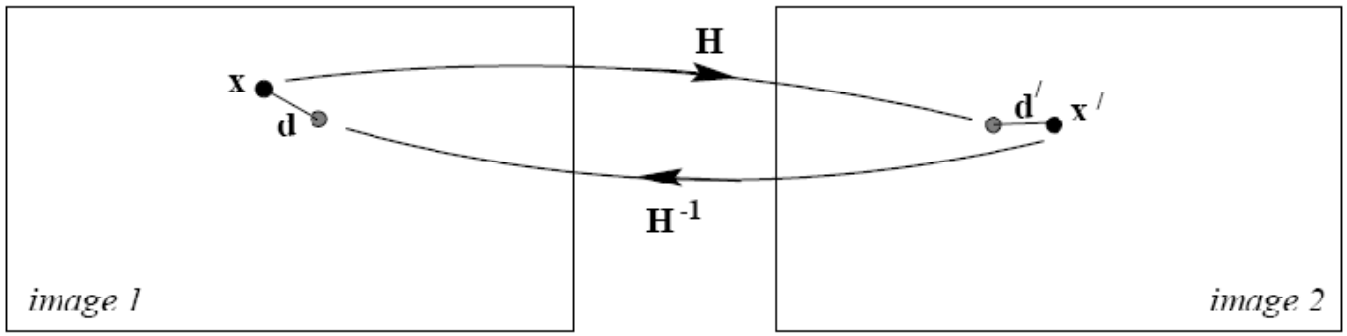
	proportion of outliers *e*						
s	5%	10%	20%	30%	40%	50%	90%
1	2	2	3	4	5	6	43
2	2	3	5	7	11	17	458
3	3	4	7	11	19	35	4603
4	3	5	9	17	34	72	4.6e4
5	4	6	12	26	57	146	4.6e5
6	4	7	16	37	97	293	4.6e6
7	4	8	20	54	163	588	4.6e7
8	5	9	26	78	272	1177	4.6e8

Source: M. Pollefeys

Algorithm summary – RANSAC robust estimation of 2D affine transformation

Repeat

1. Select **1 region to region correspondence** (equivalent to 3 point correspondences)
2. Compute H (2x2 matrix) + t (2x1) vector for translation
3. Measure support (number of inliers within threshold distance, i.e. $d^2_{\text{transfer}} < t$) $d^2_{\text{transfer}} = d(\mathbf{x}, \mathrm{H}^{-1}\mathbf{x}')^2 + d(\mathbf{x}', \mathrm{H}\mathbf{x})^2$



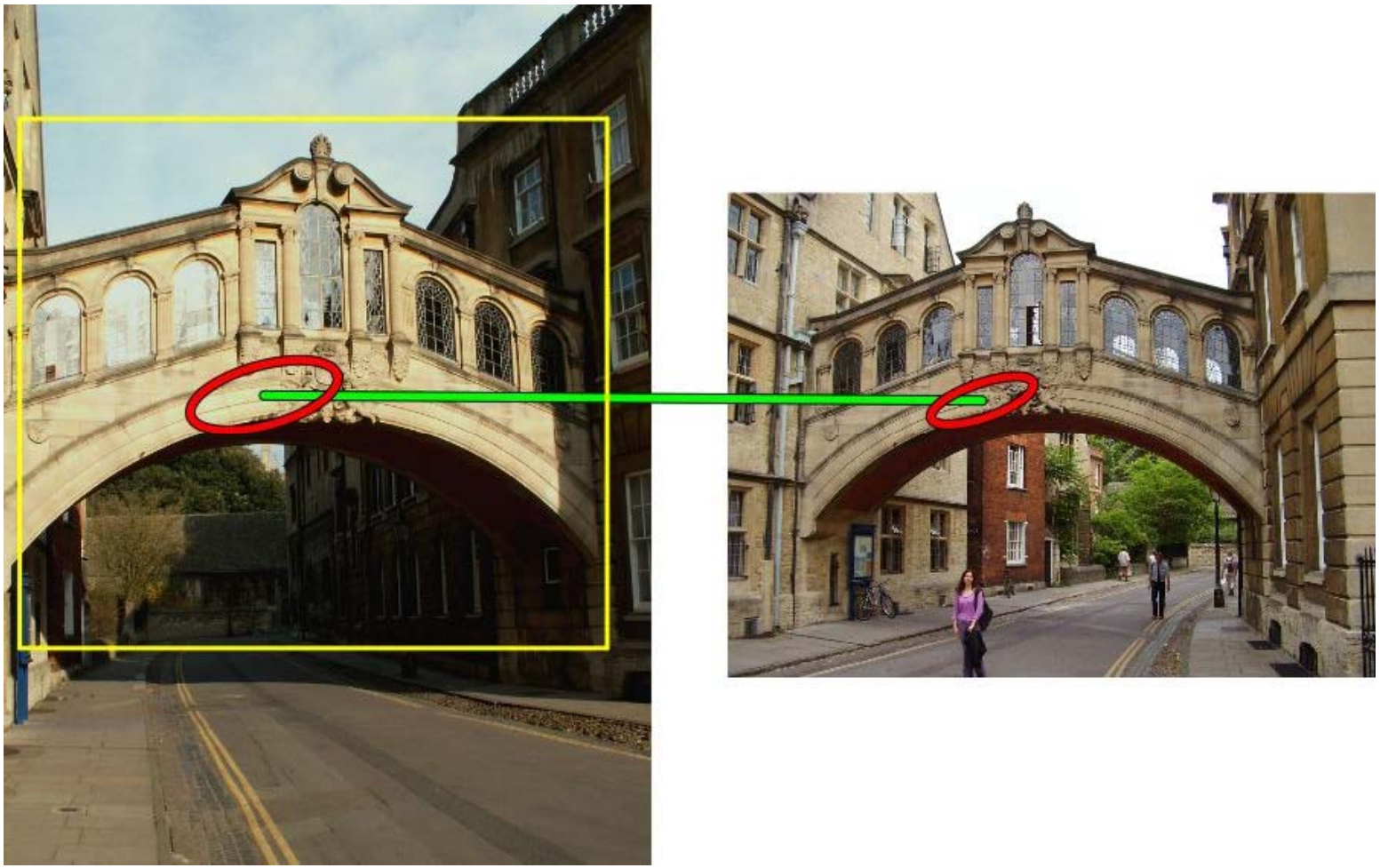
Choose the (H,t) with the largest number of inliers

(Re-estimate (H,t) from all inliers)

Slide credit: J. Sivic

Estimating spatial correspondences

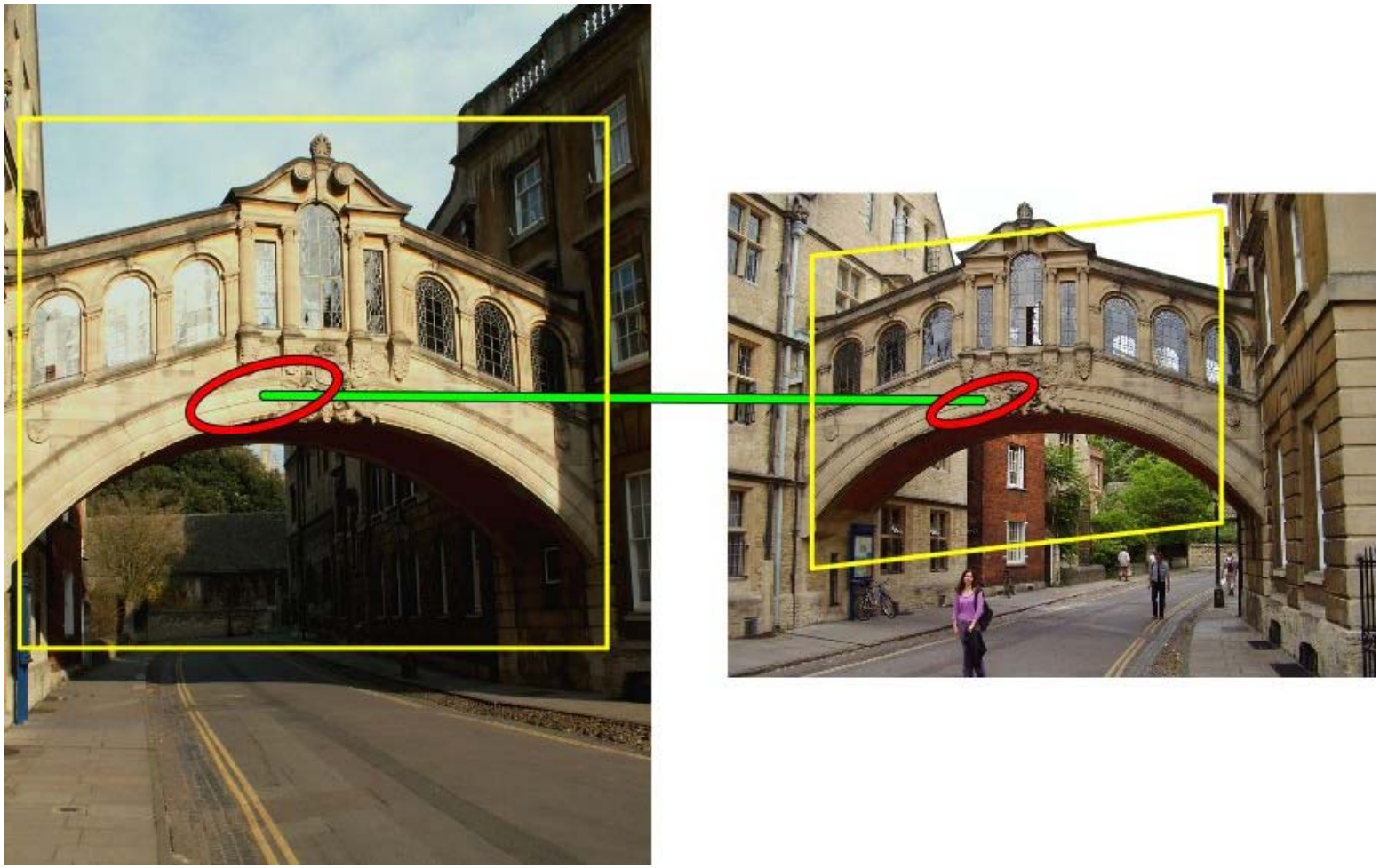
1. Test each correspondence



Slide credit: J. Sivic

Estimating spatial correspondences

2. Compute a planar affine transformation (6 dof)

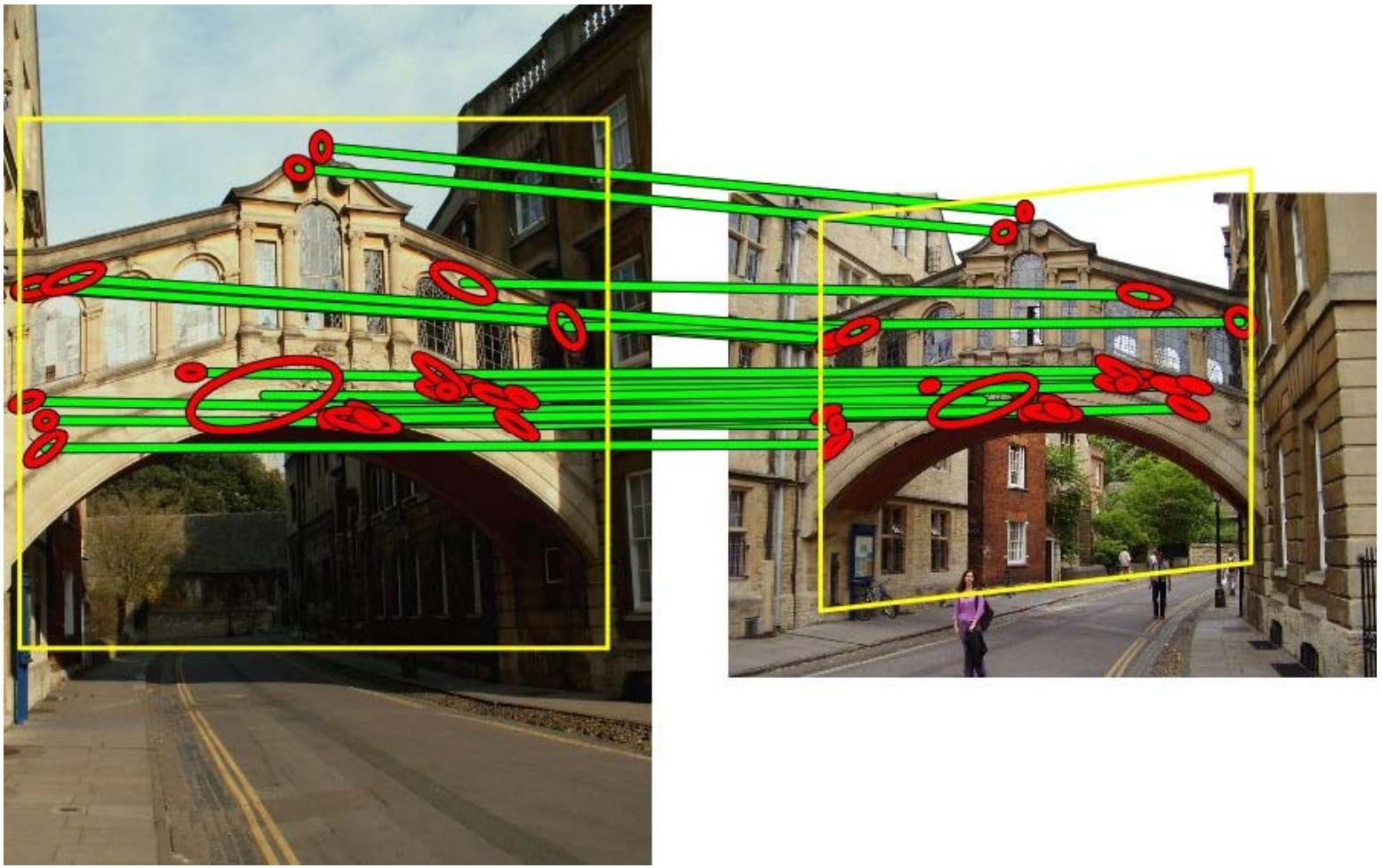


Need just one correspondence

Slide credit: J. Sivic

Estimating spatial correspondences

3. Score by number of consistent matches

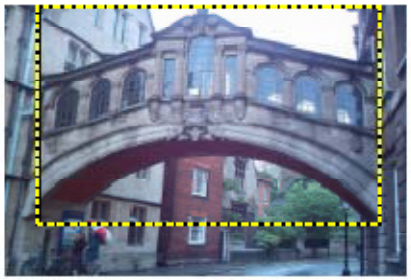


Re-estimate full affine transformation (6 dof)

