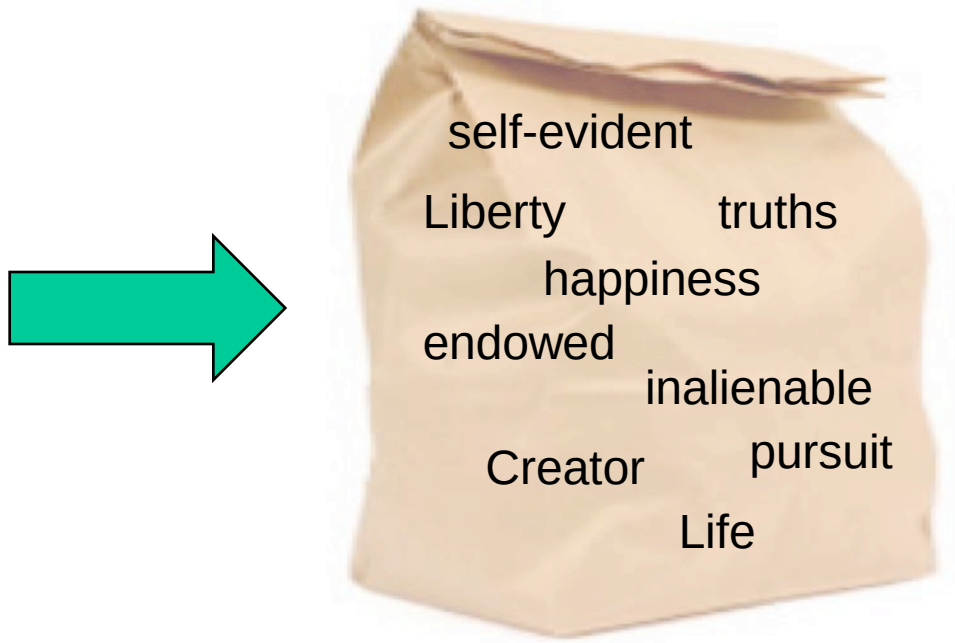
Feature-based methods for image matching

- Bag of Visual Words approach
- Feature descriptors
  - SIFT descriptor
  - SURF descriptor
- Geometric consistency check
- Vocabulary tree

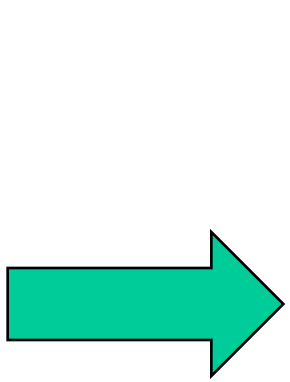
A Bag of Words

Representing a Text as a "Bag of Words"

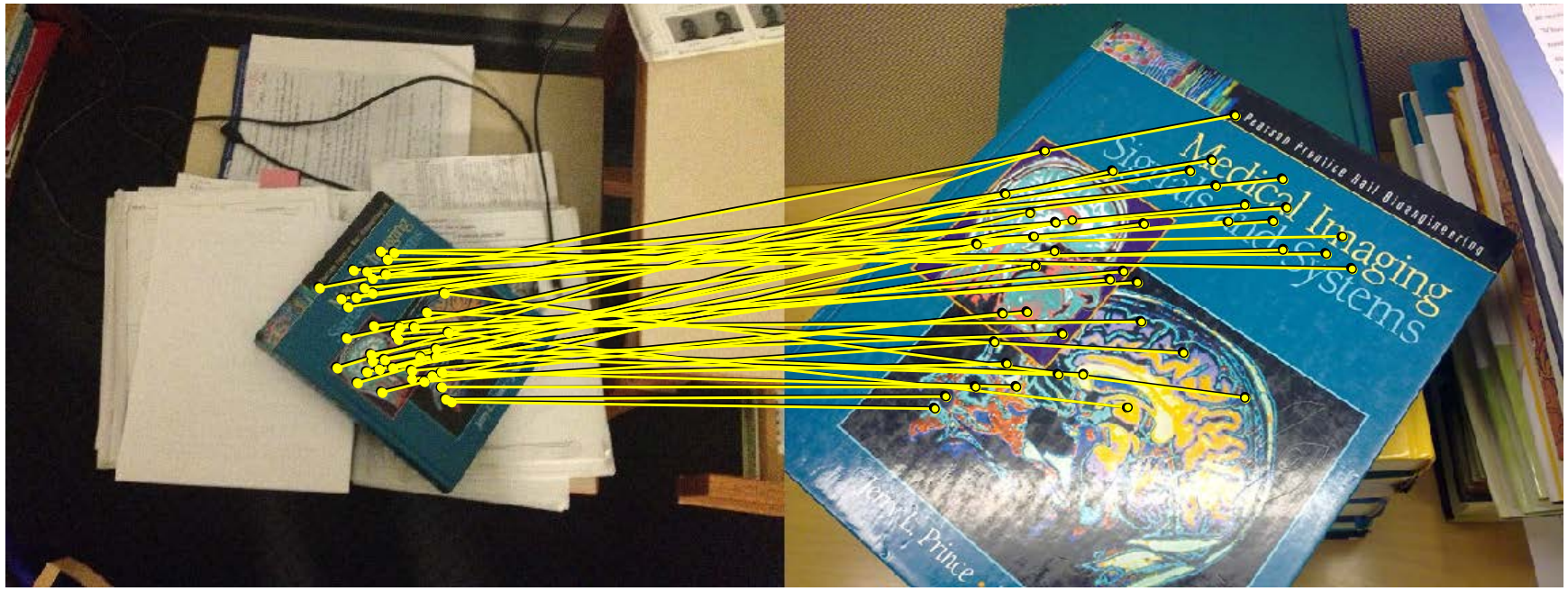
We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness. That to secure these rights, Governments are instituted among Men, deriving their just powers 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”

