

S A I R

Spatial AI & Robotics Lab

CSE 473/573-A

W10: RETRIEVAL & DETECTION

Chen Wang
Spatial AI & Robotics Lab
Department of Computer Science and Engineering



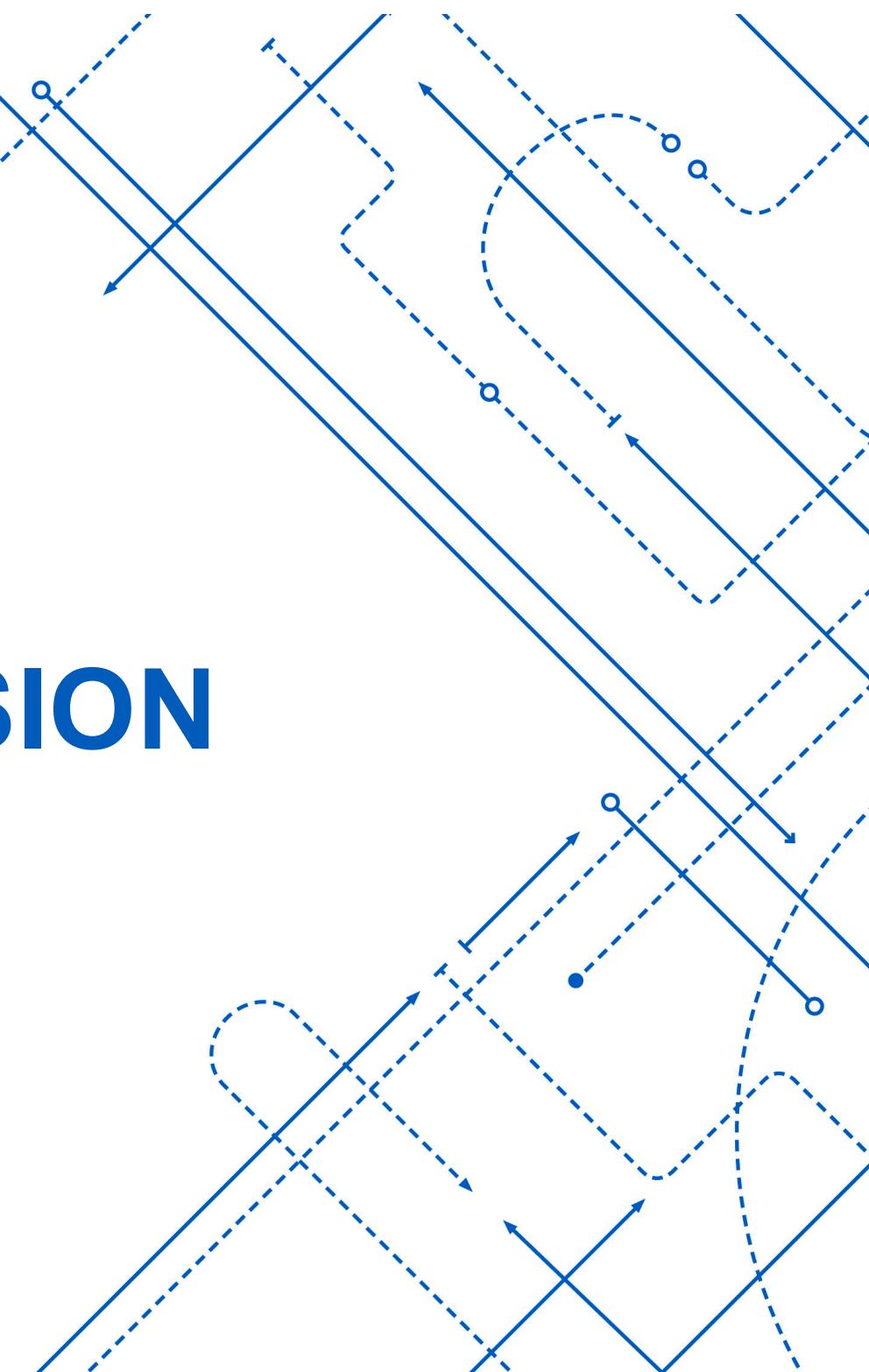
University at Buffalo The State University of New York





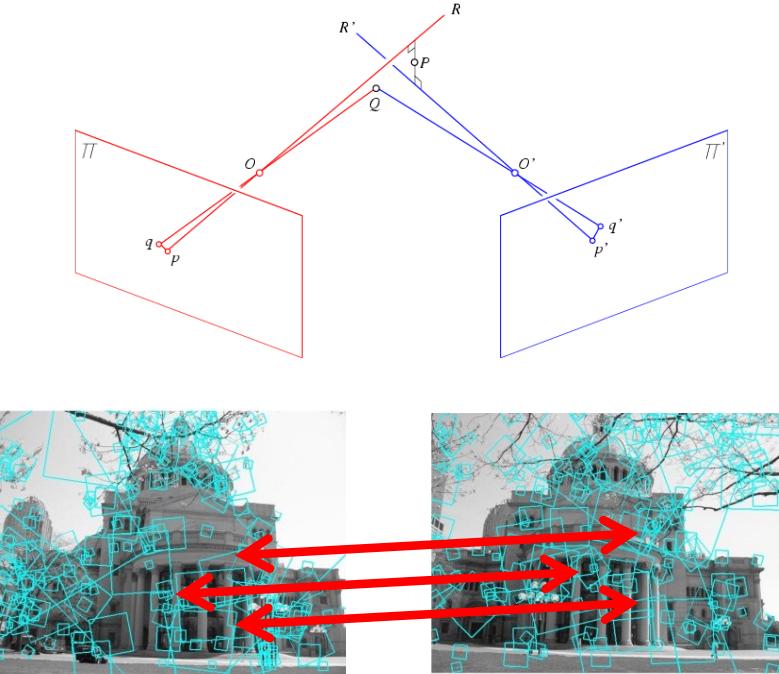
COMPUTER VISION

Instance Retrieval



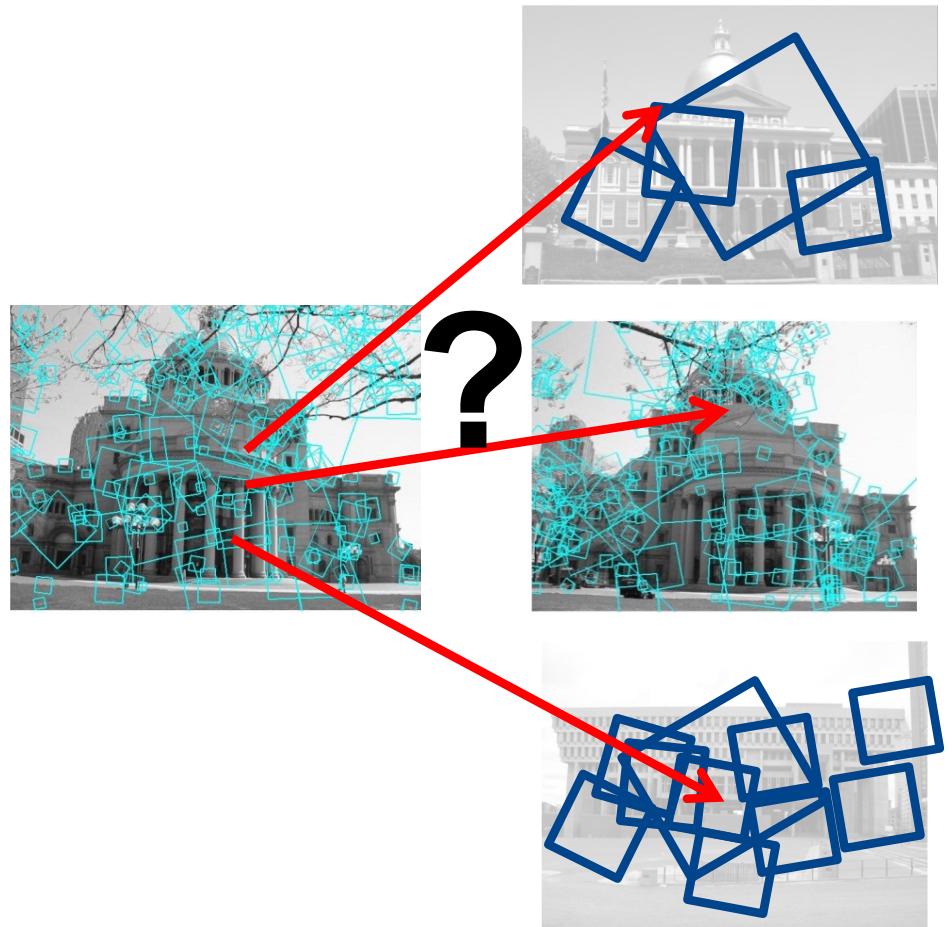
Multi-view matching

Matching two given views for depth



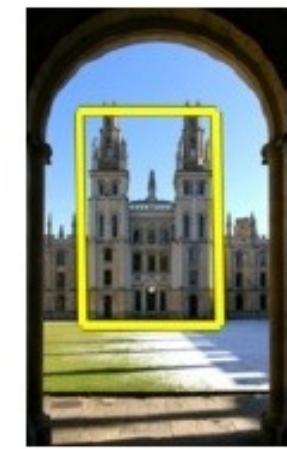
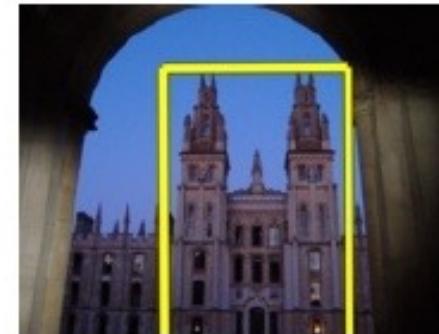
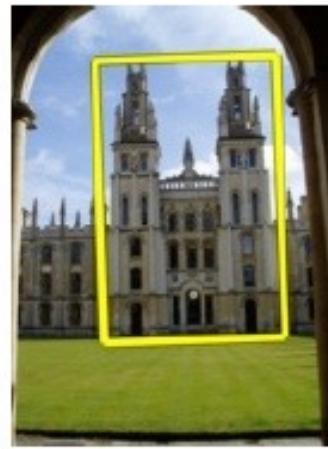
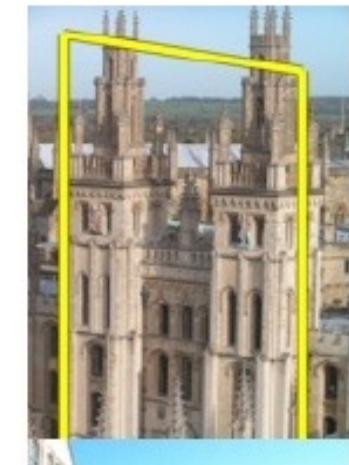
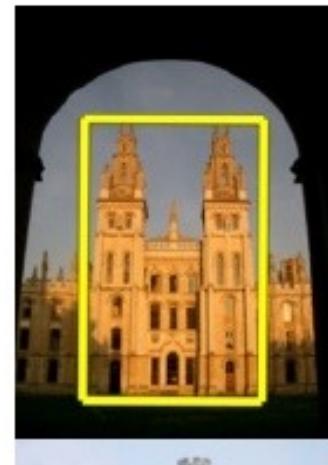
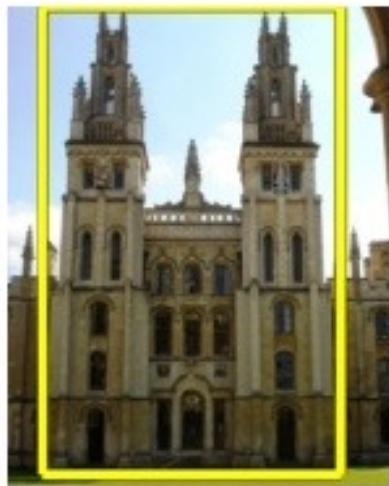
Search for a matching view for recognition

VS



Efficient Retrieval

How to quickly find images in a large database that match a given image region?

