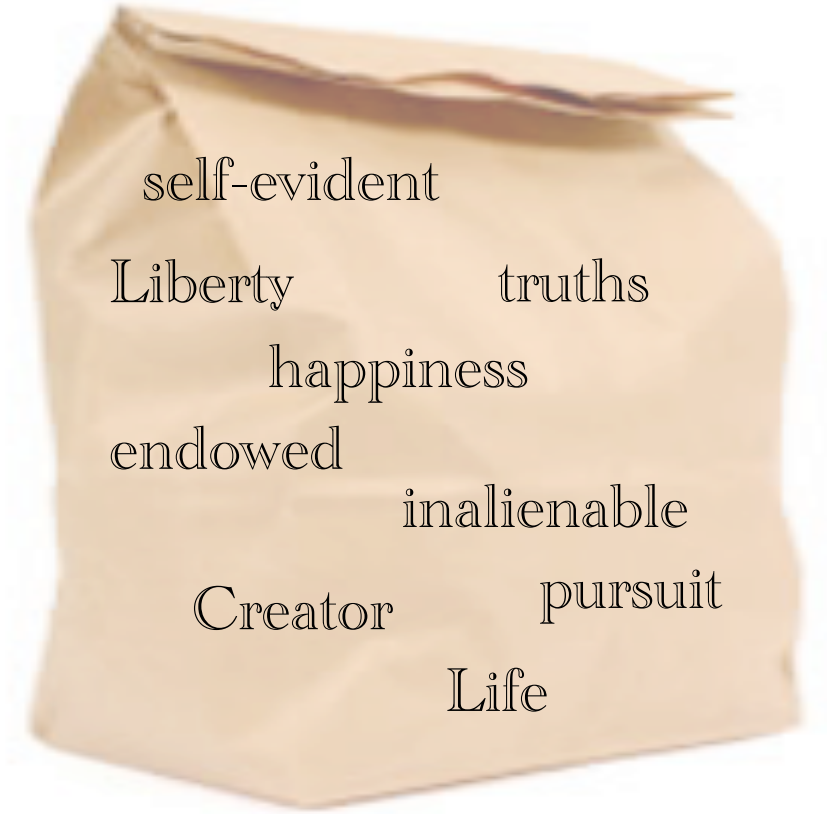
Feature-based methods for image matching

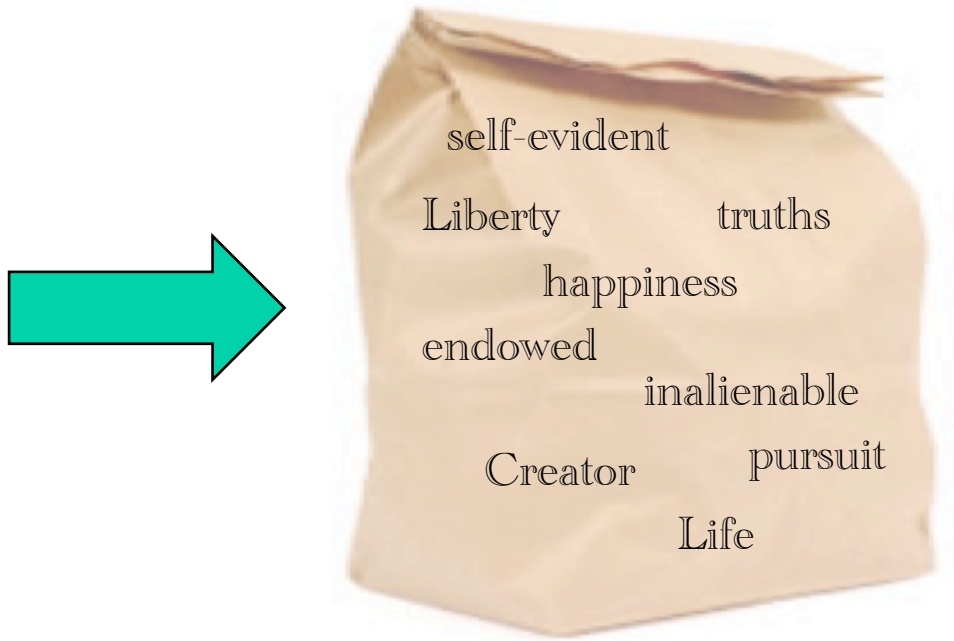
- Bag of Visual Words approach
- Feature descriptors
  - SIFT descriptor
  - SURF descriptor
- Geometric consistency check
- Aggregation of local descriptors into global descriptors
  - Vocabulary trees
  - Fisher vectors
- Image-based retrieval
  - MPEG CDVS standard
  - Mobile visual search
  - Augmented reality

A Bag of Words

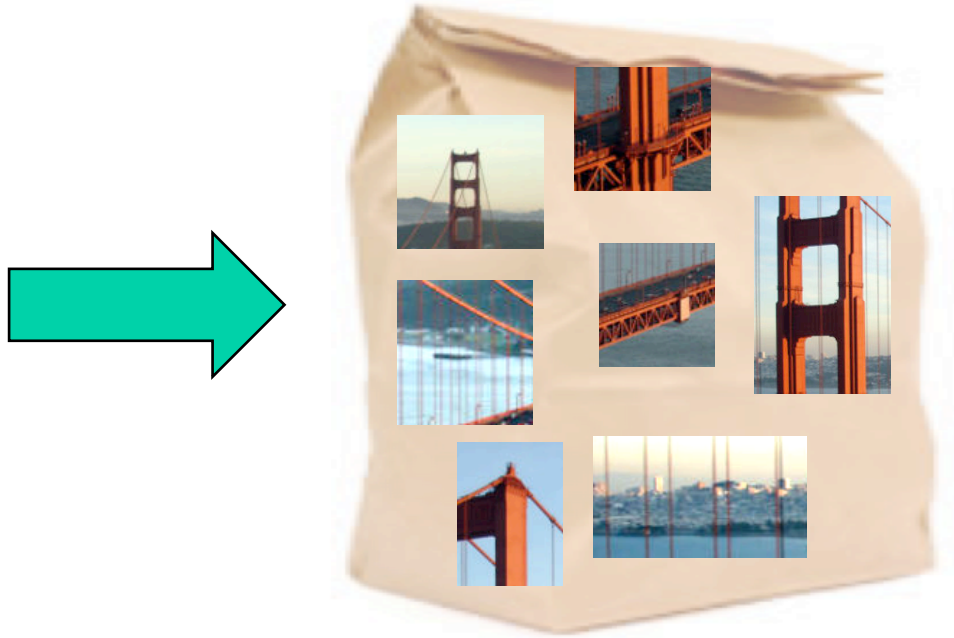


Representing a Text as a "Bag of Words"

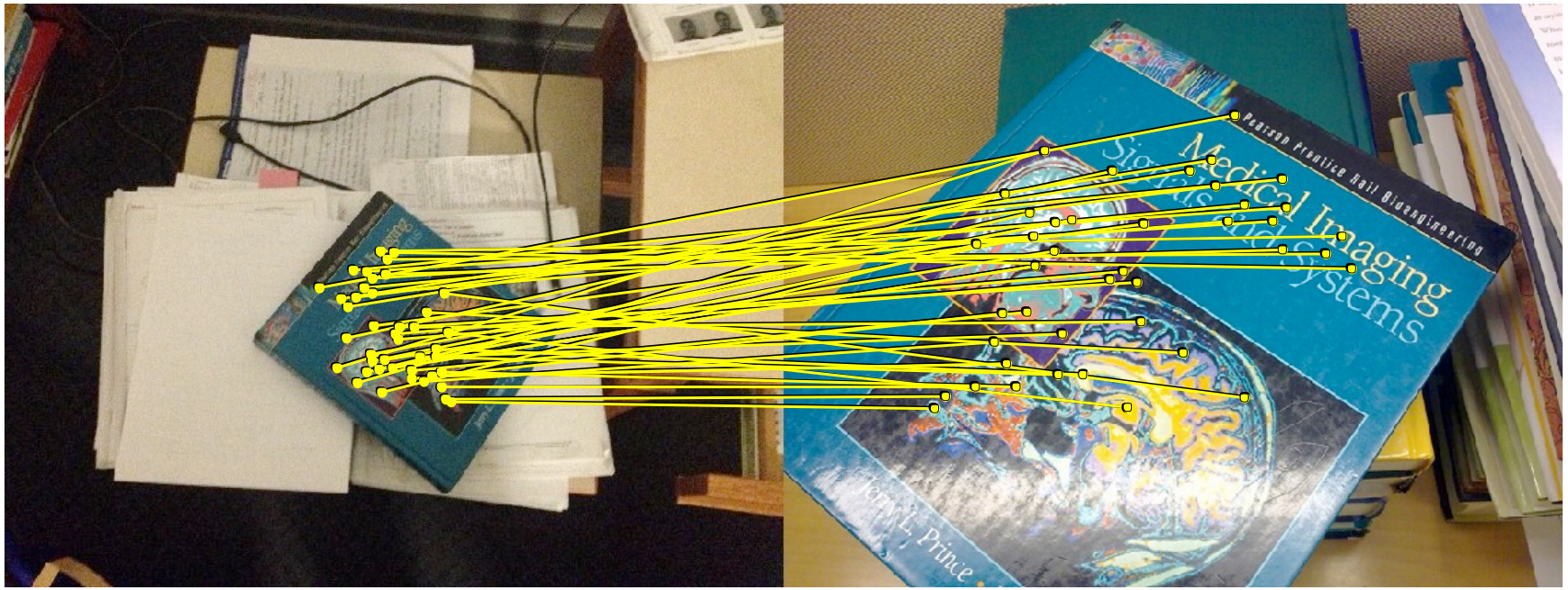
We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness. That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed, That whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to institute new Government, laying its foundation on such principles and organizing its powers in such form, as to them shall seem most likely to effect their Safety and Happiness. Prudence, indeed, will dictate that Governments long established should not be changed for light and transient causes; and accordingly all experience hath shewn, that mankind are more disposed to suffer, while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, pursuing invariably the same Object evinces a design to reduce them under absolute Despotism, it is their right, it is their duty, to throw off such Government, and to provide new Guards for their future security.