Slide credit: J. Sivic

Verification by spatial layout - overview

1. Query



2. Initial retrieval set (bag of words model)

3. Spatial verification (re-rank on # of inliers)

Slide credit: J. Sivic

Oxford buildings dataset

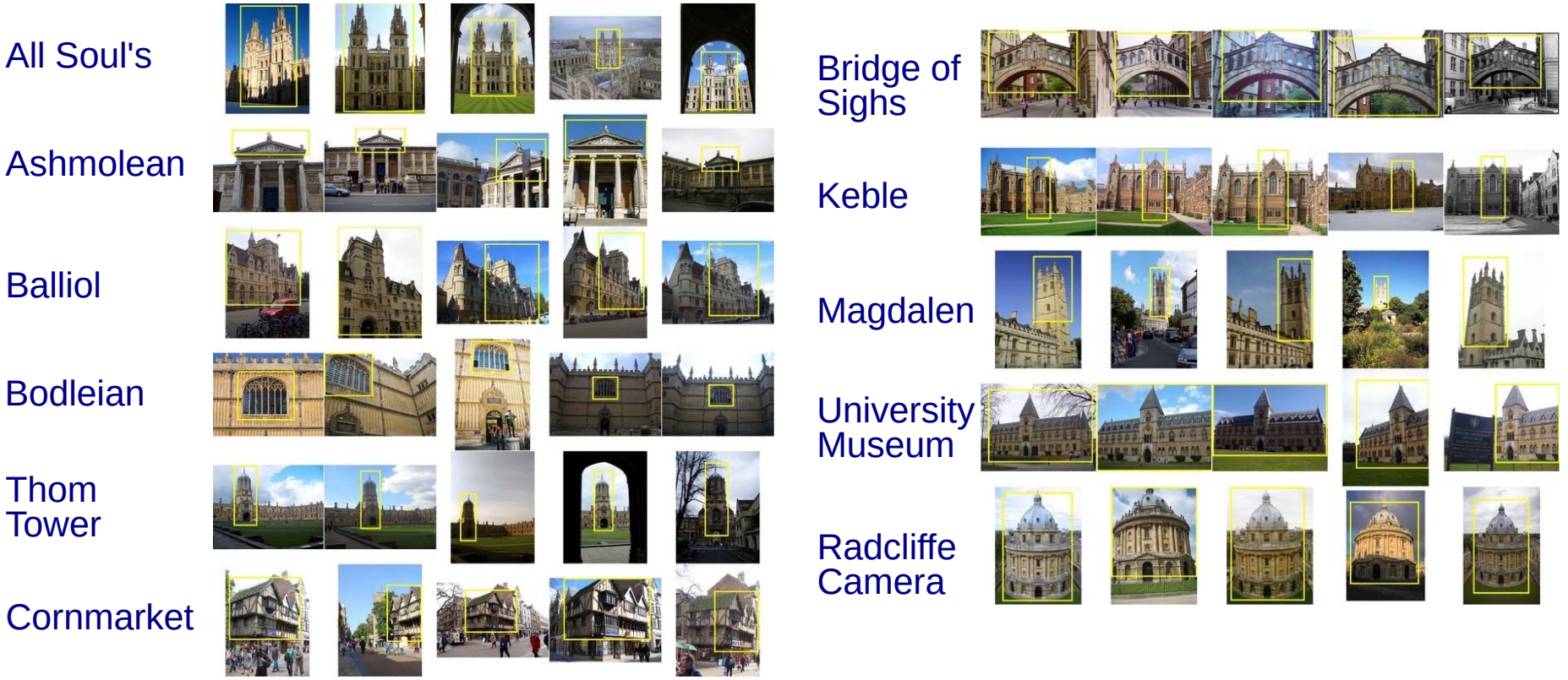
- Automatically crawled from flickr
- Consists of:

Dataset	Resolution	# images	# features	Descriptor size
i	1024×768	5,062	16,334,970	1.9 GB
ii	1024×768	99,782	277,770,833	33.1 GB
iii	500×333	1,040,801	1,186,469,709	141.4 GB
Total		1,145,645	1,480,575,512	176.4 GB

Slide credit: J. Sivic

Oxford buildings dataset

- Landmarks plus queries used for evaluation

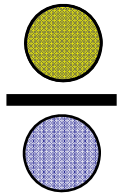


- Ground truth obtained for 11 landmarks
- Evaluate performance by mean Average Precision

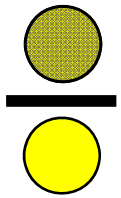
Slide credit: J. Sivic

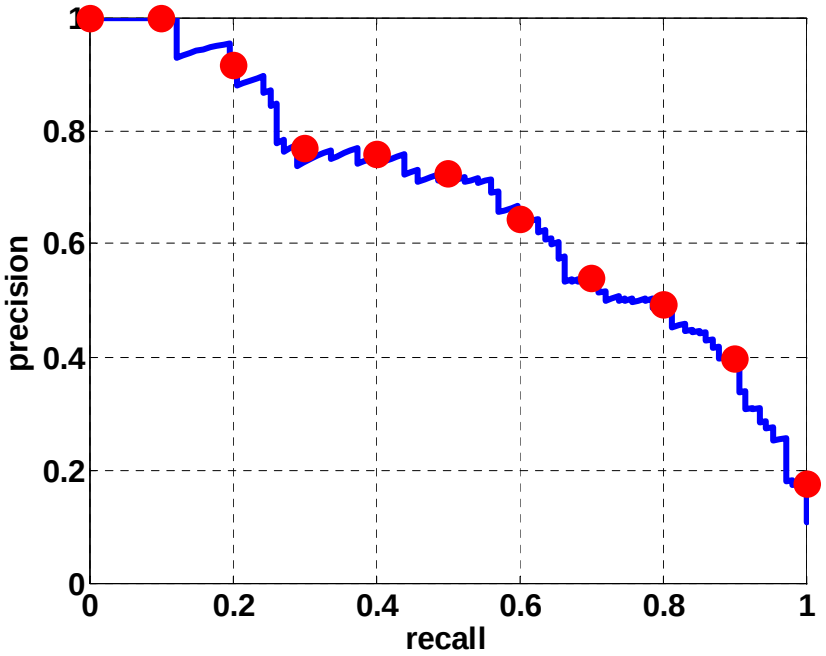
Measuring retrieval performance: Precision - Recall

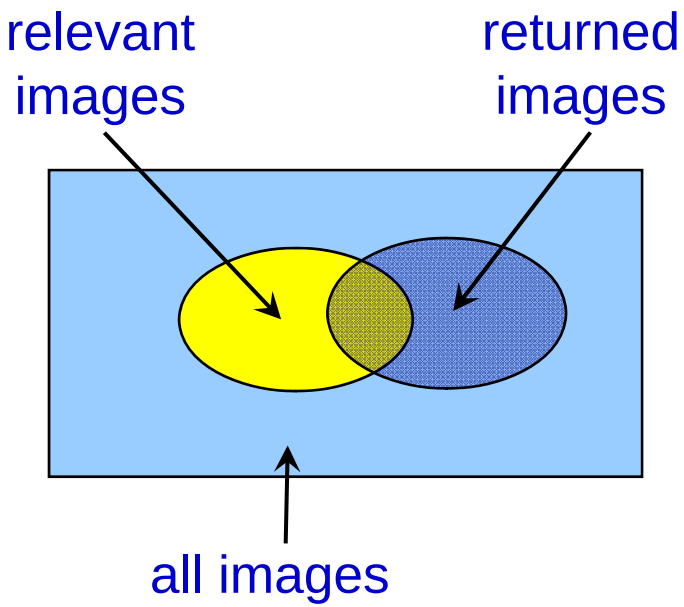
- Precision: % of returned images that are relevant