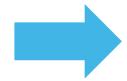


Local Retrieval

1. Collect all words within query region
2. Inverted file index to find relevant frames
3. Compare word counts
4. Spatial verification



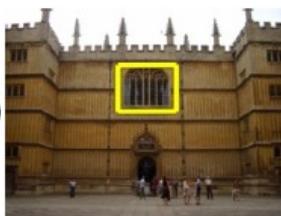
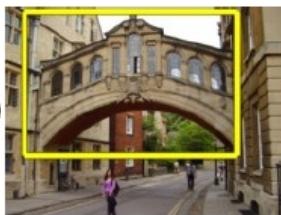
Query
region



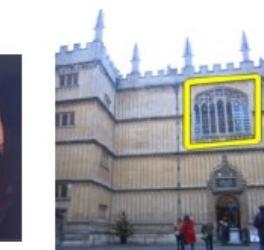
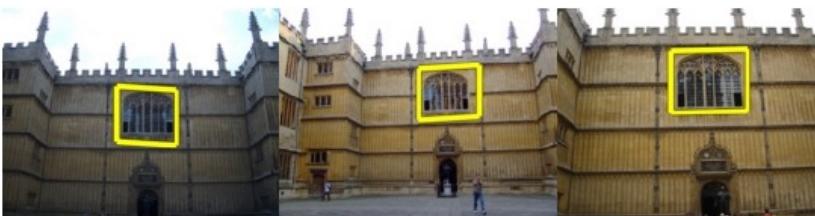
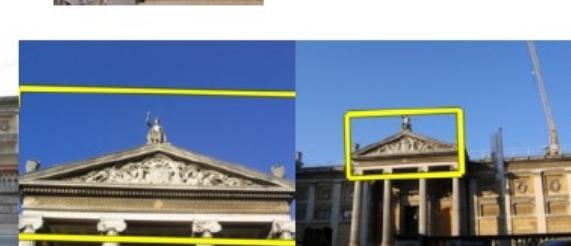
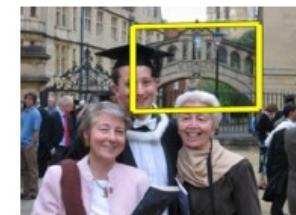
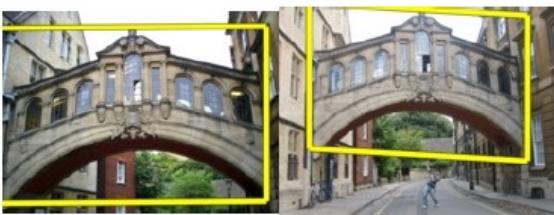
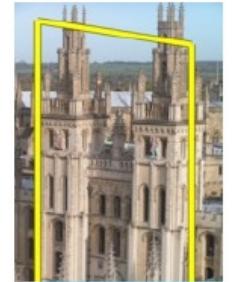
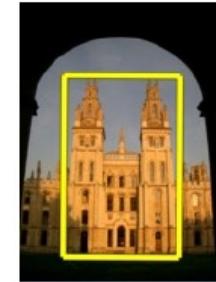
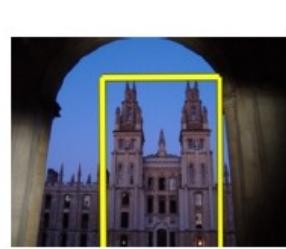
Retrieved frames

Application: Image Retrieval

Query

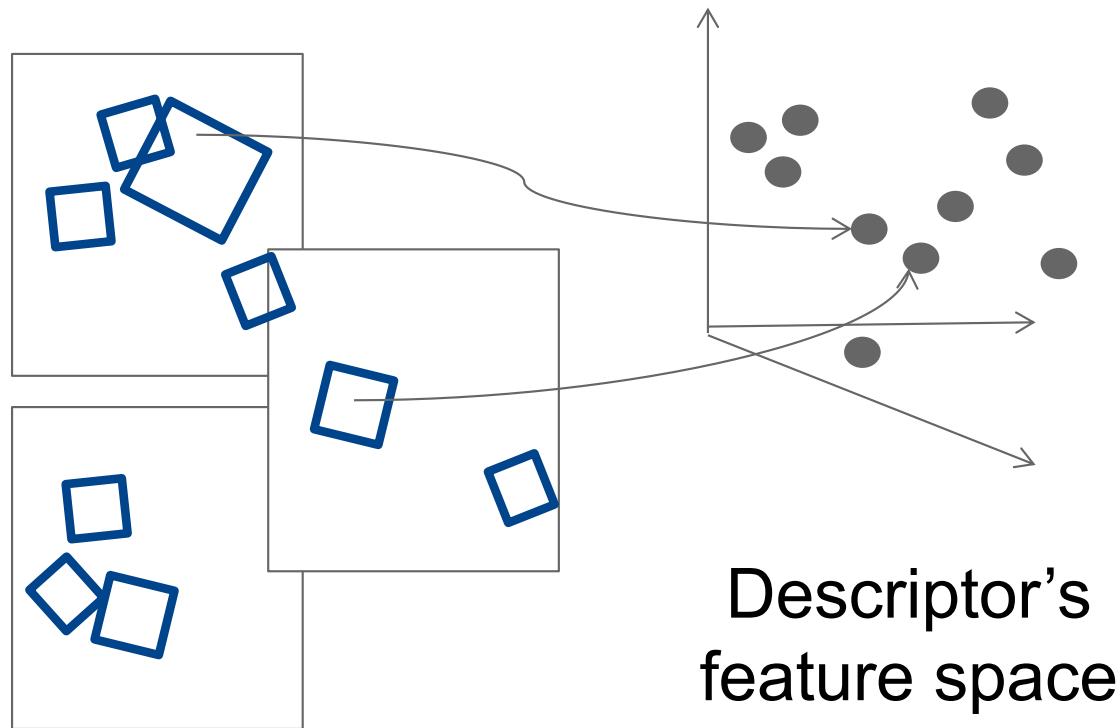


Results from 5k Flickr images (demo available for 100k set)



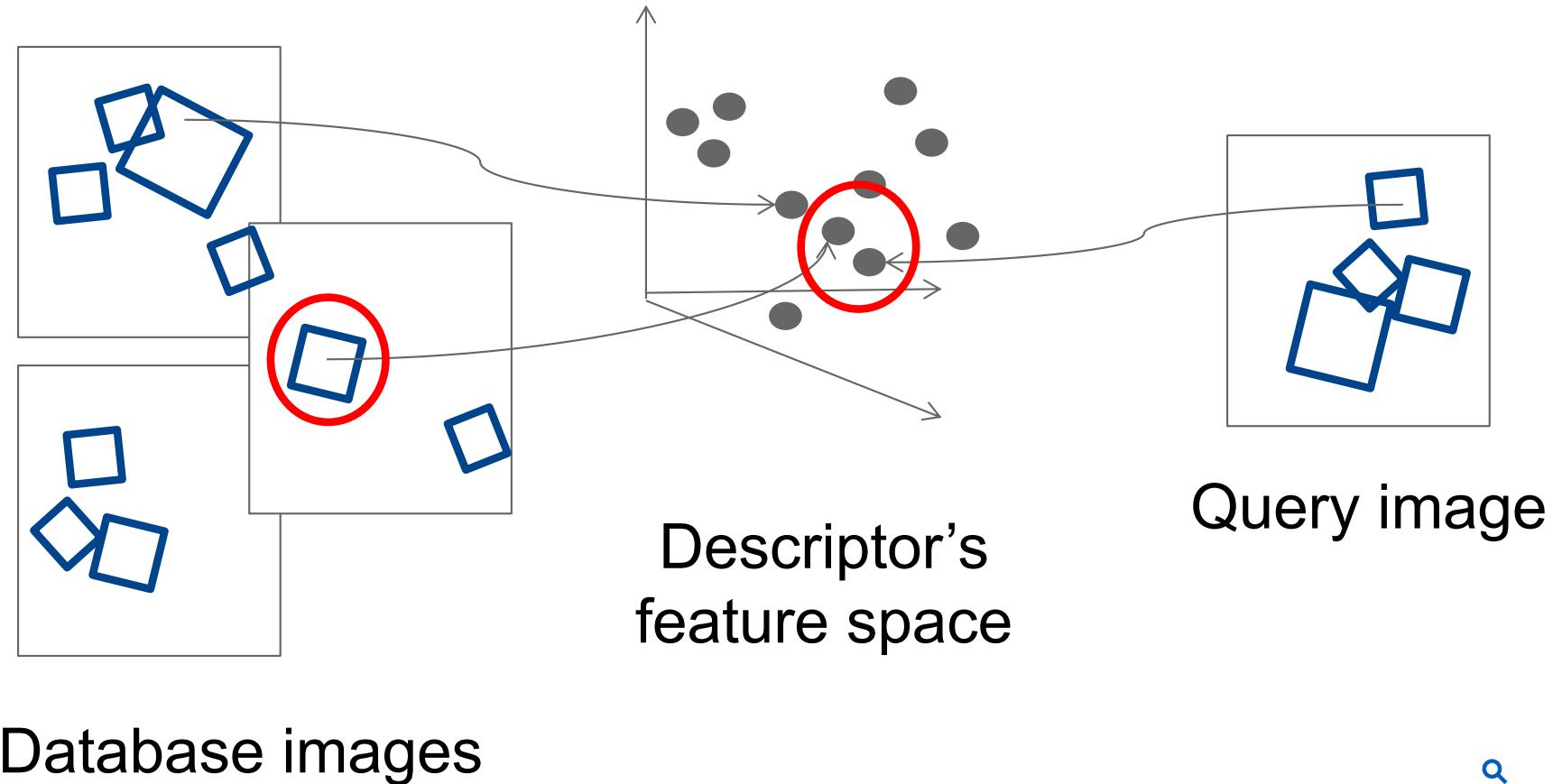
Indexing local features

- Each patch / region has a **descriptor**, which **is a point** in some high-dimensional feature space, e.g., SIFT.



Indexing local features

- When we see close points in feature space, we have similar descriptors, which indicates similar local content.



Easily have millions of features to search!

Indexing local features: inverted file index

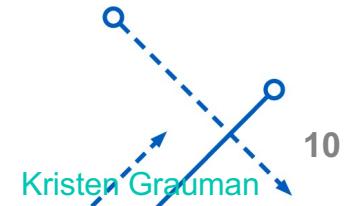
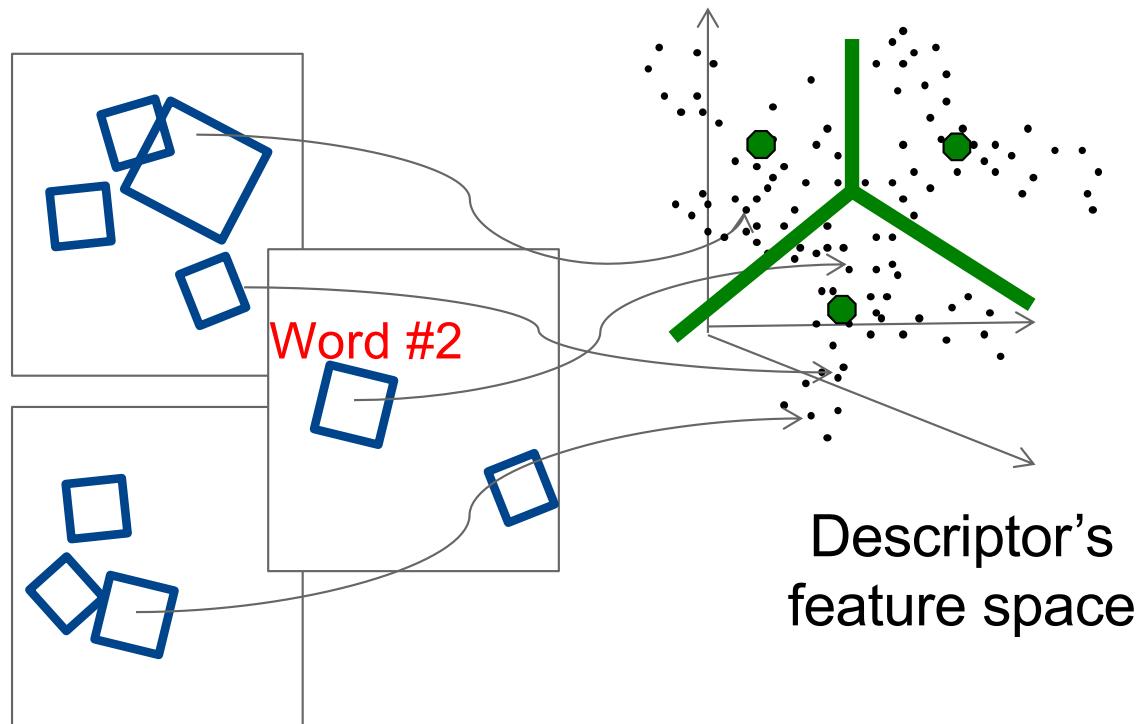
- For text documents, an efficient way to find all *pages* on which a *word* occurs is to use an index...

- We want to find all *images* in which a *feature* occurs.
- To use this idea, we map features to “visual words”.

