Patch-based Representations

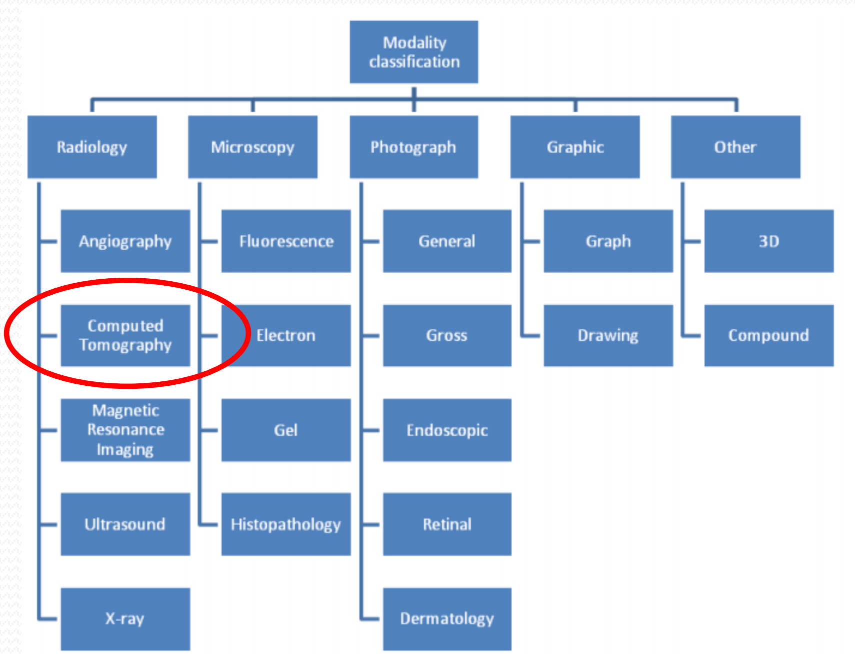
Assaf Cohen

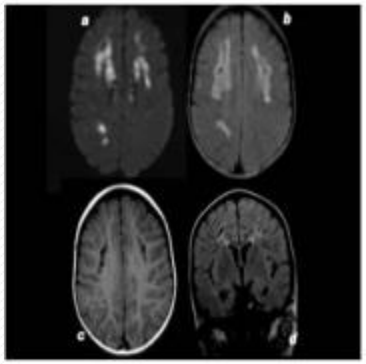
Agenda

- Introduction
  - Problem definition
  - Challenges
  - Relation to other computer vision problems
- Bag-of-words
- Implementation options
- Bag-of-words with spatial information
- Scalable vocabulary tree

Image Categorization

- **Image categorization** - labeling of images into one of a number of predefined categories.
- Example – Image categorization according to the type of the image





CT

Image Categorization (cont)

- Image categorization according to the content of the image

Dog

Cat

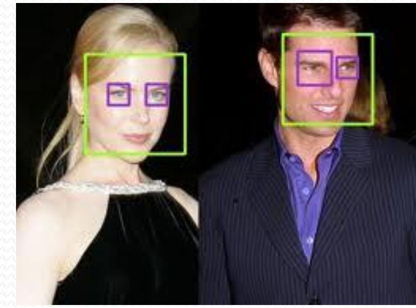
Dog?
Cat?
Animals...

Challenges in Image Categorization

- the appearance of object instances varies substantially owing to changes in:
  - Pose
  - Imaging and lighting conditions
  - Occlusions
  - Within-class shape variations

Related problems

- **Recognition** - identification of particular object instances

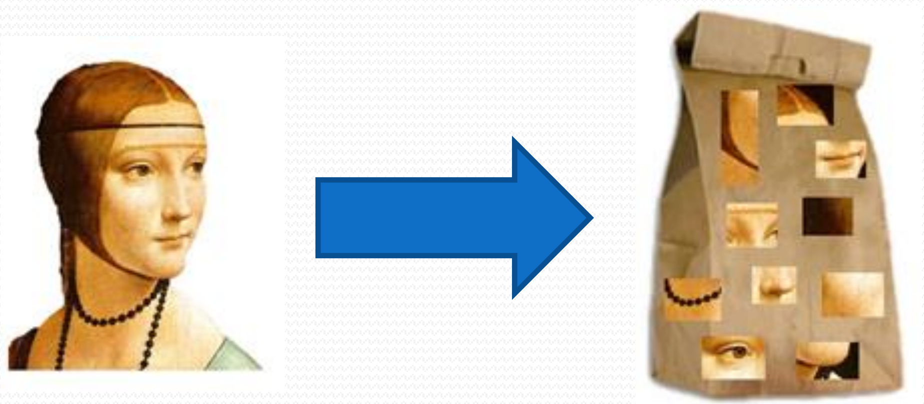


- **Content Based Image Retrieval** - retrieving images on the basis of low-level image features, given a query image.

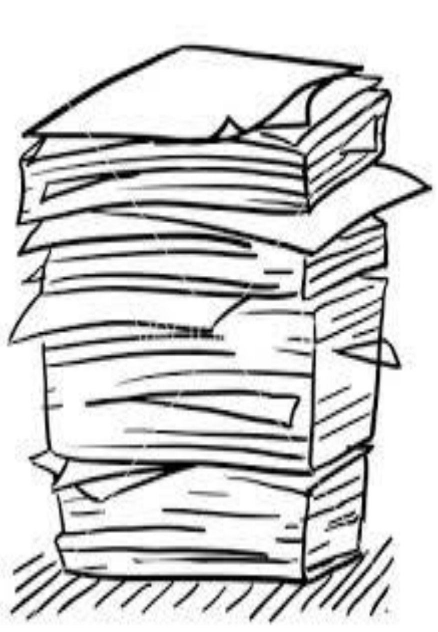
- **Detection** - deciding whether or not a member of *one visual category is* present in a given image.

Patch-based Approach

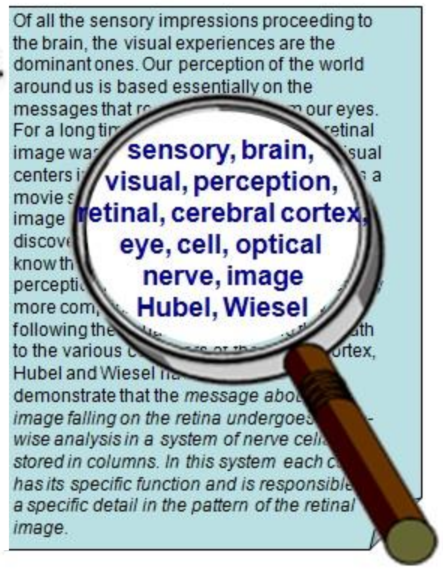
- Use **low-level features** to directly infer high-level semantic information about the scene without going through the intermediate step of segmenting the image into more “basic” semantic entities.



Motivation
Bag of words



Medicine



Business

Characterizing Colonic Detections in CT Colonography Using Curvature-Based Feature Descriptor and Bag-of-Words Model

Javed M. Aman, Ronald M. Summers, and Jianhua Yao

Radiology and Imaging Sciences Department, Clinical Center, National Institutes of Health, Bethesda, MD, USA

Abstract. We present a method based on the content-based image retrieval (CBIR) paradigm to enhance the performance of computer aided detection (CAD) in computed tomographic colonography (CTC). The method explores curvature-based feature descriptors in conjunction with bag-of-words (BoW) models to characterize colonic detections. The diffusion distance is adopted to improve feature matching and clustering. Word selection is also applied to remove non-informative words. A representative database is constructed to categorize different types of detections. Query detections are compared with the database for classification. We evaluated the performance of the system by using digital phantoms of common structures in the colon as well as real CAD detections. The results demonstrated the potential of our technique for distinguishing common structures within the colon as well as for classifying true and false-positive CAD detections.

Keywords: CAD, CT colonography, affine invariant feature, bag-of-words.

1 Introduction

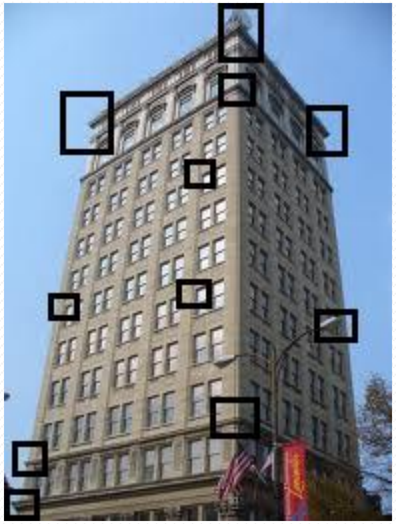
Cancer screening and early detection are an important step in colon cancer prevention. Optical colonoscopy (OC) is the traditional colon cancer screening procedure. However, because of its invasiveness, many patients forego this procedure. Computed tomographic colonography (CTC) has emerged as a minimally invasive screening procedure. CTC can benefit from CAD systems to improve the sensitivity and reduce the interpretation time [1]. Most CAD systems require post processing to reduce the number of false positives. We propose a method based on the Content-Based Image Retrieval (CBIR) paradigm to enhance the CAD performance.