Representing an Image as a "Bag of Visual Words"

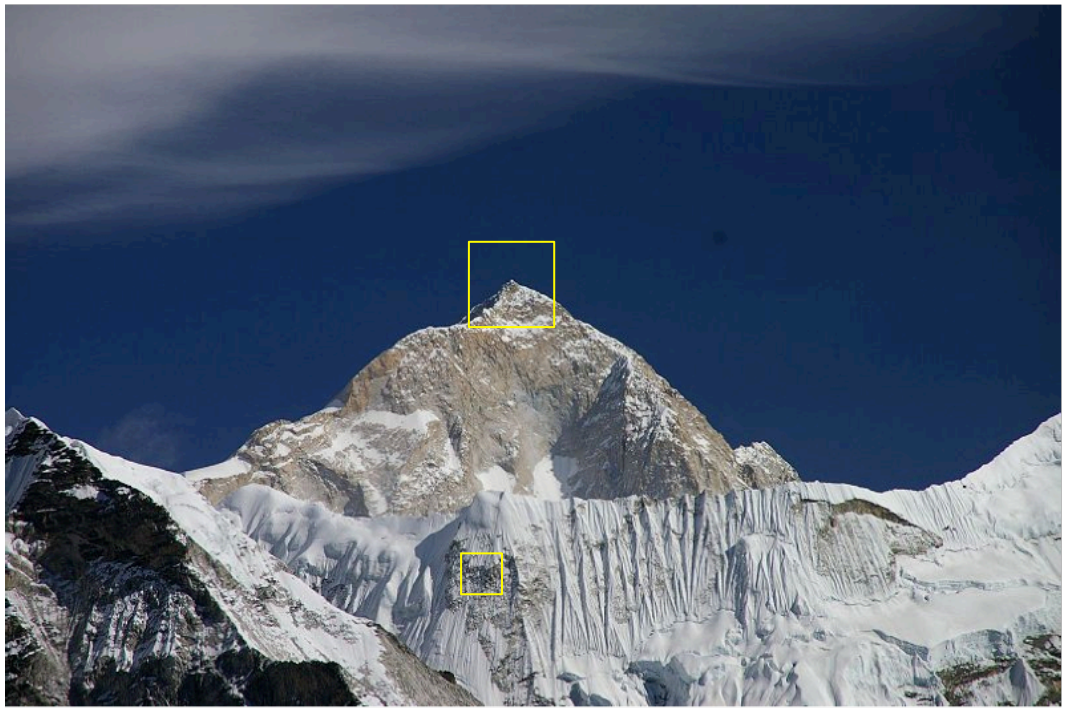


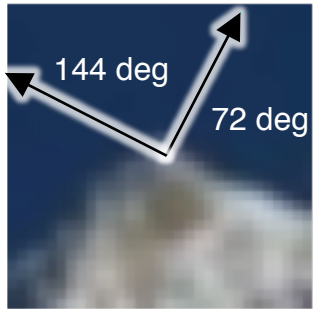
Feature descriptors

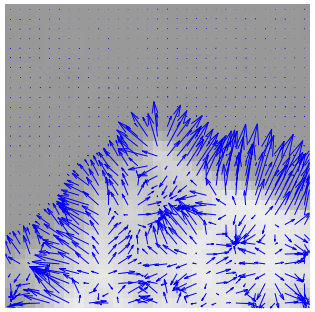


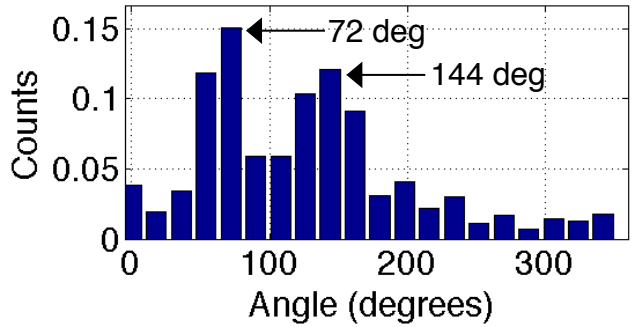
- Represent local pattern around a keypoint by a vector (“feature descriptor”)
- Establish feature correspondences by finding the nearest neighbor in descriptor space

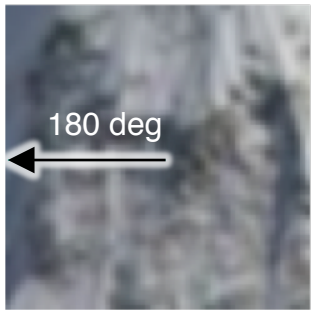
Scale/rotation invariant feature descriptors

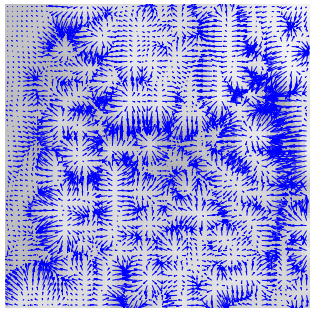


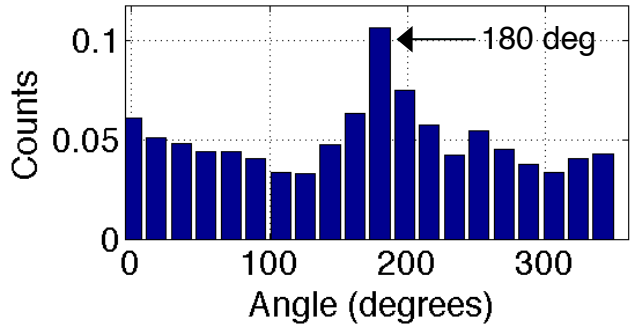










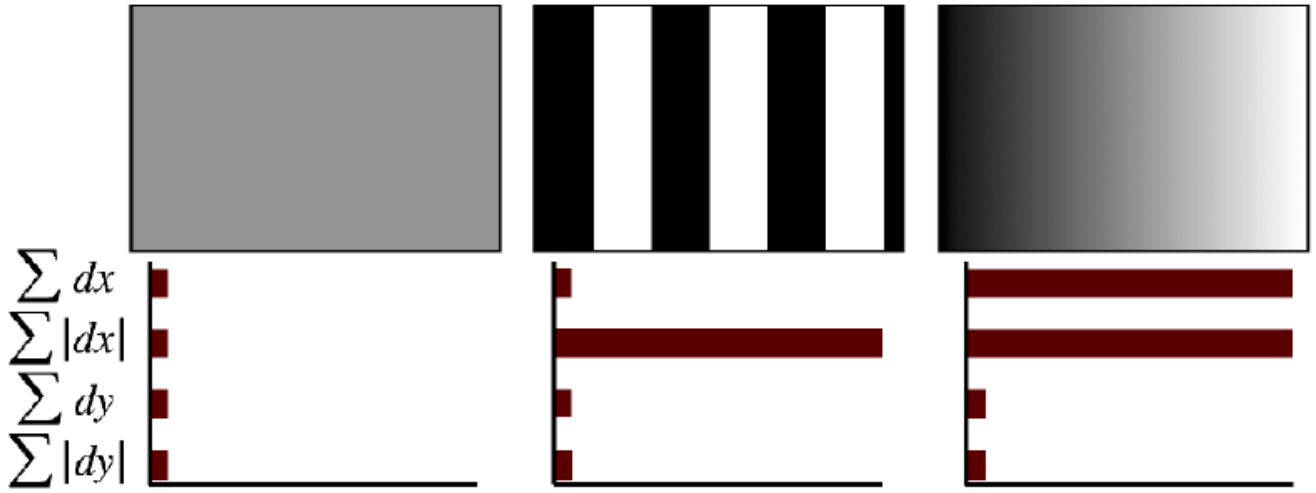


- Scale invariance: extract features at scale provided by keypoint detection
- Rotation invariance:
  - Detect dominant orientation by finding peak in orientation histogram
  - Rotate coordinate system to dominant orientation
  - Multiple strong orientation peaks: generate second feature point

SIFT descriptors

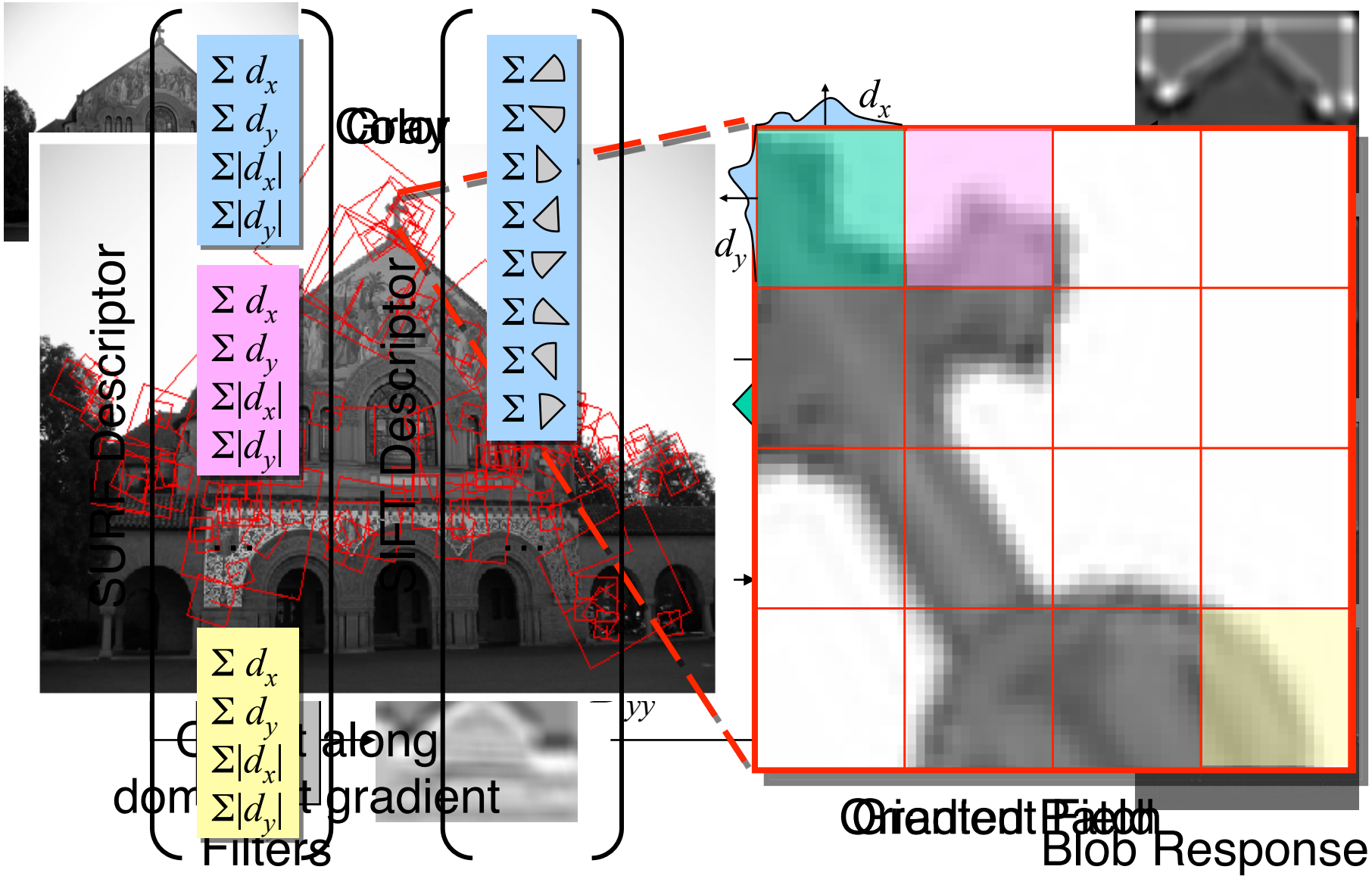
- SIFT - Scale-Invariant Feature Transform *[Lowe, 1999, 2004]*
- Sample thresholded image gradients at 16x16 locations in scale space
  (in local coordinate system for rotation and scale invariance)
- For each of 4x4 subregion, generate orientation histogram with 8 directions each; each observation weighted with magnitude of image gradient and a window function
- 128-dimensional feature vector

SURF descriptors



- SURF – Speeded Up Robust Features *[Bay et al. 2006]*
- Compute horizontal and vertical pixel differences, dx, dy (in local coordinate system for rotation and scale invariance, window size 20σ x 20σ, where σ^2 is feature scale)
- Sum dx, dy, and $|dx|, |dy|$ over 4x4 subregions (SURF-64) or 3x3 subregions (SURF-36)
- Normalize vector for gain invariance, but distinguish bright blobs and dark blobs based on sign of Laplacian (trace of Hessian matrix)