Index
"Along I-75," From Detroit to Florida; <i>inside back cover</i> "Drive I-95," From Boston to Florida; <i>inside back cover</i> 1929 Spanish Trail Roadway; 101-102,104 511 Traffic Information; 83 AIA (Barrier Isl) - I-95 Access; 86 AAA (and CAA); 83 AAA National Office; 88 Abbreviations; Colored 25 mile Maps; cover Exit Services; 196 Travelogue; 85 Africa; 177 Agricultural Inspection Stns; 126 Ah-Tah-Thi-Ki Museum; 160 Air Conditioning, First; 112 Alabama; 124 Alachua; 132 County; 131 Alafia River; 143 Alapaha, Name; 126 Alfred B Macay Gardens; 106 Alligator Alley; 154-155 Alligator Farm, St Augustine; 169 Alligator Hole (definition); 157 Alligator, Buddy; 155 Alligators; 100,135,138,147,156 Anastasia Island; 170 Anhaica; 108-109,146 Apalachee River; 112 Appleton Mus of Art; 136 Aguifer; 102 Arabian Nights; 94 Art Museum, Ringling; 147 Aruba Beach Cafe; 183 Auxilla River Project; 106 Babcock-Web WMA; 151 Bahia Mar Marina; 184 Baker County; 99 Barefoot Mailmen; 182 Barge Canal; 137 Bee Line Expy; 80 Belz Outlet Mall; 89 Bernard Castro; 136 Big "I"; 165 Big Cypress; 155,158 Big Foot Monster; 105 Billie Swamp Safari; 160 Blackwater River SP; 117 Blue Angels A4-C Skyhawk; 117 Atrium; 121 Blue Springs SP; 87 Blue Star Memorial Highway; 125 Boca Ciega; 169 Boca Grande; 150 Boca Raton; 182 Bonnie Blue Flag; 124 Boyds Hill Nature Trail; 188 Bradenton; 145-147 Breakers, The, Palm Beach; 181 Brickell Point, Miami; 185 Britton Hill; 116 Brogan Museum; 107 Bromeliads (see Epiphytes) Broward, Gov, Napoleon; 156 Bulow Plantation Ruins; 171 Bush, Gov. Jeb; 100 Butterfly Center, McGuire; 134 CAA (see AAA) CCC, The; 111,113,115,135,142 Ca d'Zan; 147 Caloosahatchee River; 152 Name; 150 Canaveral Natnl Seashore; 173 Cannon Creek Airpark; 130 Canopy Road; 106,169 Cape Canaveral; 174 Castillo San Marcos; 169 Cave Diving; 131 Cayo Costa, Name; 150 Celebration; 93 Charlotte County; 149 Charlotte Harbor; 150 Chautauqua; 116 Chipley; 114 Name; 115 Choctawatchee, Name; 115 Circus Museum, Ringling; 147 Citrus; 88,97,130,136,140,180 CityPlace, W Palm Beach; 180 City Maps, Ft Lauderdale Expwy; 194-195 Jacksonville; 163 Kissimmee Expwy; 192-193 Miami Expressways; 194-195 Orlando Expressways; 192-193 Pensacola; 26 Tallahassee; 191 Tampa-St. Petersburg; 63 St. Augustine; 191 Civil War; 100,108,127,138,141 Clearwater Marine Aquarium; 187 Collier County; 154 Collier, Barron; 152 Colonial Spanish Quarters; 168 Columbia County; 101,128 Coquina Building Material; 165 Corkscrew Swamp, Name; 154 Cowboys; 95 Crab Trap II; 144 Cracker, Florida; 88,95,132 Crosstown Expy; 11,35,98,143 Cuban Bread; 184 Dade Battlefield; 140 Dade, Maj, Francis; 139-140,161 Danie Beach Hurricane; 184 Daniel Boone, Florida Walk; 117 Daytona Beach; 172-173 De Land; 87 De Soto, Hernando, Anhaica; 108-109,146 County; 149 Explorer; 146 Landing; 146 Napitaca; 103 National Park; 147 Tallahassee; 108 DeFunak Springs; 116 Name; 115 Delnor-Wiggins Pass SP; 155 Denoeil Cafe, St Augustine; 169 Devil's Millhopper; 132 Dickson Azalea Park; 89 Dinosaur World; 98 Discovery Cove; 90 Dixie Highway; 186 Don Garits Drag Racing Mus; 138 Douglas, Marjory Stoneman; 159 Driving Lanes; 85 Duval County; 163 Eau Gallie; 175 Edison, Thomas; 152 Eglin AFB; 116-118 Eight Reale; 176 Ellenton; 144-145 Emanuel Point Wreck; 120 Emergency Callboxes; 83 Epiphytes; 142,148,157,159 Escambia Bay; 119 Bridge (I-10); 119 County; 120 Estero; 153 Everglade, 90,95,139-140,154-160 Draining of; 156,181 Wildlife MA; 160 Wonder Gardens; 154 Falling Waters SP; 115 Fantasy of Flight; 95 Fayer Dykes SP; 171 Fires, Forest; 166 Fires, Prescribed ; 148 Fisherman's Village; 151 Flagler County; 171 Flagler, Henry; 97,165,167,171 Florida Aquarium; 186 Florida, 12,000 years ago; 187 Cavern SP; 114 Map of all Expressways; 2-3 Mus of Natural History; 134 National Cemetery ; 141 Part of Africa; 177 Platform; 187 Sheriff's Boys Camp; 126 Sports Hall of Fame; 130 Sun 'n Fun Museum; 97 Supreme Court; 107 Florida's Turnpike (FTP); 178,189 25 mile Strip Maps; 66 Administration; 189 Coin System; 190 Exit Services; 189 HEFT; 76,161,190 History; 189 Names; 189 Service Plazas; 190 Spur SR91; 76 Ticket System; 190 Toll Plazas; 190 Ford, Henry; 152 Fort Barrancas; 122 Buried Alive; 123 Fort Caroline; 164 Fort Clinch SP; 161 Fort De Soto & Egmont Key; 188 Fort Lauderdale; 161,182-184 Fort Myers; 152-153 Fort Pierce; 177-178 Farmers Market; 178 Fountain of Youth; 170 Frank Lloyd Wright Center; 97 Gadsden County; 110 Gainesville; 99,104,131-135,146 Gamble Plantation; 145 Garden of Eden; 112 Gasparilla, Pirate; 150 Gatorade; 134 Gaylord Palms; 90 Geology;102-103,110,131-132

Visual Words

- Map descriptors to “words” by quantizing feature space
 - Quantize via clustering
 - Cluster centers are the prototype “words”
 - Determine which word to assign to each new image region by finding the closest cluster center.



Visual words

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

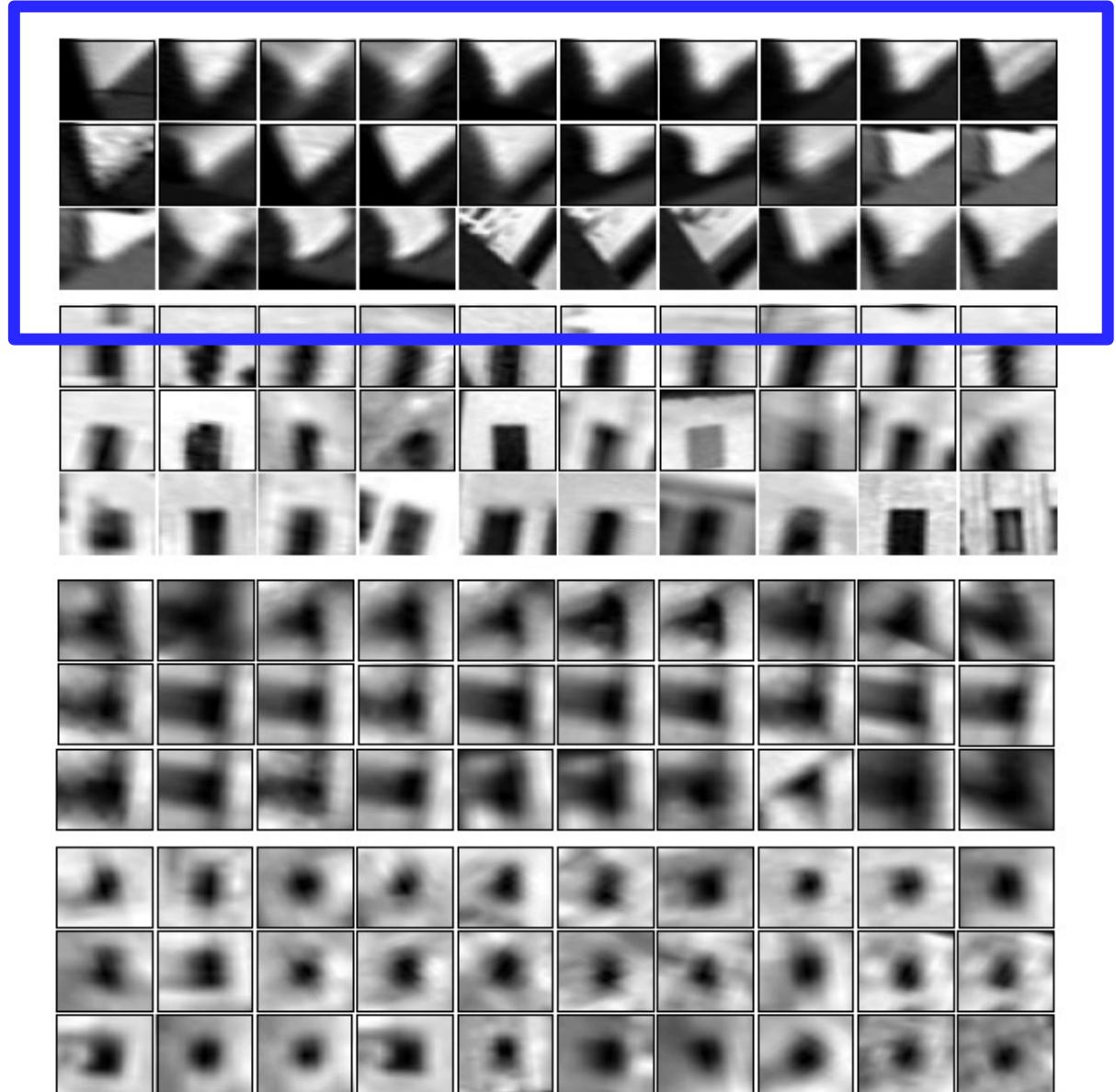
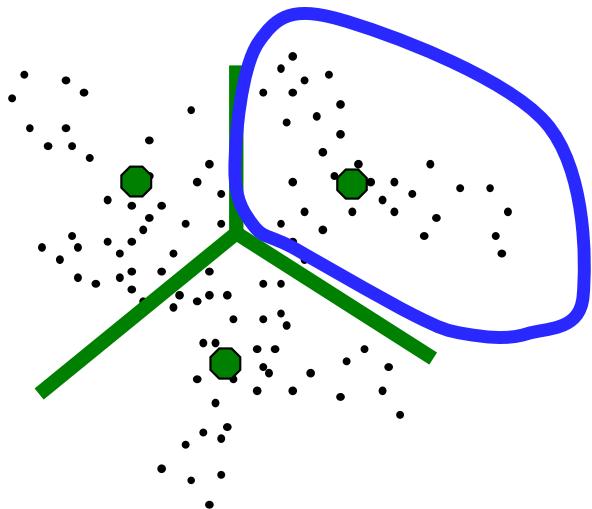


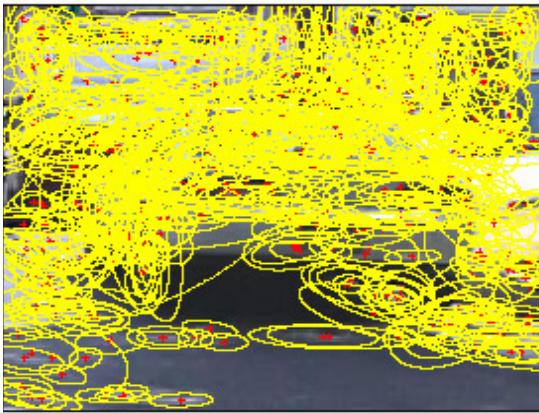
Figure from Sivic & Zisserman, ICCV 2003

Visual vocabulary formation

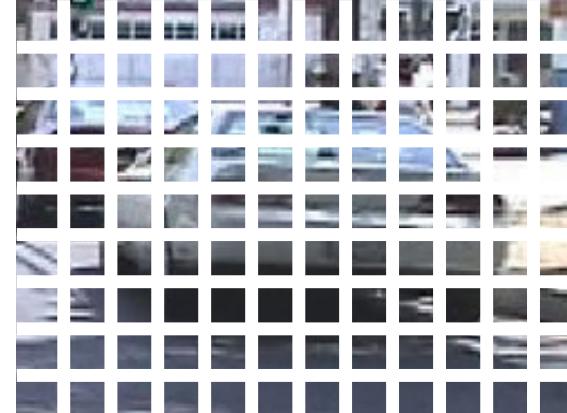
Things need to consider:

- Sampling strategy: where to extract features?
- Clustering / quantization algorithm
- Unsupervised vs. supervised
- What corpus provides features, vocabulary size, number of words (universal vocabulary?)

Sampling strategies



Sparse, at interest points



Dense, uniformly



Randomly



Multiple interest operators

- To find **specific, textured objects, sparse sampling** from interest points often more reliable.
- **Multiple** complementary interest **operators** offer more image coverage.
- For object **categorization, dense sampling** offers better coverage.

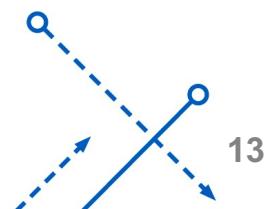


Image credits: F-F. Li, E. Nowak, J. Sivic

Inverted file index

- Database images are loaded into the index mapping words to image numbers



Word #	Image #
1	3
2	
...	
7	1, 2
8	3
9	
10	
...	
91	2

⋮ ⋮ ⋮

Inverted file index

- New query image is mapped to indices of database images that share a word.