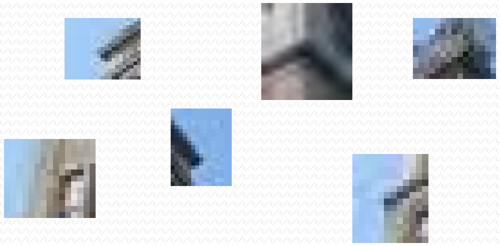
CBIR is a computer vision technique for searching for similar images within an image database. It has been used in applications such as medical image searching [2] and artwork retrieval [3]. The images in a CBIR system are characterized as a set of feature descriptors computed directly from the images. Detecting affine transformation-invariant salient feature points in an image is important for the success of a CBIR system. The scale-invariant feature transform (SIFT) proposed by Lowe [6] is one of such feature descriptors. However, images vary greatly in the number of

H. Yoshida et al. (Eds.): Virtual Colonoscopy and Abdominal Imaging 2010, LNCS 6668, pp. 15–23, 2011.
© Springer-Verlag Berlin Heidelberg 2011

Visual Categorization with Bags of Keypoints

Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, Cédric Bray (2004)

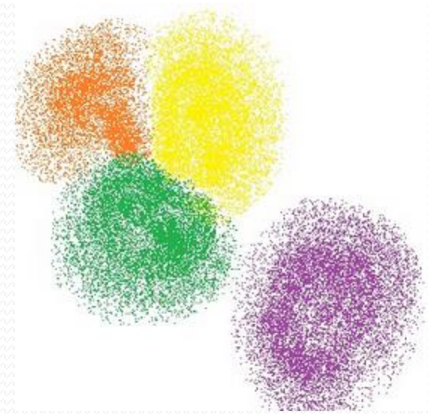


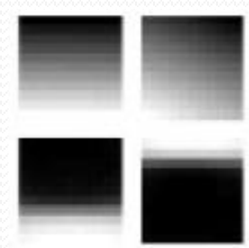


Method Overview

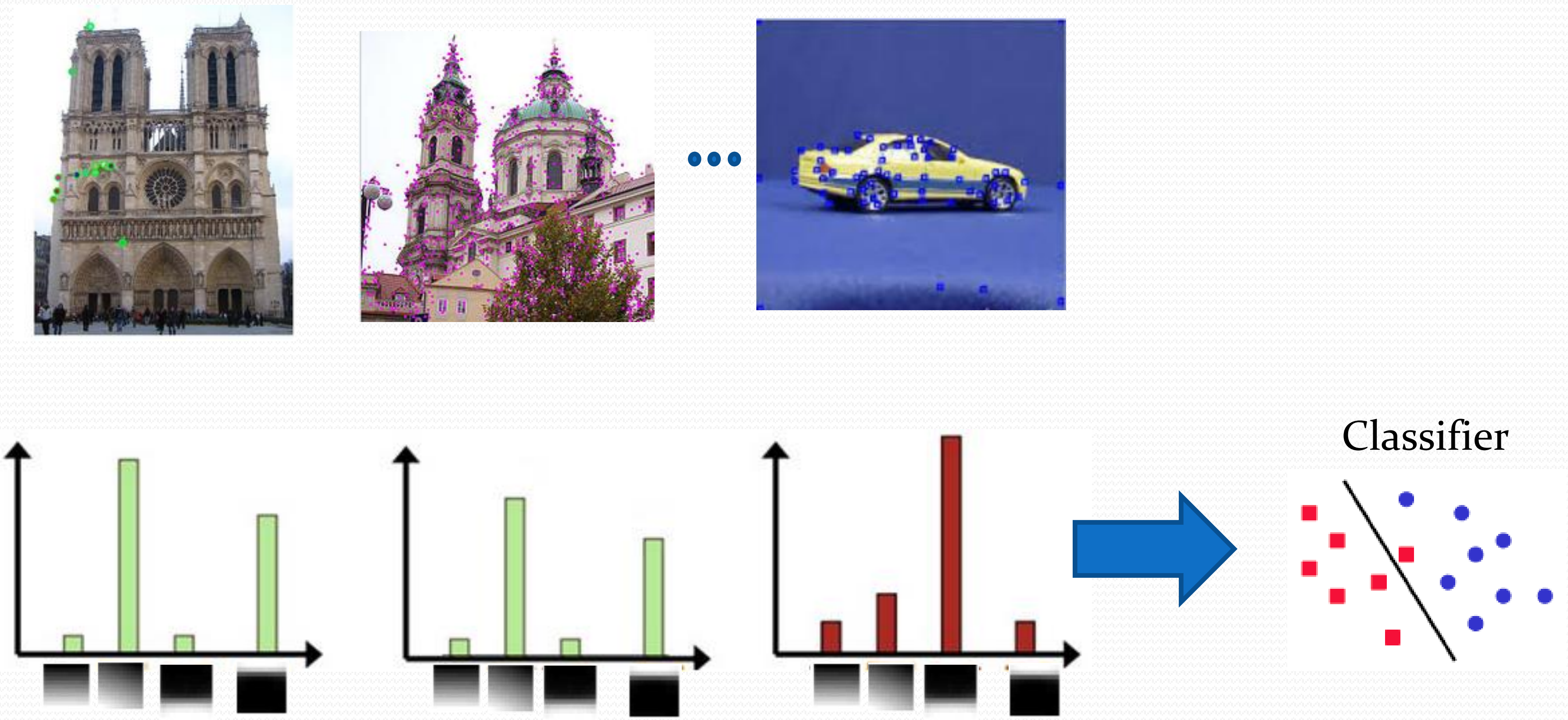
Learning – Vocabulary construction

Interesting point detection → Patch description → Bag of keypoints

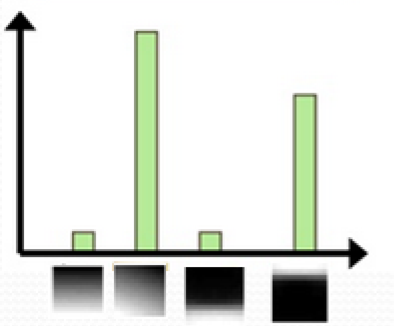




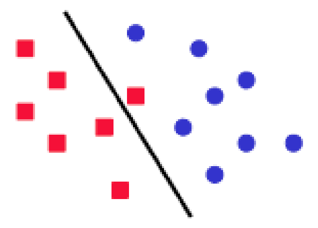
Method Overview
Learning – Training a classifier



Method Overview
Classification



Classifier



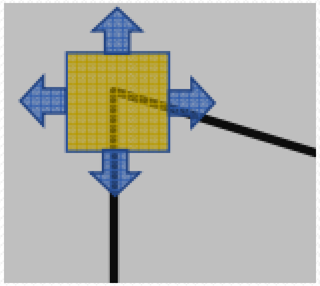
Building!!!

Bag of visual keypoints

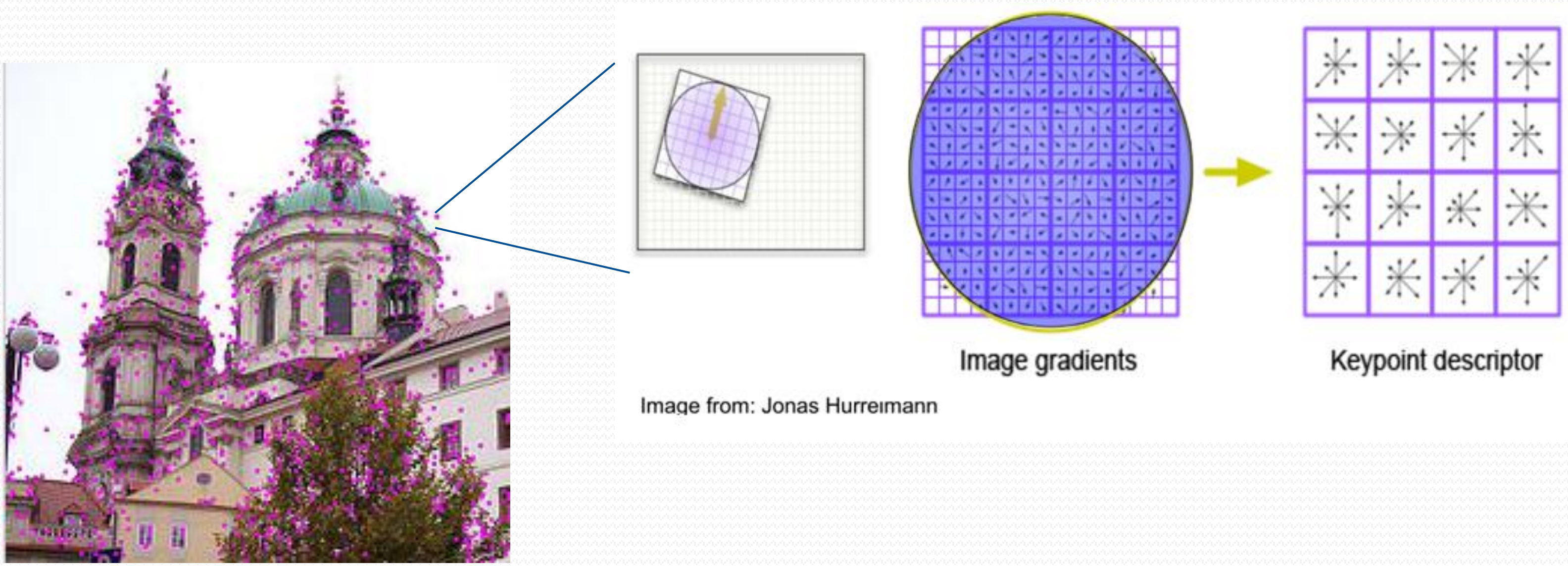
- Implementation choices:
  - How to sample the patches?
  - How to describe them?
  - How to quantify the resulting descriptor space distribution?
  - How to classify images based on the result?