Computing feature descriptors



“Bag of Visual Words” Matching




Pairwise Comparison

Geometric mapping

- Notation:
  - Homogeneous coordinates; reference image $\underline{\mathbf{x}} = \begin{pmatrix} x & y & 1 \end{pmatrix}^T$
  - Inhomogeneous coordinates; target image $\mathbf{x}' = \begin{pmatrix} x' & y' \end{pmatrix}^T$
- Translation

$$\mathbf{x}' = \mathbf{x} + \mathbf{t} \quad \text{or} \quad \mathbf{x}' = \begin{bmatrix} \mathbf{I} & \mathbf{t} \end{bmatrix} \underline{\mathbf{x}}$$

- Euclidean transformation (rotation and translation)

$$\mathbf{x}' = \begin{bmatrix} \cos\theta & -\sin\theta & t_x \\ \sin\theta & \cos\theta & t_y \end{bmatrix} \underline{\mathbf{x}}$$

- Scaled rotation (similarity transform)

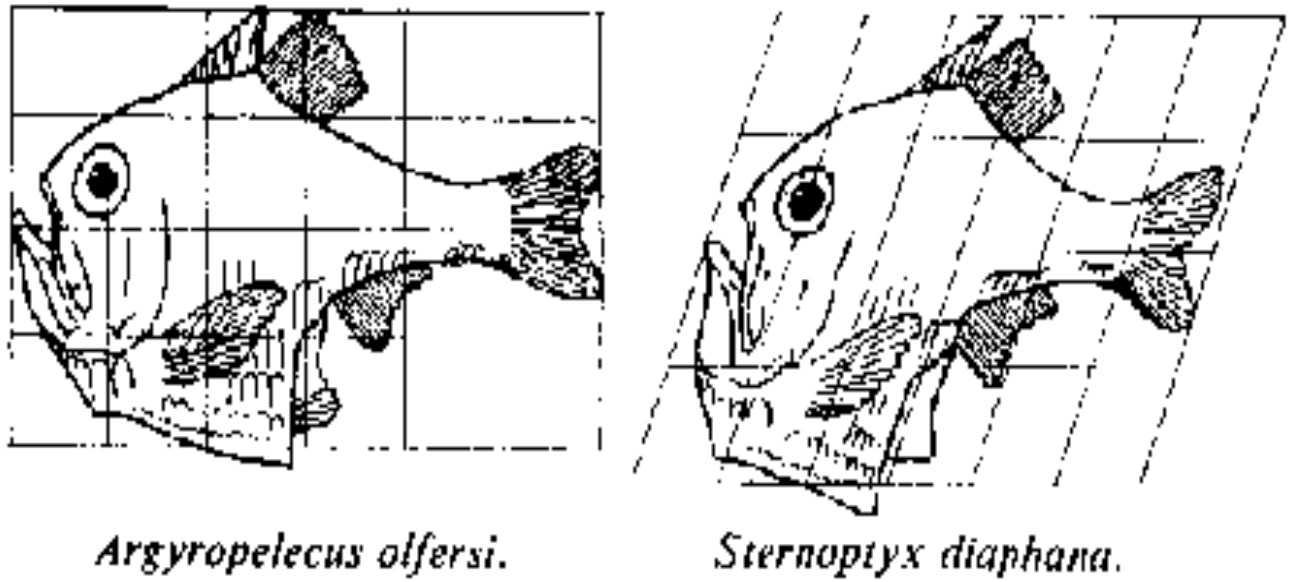
$$\mathbf{x}' = \begin{bmatrix} s\cdot\cos\theta & -s\cdot\sin\theta & t_x \\ s\cdot\sin\theta & s\cdot\cos\theta & t_y \end{bmatrix} \underline{\mathbf{x}}$$

Geometric mapping

- Affine transformation

$$\mathbf{x}' = \begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \end{bmatrix} \underline{\mathbf{x}}$$

- Motion of planar surface in 3d under orthographic projection
- Parallel lines are preserved

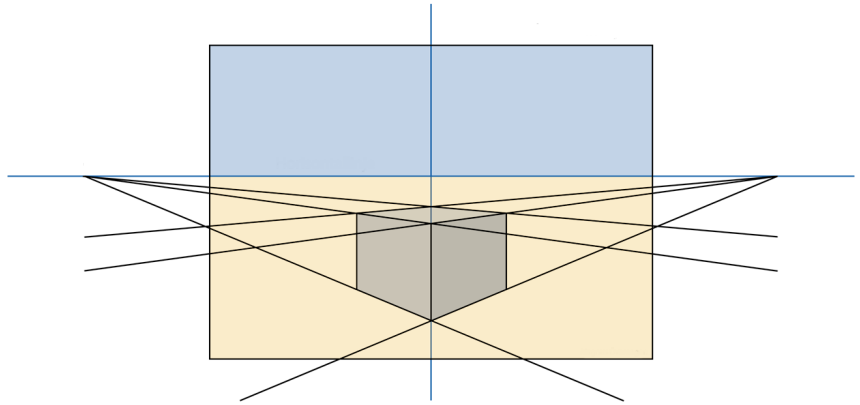


Argyropelecus olfersi. *Sternoptyx diaphana.*

Geometric mapping

- Motion of planar surface in 3d under perspective projection
- Homography

$$\underline{\mathbf{x}}' : \begin{pmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{pmatrix} \underline{\mathbf{x}}$$



- Inhomogeneous coordinates (after normalization)

$$x' = \frac{h_{00}x + h_{01}y + h_{02}}{h_{20}x + h_{21}y + h_{22}} \qquad y' = \frac{h_{10}x + h_{11}y + h_{12}}{h_{20}x + h_{21}y + h_{22}}$$

- Straight lines are preserved

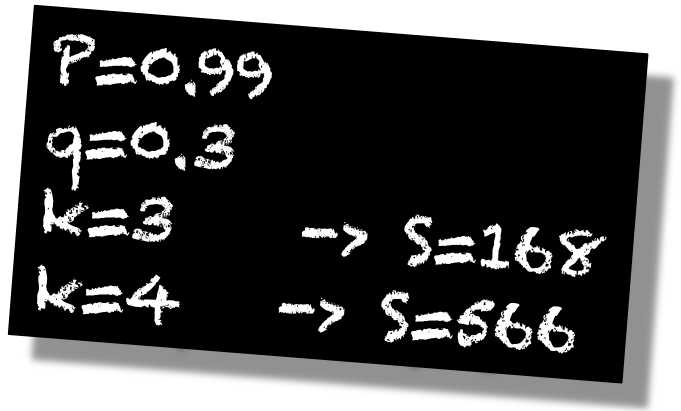
RANSAC

- RANdom Sample Consensus *[Fischer, Bolles, 1981]*
- Randomly select subset of k correspondences
- Compute geometric mapping parameters by linear regression
- Apply geometric mapping to all keypoints
- Count no. of inliers (closer than ε from the corresponding keypoint, typical ε = 1…3 pixels)
- Repeat process S times, keep geometric mapping with largest no. of inliers
- Required number of trials

$$S = \frac{\log\left(1-P\right)}{\log\left(1-q^k\right)}$$

Total probability of success

Probability of valid correspondence

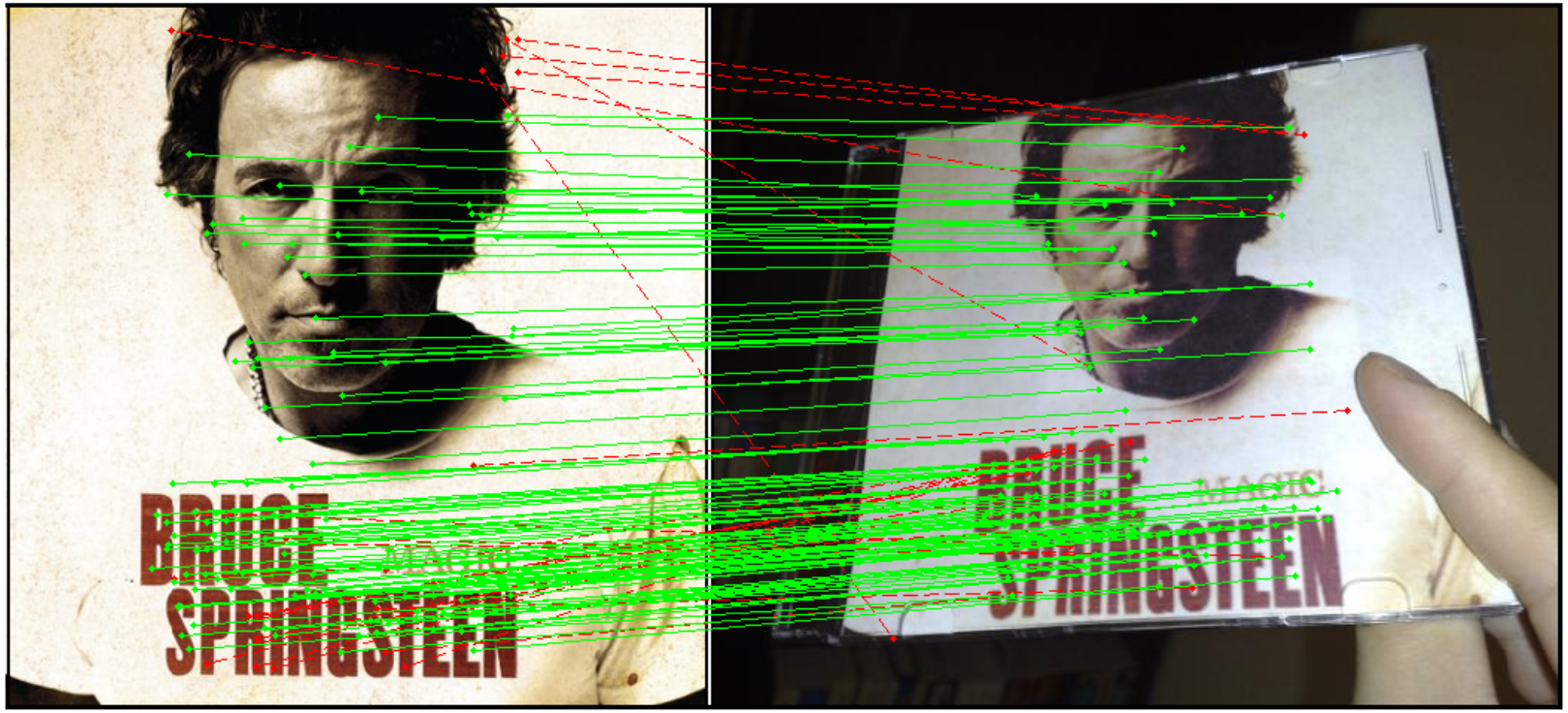


- Use small number of correspondences

RANSAC with Affine Model

RANSAC with Homography

SURF features & affine RANSAC

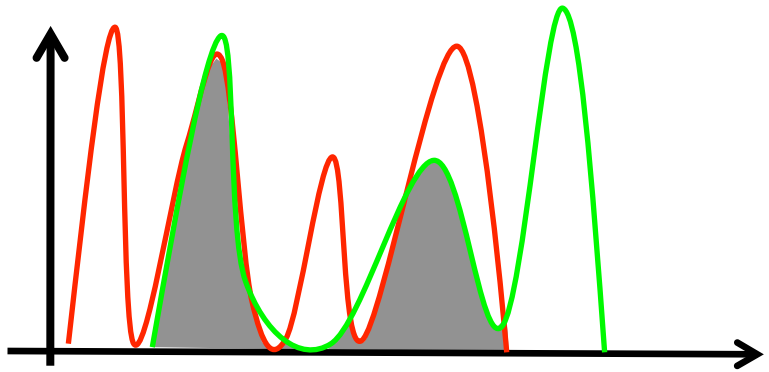


Local Feature Descriptor Aggregation

- Nearest-neighbor matching of variable-size sets of local features is costly
- Compare images based on a global binary signature of constant size ("hash") instead
- **Simple:** VQ of feature vectors to generate histogram, compare non-empty histogram bins ("bag of features," "bag of visual words")
- **Better:** binarize gradient of log likelihood of w.r.t. to parameter vector ("Fisher vector")