New query image

Word #	Image #
1	3
2	
7	1, 2
8	3
9	
10	
...	
91	2



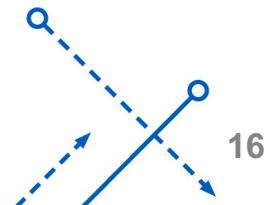
Image #1



Image #2

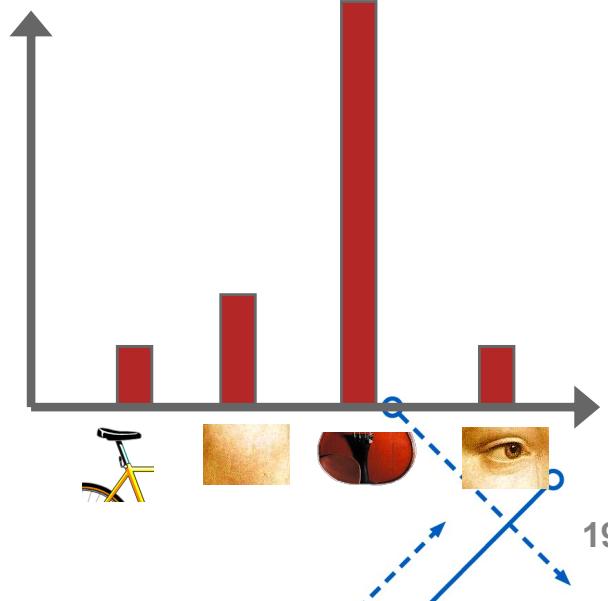
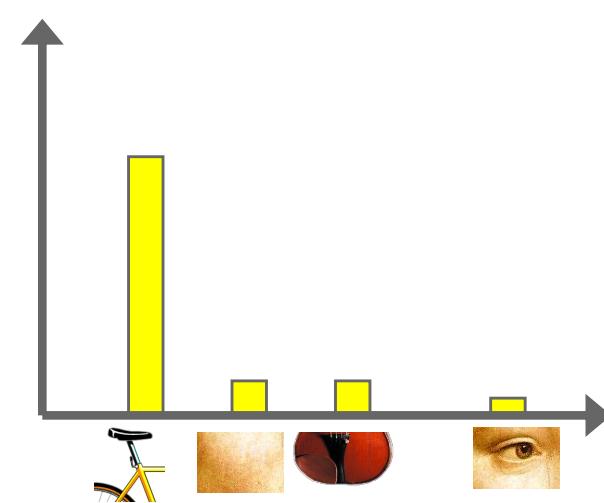
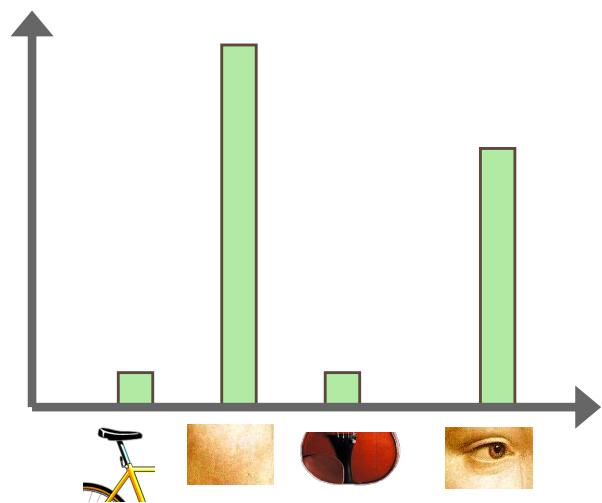
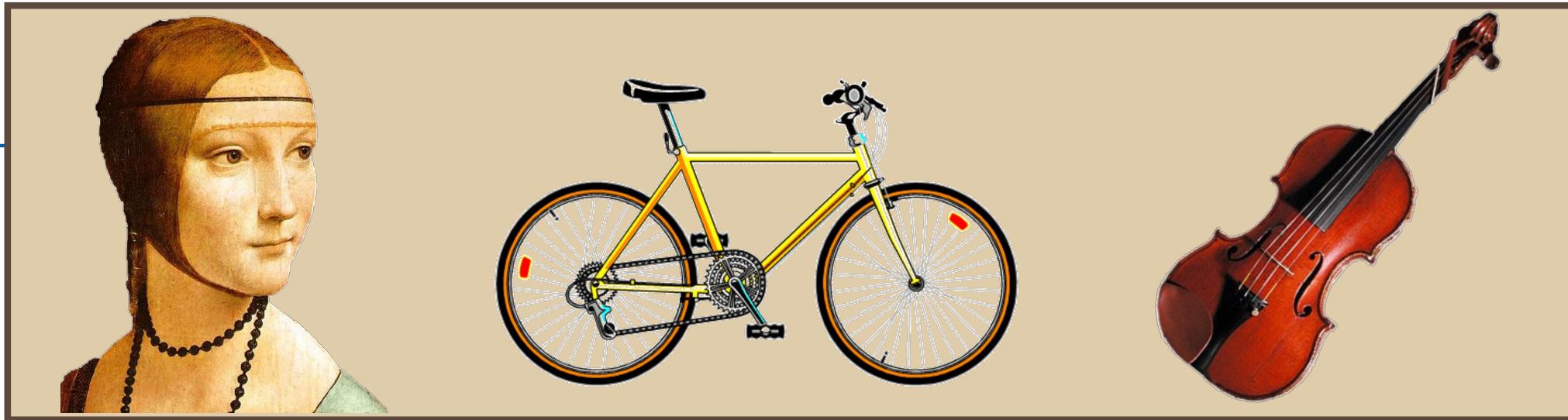
Inverted file index

- Key requirement for inverted file index to be efficient:
 - Sparsity
 - If most pages/images contain most words, then it's no better off than exhaustive search.
 - Exhaustive search would mean comparing the word distribution of a query versus every page.



Instance recognition: remaining issues

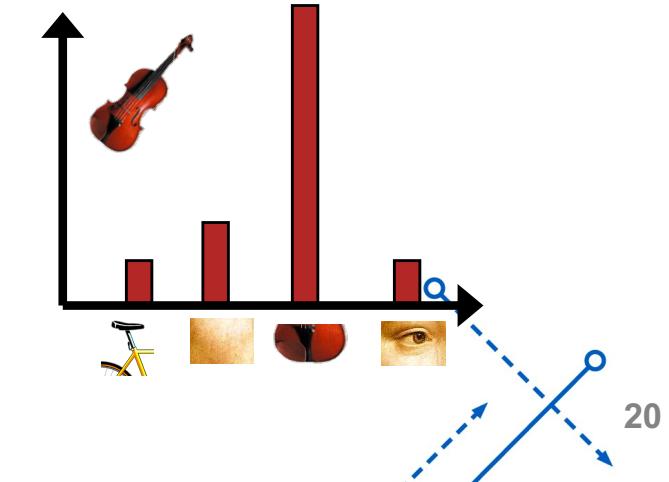
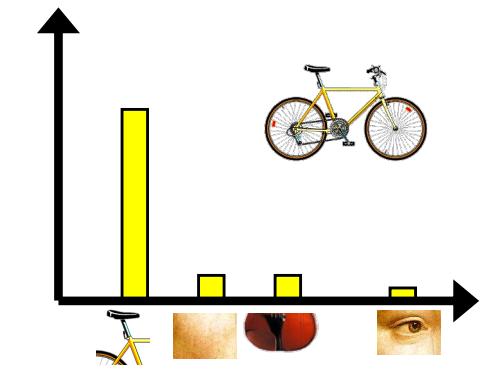
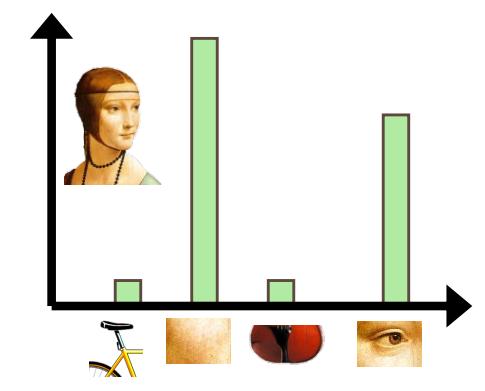
- How to summarize the content of an entire image? And estimate overall similarity?
- How large should the vocabulary be? How to perform quantization efficiently?
- Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?
- How to score the retrieval results?



Bags of visual words

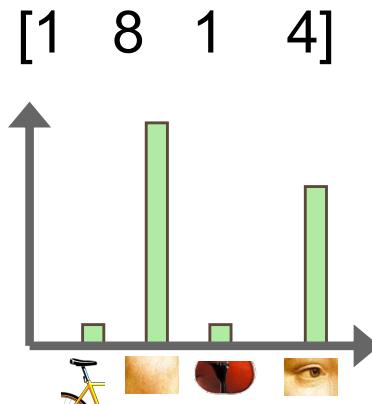
- Summarize entire image based on its distribution (histogram) of word occurrences.

- Analogous to bag of words representation commonly used for documents.

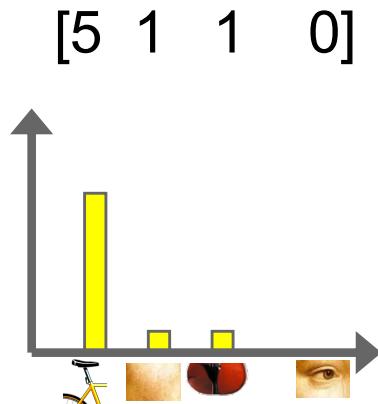


Comparing bags of words

- Rank frames by normalized inner product between their (possibly weighted) occurrence counts---*nearest neighbor* search for similar images.



\vec{d}_j



\vec{q}

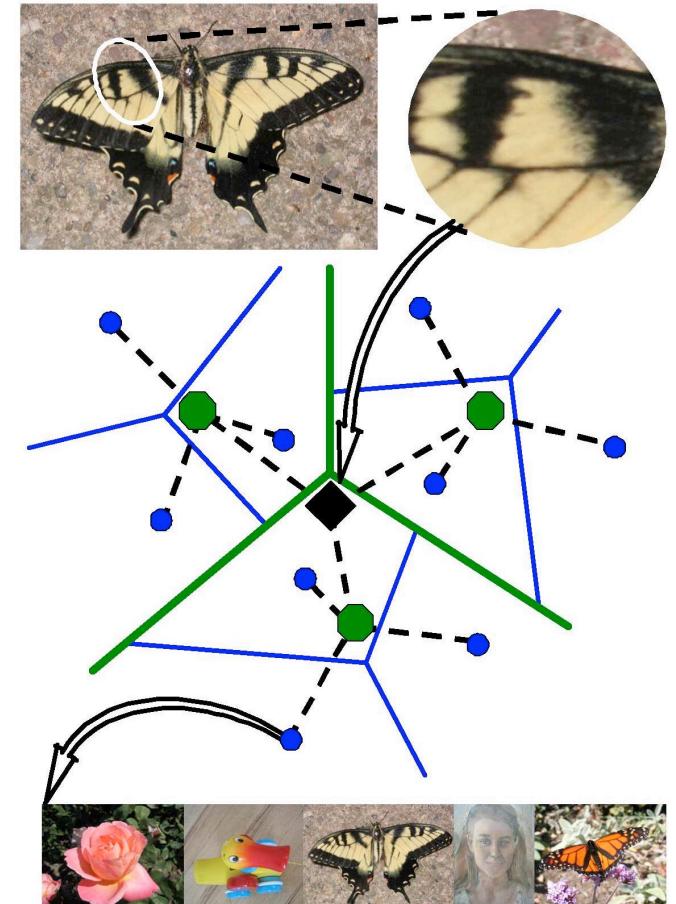
$$sim(d_j, q) = \frac{\langle d_j, q \rangle}{\|d_j\| \|q\|}$$