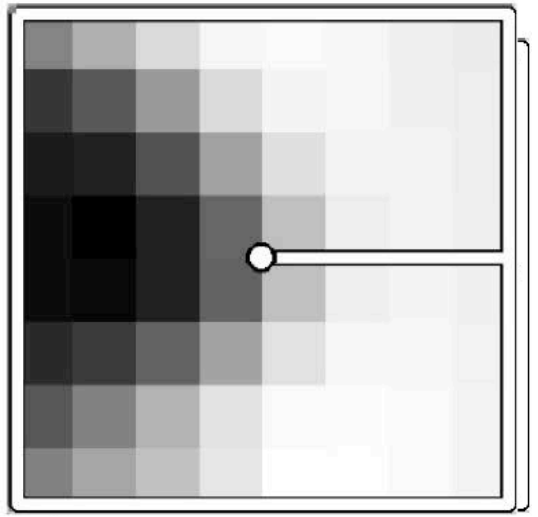


Feature descriptors



- 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



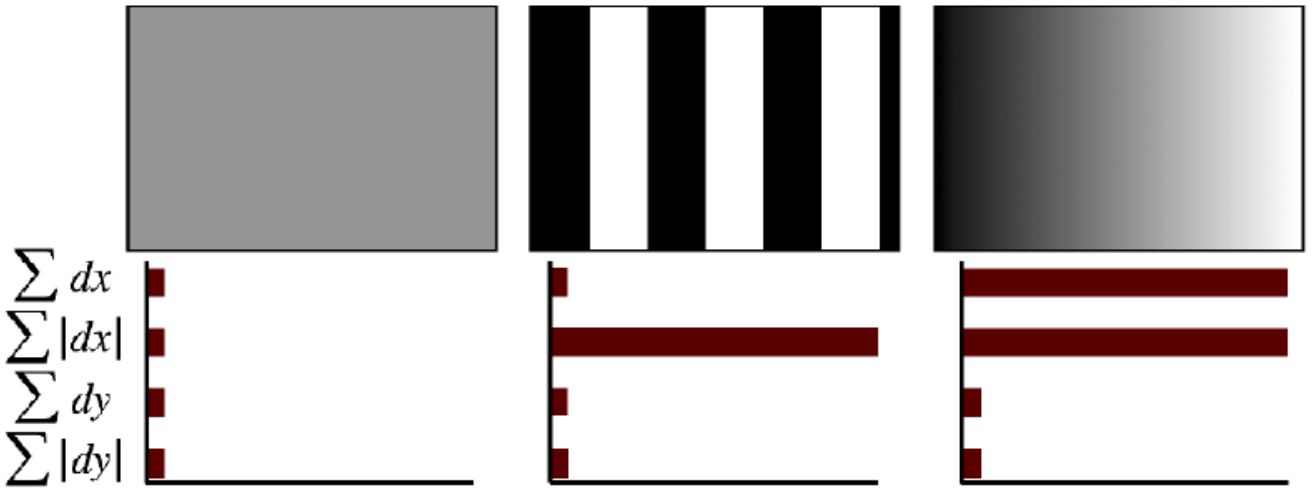
- Scale invariance: extract features at scale provided by keypoint detection
- Rotation invariance:
  - Detect dominant orientation
    - Average gradient direction
    - Peak detection in gradient direction histogram
  - Rotate coordinate system to dominant orientation
  - Multiple strong orientation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

SIFT descriptors

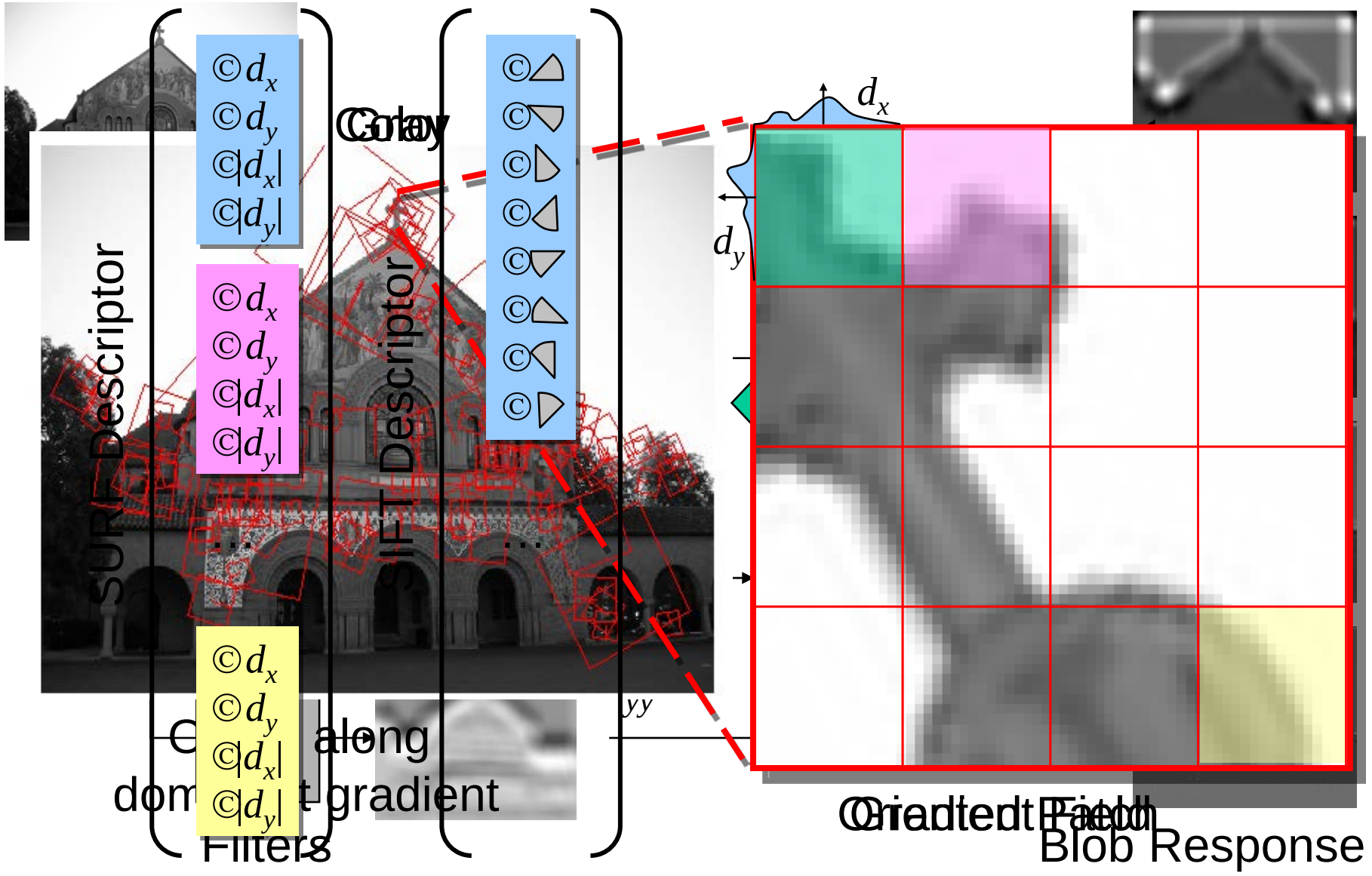
- SIFT - Scale-Invariant Feature Transform *[Lowe,1999, 2004]*
- Sample thresholded image gradients at 16x16 locations in scale space
  (in local coordinate system for rotation and scale invariance)
- For each of 4x4 subregion, generate orientation histogram with 8 directions each; each observation weighted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

SURF descriptors

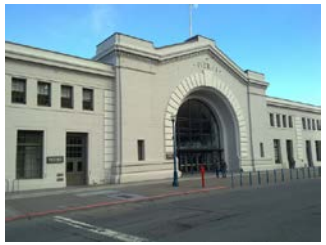
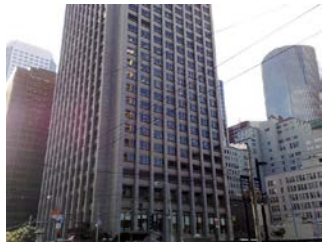
- SURF – Speeded Up Robust Features *[Bay et al. 2006]*
- Compute horizontal and vertical pixel differences, dx, dy (in local coordinate system for rotation and scale invariance, window size 20σ x 20σ, where σ^2 is feature scale)
- Sum dx, dy, and $|dx|$,$|dy|$ over 4x4 subregions (SURF-64) or 3x3 subregions (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

Which of the following statements are true?

(a) A bag of visual words representation is robust against partial occlusions of an object.

(a) The SIFT descriptor can only be calculated for SIFT keypoints. Similarly, the SURF descriptor can only be calculated for SURF keypoints.

(b) Both SIFT and SURF descriptors only depend on image gradients.

(c) The SIFT descriptor is more robust against image rotation since it uses an orientation histogram.