- Recall: % of relevant images that are returned

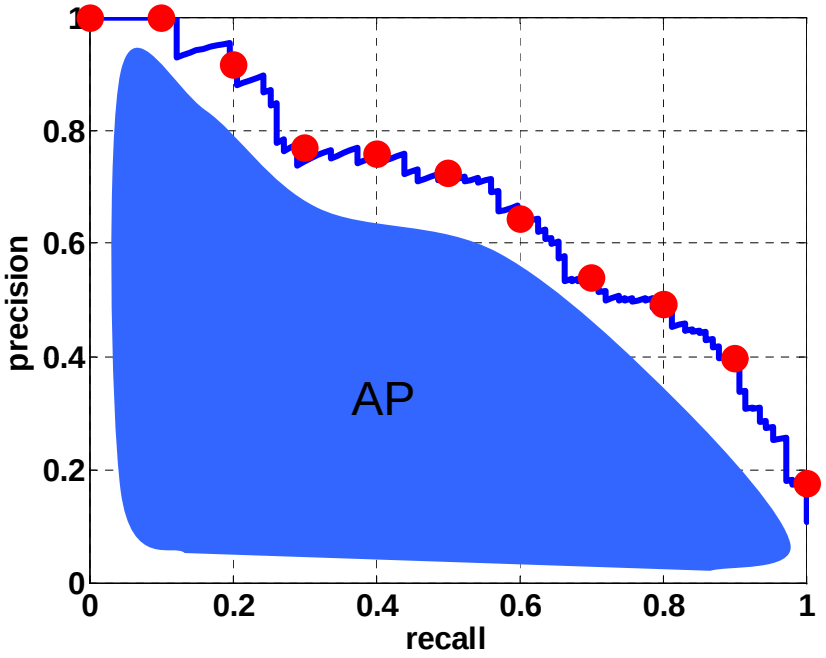






Slide credit: J. Sivic

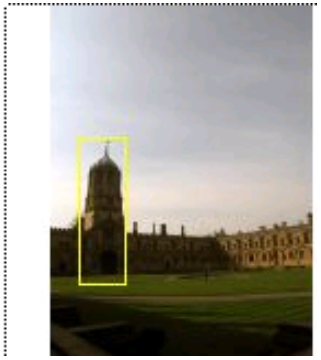
Average Precision

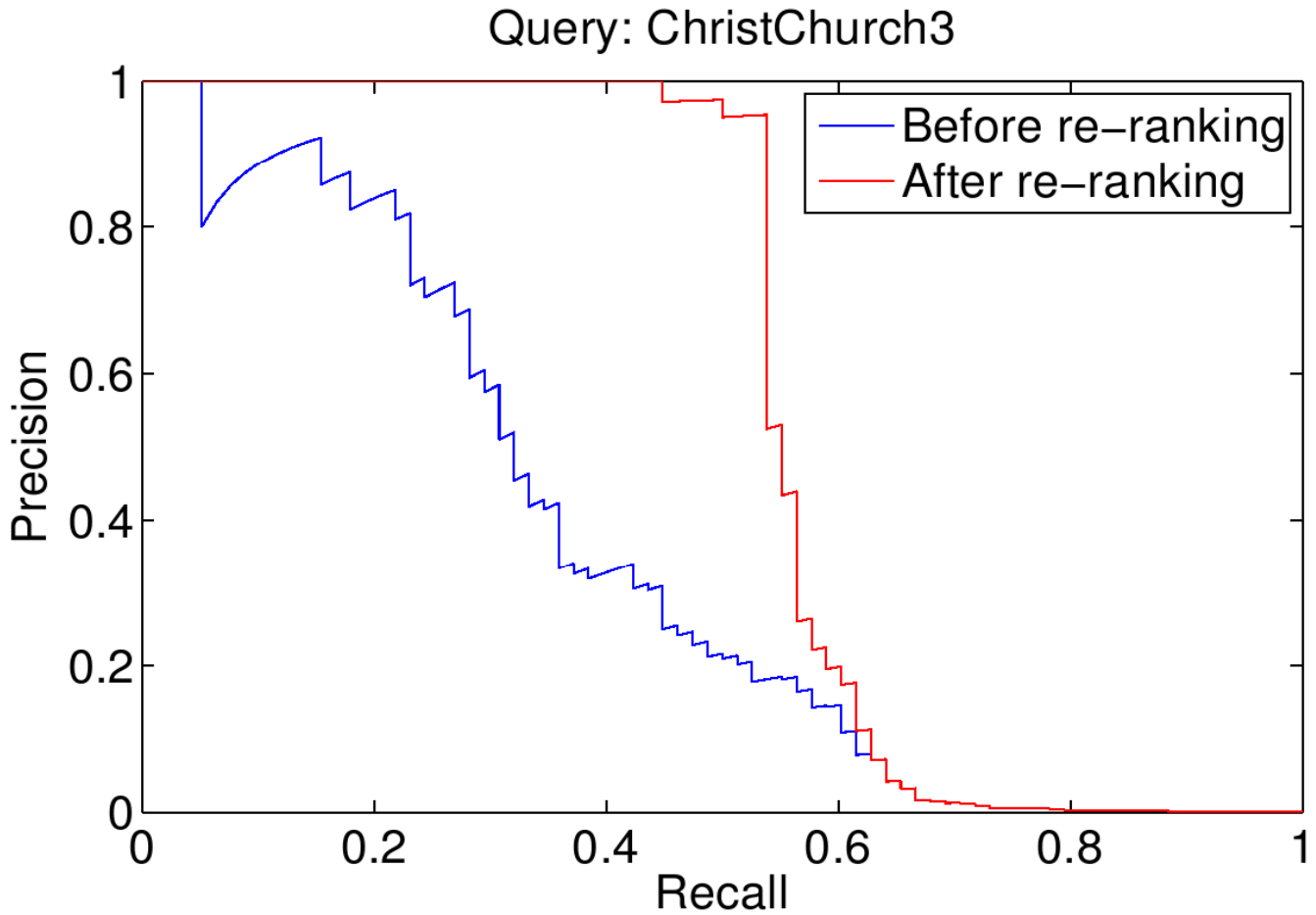


- A good AP score requires both high recall and high precision
- Application-independent

Performance measured by mean Average Precision (mAP) over 55 queries on 100K or 1.1M image datasets

Slide credit: J. Sivic

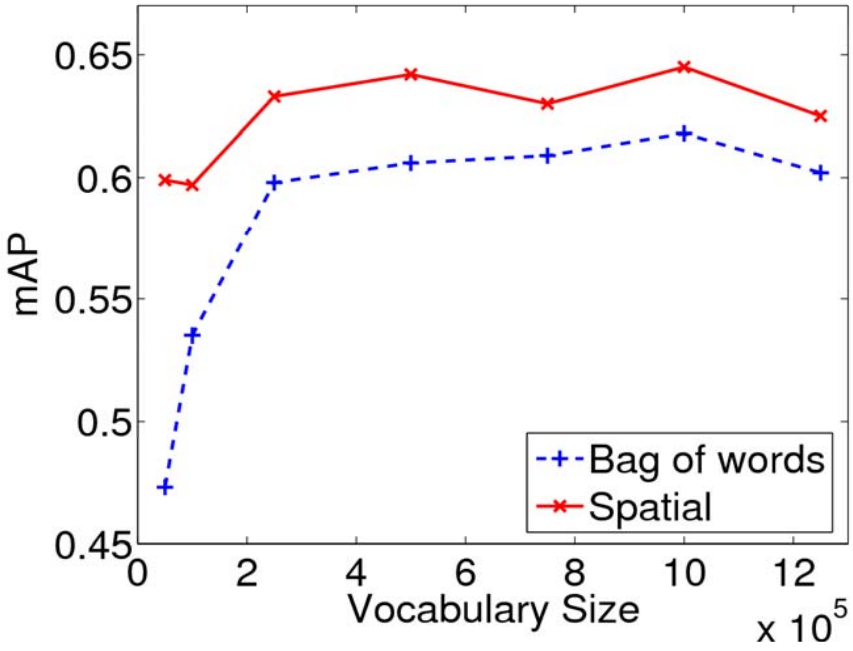




Slide credit: J. Sivic

Mean Average Precision variation with vocabulary size

vocab size	bag of words	spatial
50K	0.473	0.599
100K	0.535	0.597
250K	0.598	0.633
500K	0.606	0.642
750K	0.609	0.630
1M	0.618	0.645
1.25M	0.602	0.625



Slide credit: J. Sivic

Query Expansion in text

In text :

- Reissue top n responses as queries
- Pseudo/blind relevance feedback
- Danger of topic drift

In vision:

- Reissue spatially verified image regions as queries

Slide credit: J. Sivic

Query expansion in text - example

Original query: Hubble Telescope Achievements

Query expansion: Select top 20 terms from top 20 documents

Added terms:

telescope
hubble
space
nasa
ultraviolet
shuttle
mirror
telescopes
earth
discovery
orbit
flaw
scientists
launch
stars
universe
mirrors
light
optical
species

Example from: Jimmy Lin, University of Maryland

Slide credit: J. Sivic

Automatic query expansion

Visual word representations of two images of the same object may differ (due to e.g. detection/quantization noise) resulting in missed returns

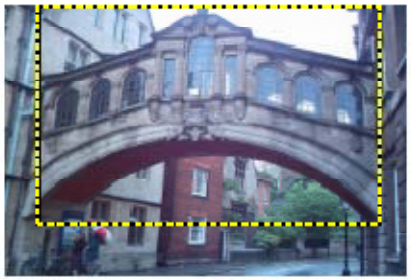
Initial returns may be used to add new relevant visual words to the query

Strong spatial model prevents ‘drift’ by discarding false positives

[Chum, Philbin, Sivic, Isard, Zisserman, ICCV’07]