for vocabulary of V words

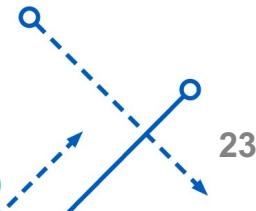
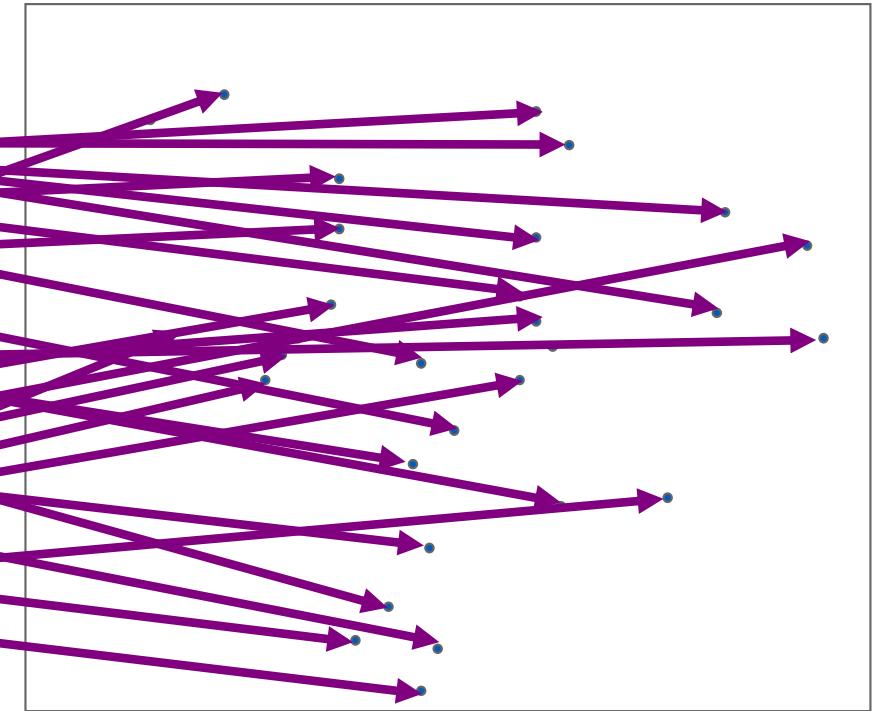
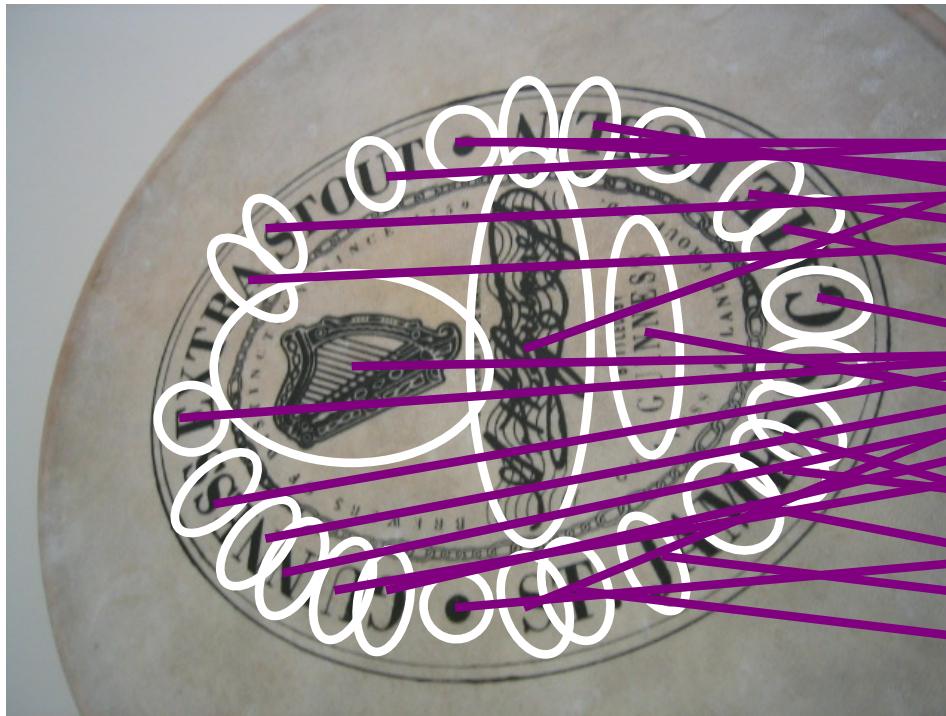


Visual vocabularies: Issues

- How to choose vocabulary size?
 - Too small: visual words not representative of all patches
 - Too large: quantization artifacts, overfitting
- Computational efficiency
 - **Vocabulary trees**

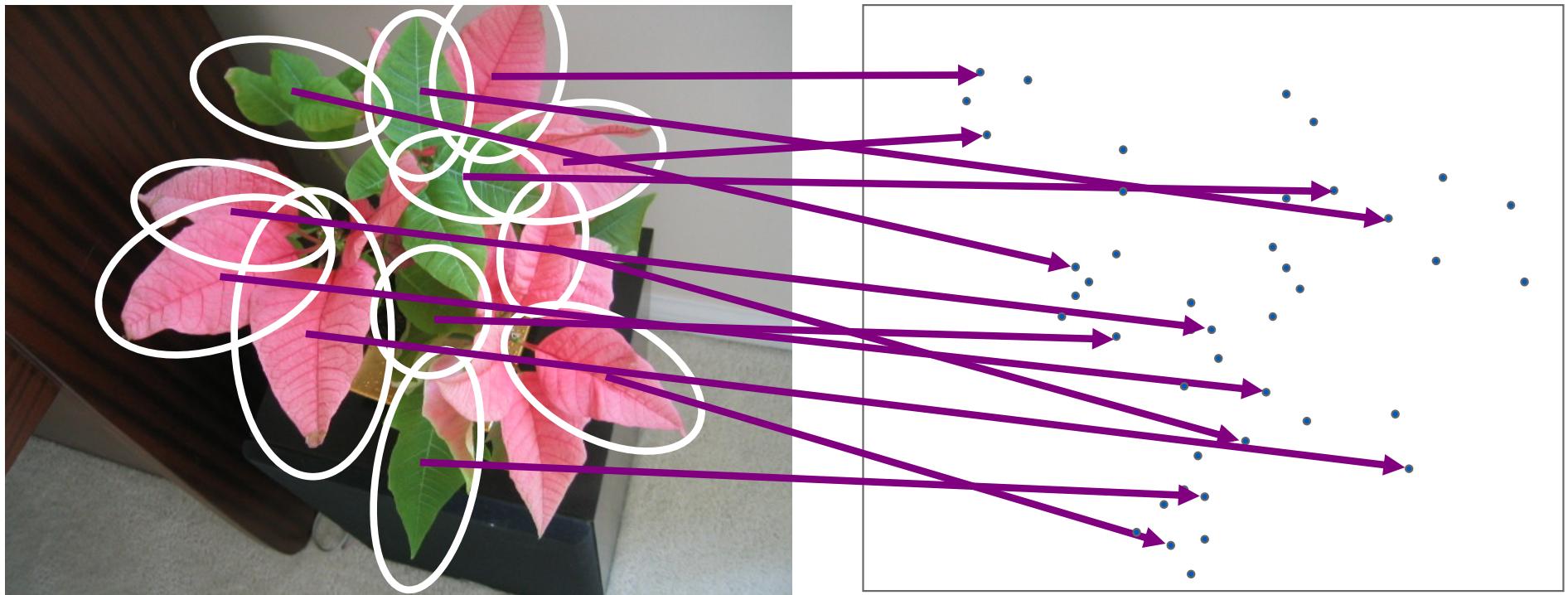


Recognition with K-tree



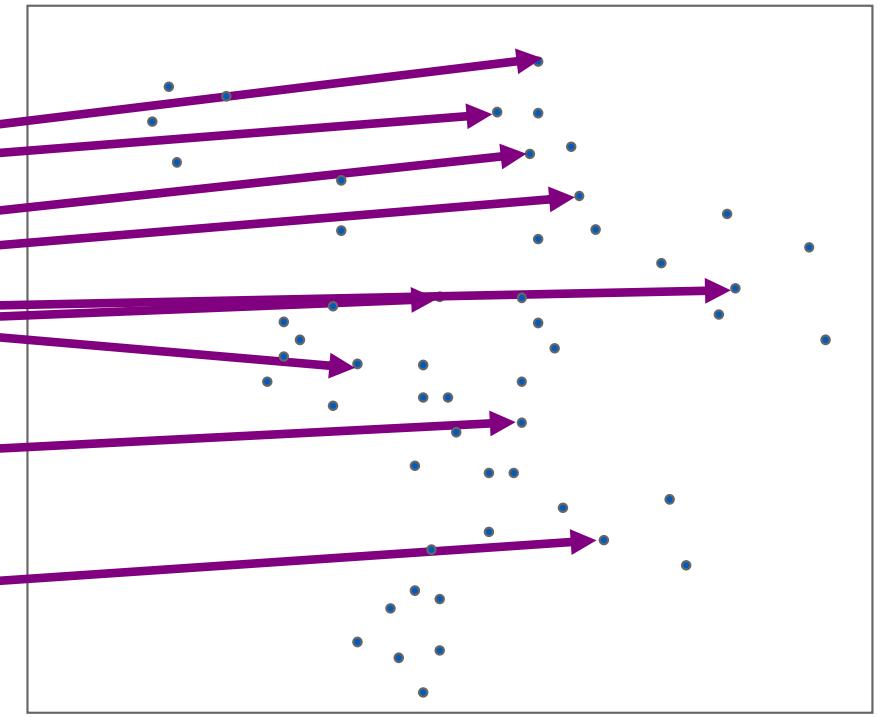
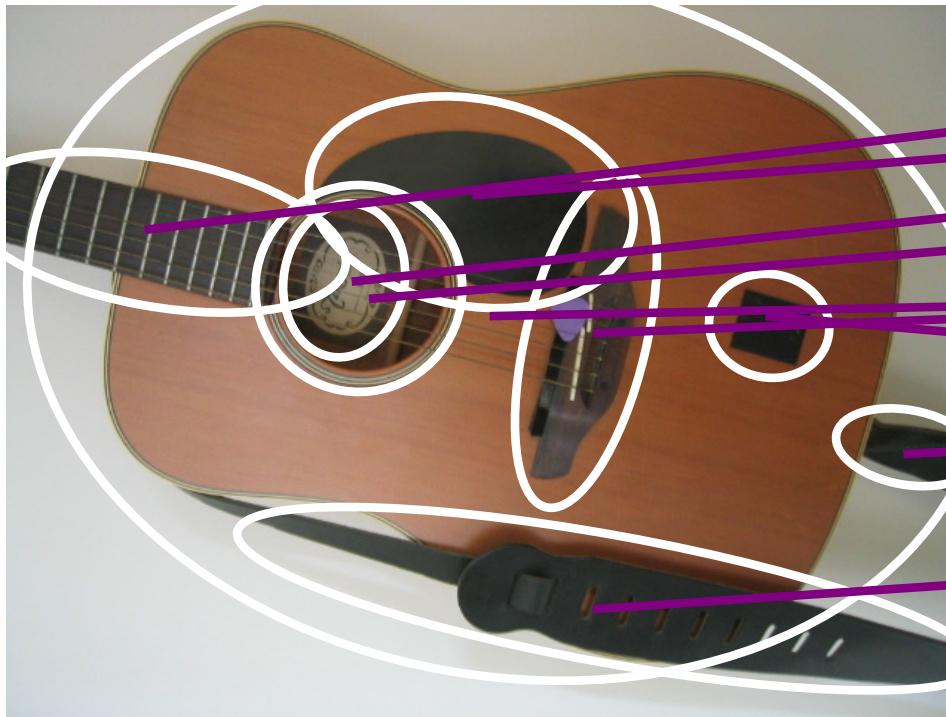
Following slides by David Nister (CVPR 2006)

Recognition with K-tree

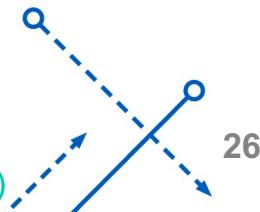
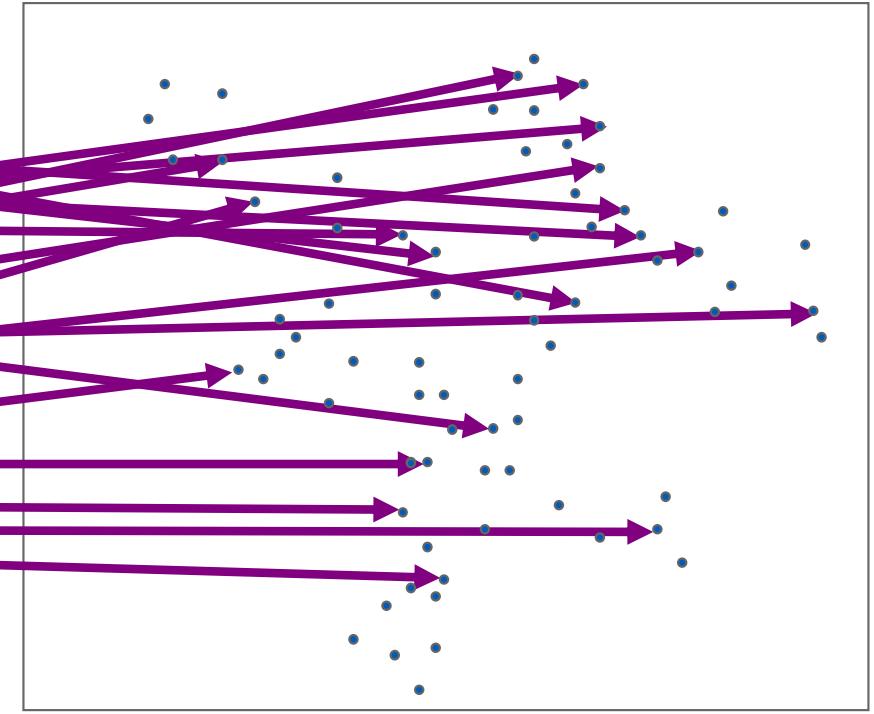
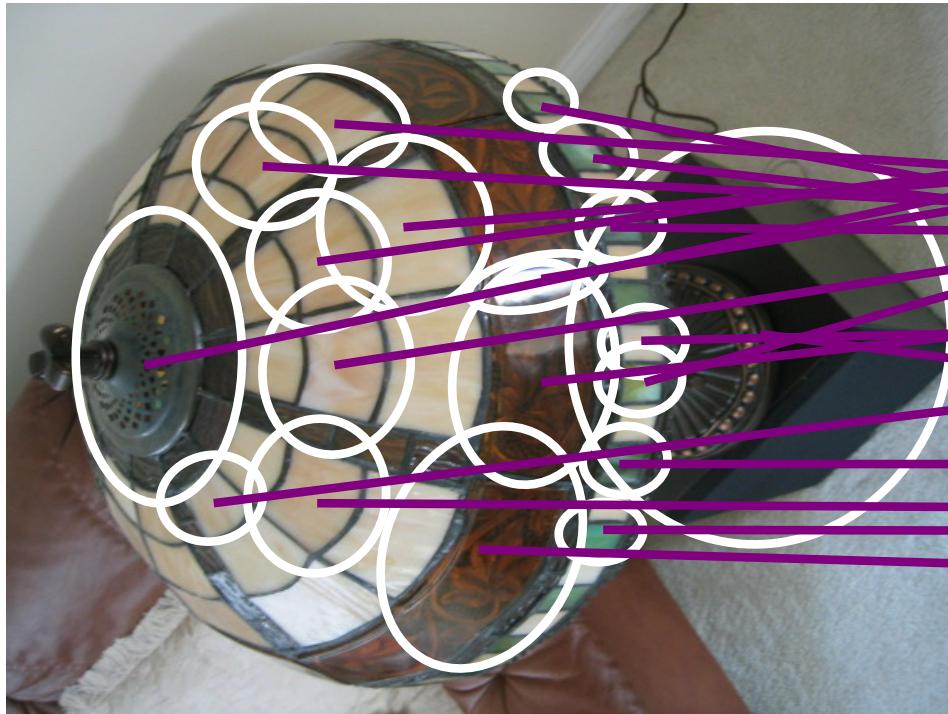