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)

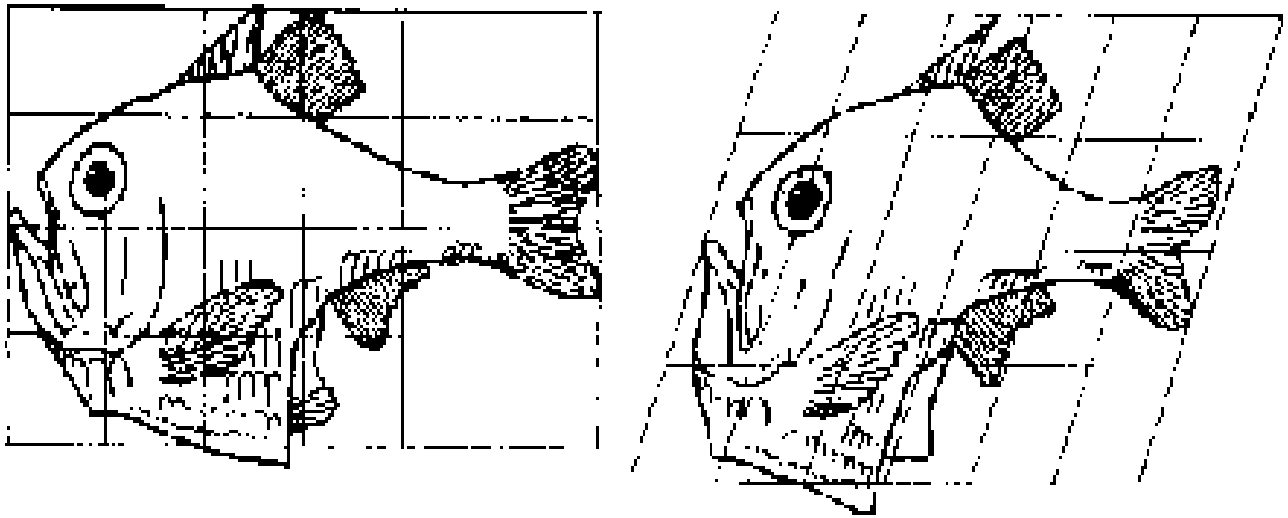
$$\mathbf{x}' = \begin{bmatrix} s\cdot\cos\theta & -s\cdot\sin\theta & t_x \\ s\cdot\sin\theta & s\cdot\cos\theta & t_y \end{bmatrix} \underline{\mathbf{x}}$$

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

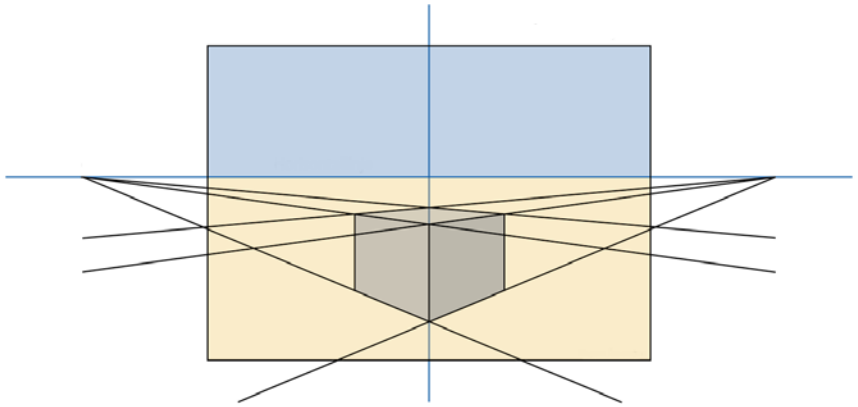


Argyropelecus olfersi. *Sternoptyx diaphana.*

Geometric mapping

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

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



- 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 $\sum$ from the corresponding keypoint, typical $\sum = 1 \ldots 3$ pixels)
- Repeat process S times, keep geometric maping with largest no. of inliers
- Required number of trials

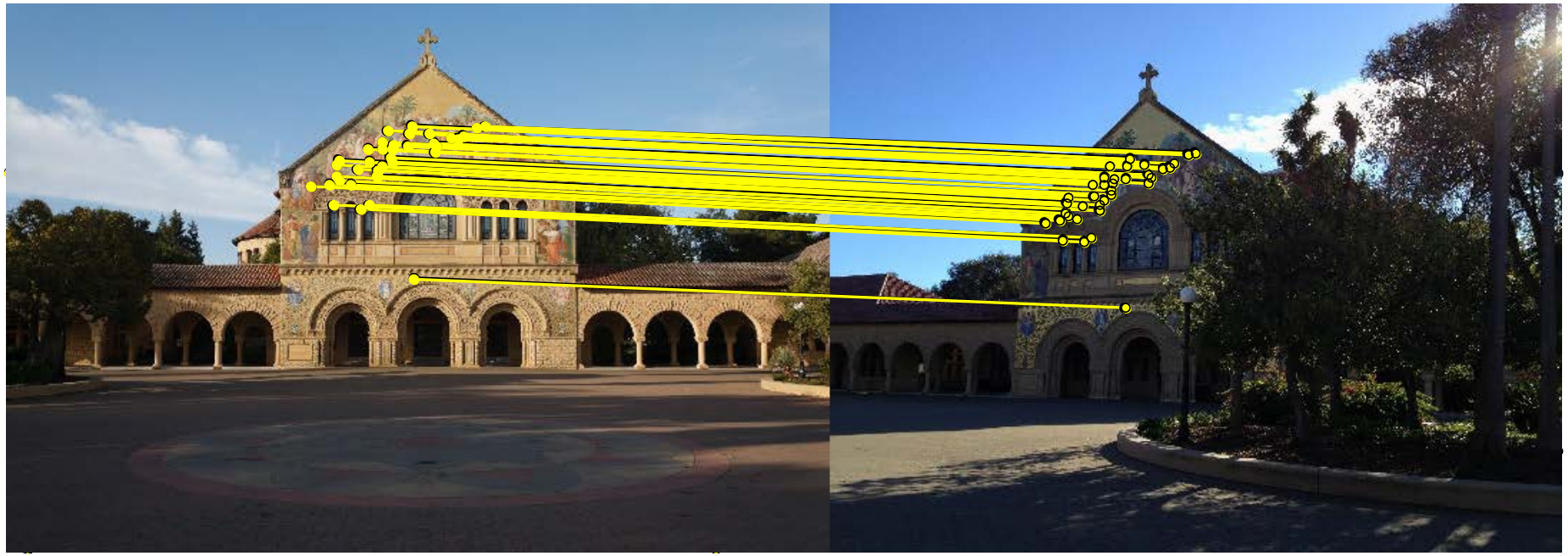
$$S = \frac{\log\left(1 - P\right)}{\log\left(1 - q^k\right)}$$