Visual query expansion - overview

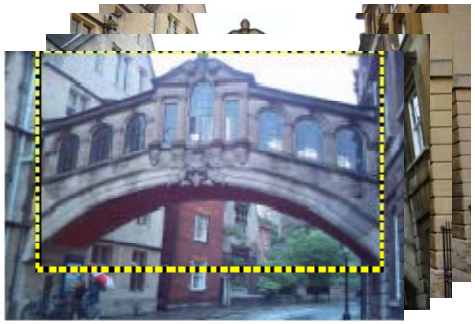
1. Original query



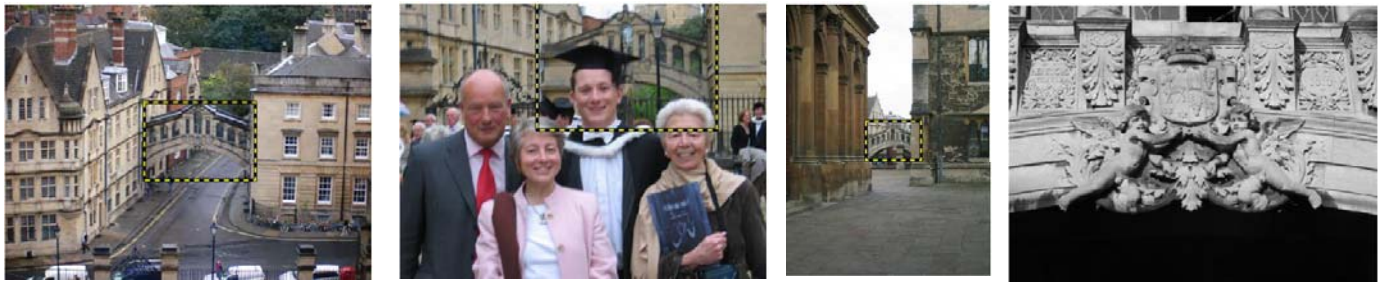
2. Initial retrieval set

3. Spatial verification

4. New enhanced query

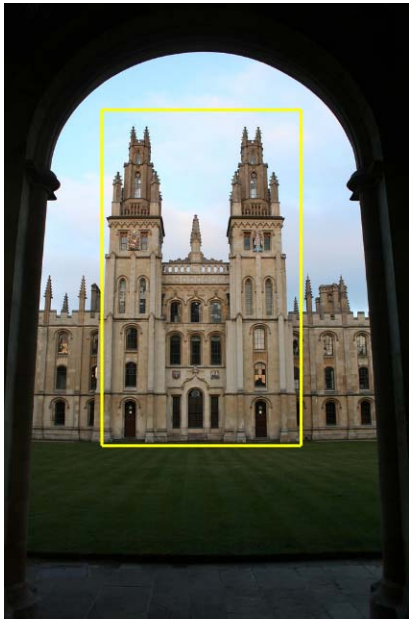


5. Additional retrieved images

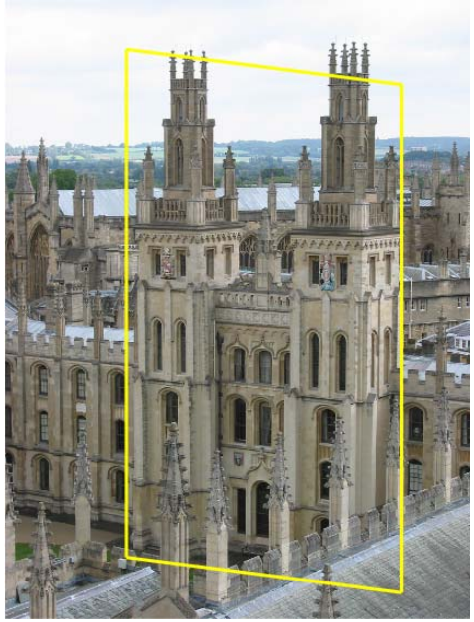


Slide credit: J. Sivic

Query Expansion



Query Image

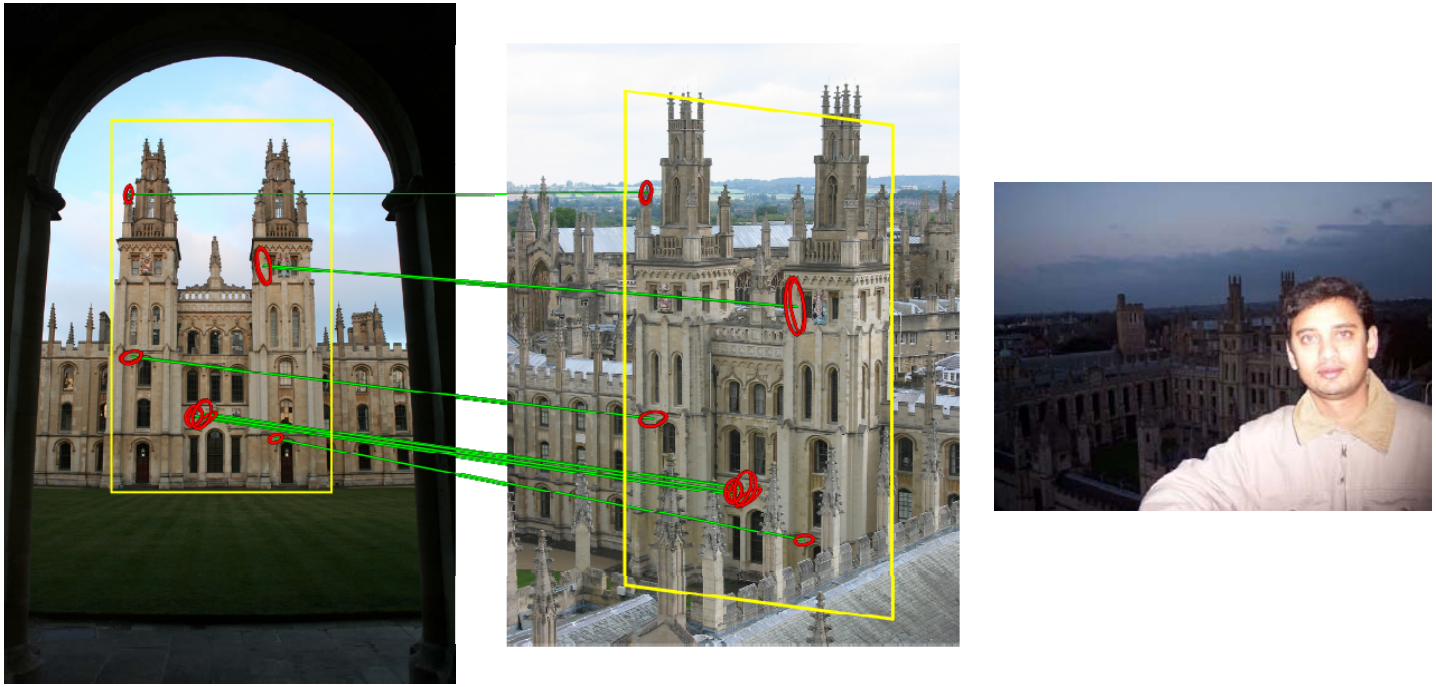


Originally retrieved image

Originally not retrieved

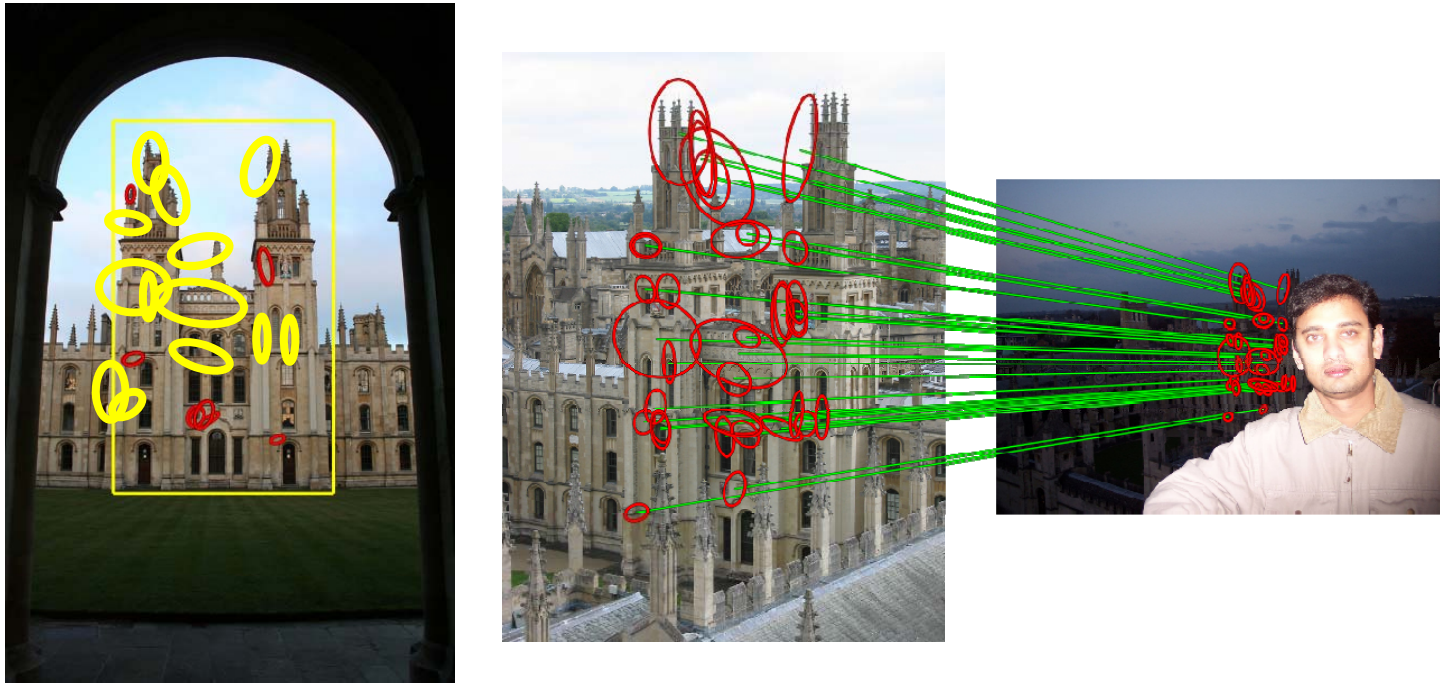
Slide credit: J. Sivic

Query Expansion



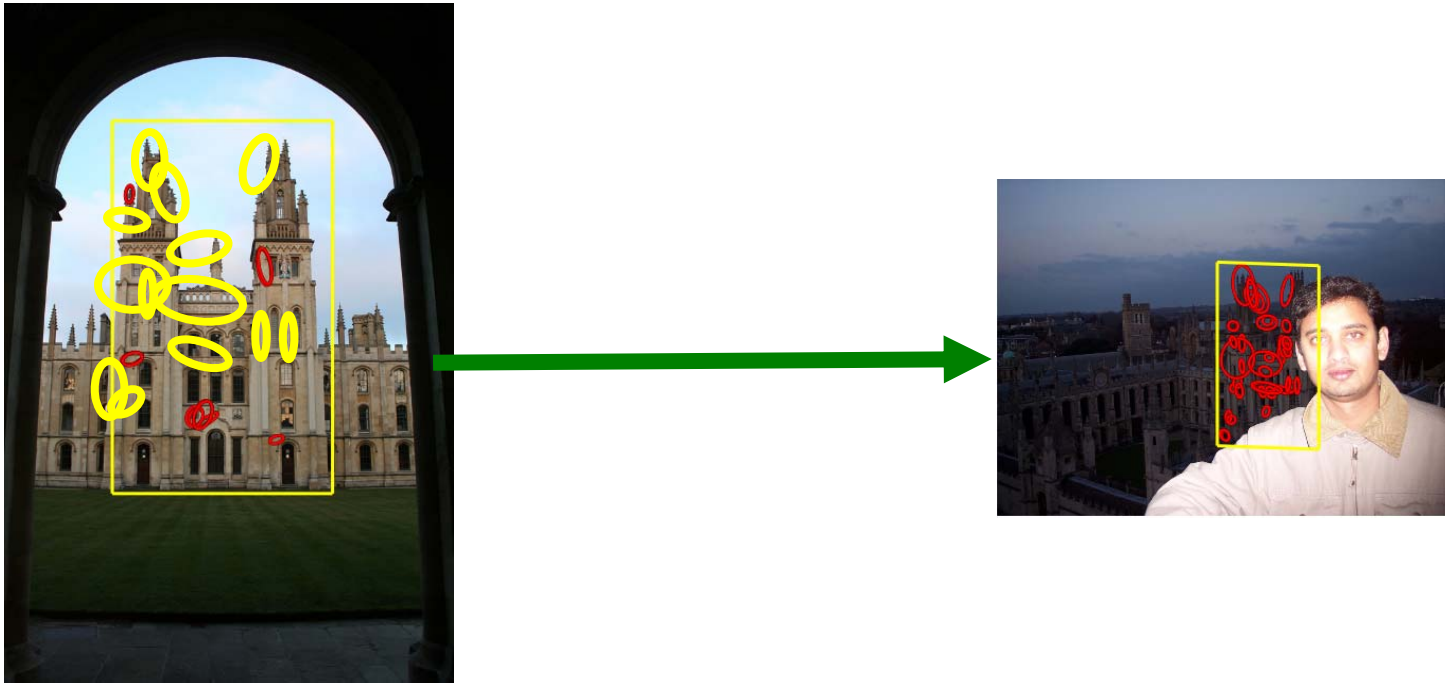
Slide credit: J. Sivic

Query Expansion



Slide credit: J. Sivic

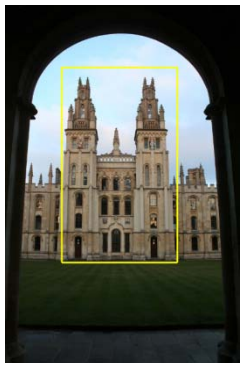
Query Expansion



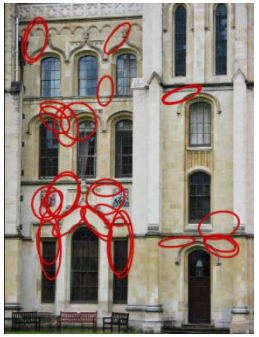
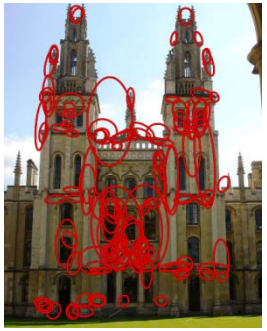
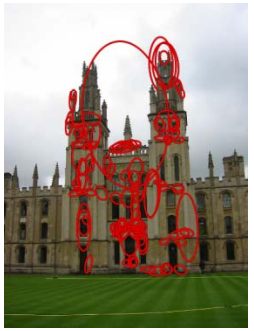
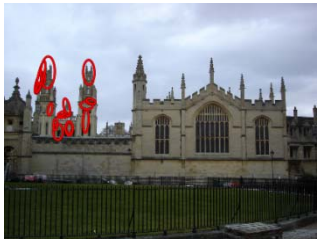
Slide credit: J. Sivic

Query Expansion

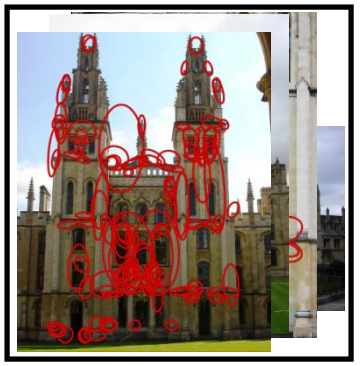
Query Image



Spatially verified retrievals with matching regions overlaid

...



New expanded query

New expanded query is formed as

- the average of visual word vectors of spatially verified returns
- only inliers are considered
- regions are back-projected to the original query image

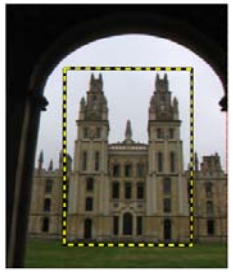
Query Expansion

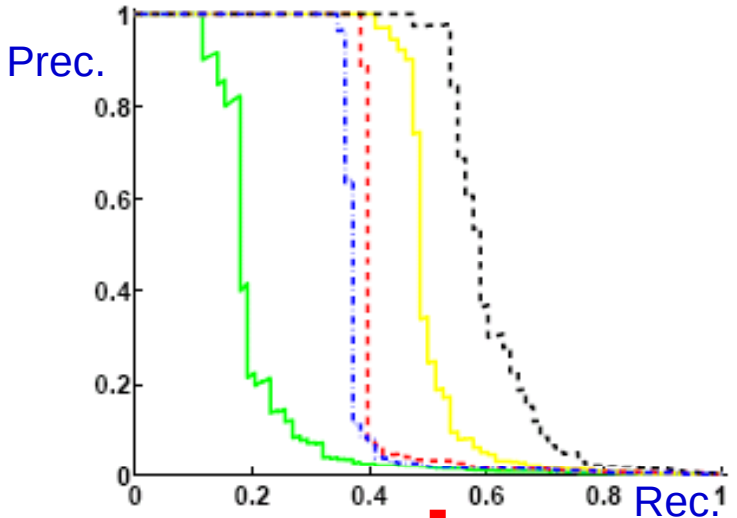
Query image | Originally retrieved | Retrieved only after expansion

Slide credit: J. Sivic

Query image

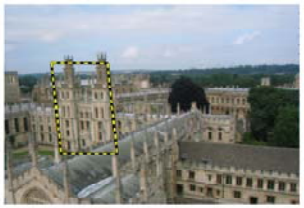
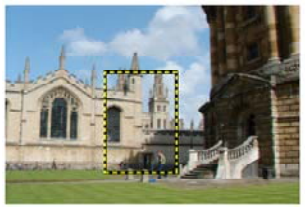
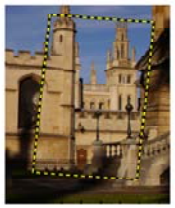
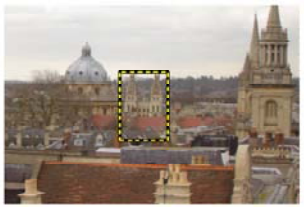
Original results (good)

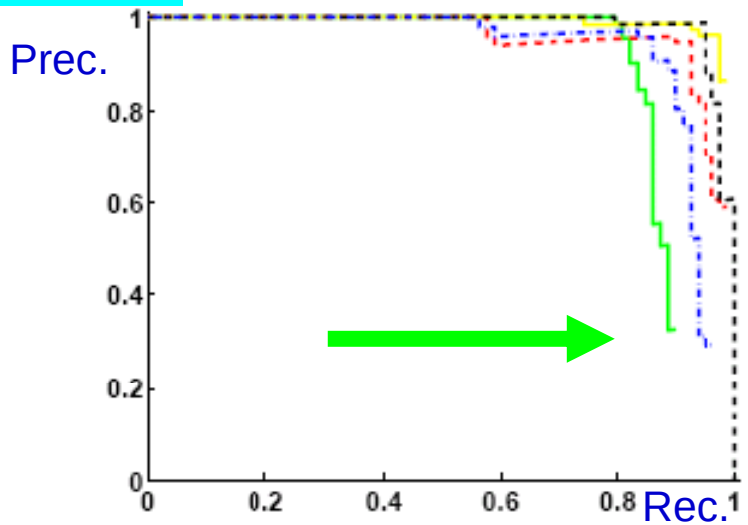





Expanded results (better)