Interesting point detection
Harris affine detector

- A descriptor that is **repeatable** – same descriptor to same patch invariant to scaling, rotation and shearing.
  - Finds Corners (significant intensity variation in every direction)
  - Scale invariant automatic scale selection
    by laplacian filter



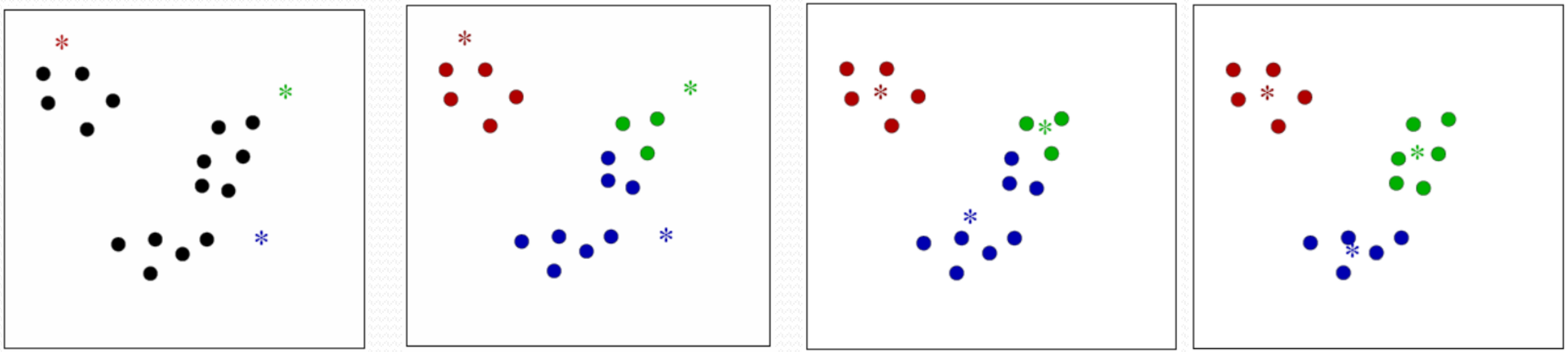
Patch descriptor
SIFT



- SIFT parameters:
  - Size of patch
  - Number of bines
  - Number of gradients per bin

Vocabulary construction *k-means*

- Clustering to k clusters.
  1. Select initial k centroids at random.
  2. Assign each object to the cluster with the nearest centroid.
  3. Compute each centroid as the mean of the objects assigned to it.
  4. Repeat previous 2 steps until no change.



Experiments

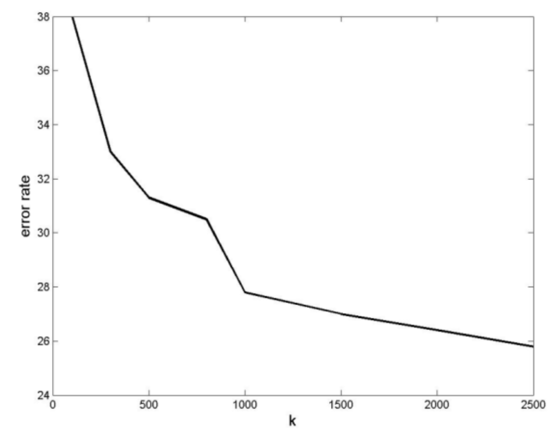
- Performance measures for classifiers:
  - **Confusion matrix** $$M_{ij} = \frac{\left|\left\{\mathbf{I_k} \in \mathbf{C_j} : \mathbf{h}(\mathbf{I_k}) = \mathbf{i}\right\}\right|}{\left|\mathbf{C_j}\right|}$$
  - **Overall error rate** $$R = \mathbf{1} - \frac{\sum_{\mathbf{j=1}}^{\mathbf{N_c}} \left|\mathbf{C_j}\right| M_{jj}}{\sum_{\mathbf{j=1}}^{\mathbf{N_c}} \left|\mathbf{C_j}\right|}$$
  - **Mean ranks** - mean position of the correct labels when labels output by the multi-class classifier are sorted by the classifier score.

Experiments

- Dataset
  - 7 classes
  - 1776 images
  - highly variable poses and significant amounts of background clutter

Experiments
Naïve Bayes

- Influence of the size of the vocabulary



True classes →	*faces*	*buildings*	*trees*	*cars*	*phones*	*bikes*	*books*
faces	**76**	4	2	3	4	4	13
buildings	2	**44**	5	0	5	1	3
trees	3	2	**80**	0	0	5	0
cars	4	1	0	**75**	3	1	4
phones	9	15	1	16	**70**	14	11
bikes	2	15	12	0	8	**73**	0
books	4	19	0	6	7	2	**69**
Mean ranks	1.49	1.88	1.33	1.33	1.63	1.57	1.57

Experiments
SVM

- Overall error dropped from 28% (by naïve bayes) to 15%
- Mean ranks approved

True classes →	*faces*	*buildings*	*trees*	*cars*	*phones*	*bikes*	*books*
faces	**98**	14	10	10	34	0	13
buildings	1	**63**	3	0	3	1	6
trees	1	10	**81**	1	0	6	0
cars	0	1	1	**85**	5	0	5
phones	0	5	4	3	**55**	2	3
bikes	0	4	1	0	1	**91**	0
books	0	3	0	1	2	0	**73**
Mean ranks	1.04	1.77	1.28	1.30	1.83	1.09	1.39

Challenging examples

- Images were classified correctly while background clutter is in a higher percentage than interest points on the object.



- Images where multiple objects were present

Examples of Failed Images

face(1)… book(2)	trees(1)…face(5)	phones(1)…cars(2)

Bag-of-keypoints

- Advantages:
  - Simplicity
  - No need for segmentation
  - computational efficiency
  - invariance to affine transformations, as well as occlusion, lighting and intra-class variations

bikes, buildings, cars

- Disadvantages:
  - No rigorous geometric information of the object components
  - Not robust when the object of interest is occupying a small fraction of the image.

Influence of the Sampling Method

Sampling Strategies for Bag-of-Features Image Classification (2006)

Interest Points

Harris-Laplace (HL)

Laplacian of Gaussian (LoG)

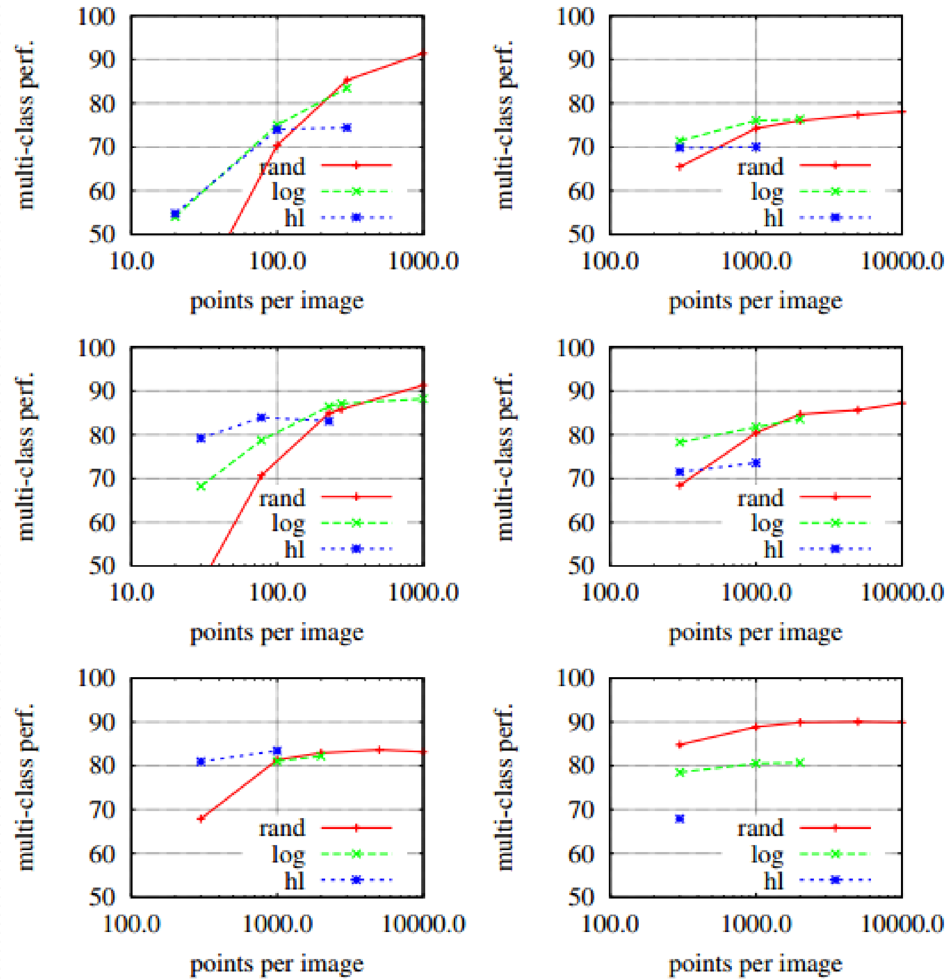
Random sampling

Influence of the Sampling Method

- Dense sampling –
  - Capture the most of the image information
  - Memory and compotation intensive
- Interest points -
  - Attractive because they are repeatable and invariant translations.
- Classification performance using the following methods:
  - LoG: Multi-scale keypoint detector
  - HL: The (non-affine) multi-scale keypoint detector
  - Random: patches are selected from a pyramid with regular grids in position and densely sampled scales.