Total probability of success

Probability of valid correspondence

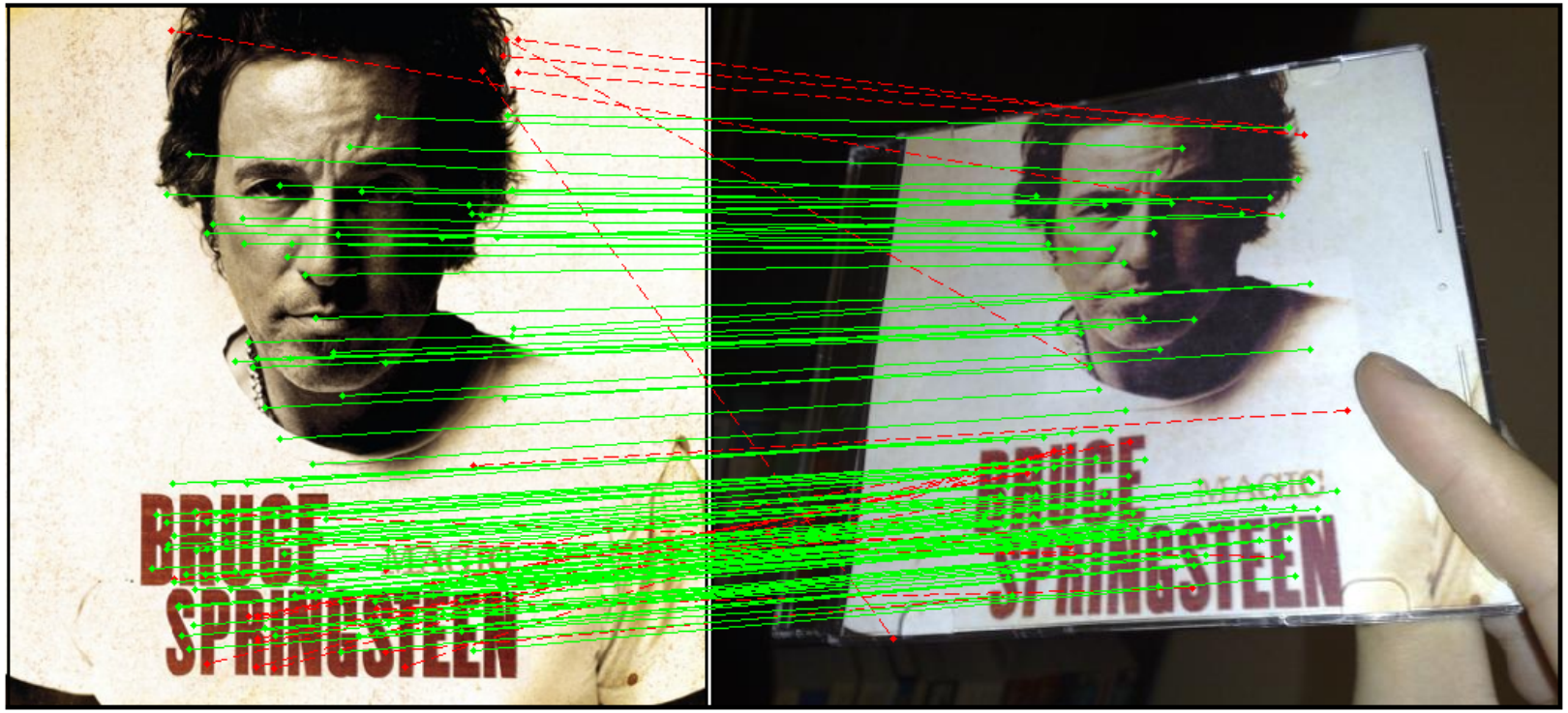
- Use small number of correspondences

RANSAC with Affine Model



RANSAC with Homography

SURF features & affine RANSAC



Which of the following statements are true?

(a) RANSAC is resilient against missing features, extraneous features, and noisy correspondences in a bag of visual words matching scenario.

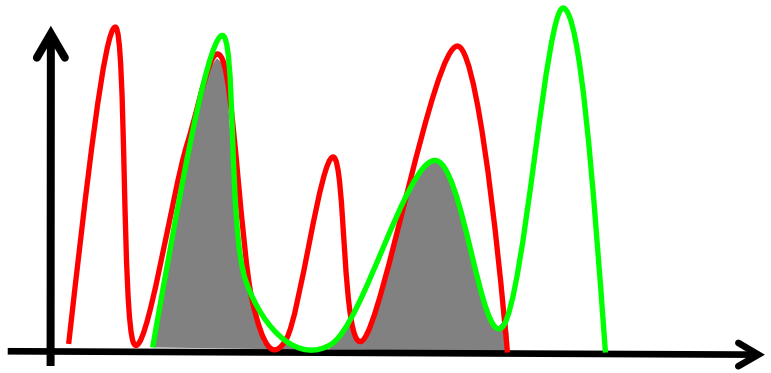
(b) An affine model contains a homography as a special case.

(c) RANSAC can only be applied if the number of inliers is larger than the number of outliers.

(d) For a fixed number of iterations in RANSAC, using a model with a larger number of parameters always increases the probability of success.

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

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

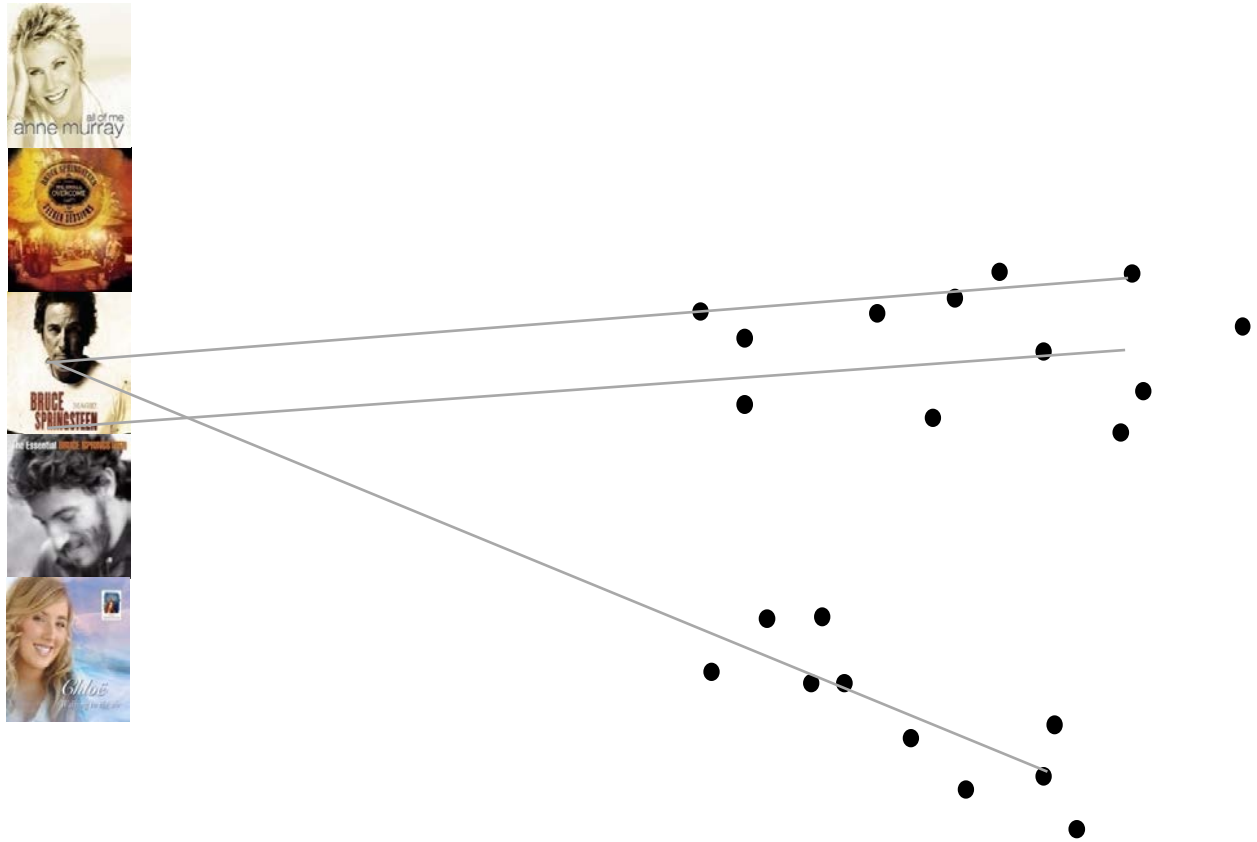
[Swain, Ballard 1991]