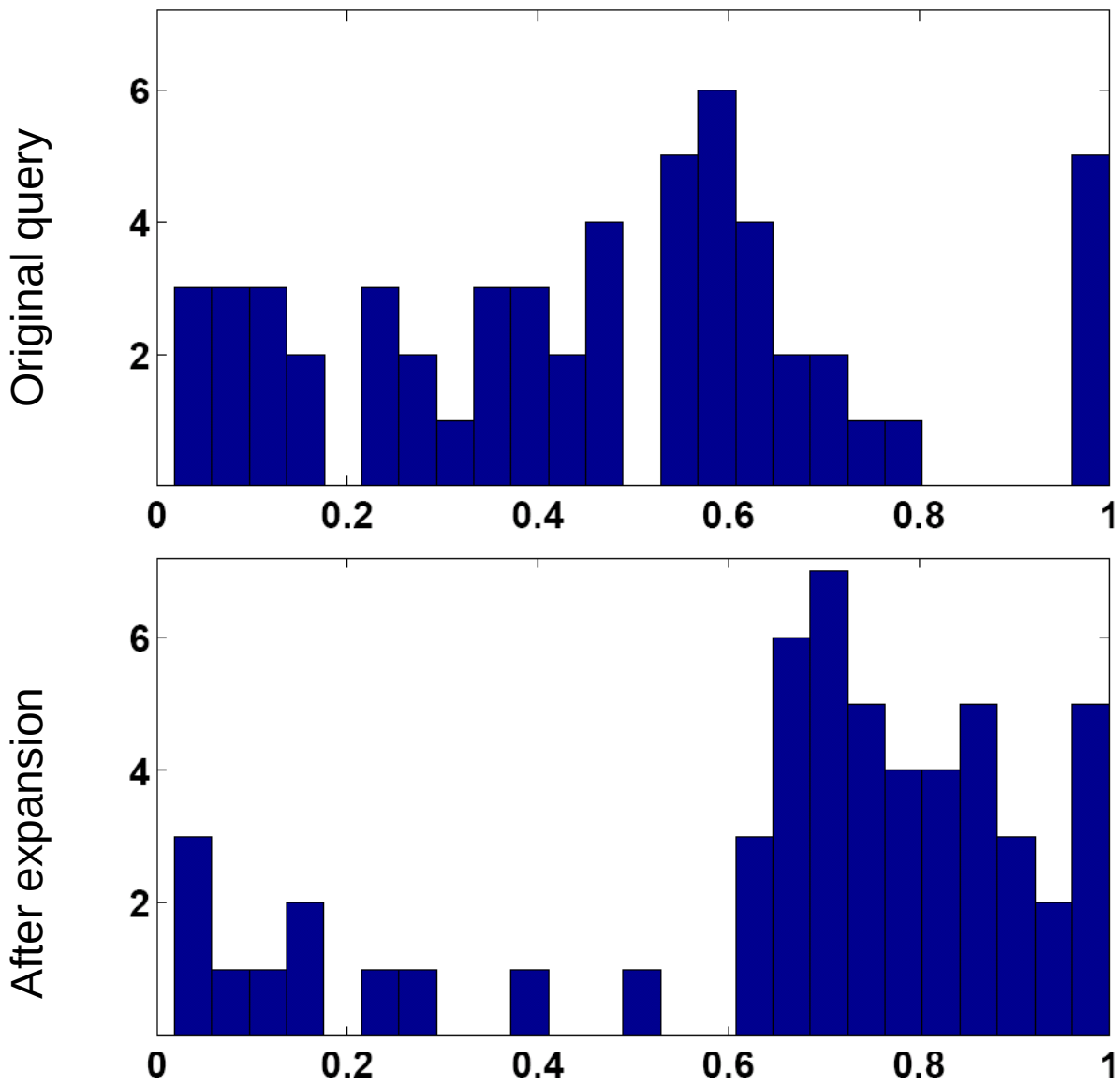
Slide credit: J. Sivic

Search in 100K images

Search in 1M images

	Ground truth		*Oxford* + *Flickr1* dataset						*Oxford* + *Flickr1* + *Flickr2* dataset					
	OK	Junk	ori	qeb	trc	avg	rec	sca	ori	qeb	trc	avg	rec	sca
All Souls	78	111	41.9	49.7	85.0	76.1	85.9	**94.1**	32.8	36.9	80.5	66.3	73.9	**84.9**
Ashmolean	25	31	53.8	35.4	51.4	66.4	74.6	**75.7**	41.8	25.9	45.4	57.6	**68.2**	65.5
Balliol	12	18	50.4	52.4	44.2	63.9	**74.5**	71.2	40.1	39.4	39.6	55.5	**67.6**	60.0
Bodleian	24	30	42.3	47.4	49.3	**57.6**	48.6	53.3	32.3	36.9	43.5	**46.8**	43.8	44.9
Christ Church	78	133	53.7	36.3	56.2	63.1	**63.3**	63.1	52.6	18.9	55.2	**61.0**	57.4	57.7
Cornmarket	9	13	54.1	60.4	58.2	74.7	74.9	**83.1**	42.2	53.4	56.0	65.2	68.1	**74.9**
Hertford	24	31	69.8	74.4	77.4	89.9	90.3	**97.9**	64.7	70.7	75.8	87.7	87.7	**94.9**
Keble	7	11	79.3	59.6	64.1	90.2	**100**	97.2	55.0	15.6	57.3	**67.4**	65.8	65.0
Magdalen	54	103	9.5	6.9	25.2	28.3	**41.5**	33.2	5.4	0.2	16.9	15.7	**31.3**	26.1
Pitt Rivers	7	9	**100**	**100**	**100**	**100**	**100**	**100**	**100**	90.2	**100**	**100**	**100**	**100**
Radcliffe Cam.	221	348	50.5	59.7	88.0	71.3	73.4	**91.9**	44.2	56.8	86.8	70.5	72.5	**91.3**
Total	539	838	55.0	52.9	63.5	71.1	75.2	**78.2**	46.5	40.5	59.7	63.1	67.0	**69.6**

Slide credit: J. Sivic



Average Precision histogram for 55 queries

Slide credit: J. Sivic

Other applications of local invariant features

Sony Aibo (Evolution Robotics)

SIFT usage

- Recognize docking station
- Communicate with visual cards

Other uses

- Place recognition
- Loop closure in SLAM

Slide credit: David Lowe

Example Applications

Mobile tourist guide

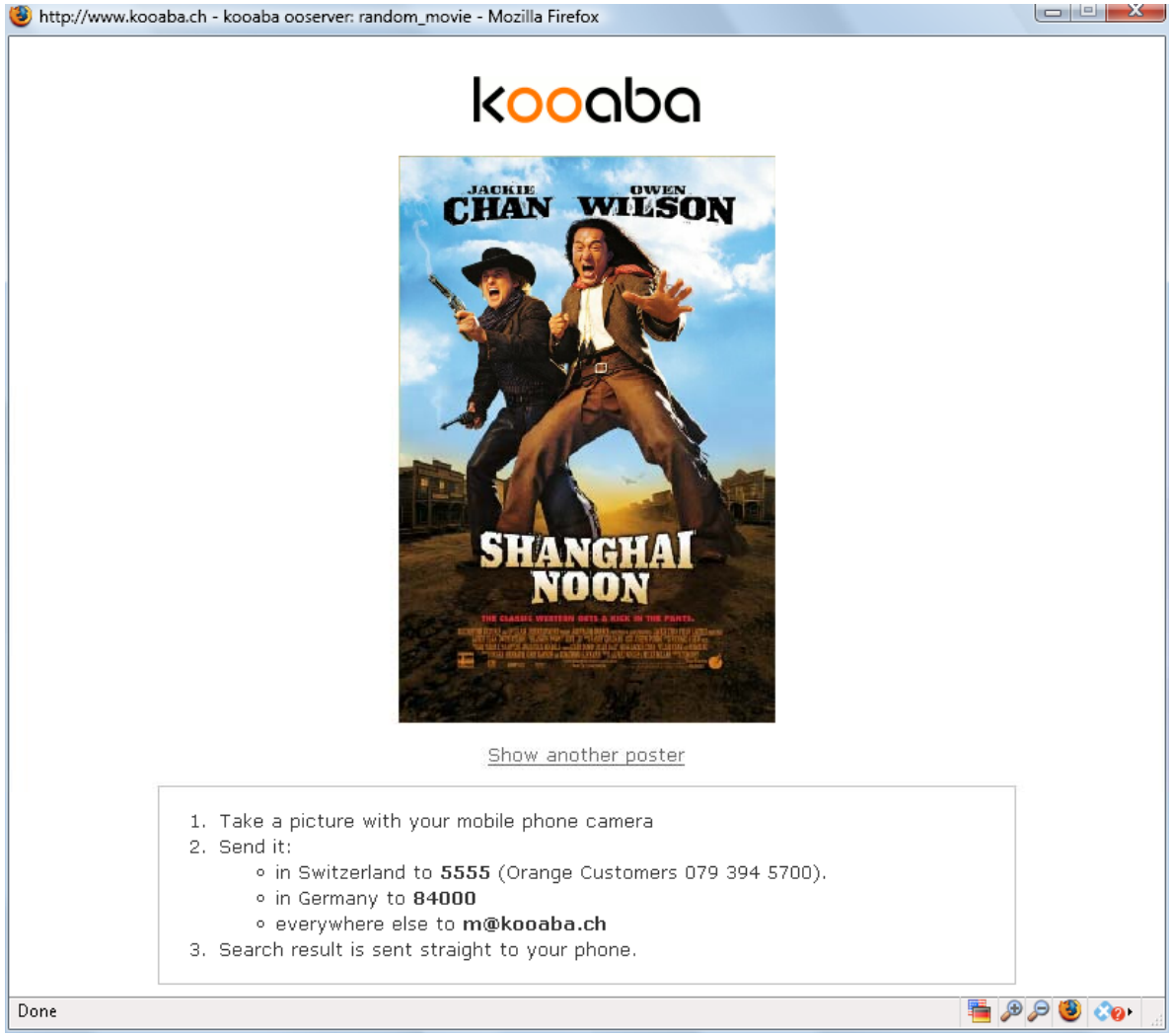
- **Self-localization**
- **Object/building recognition**
- **Photo/video augmentation**

[Quack, Leibe, Van Gool, CIVR'08]