Comparing Feature Histograms

- Speed up by comparing histograms of features: pairwise image comparison only for similar histograms
- Histogram intersection



Query histogram

Histogram of database entry

$$\rho = \frac{\sum_{i=1}^{n} \min(Q_i, D_i)}{\sum_{i=1}^{n} D_i}$$

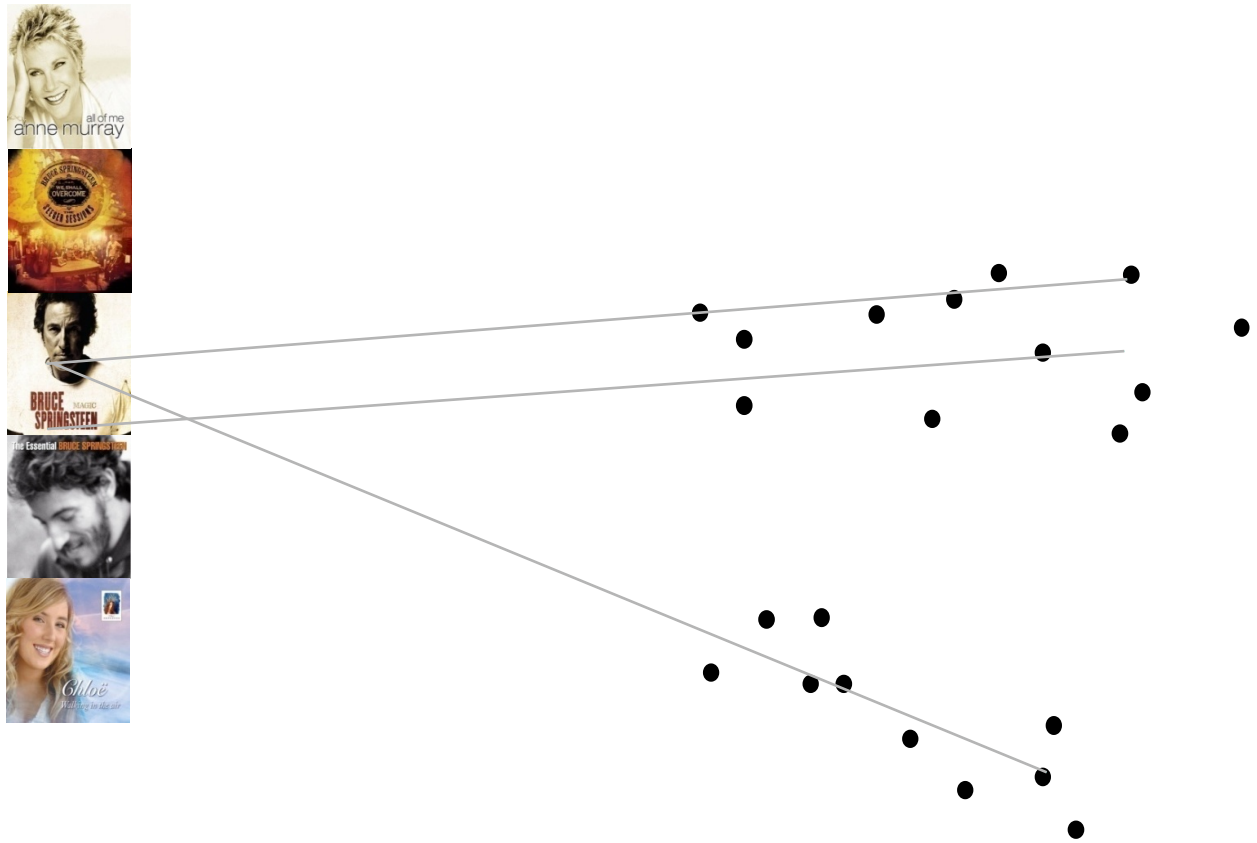
[Swain, Ballard 1991]

- Equivalent to mean absolute difference, if both histograms contain same number of samples

Growing Vocabulary Tree

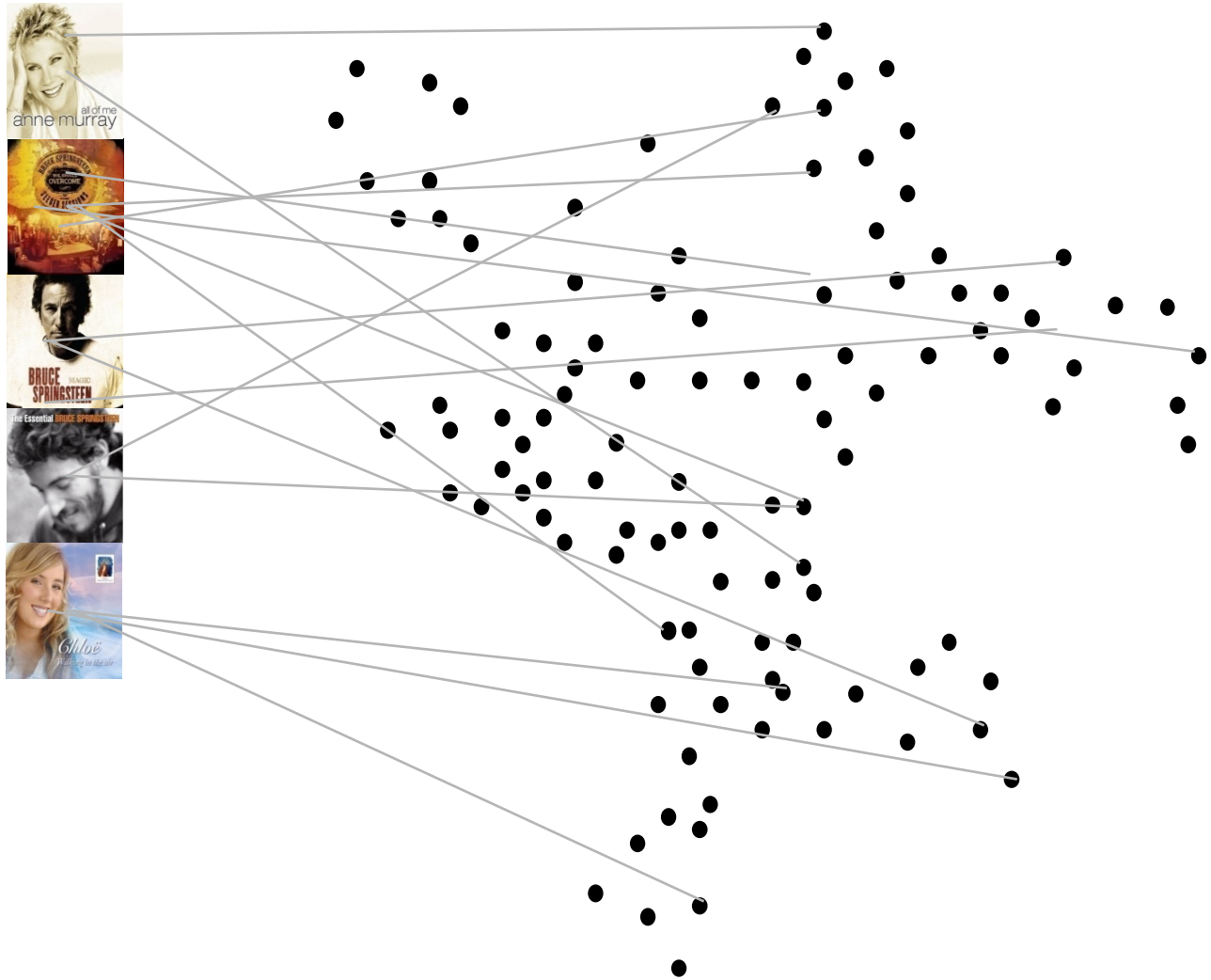
[Nistér and Stewenius, 2006]

Growing Vocabulary Tree



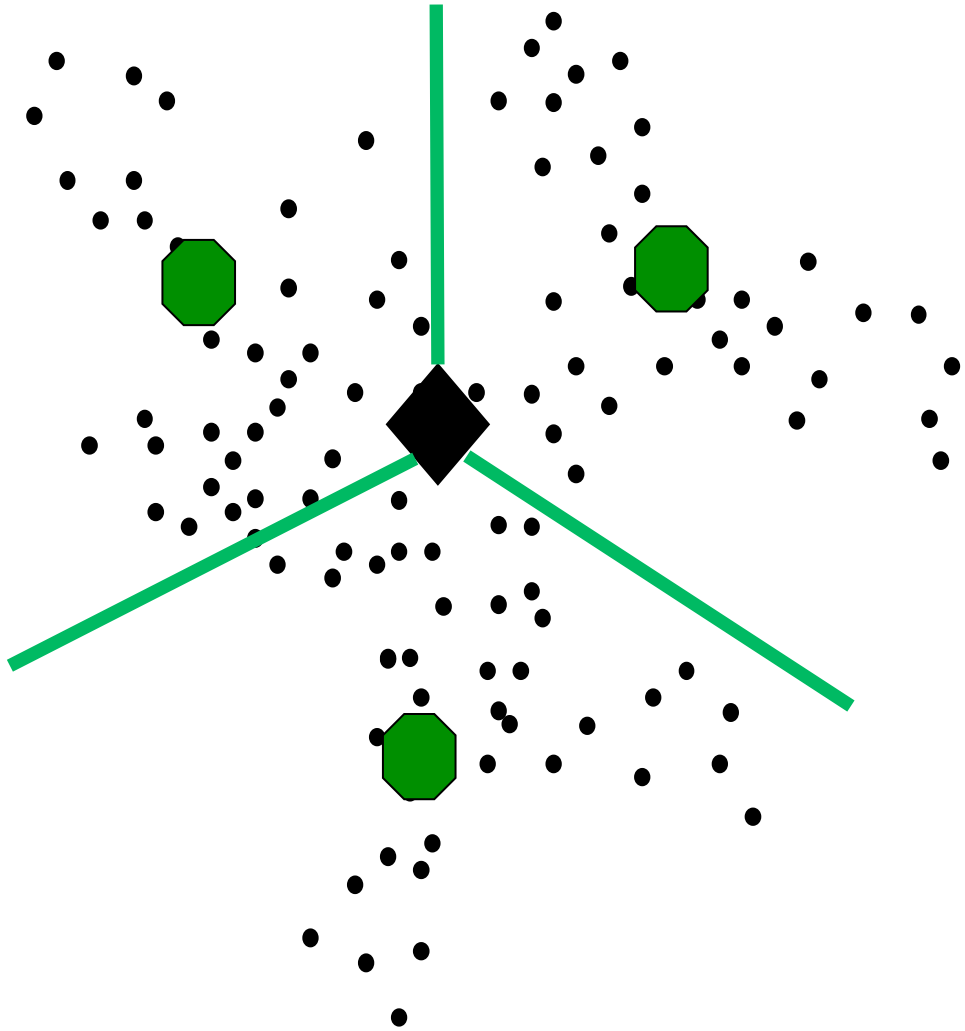
[Nistér and Stewenius, 2006]