Influence of the Sampling Method

- Test settings:
  - Vocabulary size of 1000 words using k-means
  - Linear SVM classifier
  - Datasets: Brodatz, Graz01; KTH-TIPS, Pascal-01, UIUCTex and Xerox7.



Influence of the Sampling Method Conclusion

- The number of returned keypoints is limited
  - The number of keypoints from the detectors is controlled using their ‘cornerness’ thresholds.
  - Even when setting the threshold to zero there is a limit to the returned keypoints
- **They simply can not sample densely enough to produce leading-edge classification results.**

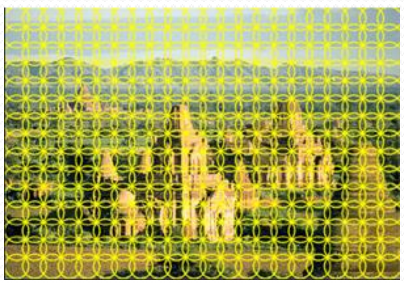
"Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories"

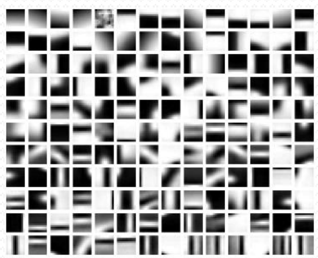
Svetlana Lazebnik, Cordelia Schmid, Jean Ponce (2006)

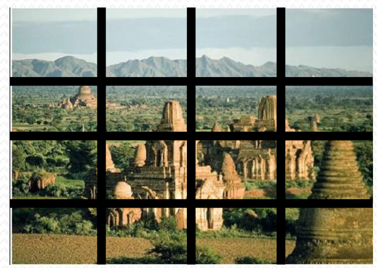
- Extension of a bag of features.
- Incorporate geometric information into bag-of-words - **Locally** orderless representation at several levels of resolution.

Algorithm Overview

Keypoint Descriptor → Word Dictionary → Spatial Histogram → Histogram Intersection → Training

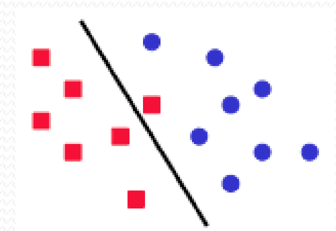


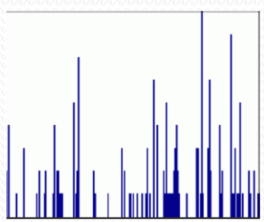


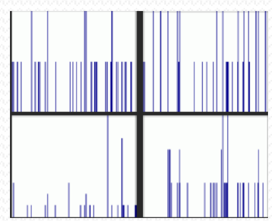


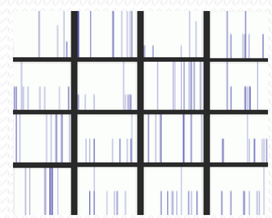
$\kappa^L(X,Y)$

SVM Classifier

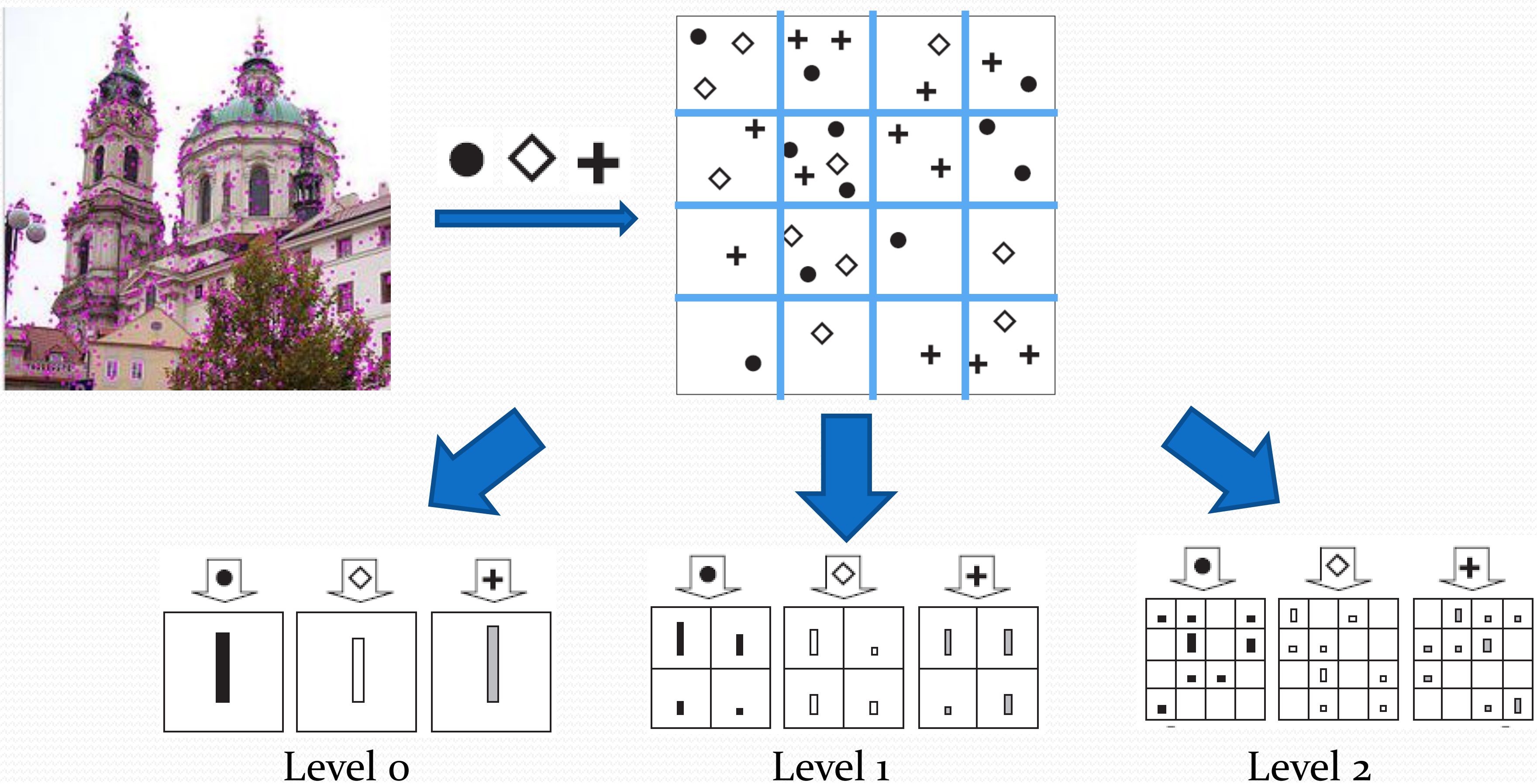






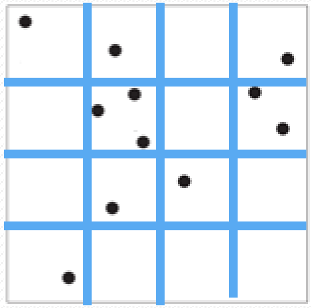


Spatial pyramid representation



Pyramid Match Kernels

- Define a kernel based on the pyramid to be used in SVM.
- X vector in *d*-dimensional feature space.
- Resolution level *l*, $D = 2^{d\ell}$ Cells.
- H_X^ℓ - histogram of X at resolution *l*.
- $H_X^\ell(i)$ - the number of points that falls into the *i*th cell.
- The number of matches at level *l*:



$$\mathcal{I}^\ell \equiv \mathcal{I}(H_X^\ell, H_Y^\ell) = \sum_{i=1}^{D} \min\left(H_X^\ell(i), H_Y^\ell(i)\right)$$

Pyramid Match Kernels

- Number of new match found in level *l*: $\mathcal{I}^{\ell} - \mathcal{I}^{\ell+1}$
- Weight of level *l*: $\frac{1}{2^{L-\ell}}$