Following slides by David Nister (CVPR 2006)

Recognition with K-tree

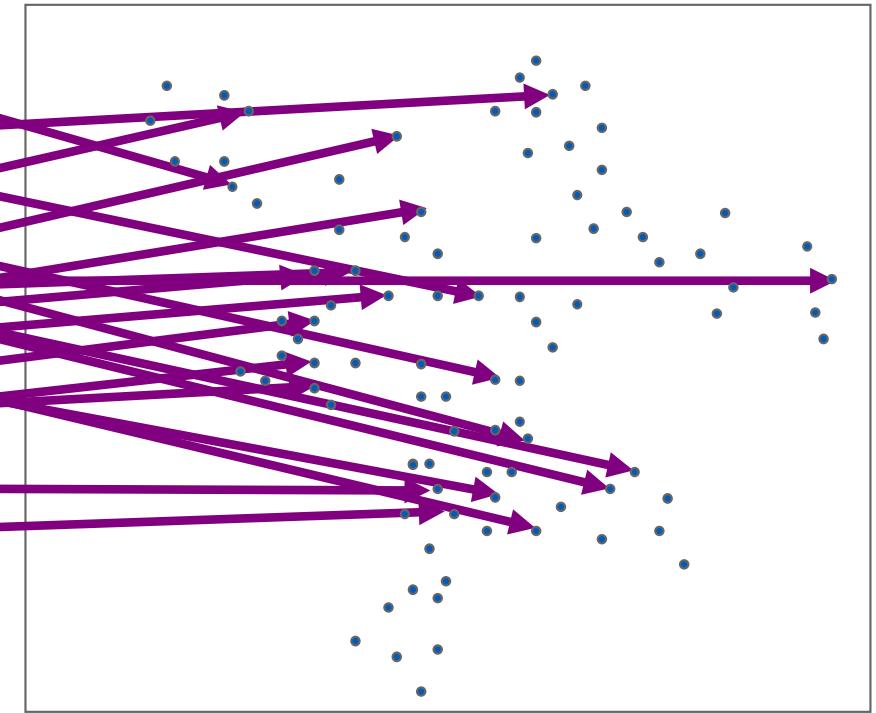
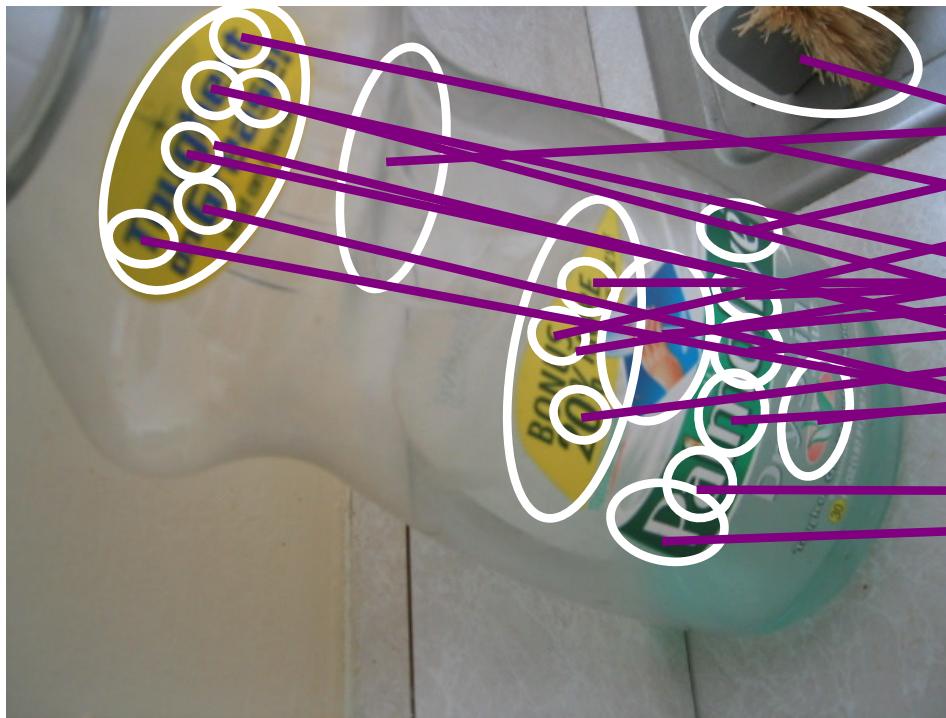


Recognition with K-tree

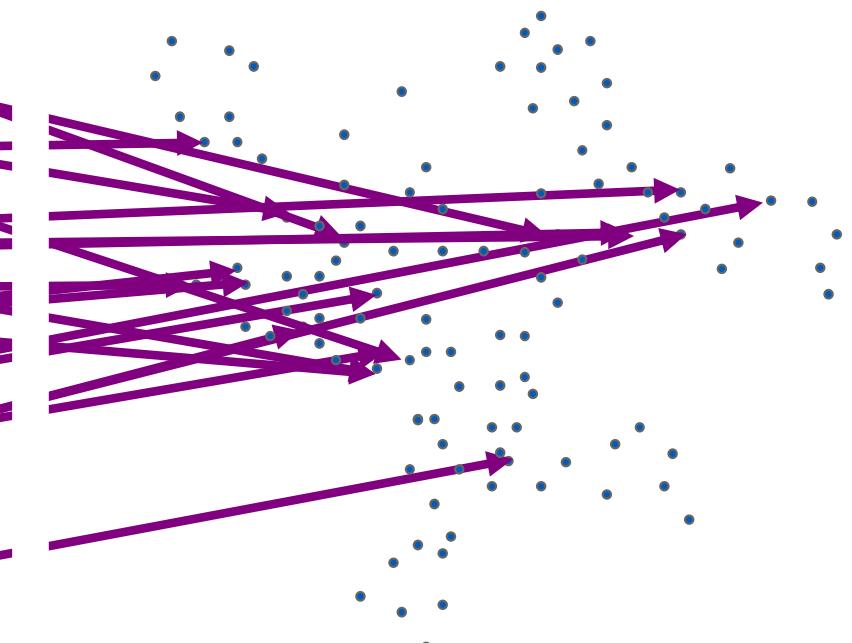


Following slides by David Nister (CVPR 2006)

Recognition with K-tree

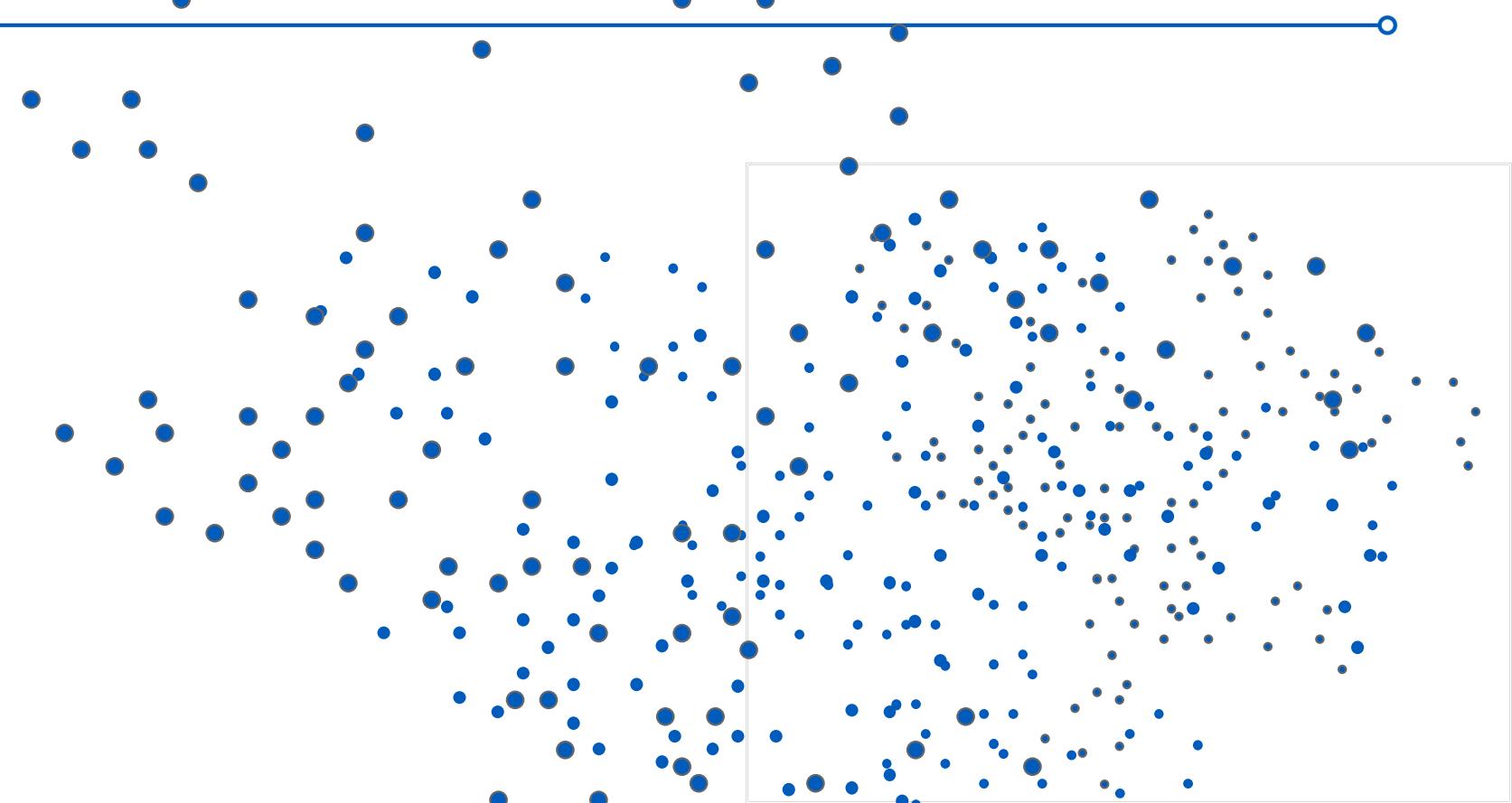


Recognition with K-tree

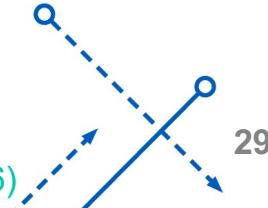


Following slides by David Nister (CVPR 2006)