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

Growing Vocabulary Tree

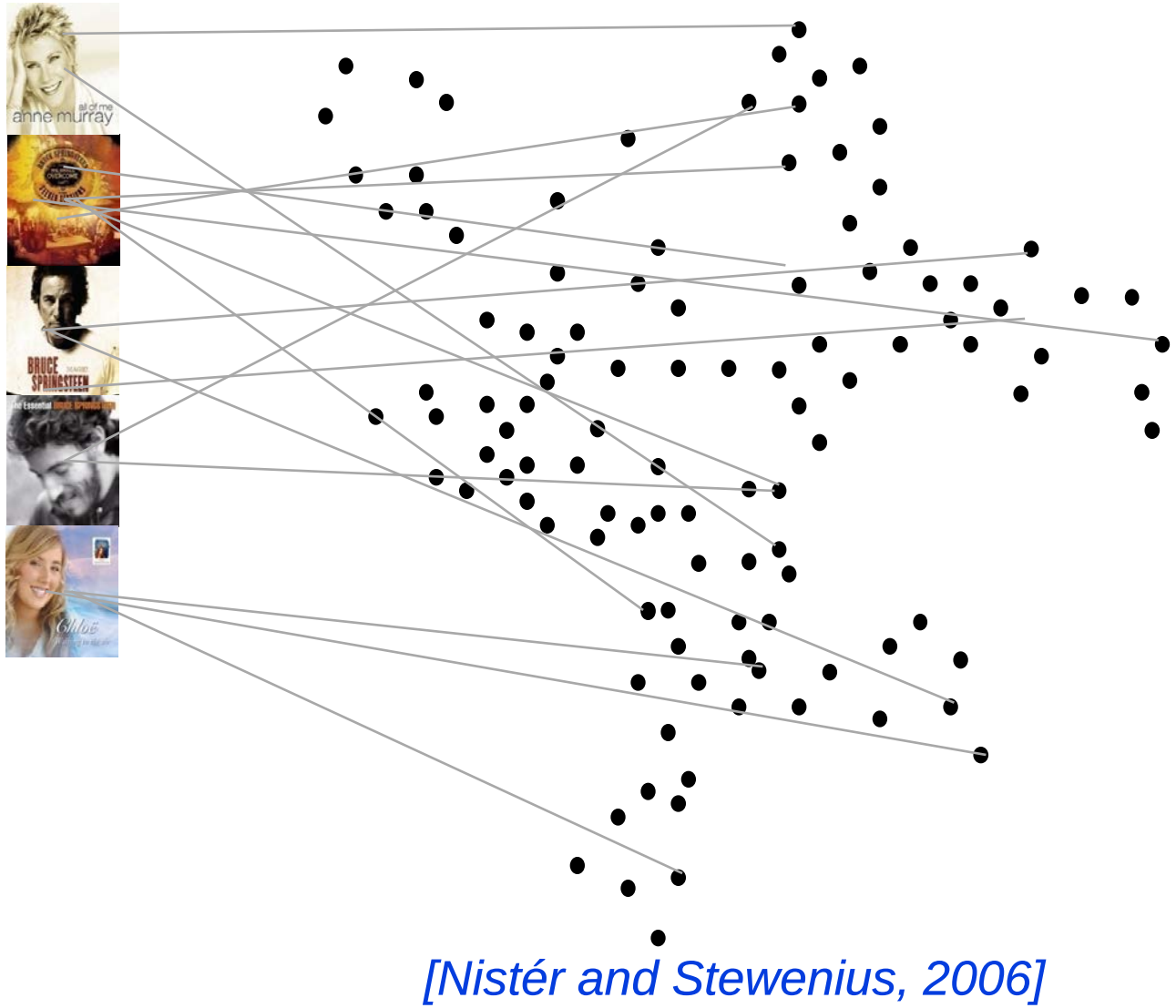
[Nistér and Stewenius, 2006]

Growing Vocabulary Tree



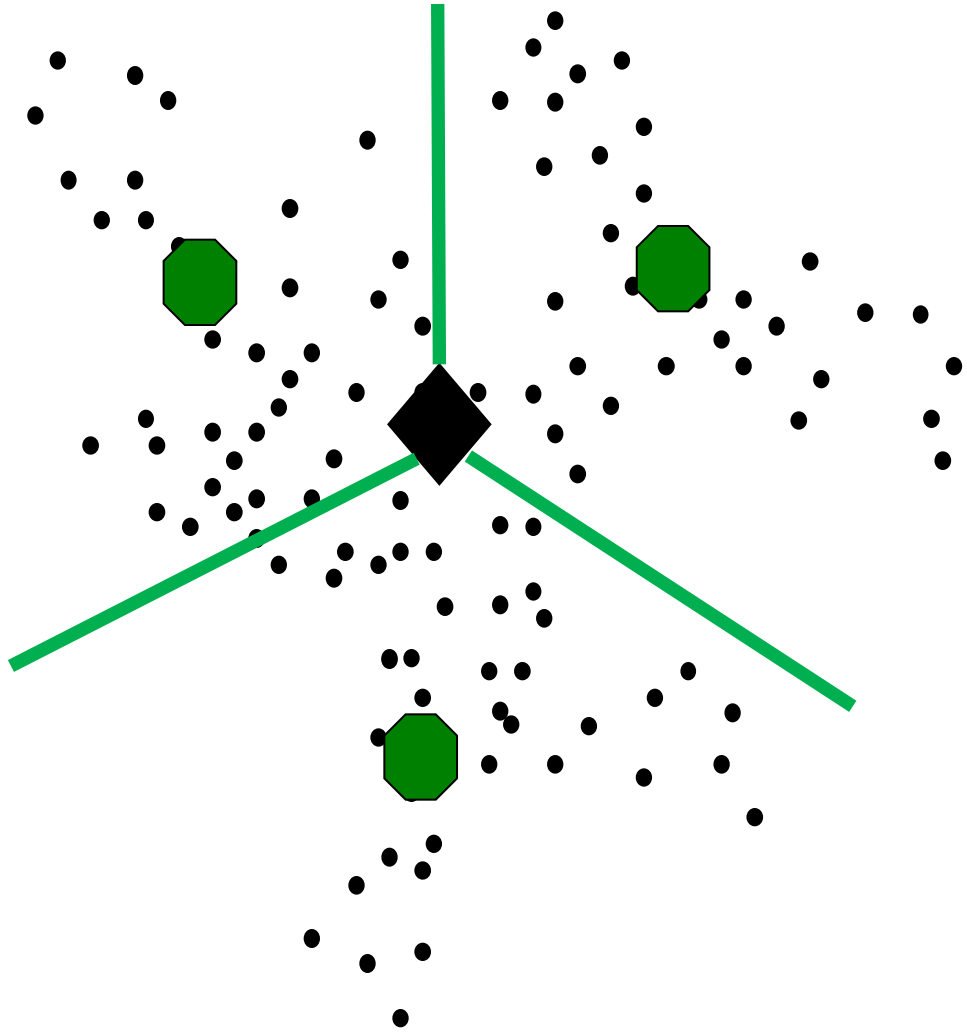
[Nistér and Stewenius, 2006]

Growing Vocabulary Tree



[Nistér and Stewenius, 2006]