Web Demo: Movie Poster Recognition

50’000 movie posters indexed

Query-by-image from mobile phone available in Switzer-land



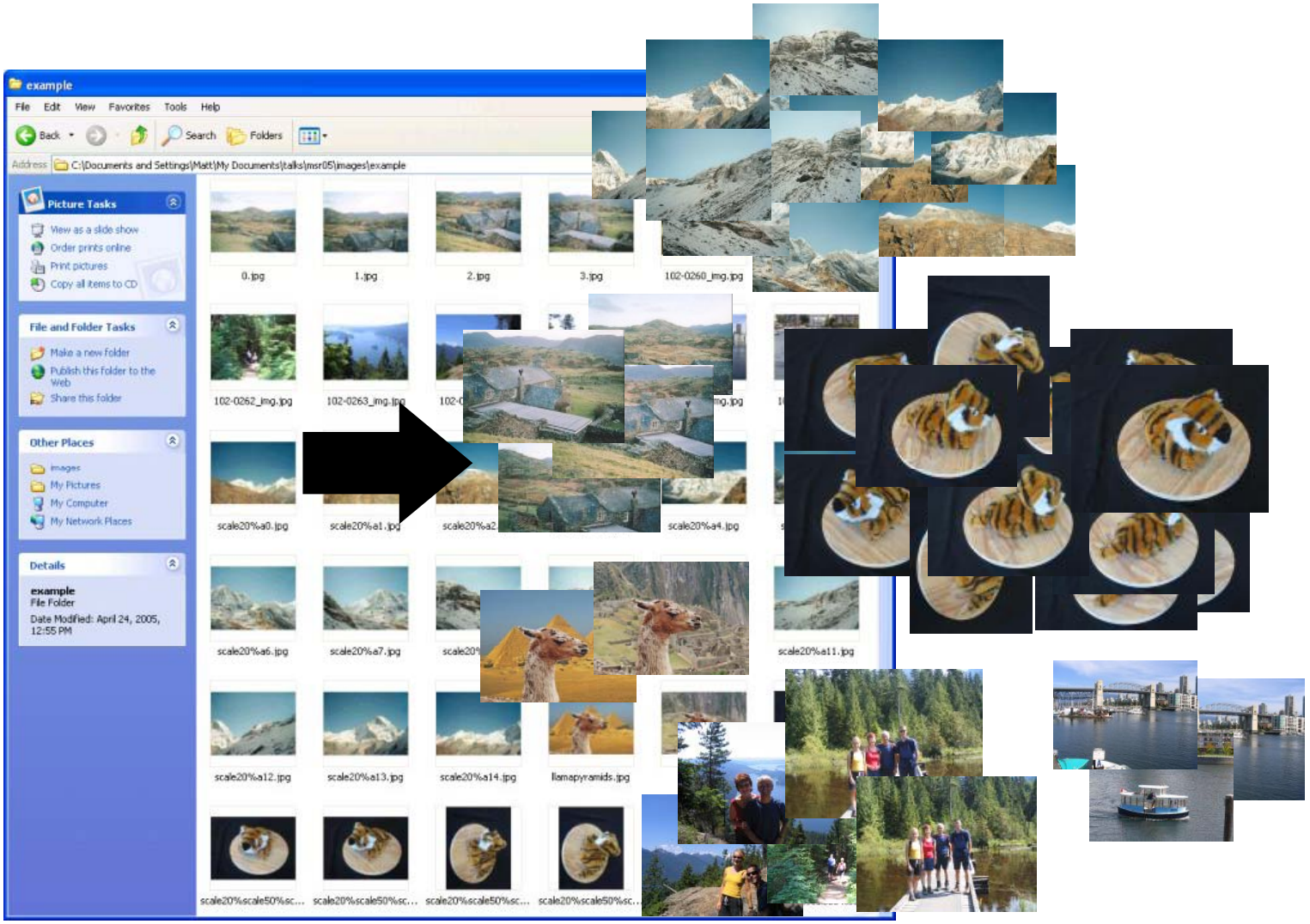
http://www.kooaba.com/en/products_engine.html#

Application: Image Auto-Annotation

Left: Wikipedia image
Right: closest match from Flickr

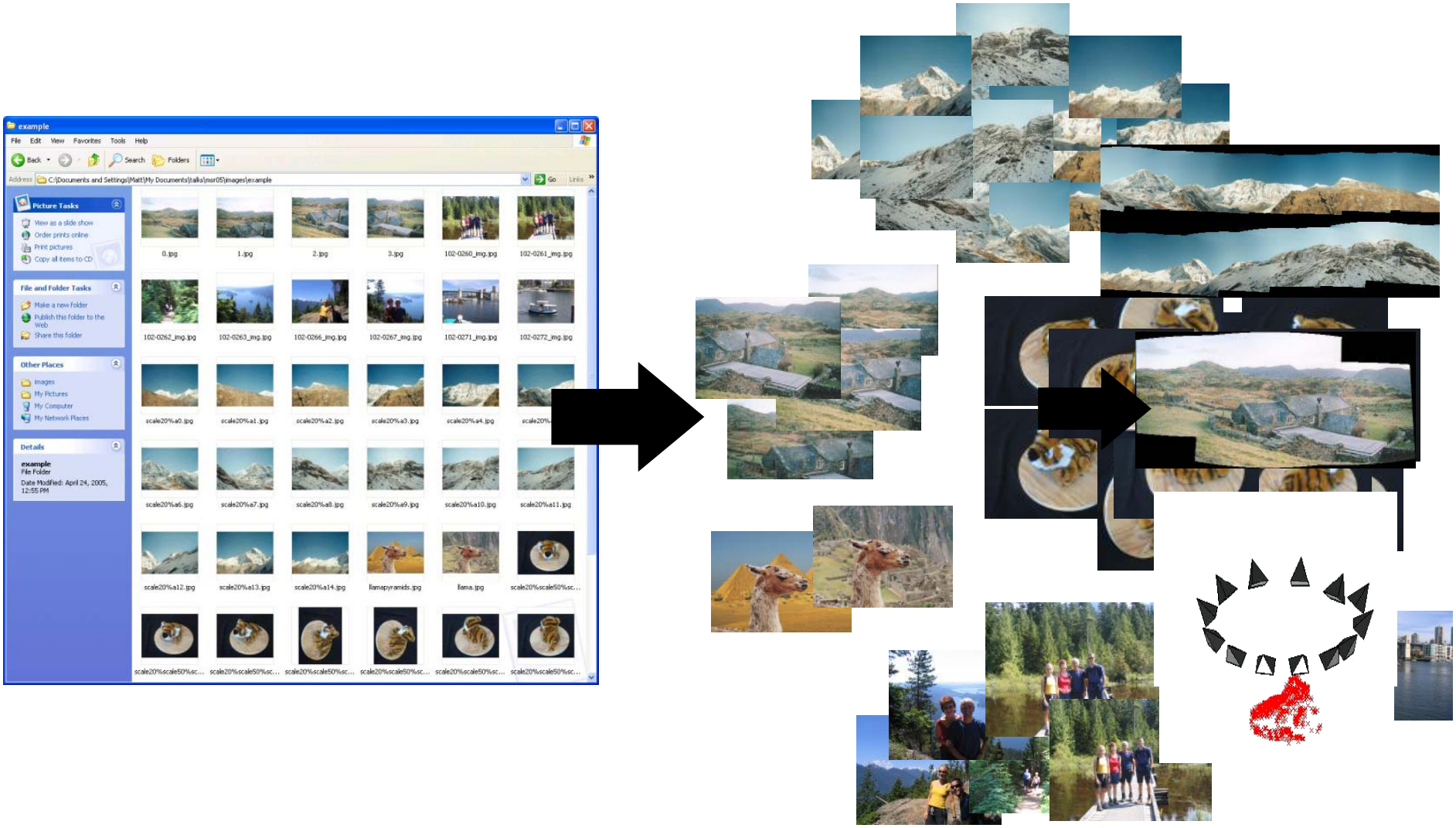
[Quack CIVR'08]

Matching in large unordered datasets



Slide credit: J. Sivic

Matching in large unordered datasets



Slide credit: J. Sivic

Photo Tourism:
Exploring Photo Collections in 3D

Noah Snavely

Steven M. Seitz

University of Washington

Richard Szeliski

Microsoft Research

Thanks to the authors for making slides available!

Photo Tourism

Photo Tourism overview

Input photographs

Scene reconstruction

Relative camera positions and orientations

Point cloud

Sparse correspondence

System for interactive browsing and exploring large collections of photos of a scene. Computes viewpoint of each photo as well as a sparse 3d model of the scene.

Input Photos

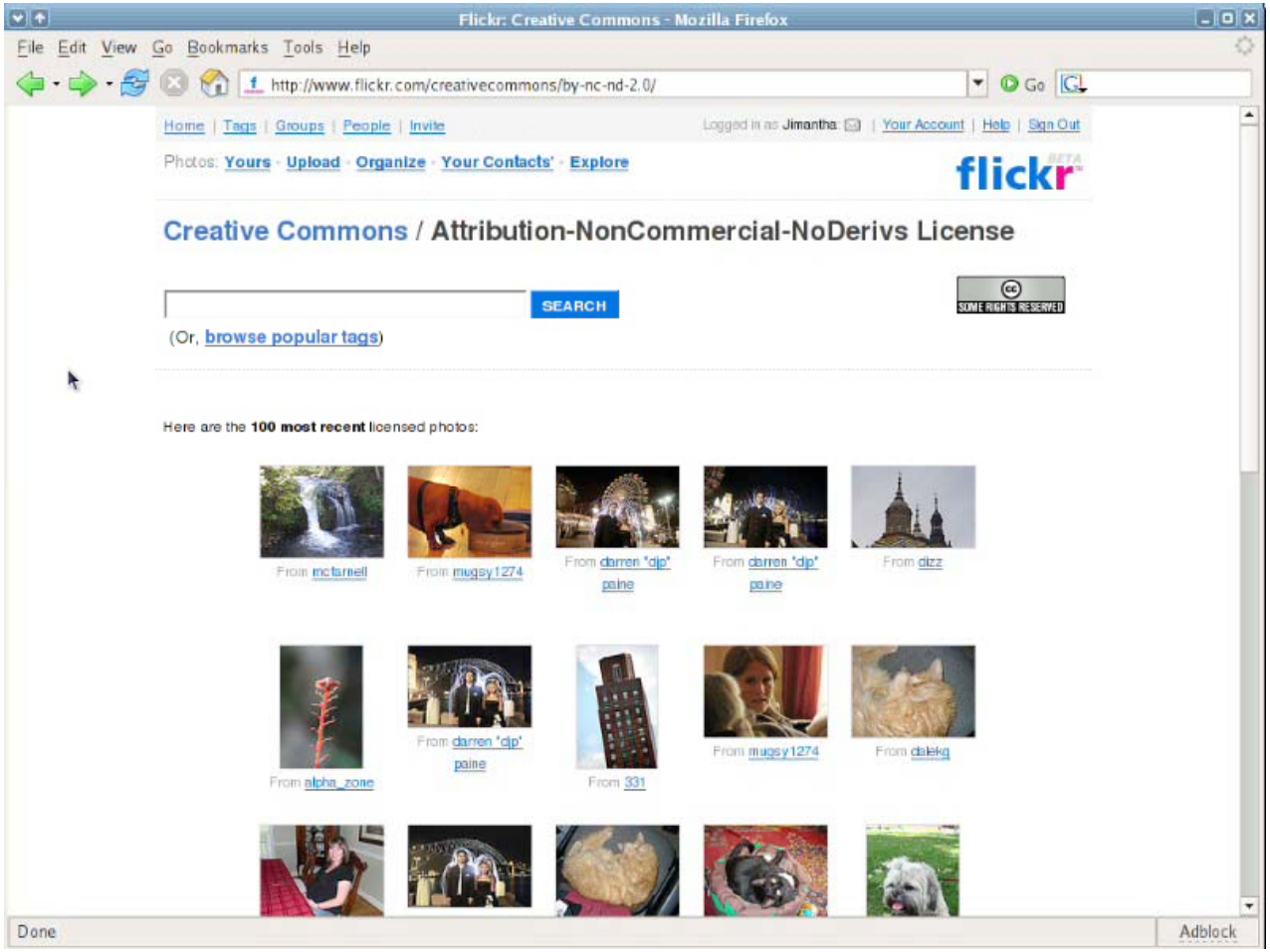


Photo Tourism overview

Input photographs

Scene reconstruction

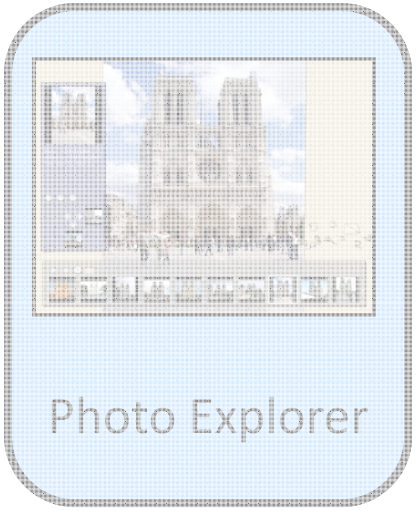
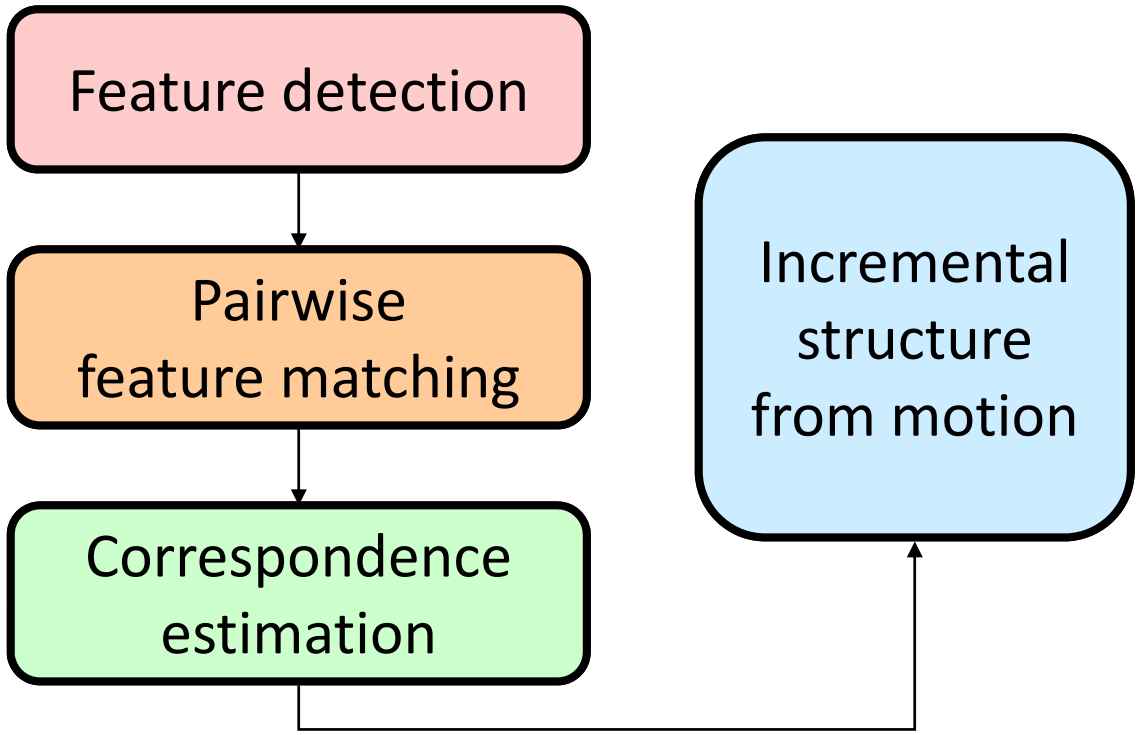