Growing Vocabulary Tree



[Nistér and Stewenius, 2006]

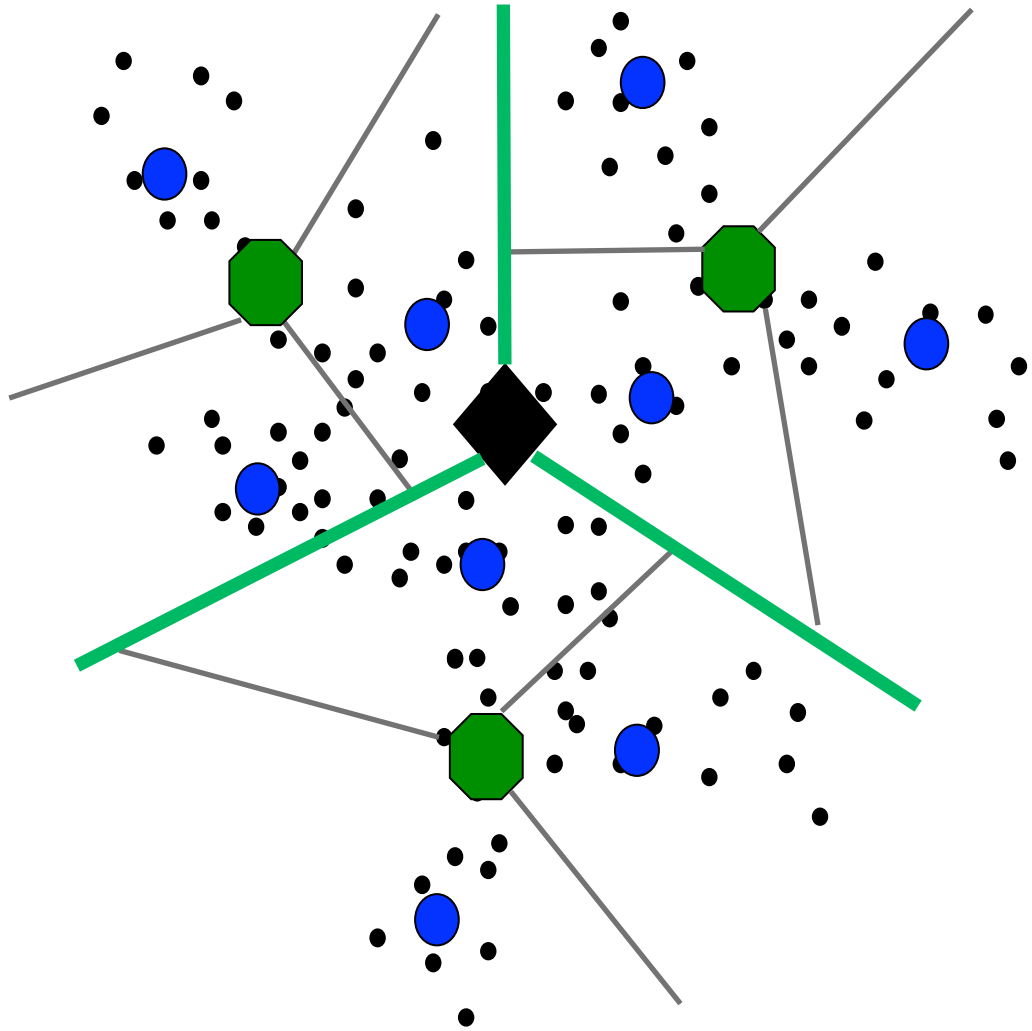
Growing Vocabulary Tree



$k = 3$

[Nistér and Stewenius, 2006]

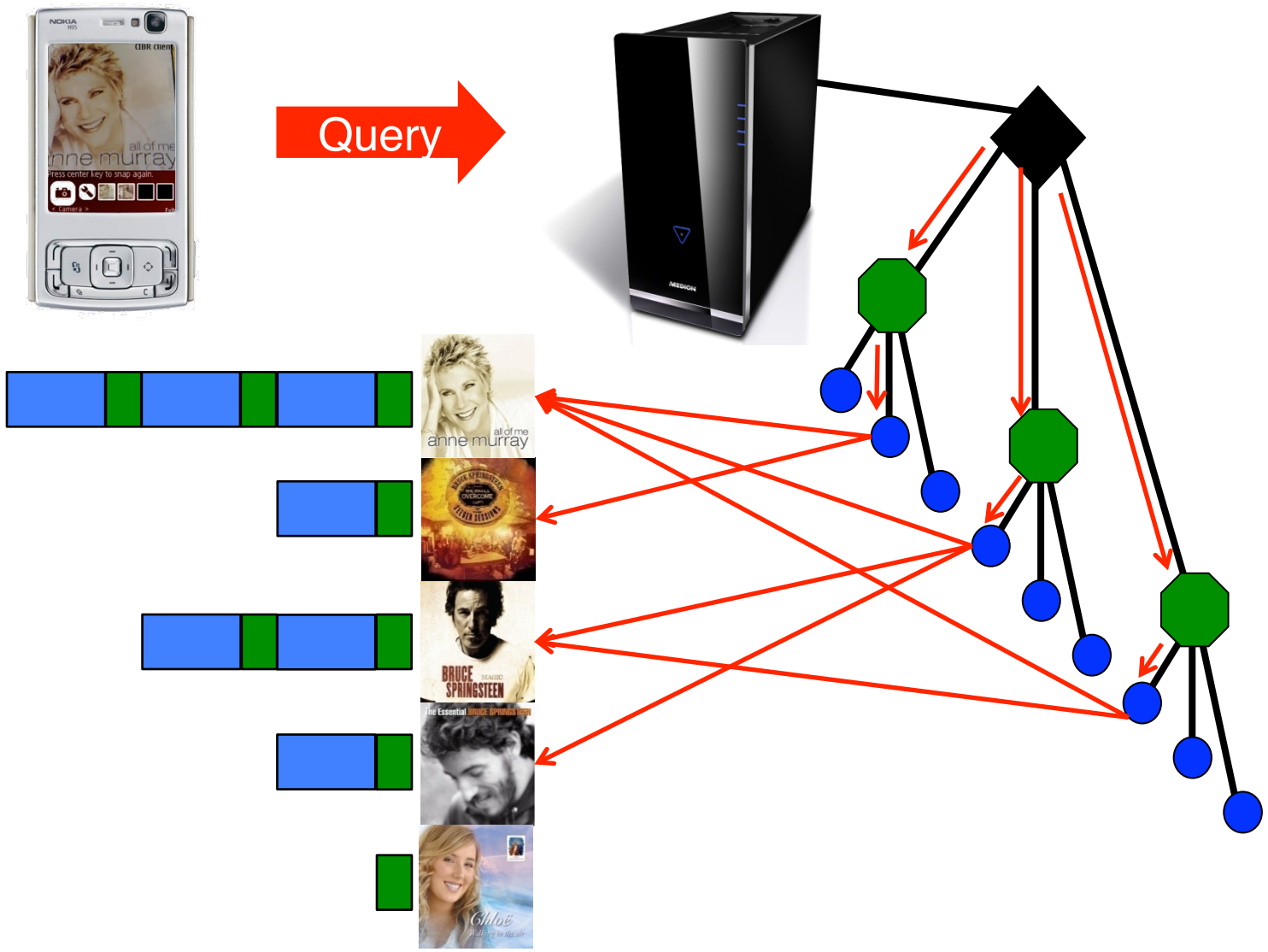
Growing Vocabulary Tree



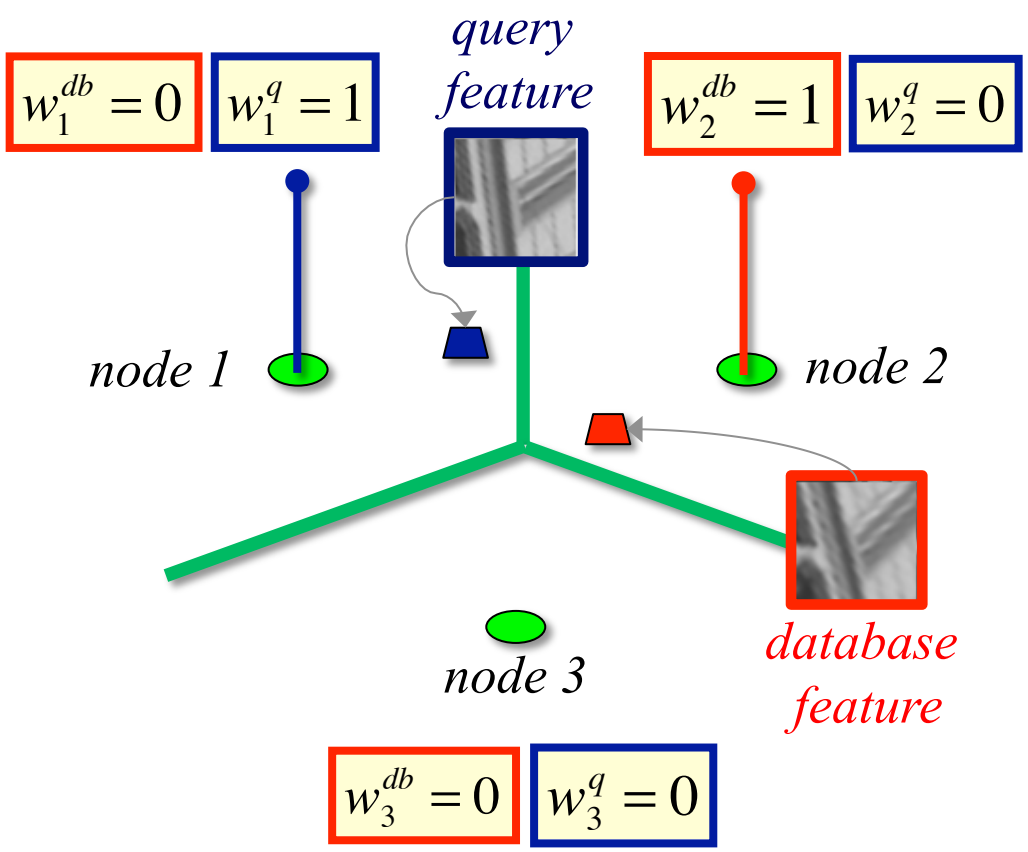
$k = 3$

[Nistér and Stewenius, 2006]