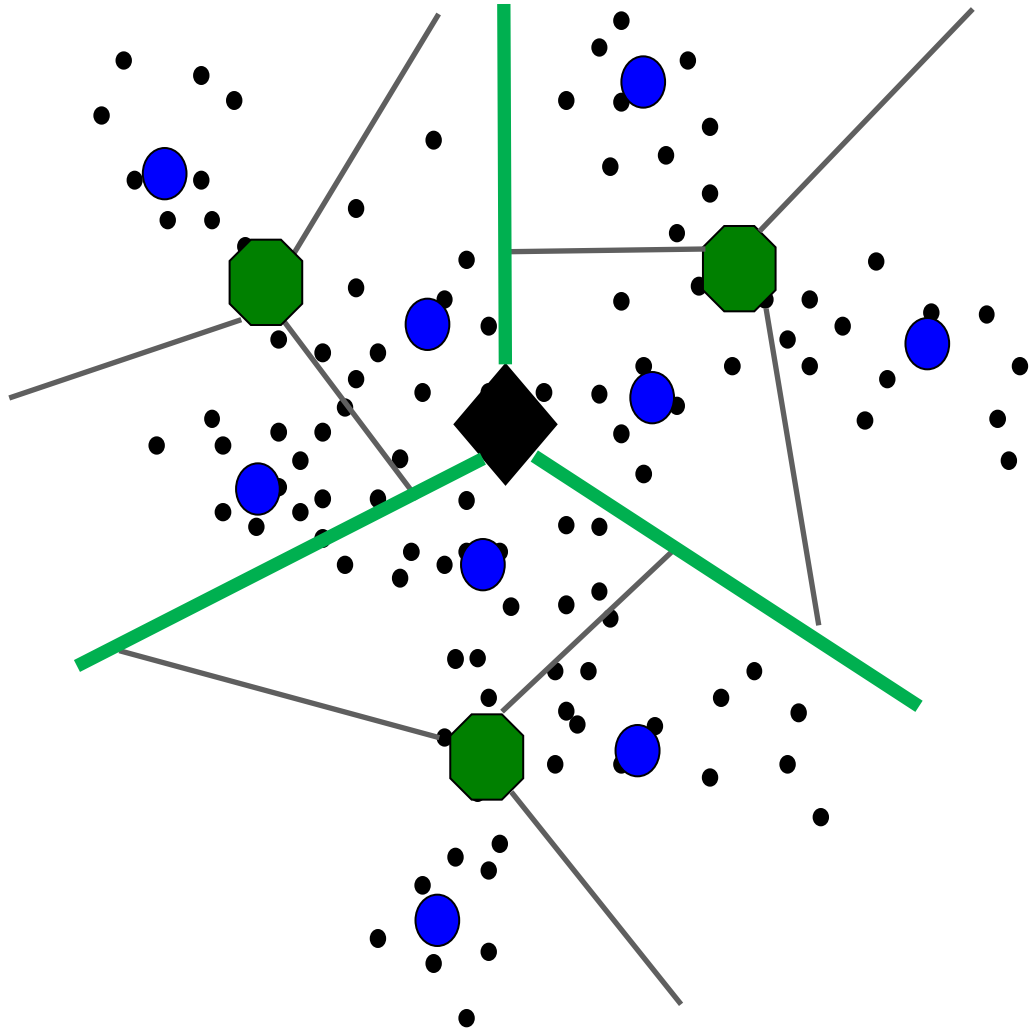
Growing Vocabulary Tree



$k = 3$

[Nistér and Stewenius, 2006]

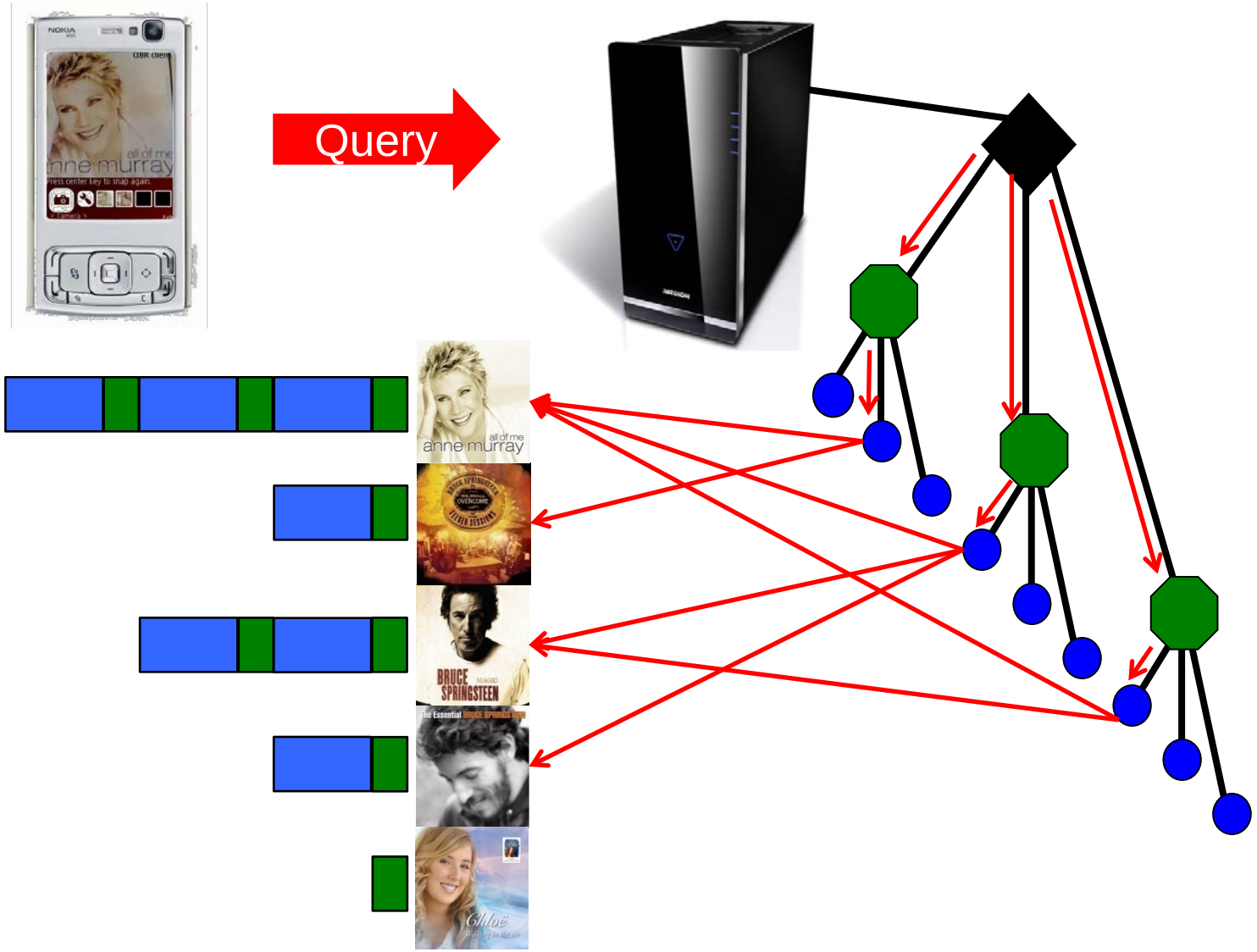
Growing Vocabulary Tree



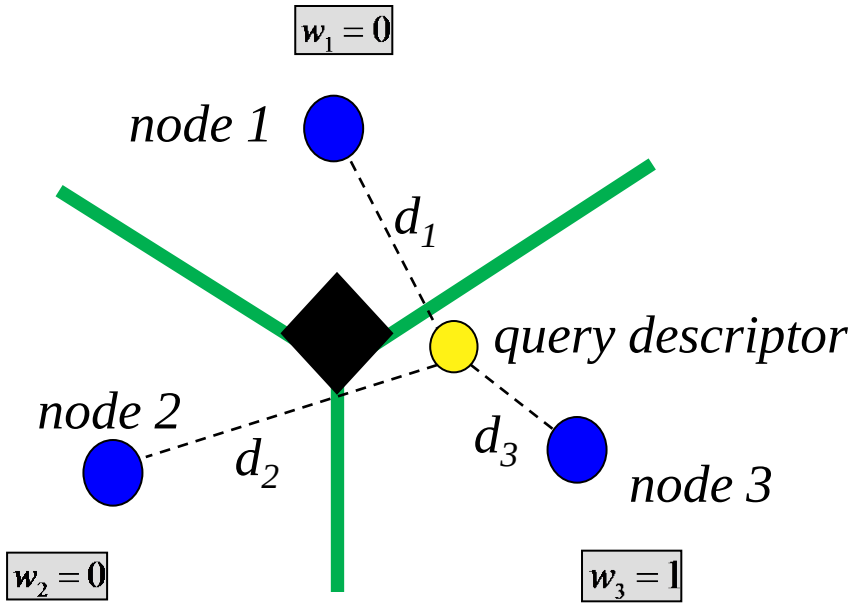
$k = 3$

[Nistér and Stewenius, 2006]