Photo Explorer

Scene reconstruction

- Automatically estimate
  - position, orientation, and focal length of cameras
  - 3D positions of feature points



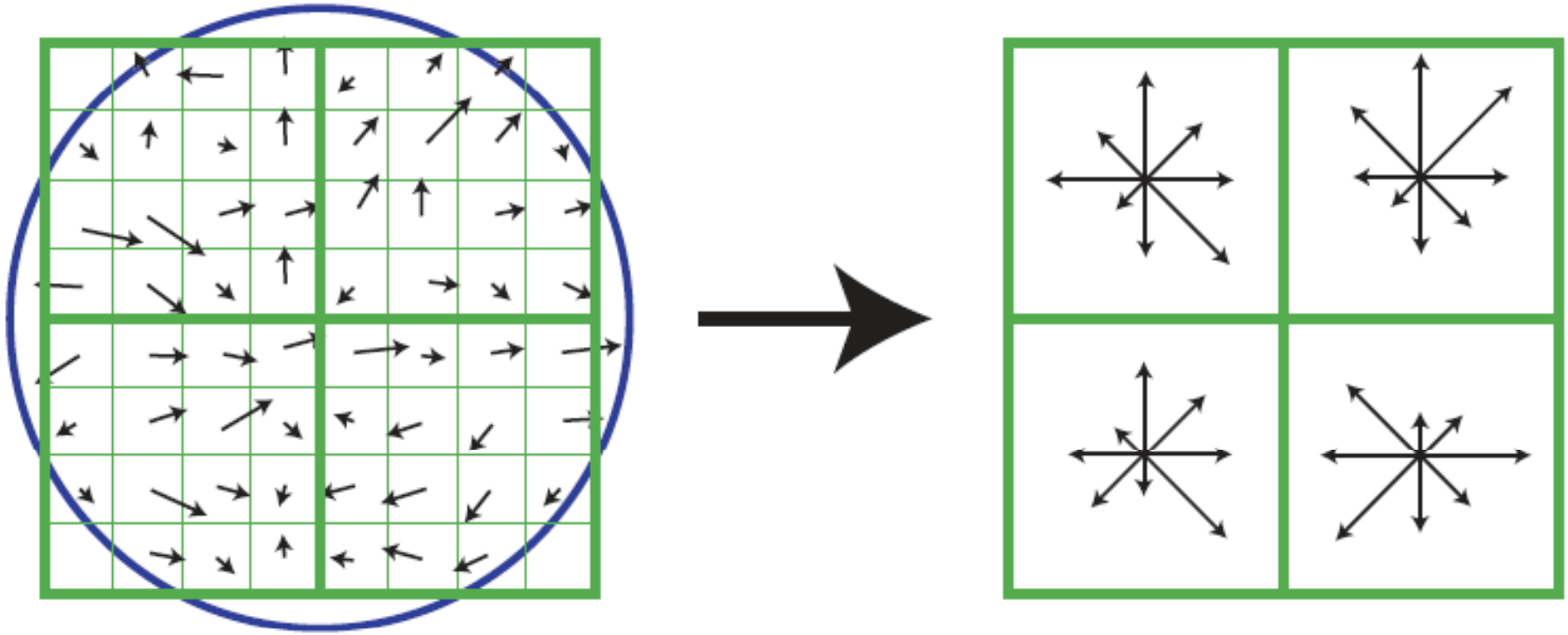
Feature detection

- Detect features using SIFT [Lowe, IJCV 2004]

Feature detection

- Detect features using SIFT [Lowe, IJCV 2004]

SIFT Reminder



Pairwise feature matching

- Match features between each pair of images

Pairwise feature matching

- Refine matching using RANSAC [Fischler & Bolles 1987] to estimate fundamental matrices between pairs

Fundamental matrix –

F is a 3x3 matrix with rank 2 such that for:

Corresponding points in stereo pair $\mathbf{y_1}$ and $\mathbf{y_2}$

$$\mathbf{y_2^T F y_1} = 0.$$

Correspondence estimation

- Link up pairwise matches to form connected components of matches across several images

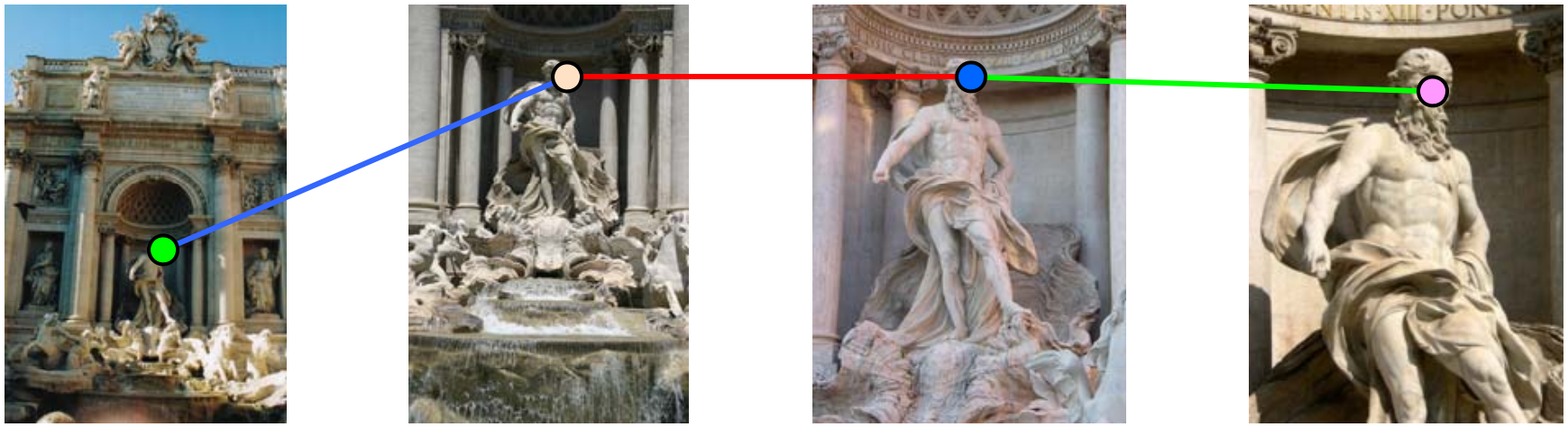
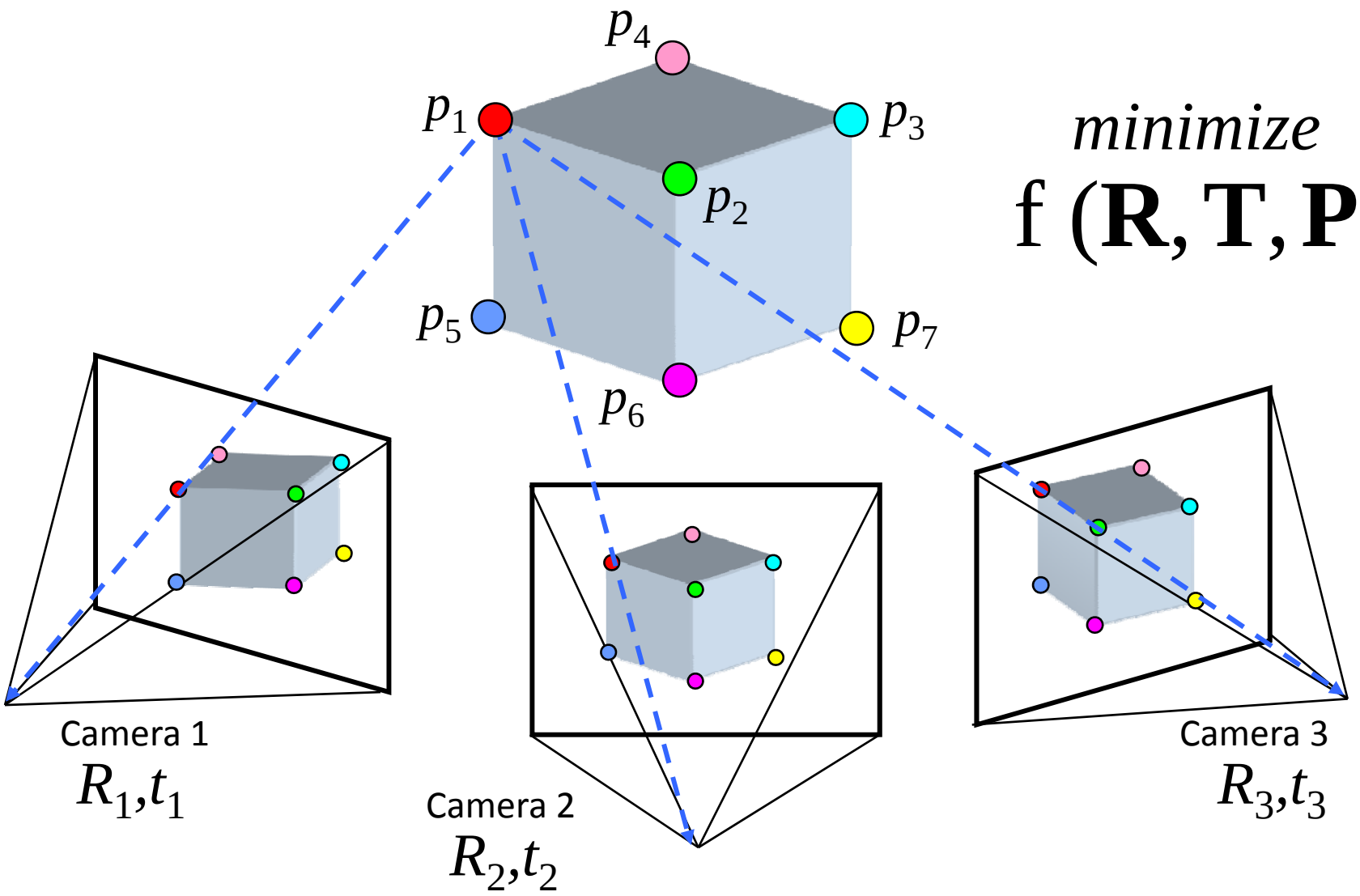