Querying Vocabulary Tree

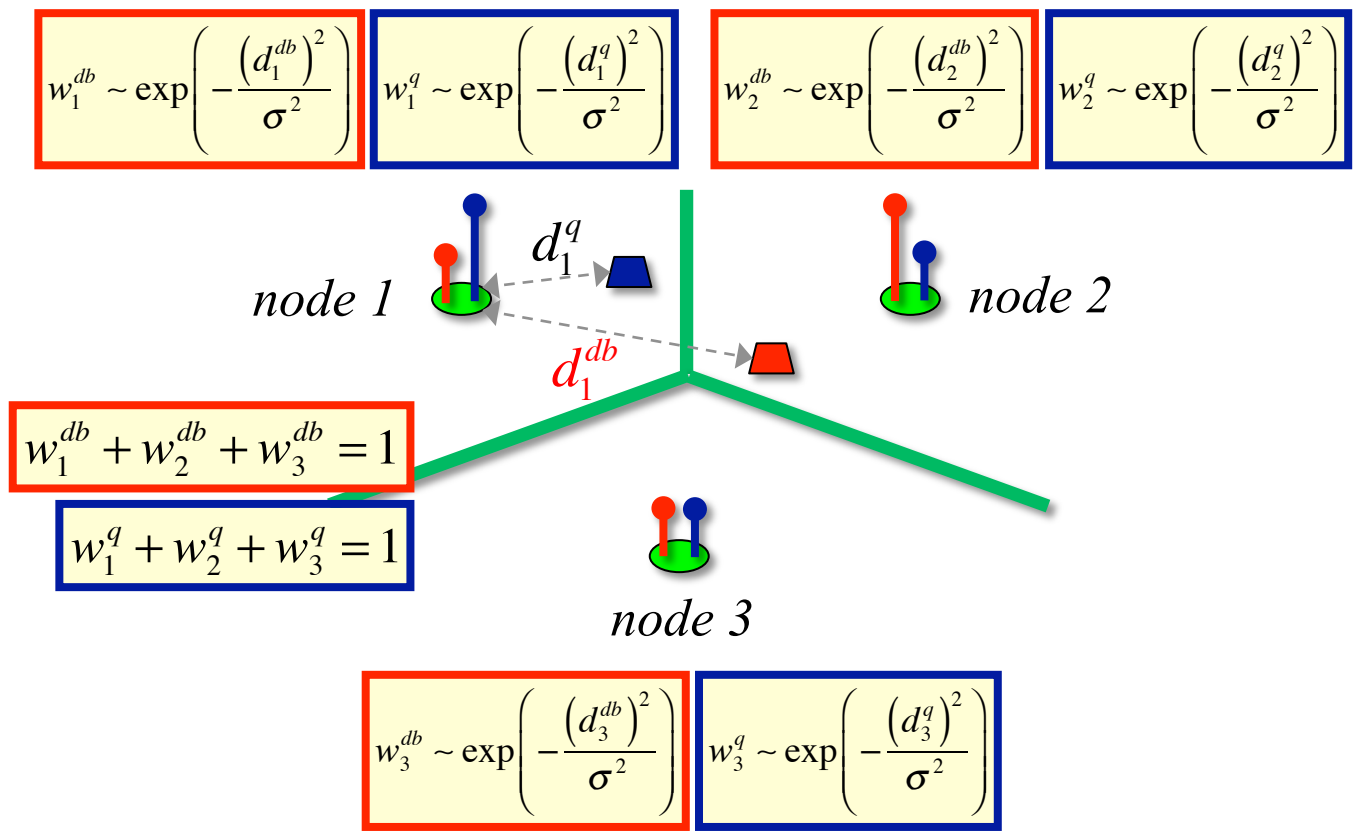


Hard Binning vs. Soft Binning



Hard Binning

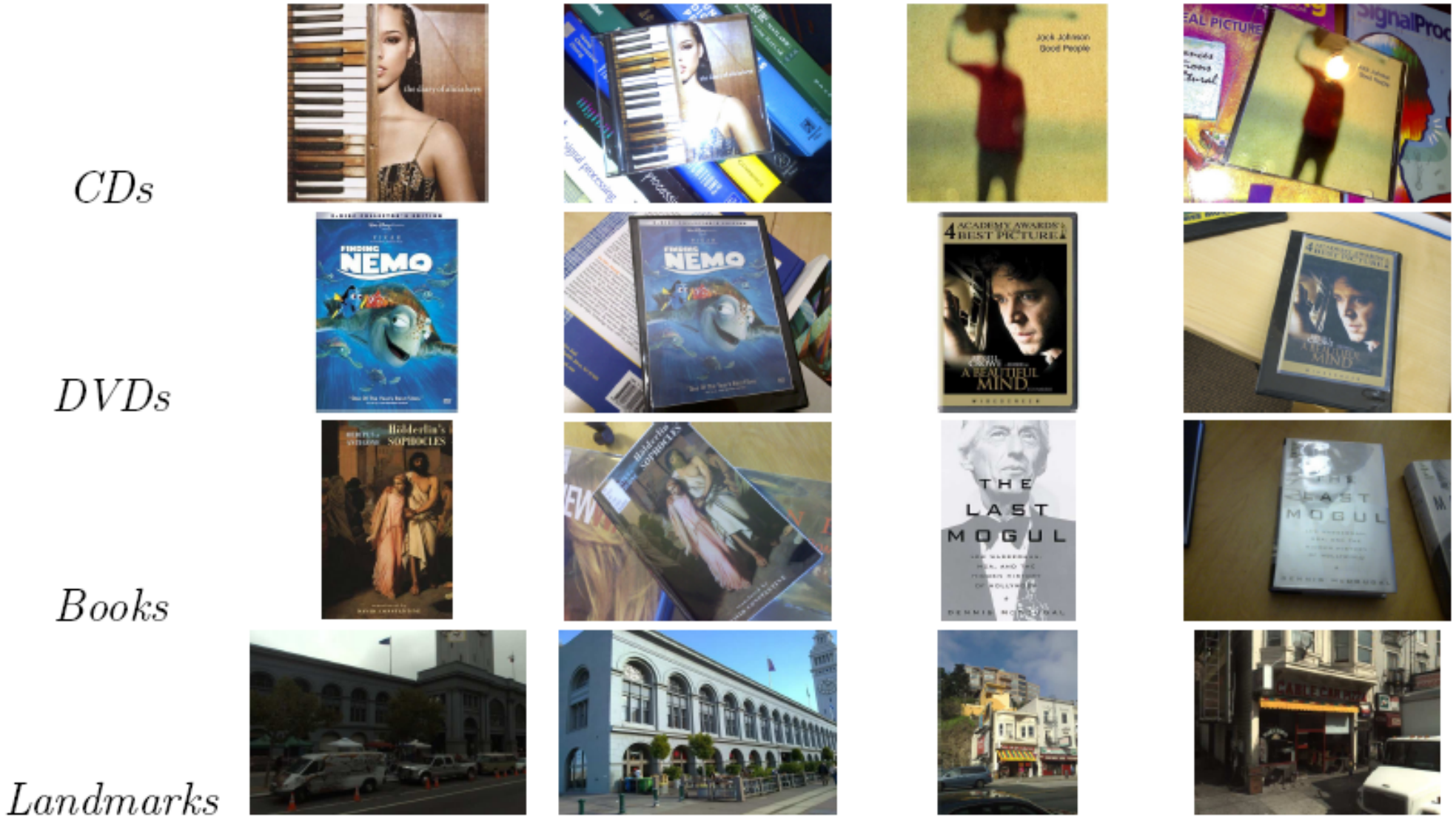
[Nistér and Stewenius, CVPR 2006]



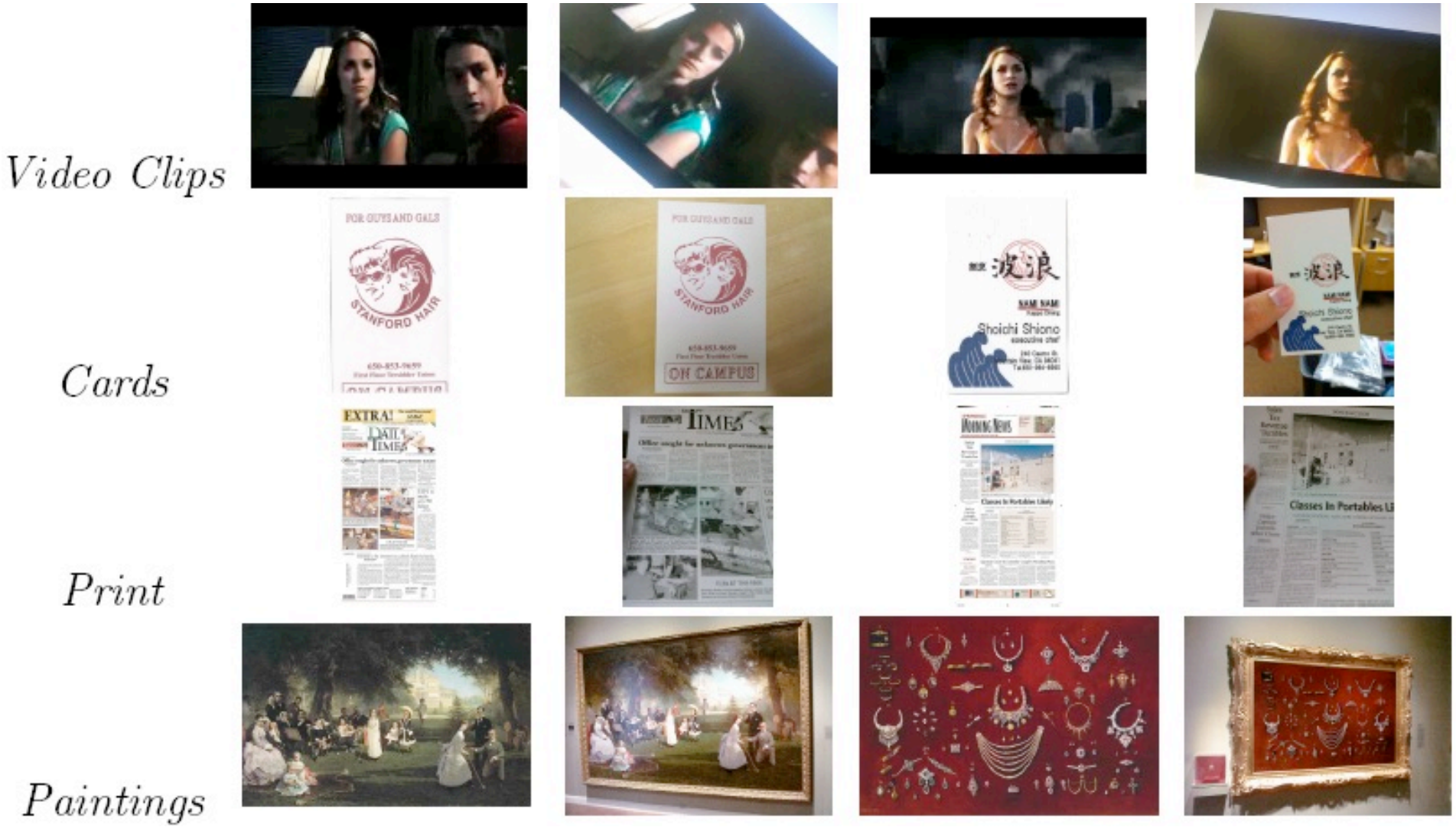
Soft Binning

[Philbin et al., CVPR 2008]