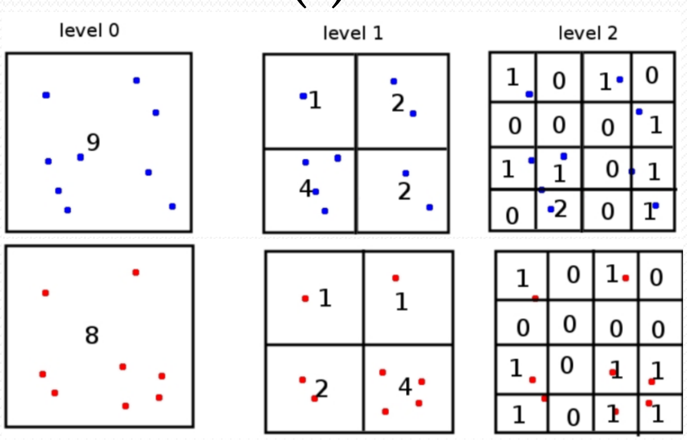
$$\begin{aligned}
\kappa^L(X,Y) &= \mathcal{I}^L + \sum_{\ell=0}^{L-1} \frac{1}{2^{L-\ell}} \left(\mathcal{I}^{\ell} - \mathcal{I}^{\ell+1}\right) \\
&= \frac{1}{2^L}\mathcal{I}^0 + \sum_{\ell=1}^{L} \frac{1}{2^{L-\ell+1}}\mathcal{I}^{\ell} .
\end{aligned}$$

Spatial Matching Scheme

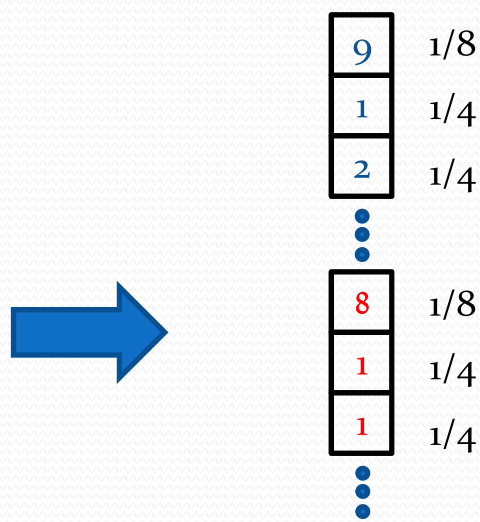
- The final kernel is the sum of separate kernels of each word *m*:

$$K^L(X,Y) = \sum_{m=1}^{M} \kappa^L(X_m, Y_m)$$

- K is a single histogram intersection of vectors formed by concatenating the histograms of all words (*m*) at all channels (*l*)

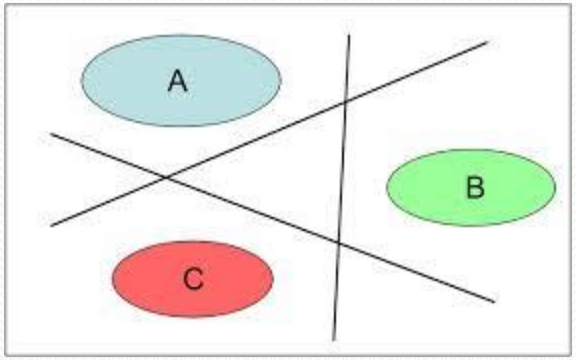


1/8 1/4 1

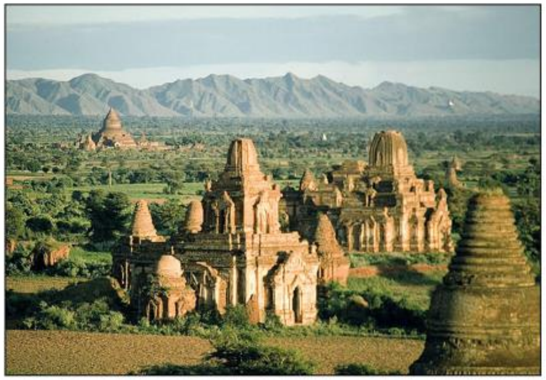


Training

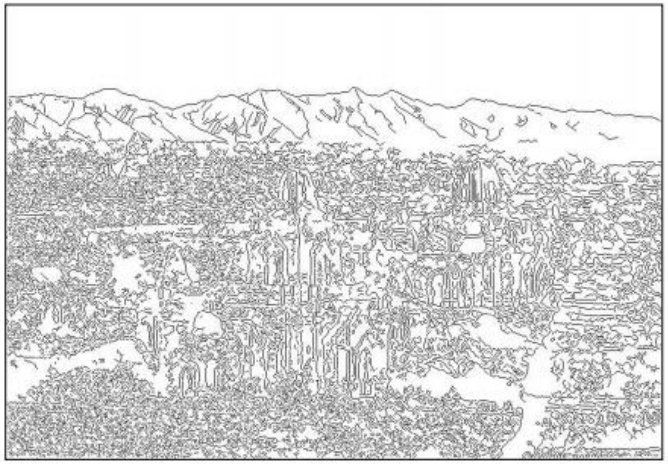
- Classes C1....Cn
- Multi-class classification: SVM one-versus-all
- Classifier for each class, Ci, trained with 2 classes
  - The class Ci
  - The union of {C1...C,}\{Ci}
- Test image is assigned with the class of the classifier with the highest response.



Evaluation
Feature Types

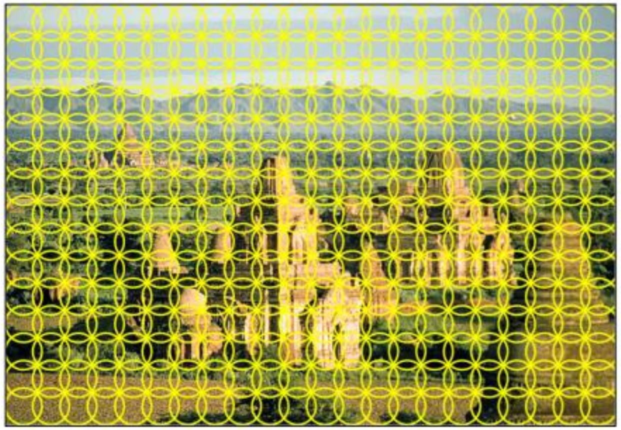


Weak features



Edge points at 2 scales and 8 orientations
(vocabulary size 16)

Strong features



SIFT descriptors of 16x16 patches sampled on a regular grid, quantized to form visual vocabulary (size 200, 400)

Feature Types

- **Caltech-101**
  - 101 categories
  - 31- 800 images per class
  - The objects occupy most of the image

	Weak features		Strong features (200)	
L	Single-level	Pyramid	Single-level	Pyramid
0	15.5 ±0.9		41.2 ±1.2	
1	31.4 ±1.2	32.8 ±1.3	55.9 ±0.9	57.0 ±0.8
2	47.2 ±1.1	49.3 ±1.4	63.6 ±0.9	**64.6** ±0.8
3	52.2 ±0.8	**54.0** ±1.1	60.3 ±0.9	64.6 ±0.7

Size of the vocabulary

- Dataset
  - 15 classes
  - 200-400 images for each class
- 100 images per class for training
- The rest for test

	Weak features ($M = 16$)		Strong features ($M = 200$)		Strong features ($M = 400$)	
L	Single-level	Pyramid	Single-level	Pyramid	Single-level	Pyramid
0 (1×1)	45.3 ±0.5		72.2 ±0.6		74.8 ±0.3	
1 (2×2)	53.6 ±0.3	56.2 ±0.6	77.9 ±0.6	79.0 ±0.5	78.8 ±0.4	80.1 ±0.5
2 (4×4)	61.7 ±0.6	64.7 ±0.7	79.4 ±0.3	**81.1** ±0.3	79.7 ±0.5	**81.4** ±0.5
3 (8×8)	63.3 ±0.8	**66.8** ±0.6	77.2 ±0.4	80.7 ±0.3	77.2 ±0.5	81.1 ±0.6

Examples

- spatial pyramids seem successful at capturing the organization of major pictorial elements or “blobs”

office

inside city

Limitations

- Textureless animals

cougar body (27.6%)

beaver (27.5%)

- Animals that camouflage well in their environment

crocodile (25.0%)

- “Thin” objects

ant (25.0%)

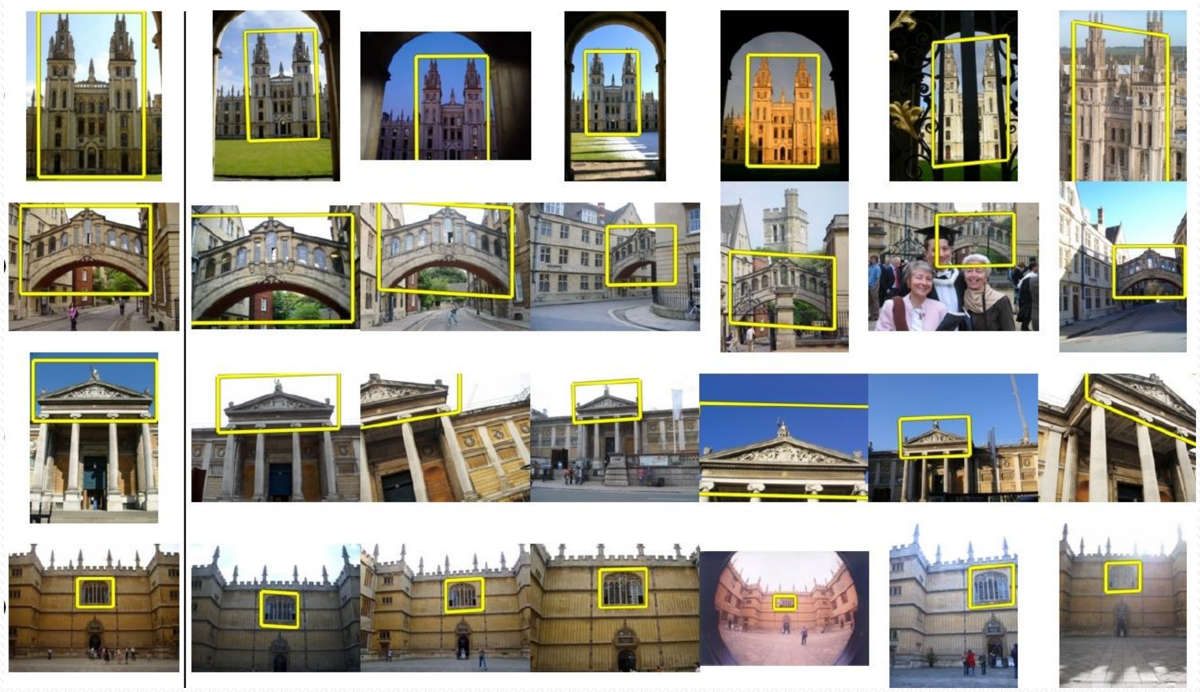
Method Conclusion

- “holistic" approach for categorization.
- Simple method.
- Gives better results than bag-of-features