Image 1 Image 2 Image 3 Image 4

Structure from motion



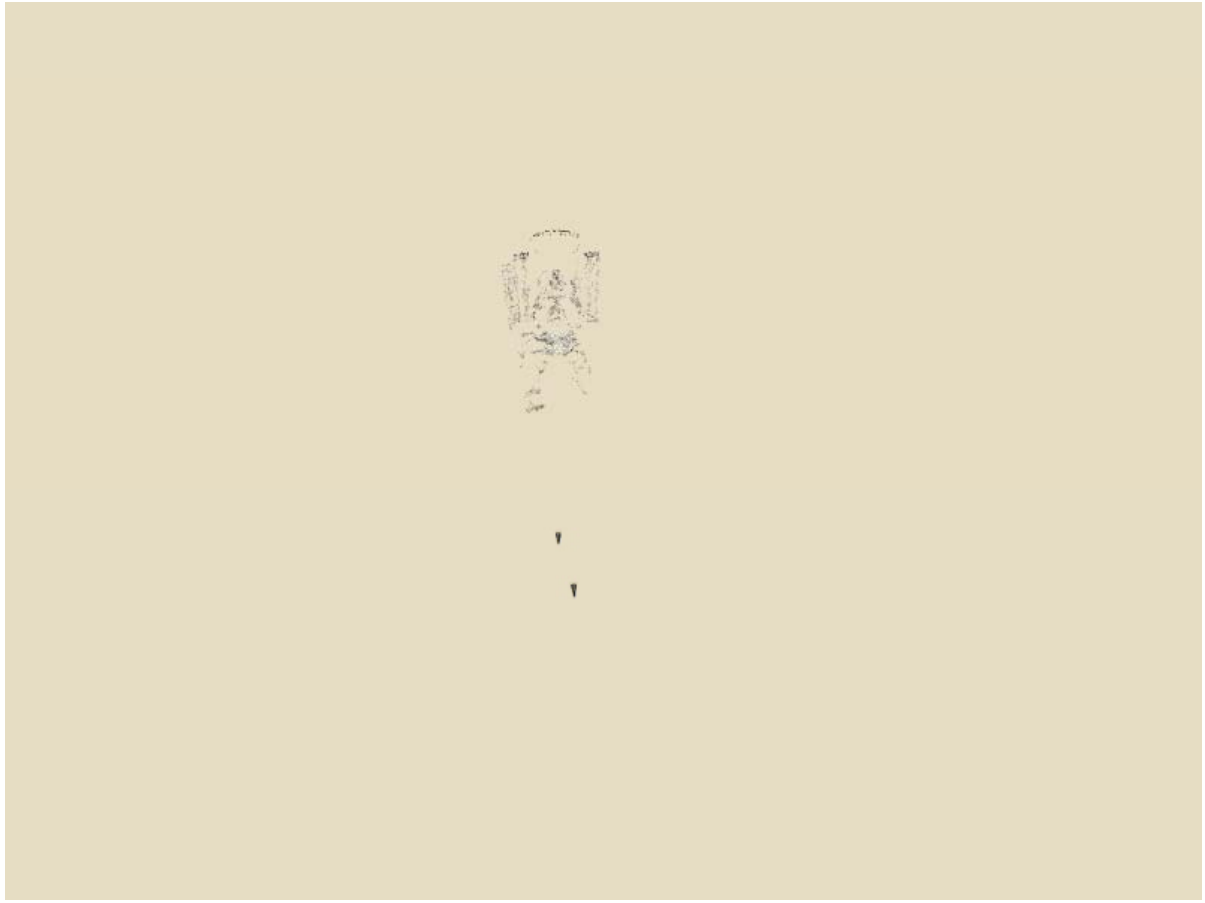
minimize

$$\mathrm{f}\,(\mathbf{R},\mathbf{T},\mathbf{P})$$

Incremental structure from motion

- Optimize parameters for two cameras and common points
- Find new image with most matches to existing points
- Initialize new camera using pose estimation
- Bundle adjust
- Add new points
- Bundle adjust

Incremental structure from motion



Reconstruction performance

- For photo sets from the Internet, 20% to 75% of the photos were registered
- Most unregistered photos belonged to different connected components

- Running time: < 1 hour for 80 photos

  > 1 week for 2600 photo

Photo Tourism overview

Input photographs

Scene reconstruction

Navigation controls

- Free-flight navigation
- Object-based browsing
- Relation-based browsing
- Overhead map

Free Flight Navigation

Object-based browsing

Relation-based browsing

Find all similar images

Find all details

Find all zoom outs

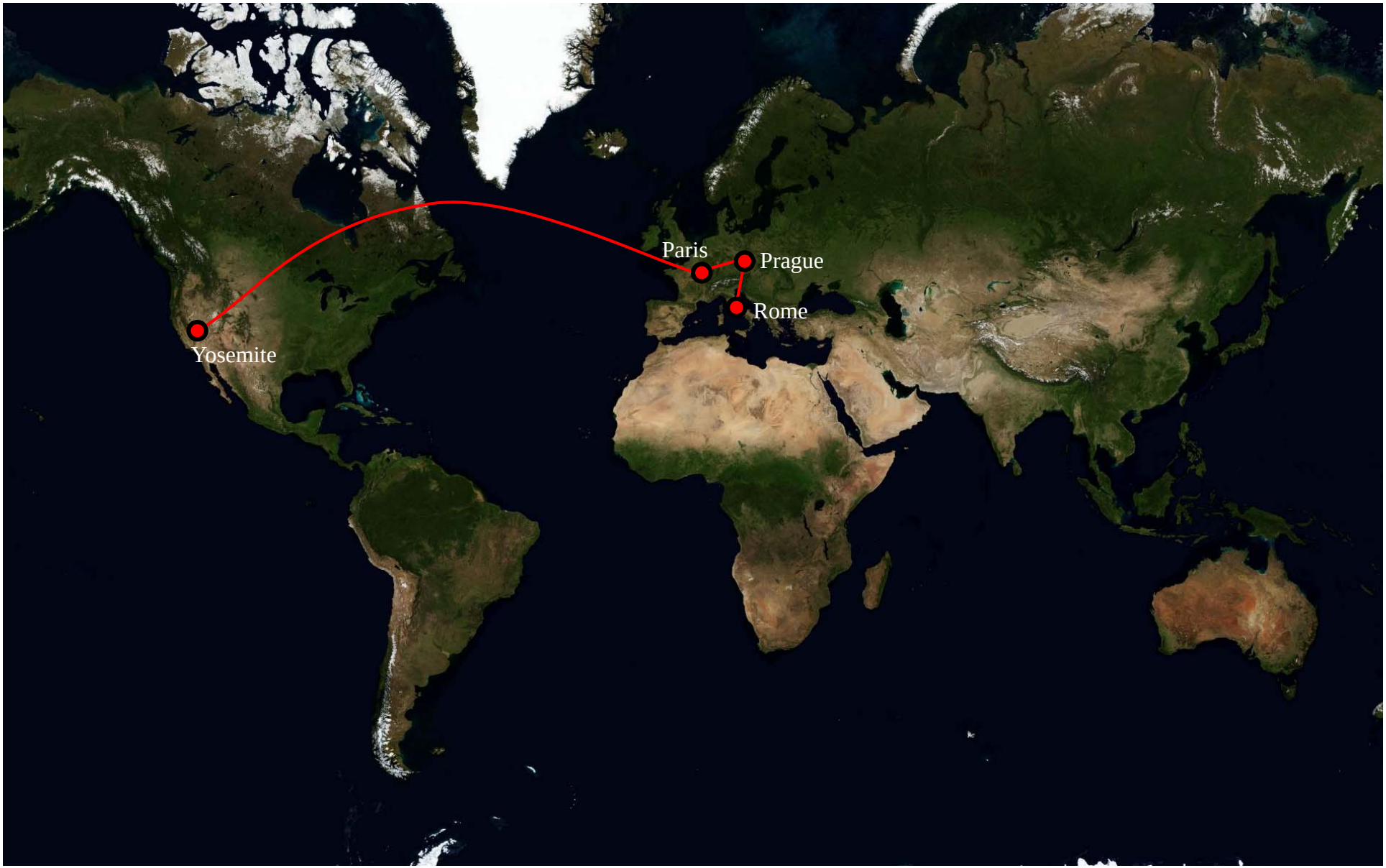
Zoom in

Move left

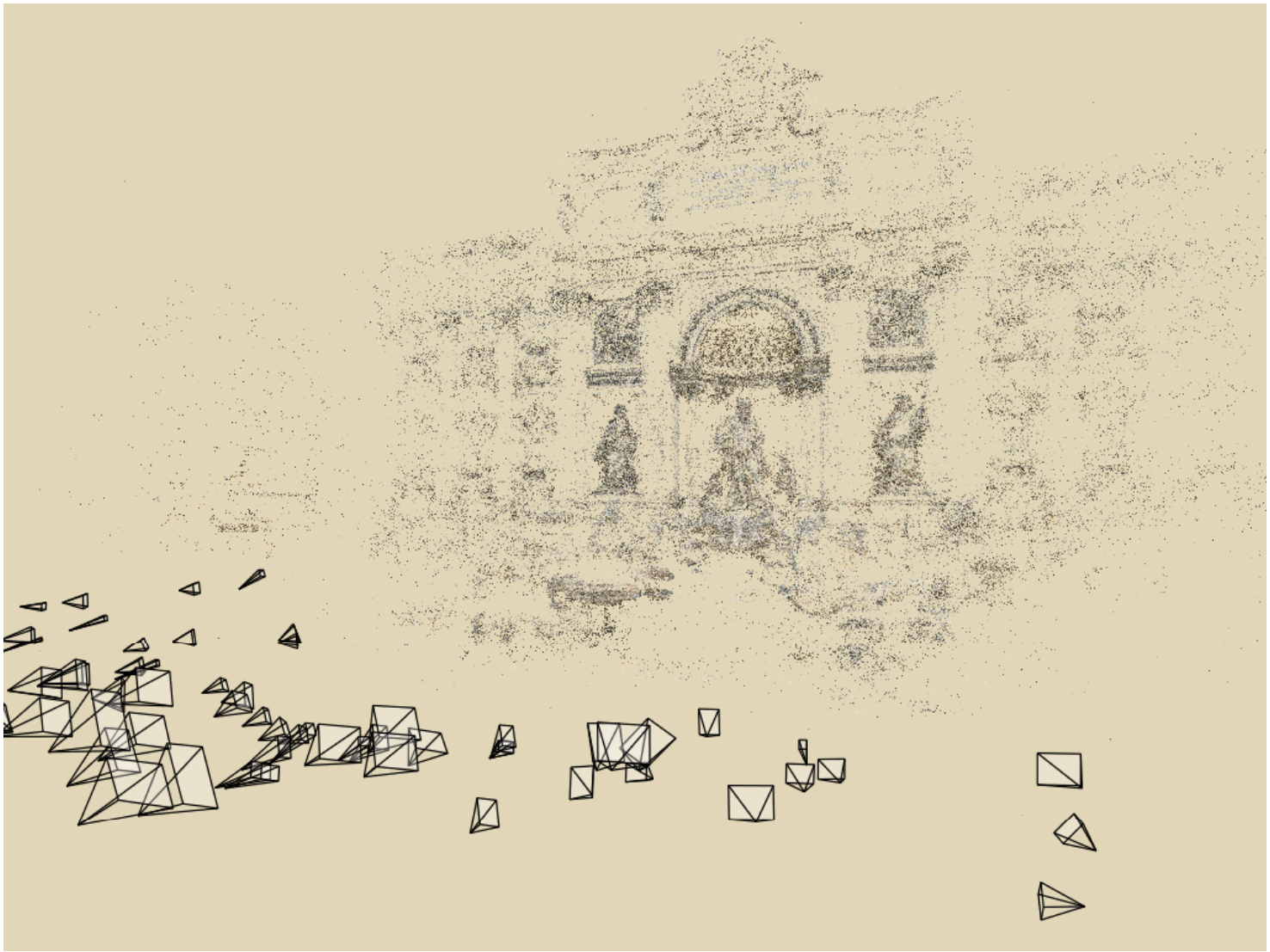
Move right

Zoom out

Overhead Map



Rendering



Rendering

Annotations

Saint Basil's Cathedral

Trafalgar Square

Rockefellar Center

Mount Rushmore

Great Wall Fly Through

Statistics

Dataset	# input	# registered
Trevi Fountain	466	360
Yosemite	325	1,893
Notre Dame	597	2,635
Prague	197	235
Great Wall	82	120
Trafalgar Square	1,893	278

Reconstruction running time

- **Great Wall**: 82 / 120 photos registered

  Running time: ~ 3 hours

- **Notre Dame**: 597 / 2,635 photos registered

  Running time: ~ 2 weeks

Advantages of 3D over 2D

- 3D geometry has multi-image consistency
- Can annotate point cloud directly
- Can import annotations from georeferenced sources (e.g., landmark databases)
- Can use depth as cue for rejecting outliers in selection

Contributions

- Automated system for registering photo collections in 3D for interactive exploration
- Structure from motion algorithm demonstrated on hundreds of photos from the Internet
- Photo exploration system combining new image-based rendering and photo navigation techniques

Limitations / Future work

- Not all photos can be reliably matched

→ Integrating GPS & other localization info.

- Structure from motion scalability

→ More efficient (sparse) algorithms

- Plane-based transitions lack parallax

Subsequent work

- Photo explorer scalability
  - Design client-server architecture for streaming images and geometry at required resolution
  - Scale to *all* of the world's photos (and videos…)
  - Photosynth project at Microsoft Live Labs (live demo)

Today

- Lowe
- Video Google
- Total Recall
- Photo Tourism

Feb 3rd – Global features (HoG, Gist, Motion History, etc.)

A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International Journal of Computer Vision, vol. 42, no. 3, pp. 145-175, May 2001. Available: http://dx.doi.org/10.1023/A:1011139631724

A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," ICCV 2003, pp. 726-733 vol.2. Available: http://dx.doi.org/10.1109/ICCV.2003.1238420

N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 886-893. Available: http://dx.doi.org/10.1109/CVPR.2005.177

A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 984-989. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1467373

Optional Readings:

B. Schiele and J. L. Crowley, "Object recognition using multidimensional receptive field histograms," in ECCV '96: Proceedings of the 4th European Conference on Computer Vision-Volume I. London, UK: Springer-Verlag, 1996, pp. 610-619. Available: http://citeseer.ist.psu.edu/schiele96object.html

A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 23, no. 3, pp. 257-267, 2001. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=910878