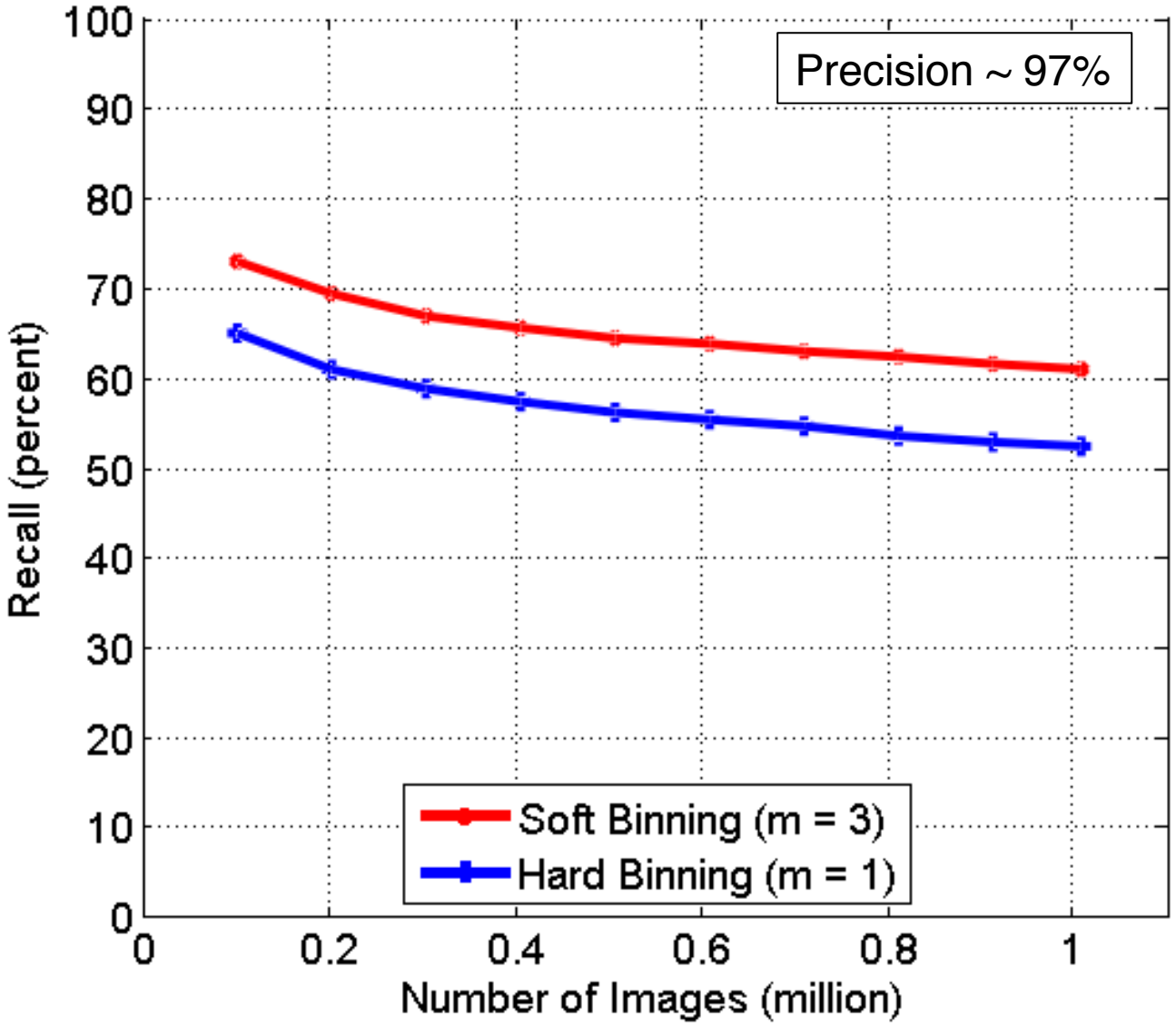
Stanford Mobile Visual Search Dataset



Stanford Mobile Visual Search Dataset



Querying: Hard Binning vs. Soft Binning



SURF features
6-level vocab tree
1M leaf nodes
Affine RANSAC
for 100 top tree results
25 inliers min.

Fisher Vector

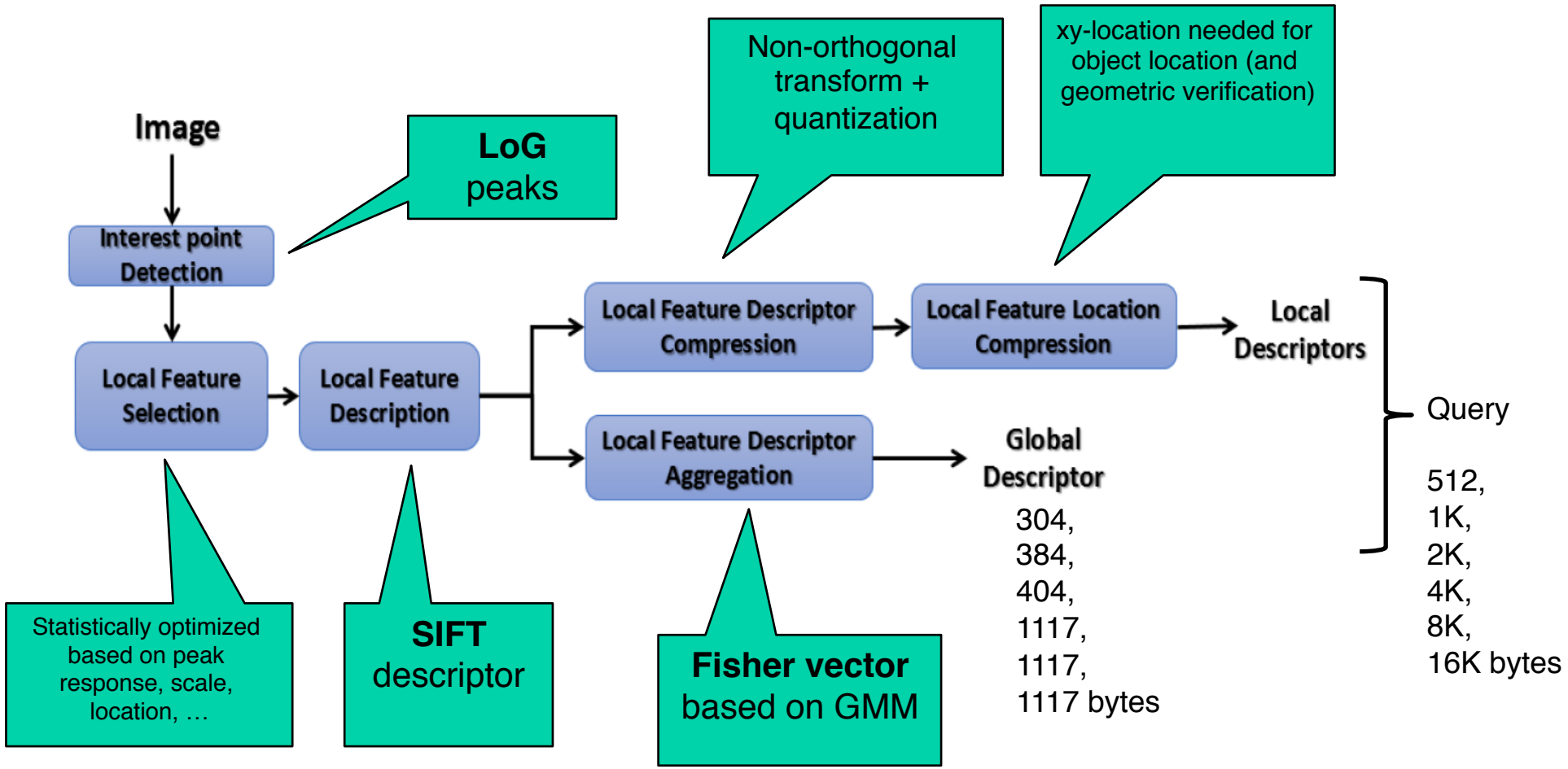
- Discriminative score function

$$U(X) = \frac{\partial}{\partial \Theta} \log p_{X|\Theta}(X|\Theta)$$

d-dimensional vector $\quad d \gg k \quad$ k-dimensional feature vector $\quad$ d Parameters

- Typical, we use Gaussian mixture model (GMM) for $p_{X|\Theta}(X|\Theta)$
- Parameters Θ: mean (and variance) of Gaussian clusters
- For GMM, feature scores $U(X)$ are soft-assigned distance vectors (and squared distance vectors) relative to cluster centers
- Sums of feature scores of an image are “Fisher vector” that can be used to compare images
- Binarization & Hamming distance comparison results in only minor performance loss (“Binarized Fisher vector”)

MPEG standard "Compact Descriptors for Visual Search" (CDVS)