Scalable Recognition with a Vocabulary Tree

David Nist´er and Henrik Stew´enius (2006)

- **Recognition** method that handles large databases (50000 images).

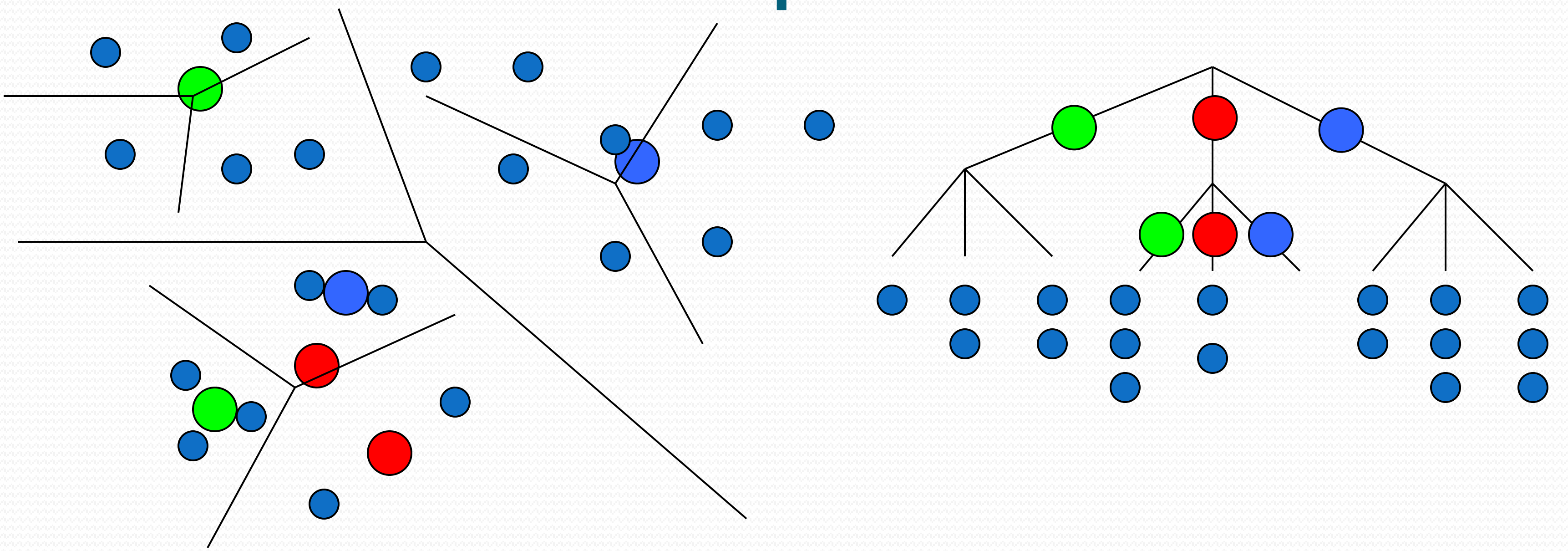


Query Results on 5K (demo available for 100K)

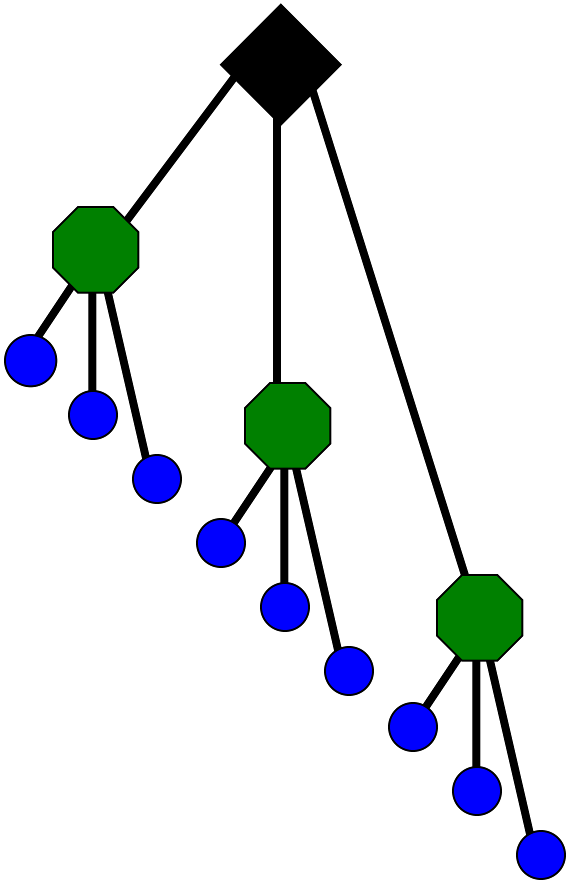
What is a Vocabulary Tree?

- Hierarchical quantization that is built by hierarchical k-means clustering.
- Instead of k defining the final number of clusters or quantization cells, k defines the branch factor of the tree.