Querying Vocabulary Tree

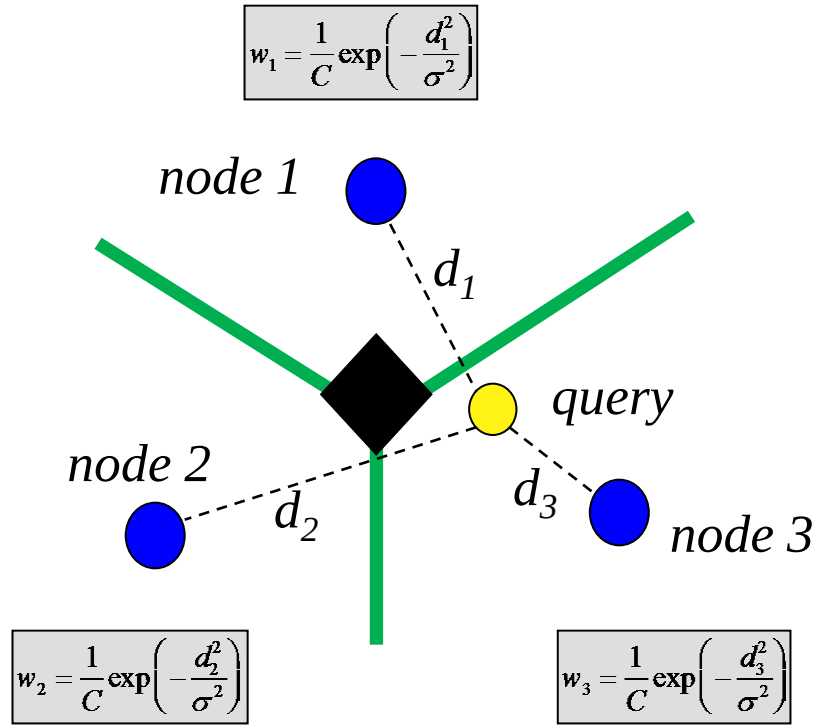


Querying: Hard Binning vs. Soft Binning



Hard Binning

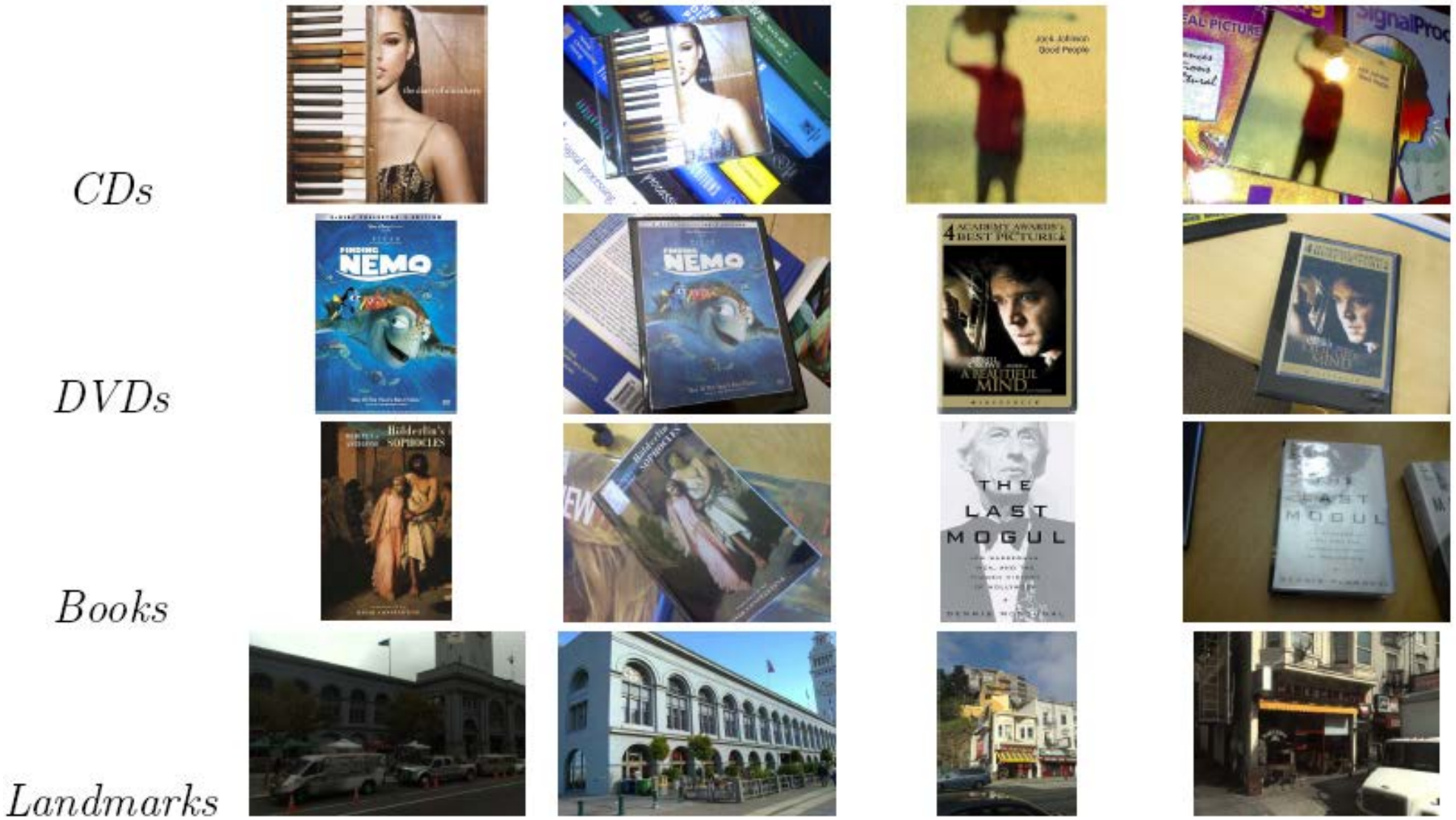
[Nistér and Stewenius, CVPR 2006]