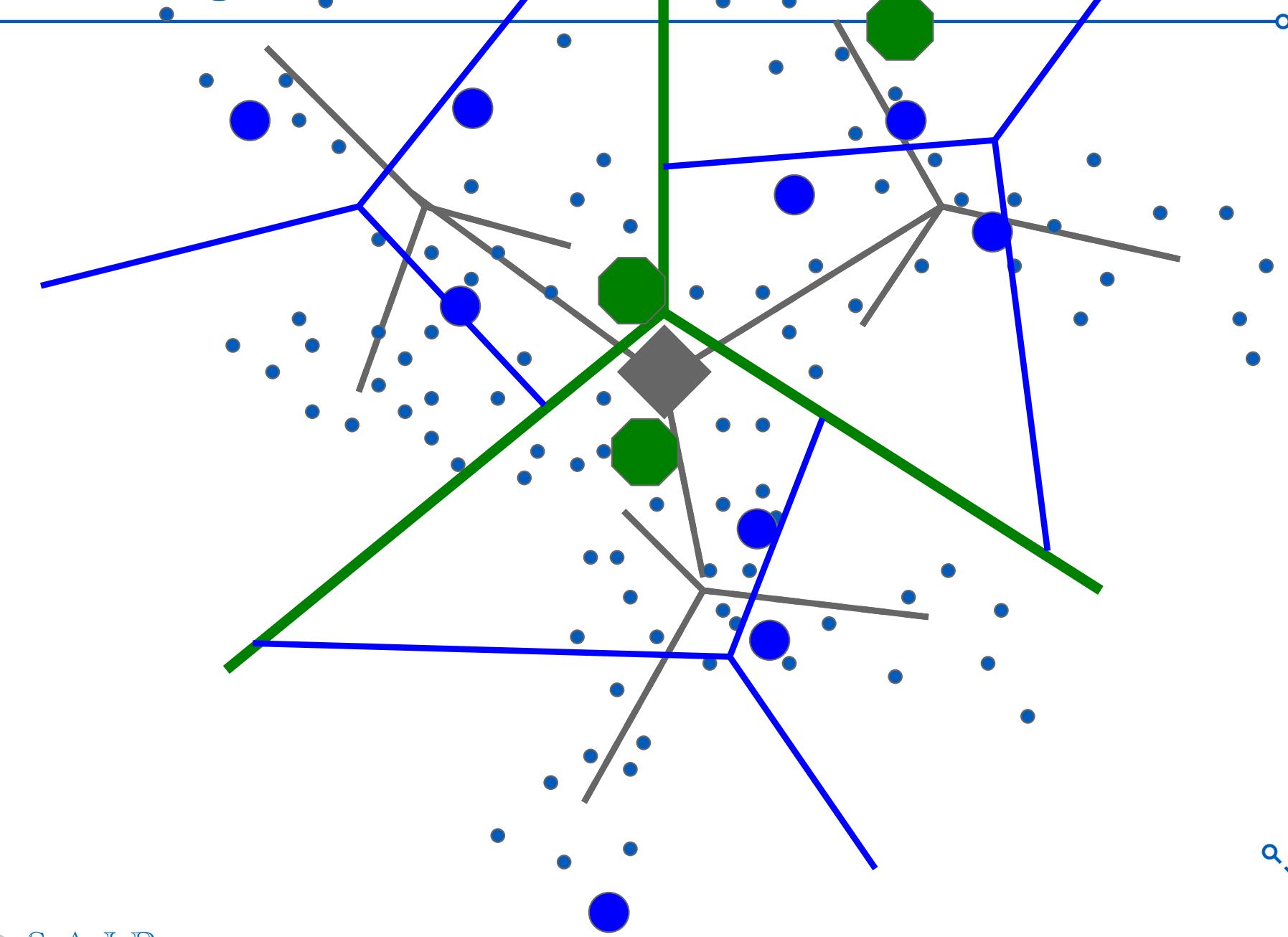
Recognition with K-tree



Following slides by David Nister (CVPR 2006)

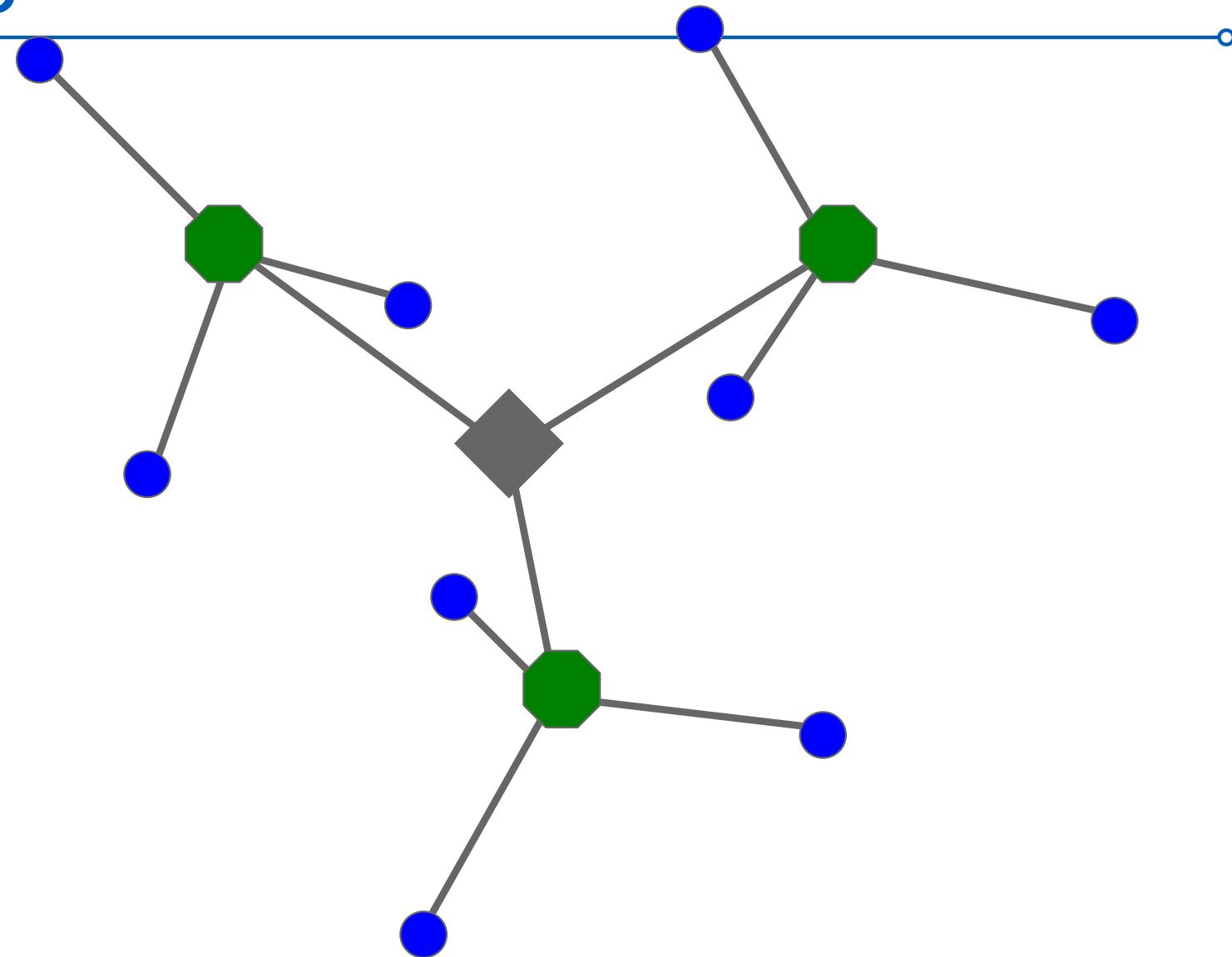


Recognition with K-tree



Following slides by David Nister (CVPR 2006)

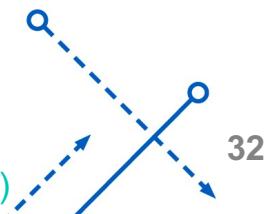
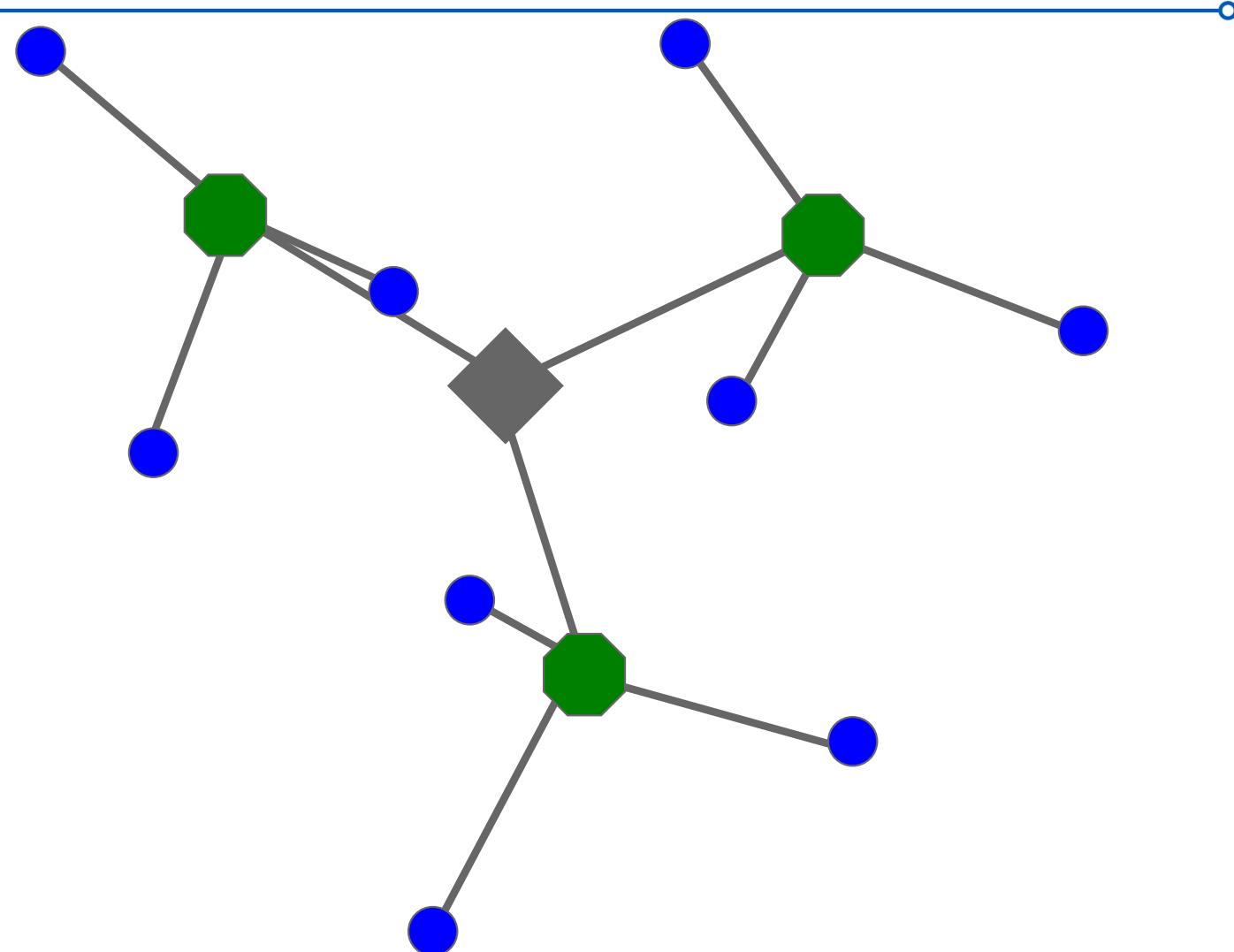
Recognition with K-tree



Following slides by David Nister (CVPR 2006)

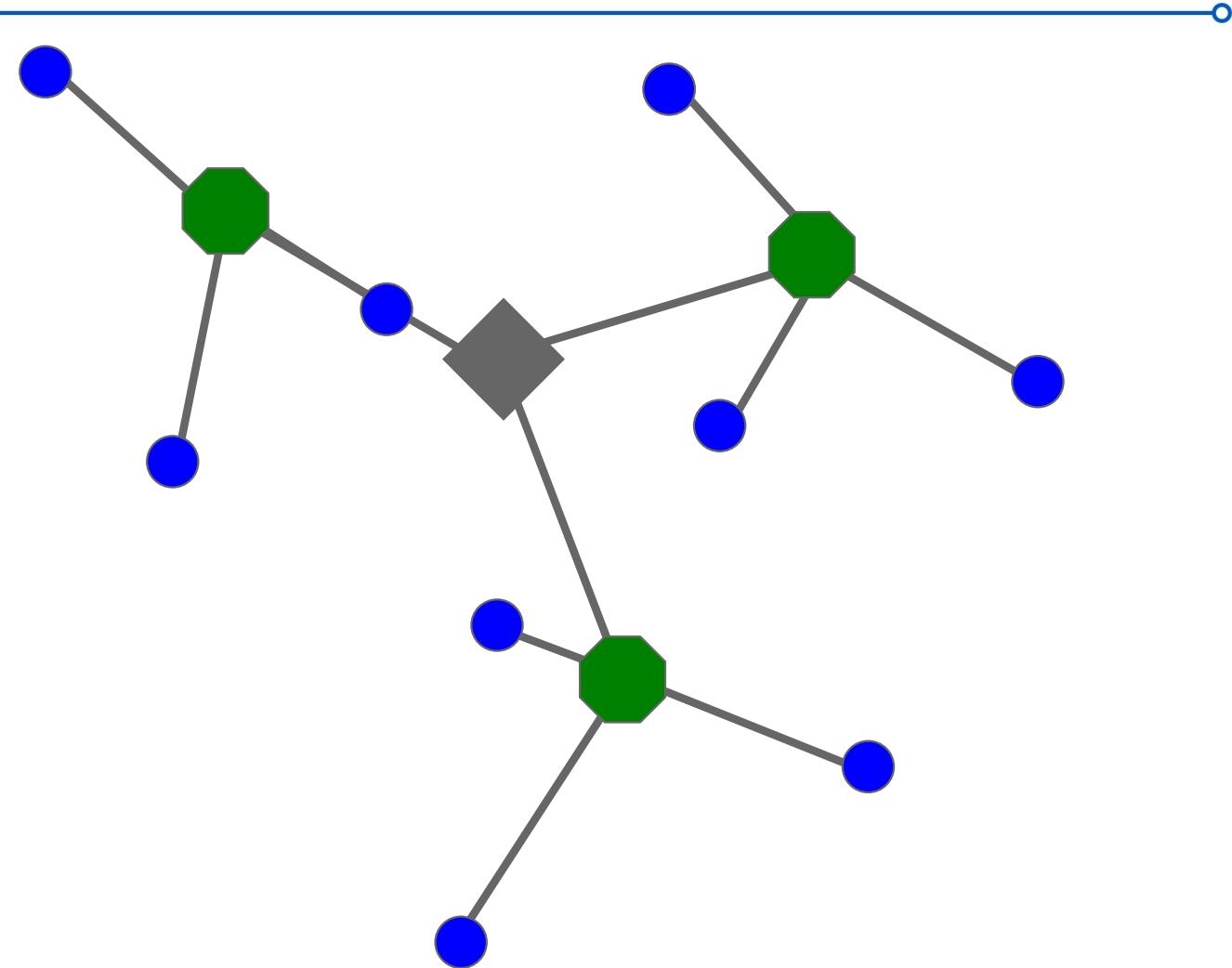


Recognition with K-tree

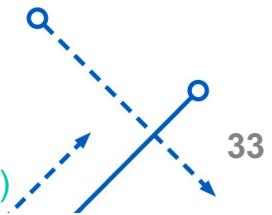


Following slides by David Nister (CVPR 2006)