What is a Vocabulary Tree? Example



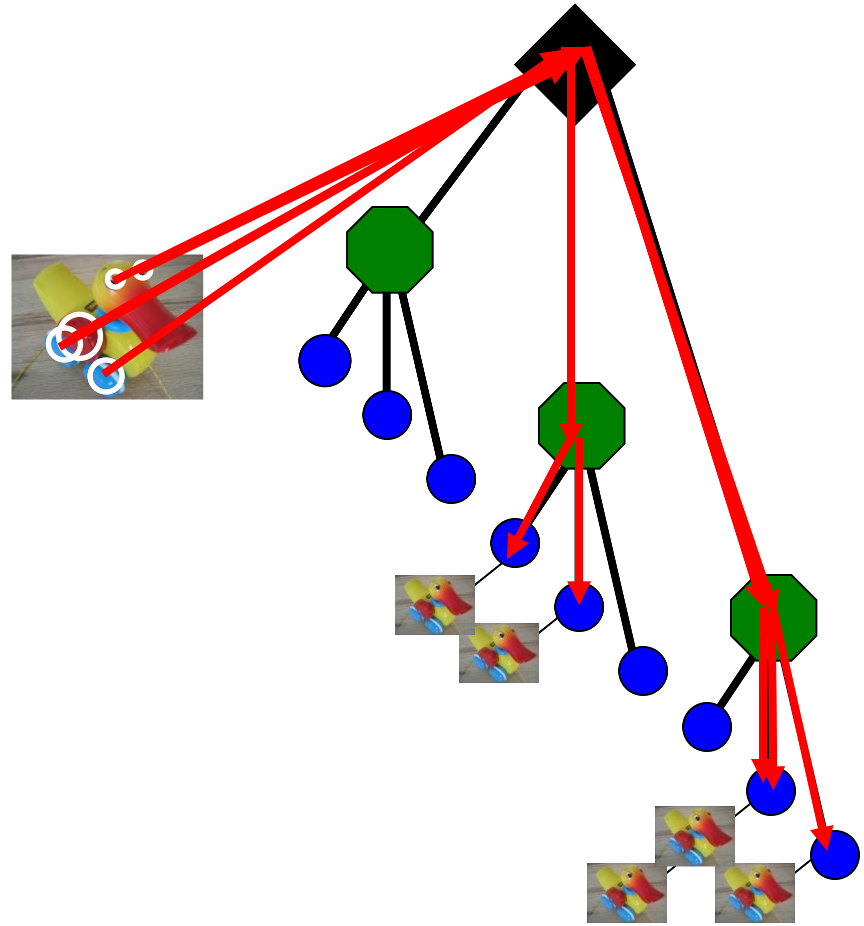
Online Phase



Vocabulary tree/inverted index

Slide credit: D. Nister

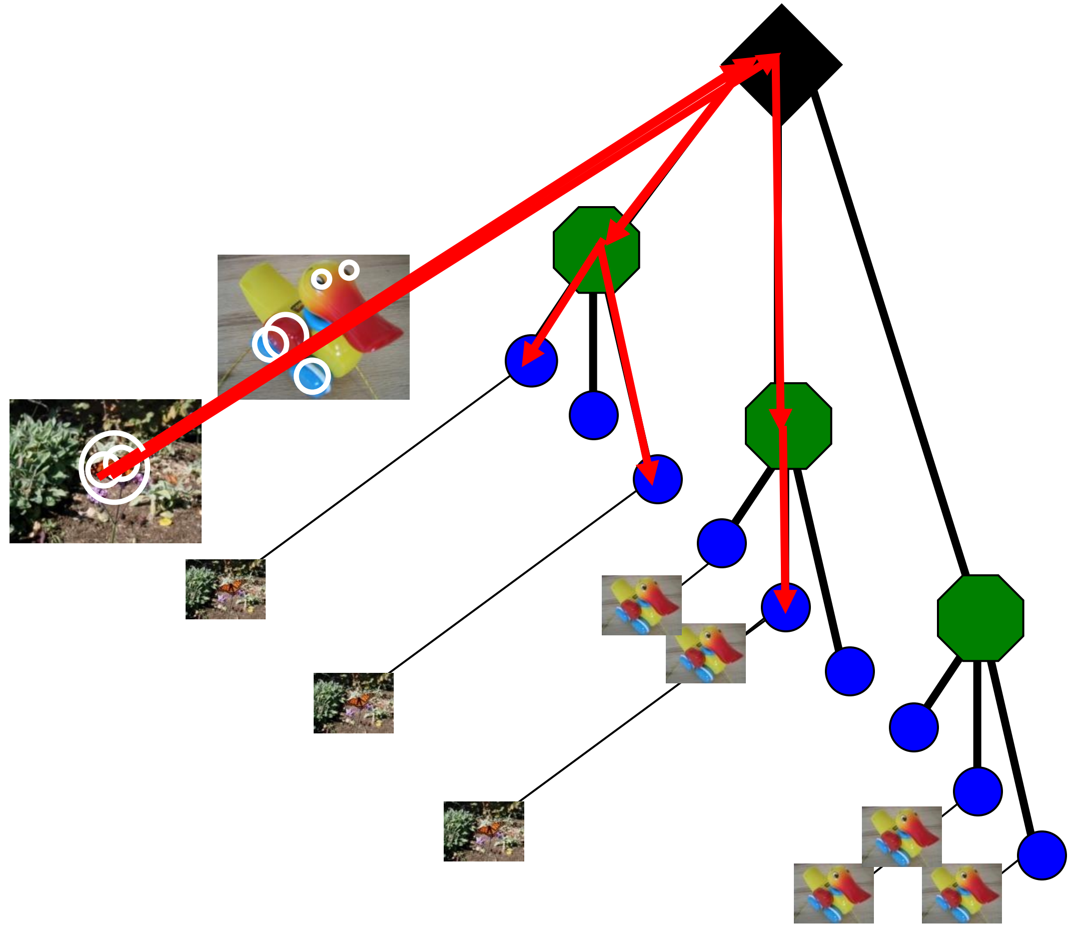
Model images



Populating the vocabulary tree/inverted index

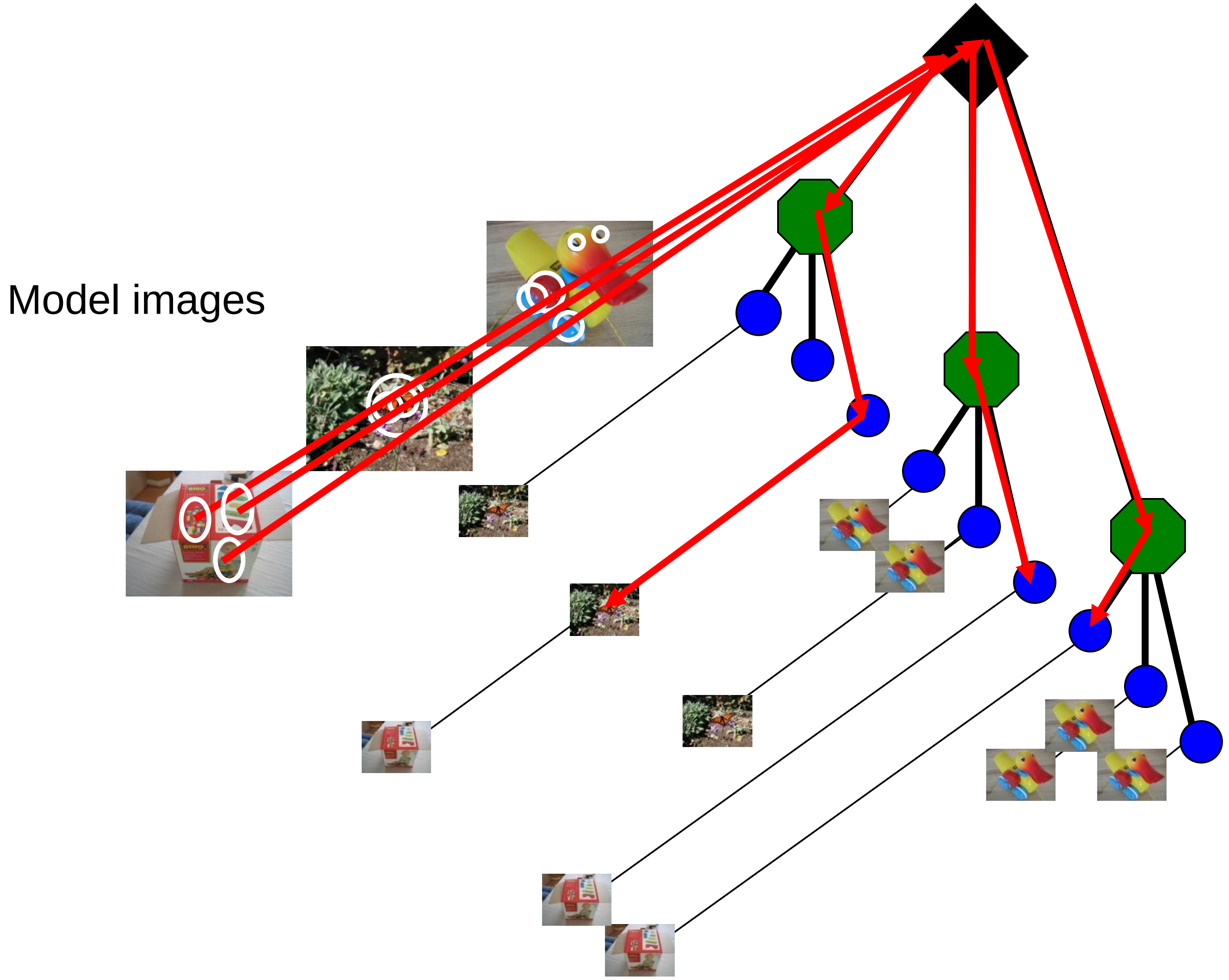
Slide credit: D. Nister

Model images



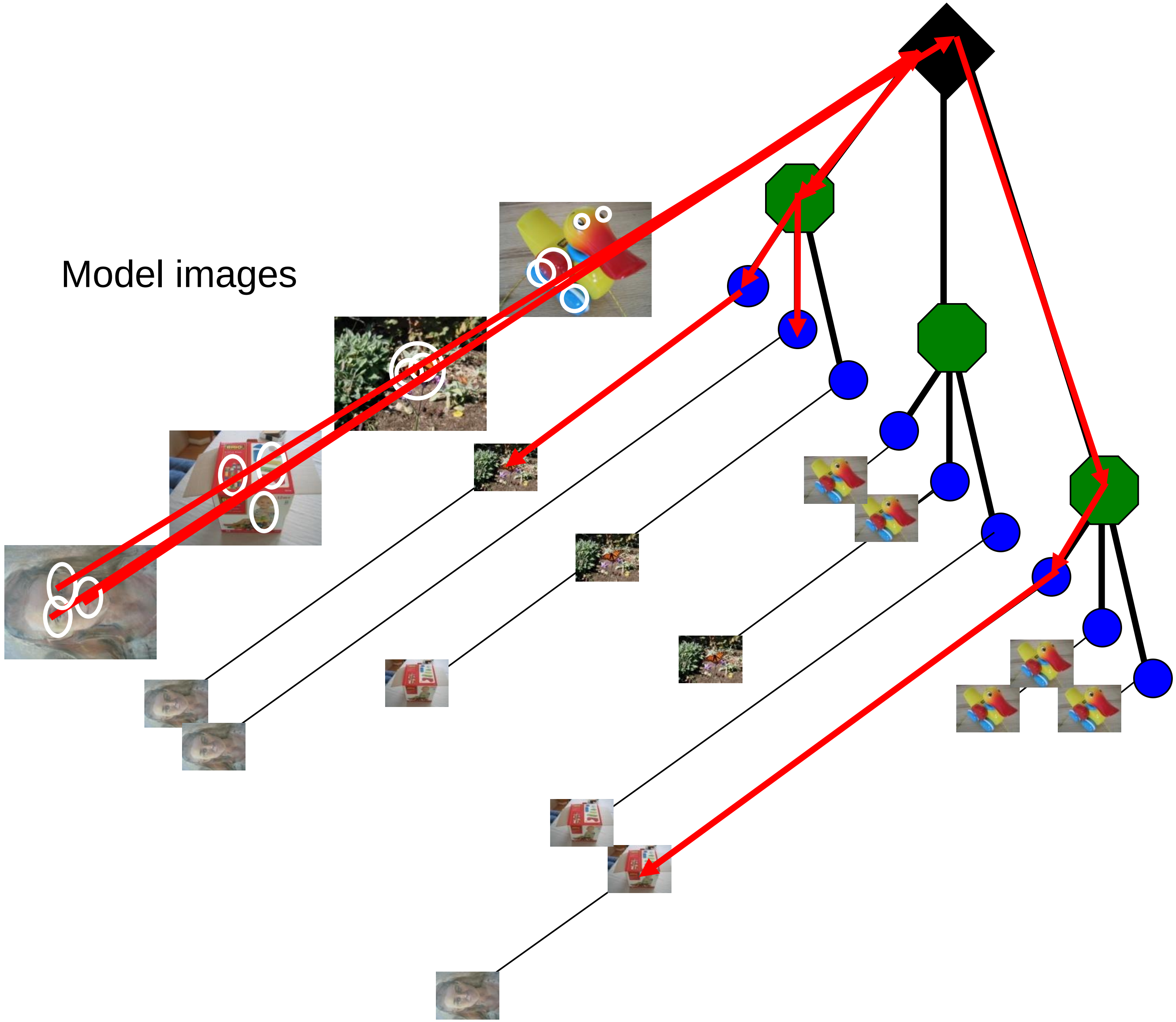
Populating the vocabulary tree/inverted index