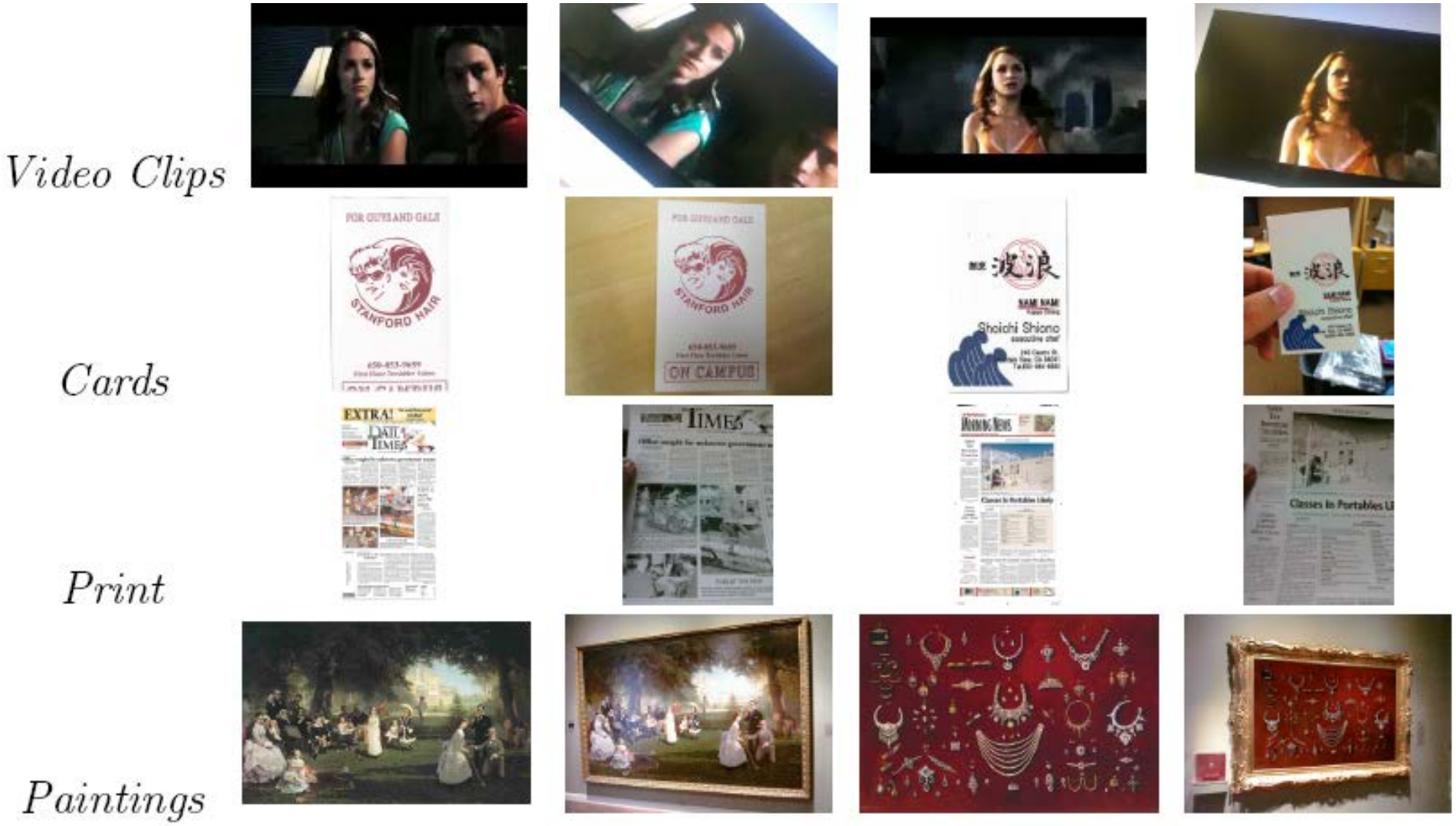
Soft Binning

[Philbin et al., CVPR 2008]

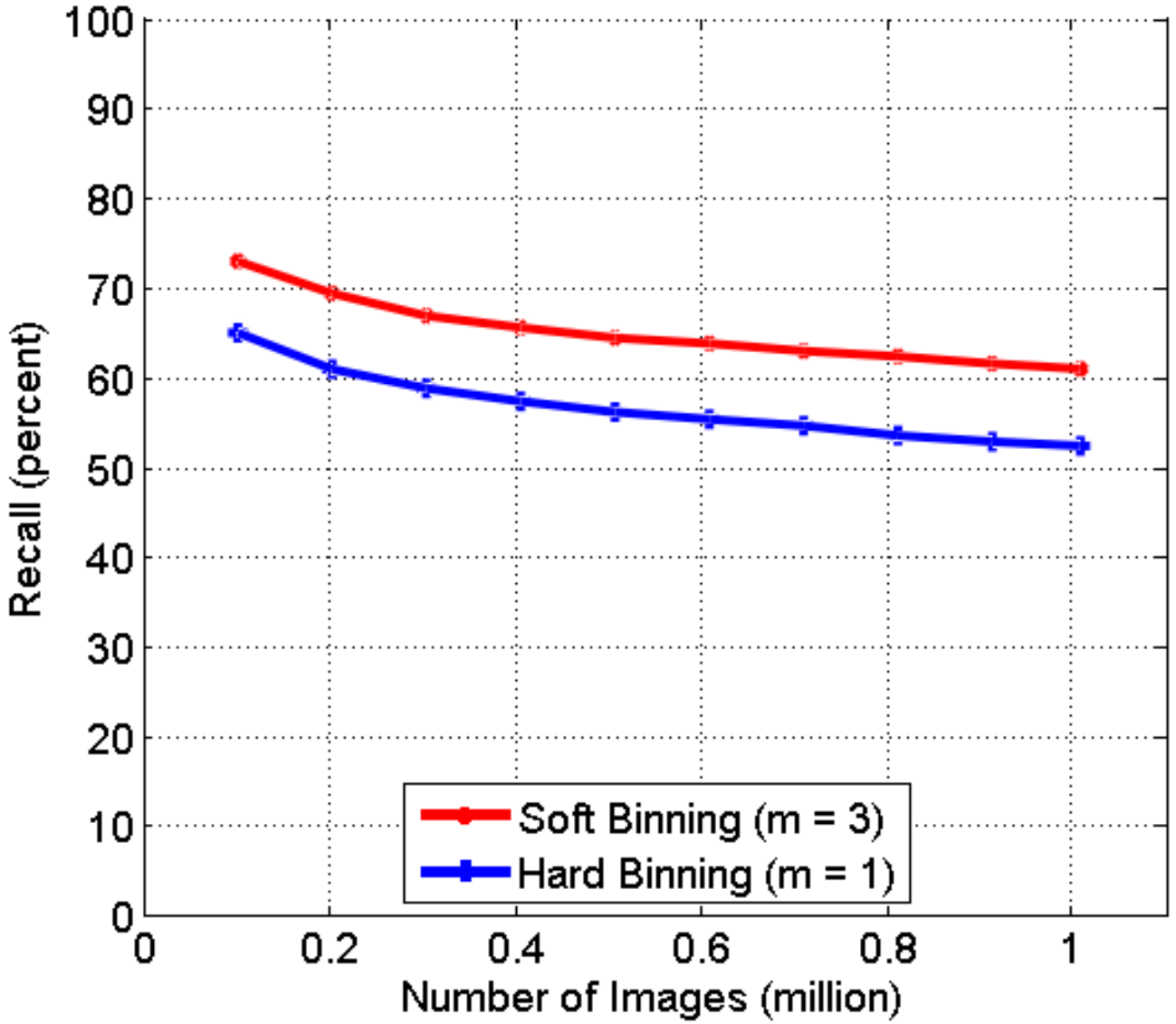
Stanford Mobile Visual Search Dataset



Stanford Mobile Visual Search Dataset



Querying: Hard Binning vs. Soft Binning



SURF features
6-level tree
1M leaf nodes
3269 query images
100 top tree results


Press select button to take photo.