CDVS Evaluation Framework

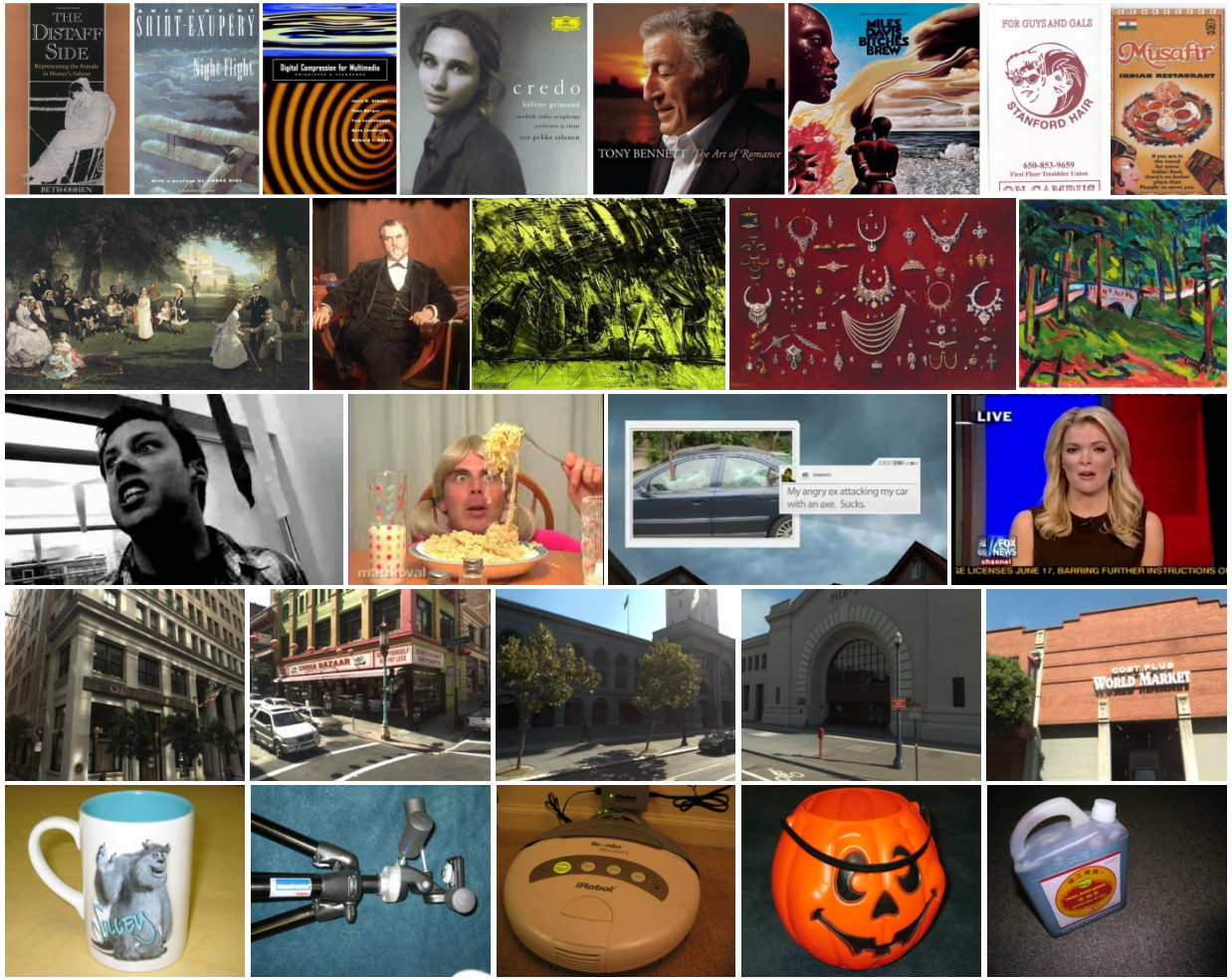
Graphics

Paintings

Video Frames

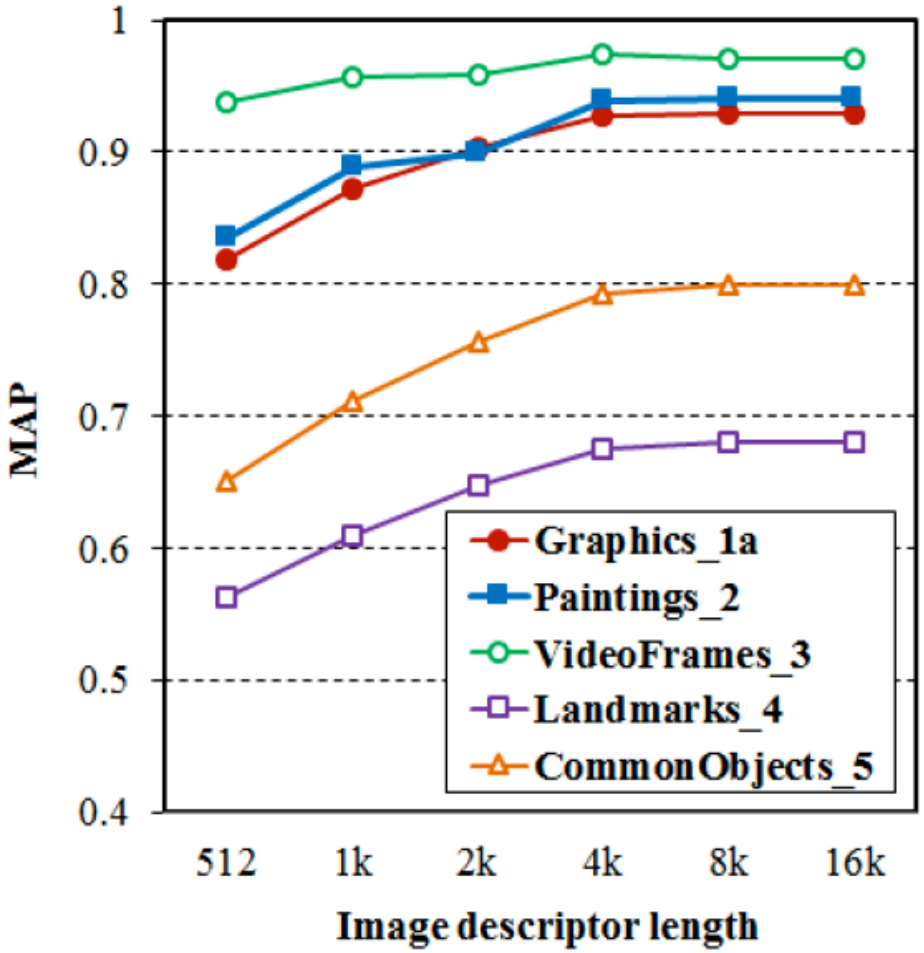
Landmarks

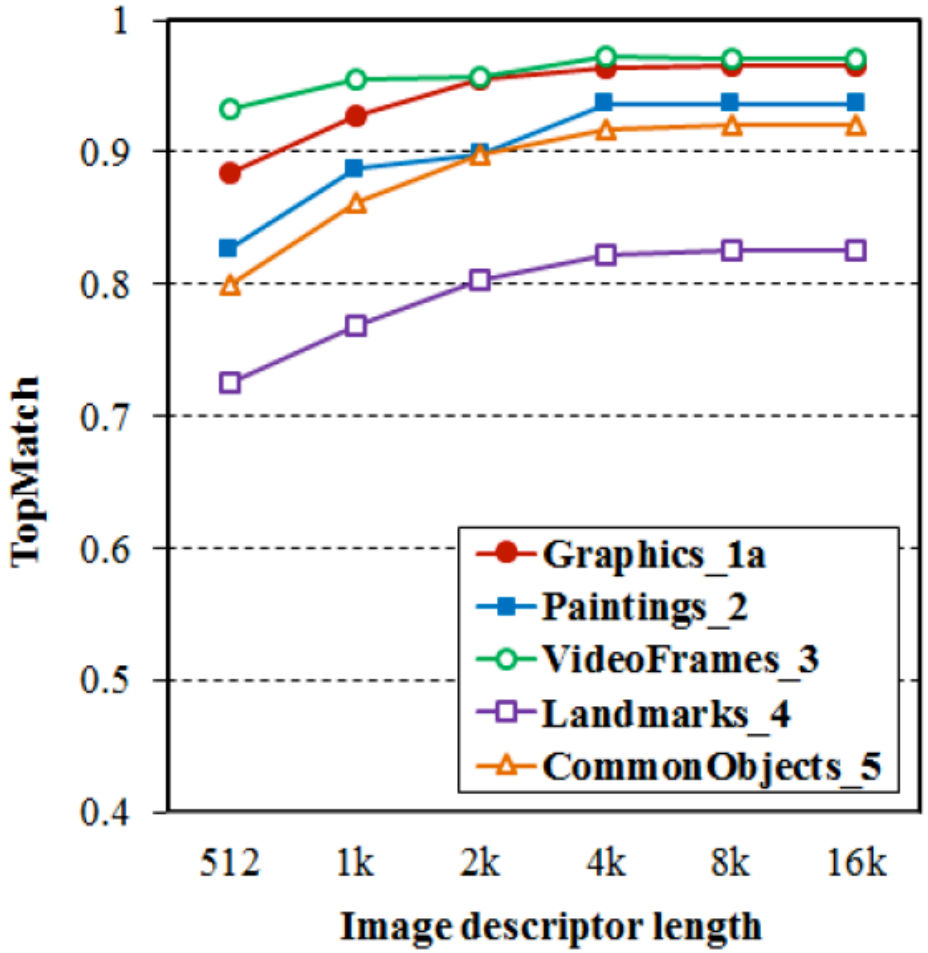
Common Objects



1M Distractor Images

MPEG CDVS Performance



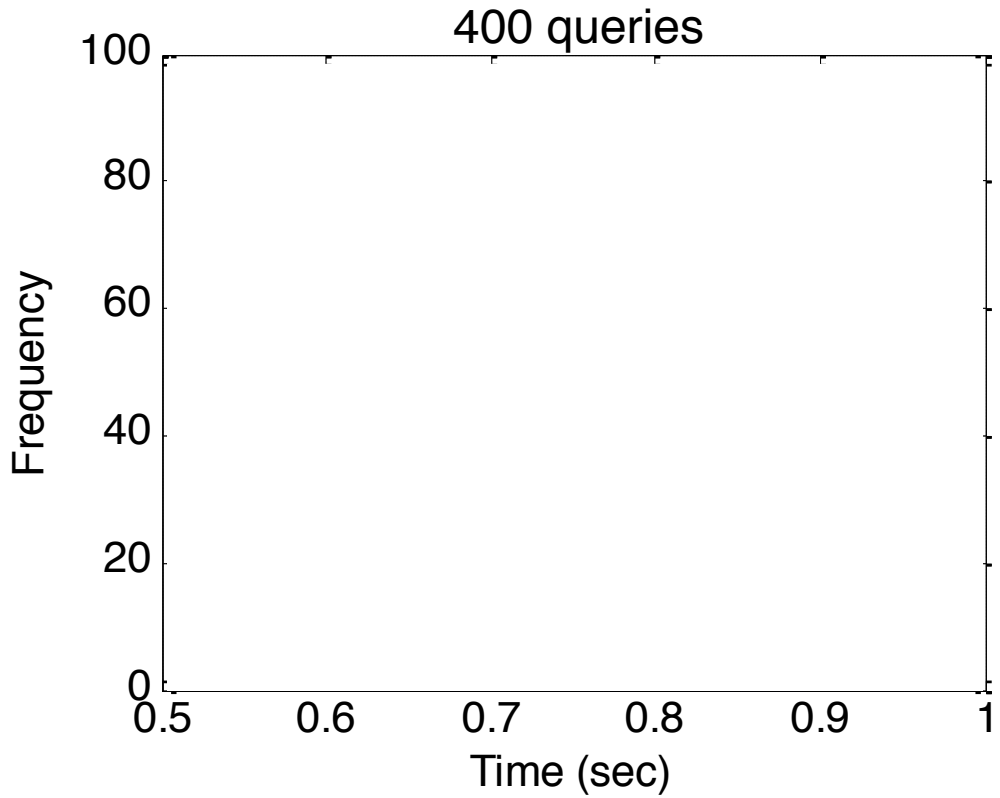


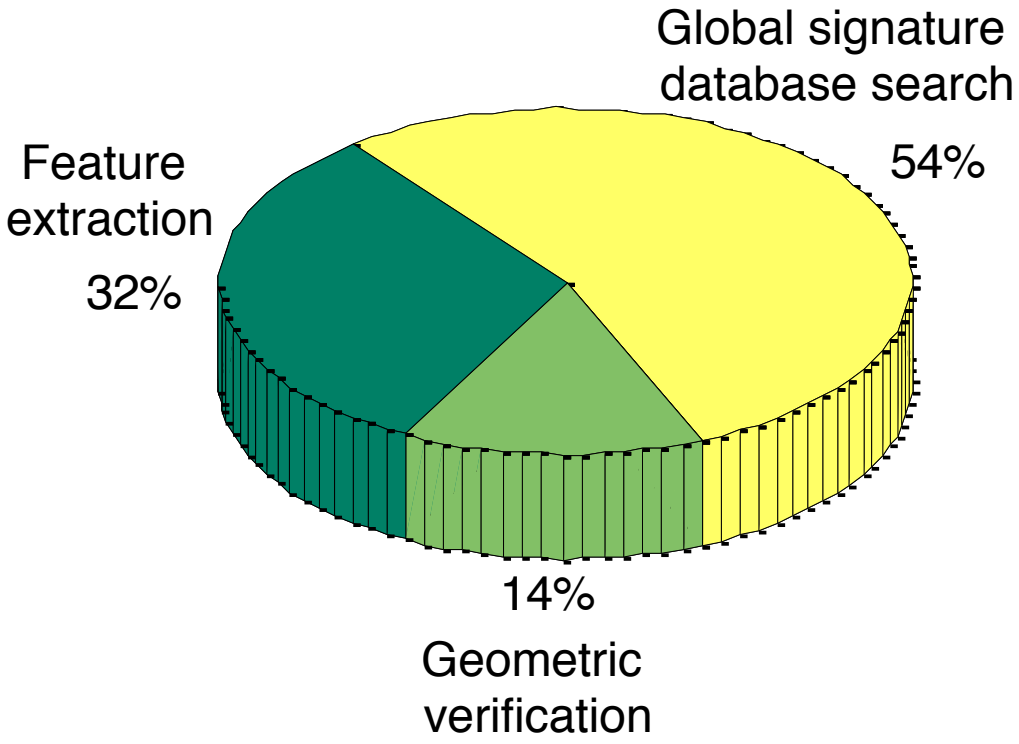
On-Device Image Matching Demo
Database of 100K Images

Samsung Galaxy S3 Smartphone

On-Device Timing Measurements

Samsung Galaxy S3 Smartphone
1.4 GHz Processor
1 GB RAM
Database of 100K Images





Augmented Reality Glasses

Augmented Reality Glasses

Augmented Reality Glasses