Slide credit: D. Nister



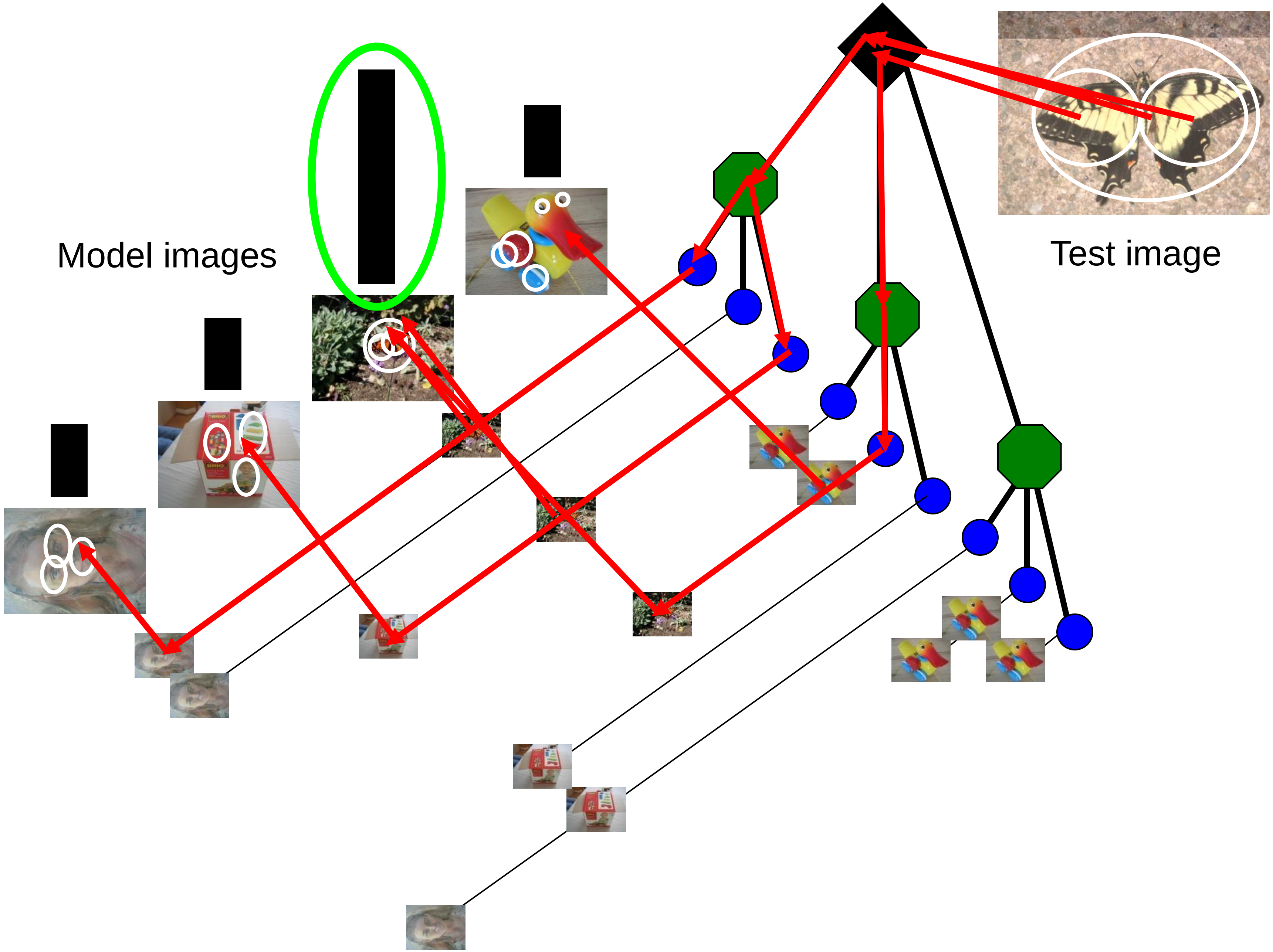
Populating the vocabulary tree/inverted index

Slide credit: D. Nister



Populating the vocabulary tree/inverted index

Slide credit: D. Nister

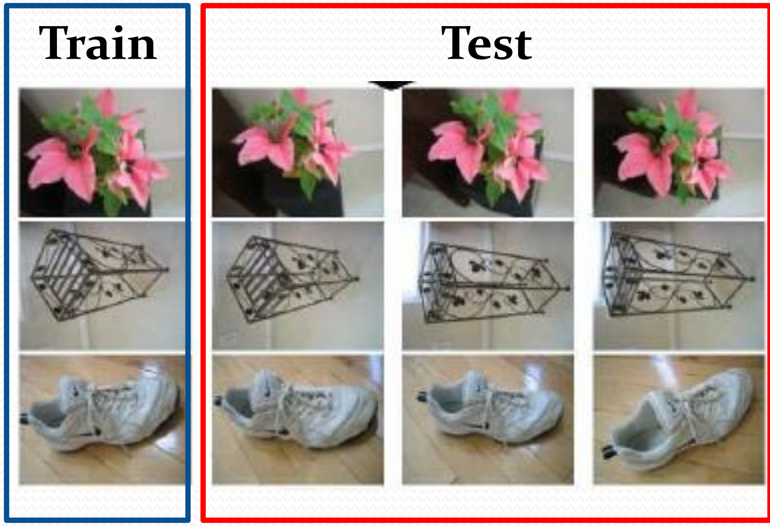


Looking up a test image

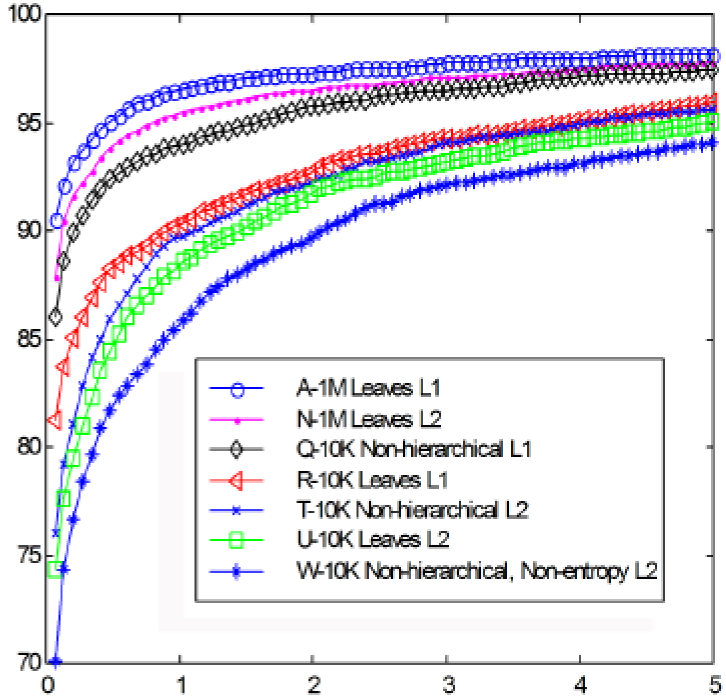
Slide credit: D. Nister

Evaluation

- 6376 images in groups of 4 (1594 objects)
- Tree built from every image and tested with the other three images in the block.



y percentage of the ground truth query images that make it into the top



x percent frames of the query for a 1400 image database.

Vocabulary Tree

- Advantages:
  - Fast image retrieval
  - Increasing the size of the vocabulary is logarithmic in the number of leaf nodes.
- Disadvantages:
  - Performance degrades as the number of leafs grows

Conclusion

- Image Categorization
- Bag-of-keypoints
  - Simple but gives good results
- Incorporate Spatial information
- Handle large database

Questions?