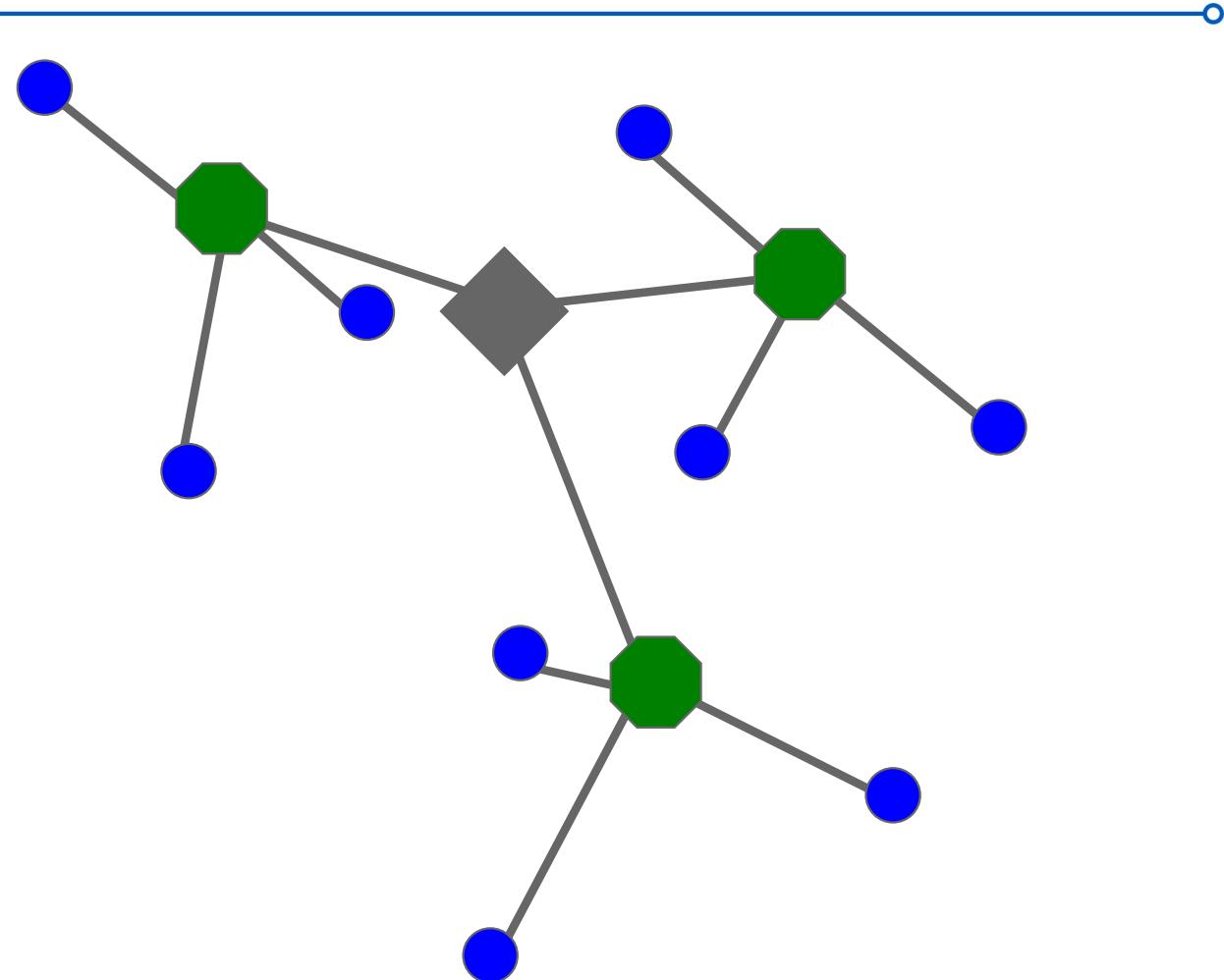
Recognition with K-tree



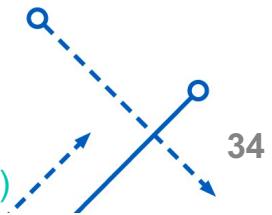
Following slides by David Nister (CVPR 2006)



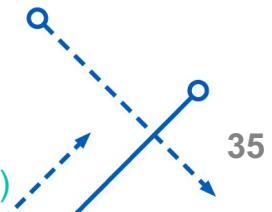
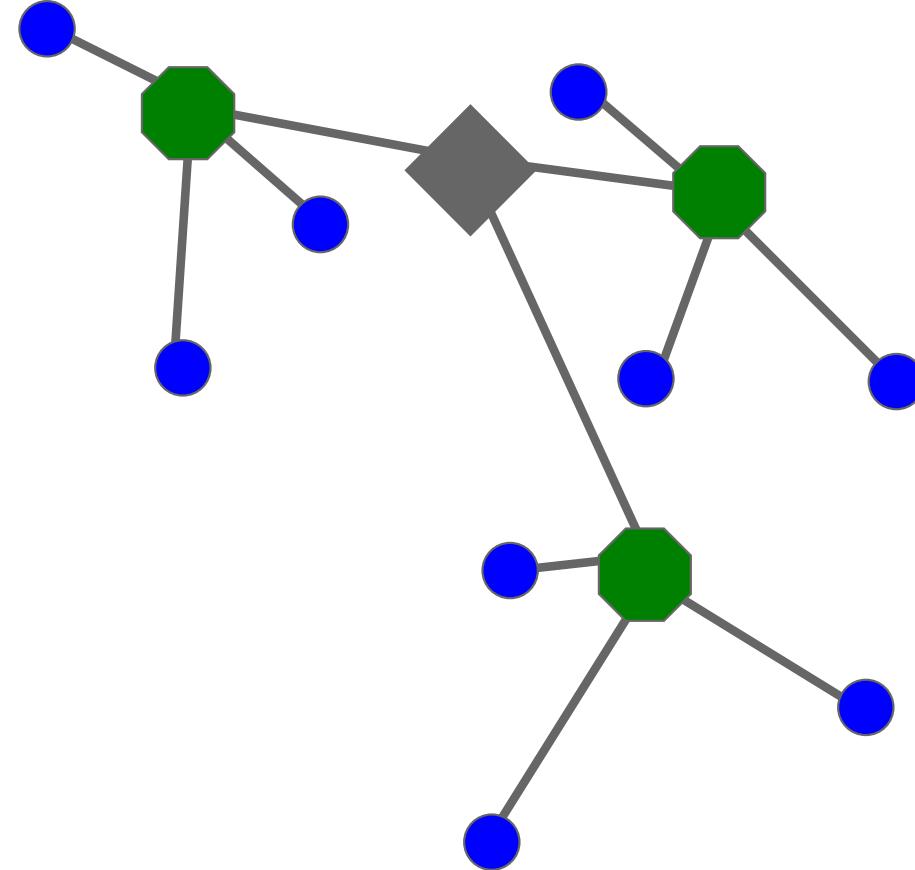
Recognition with K-tree



Following slides by David Nister (CVPR 2006)

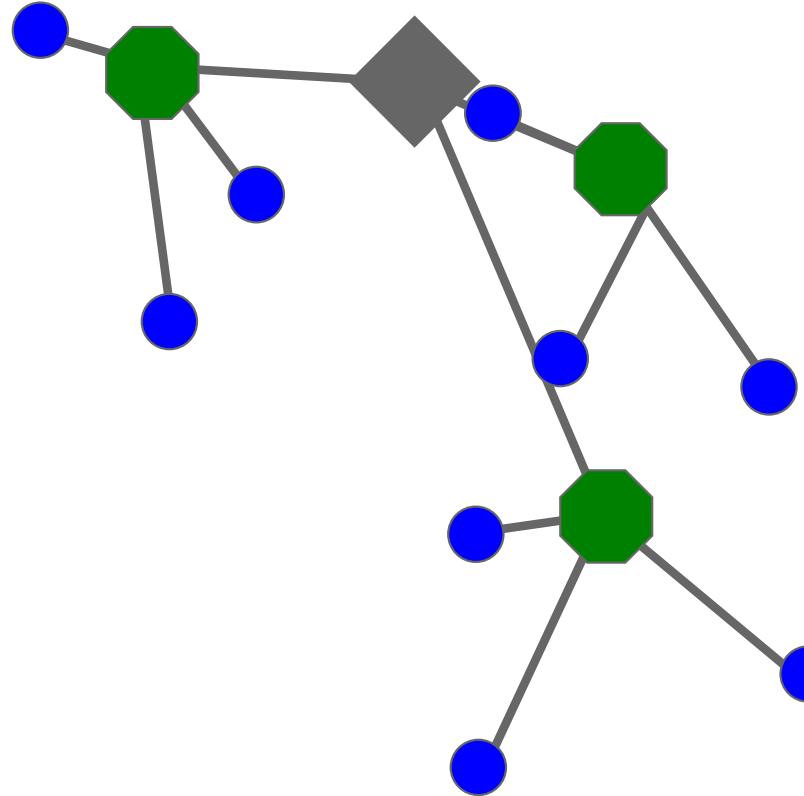


Recognition with K-tree

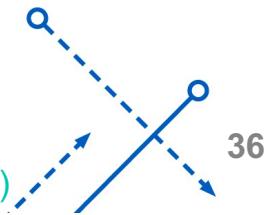


Following slides by David Nister (CVPR 2006)

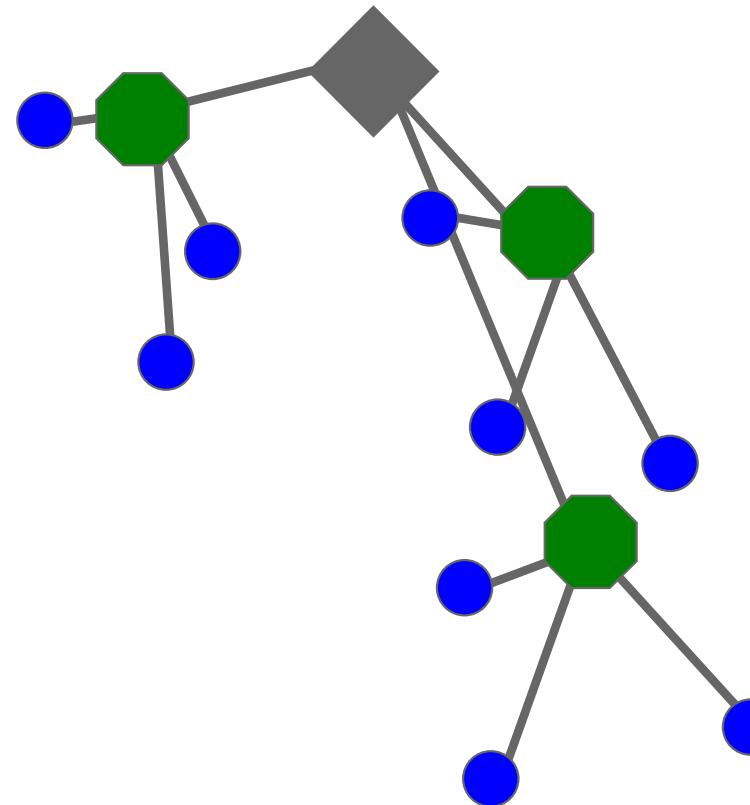
Recognition with K-tree



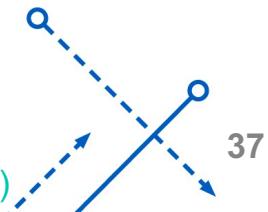
Following slides by David Nister (CVPR 2006)



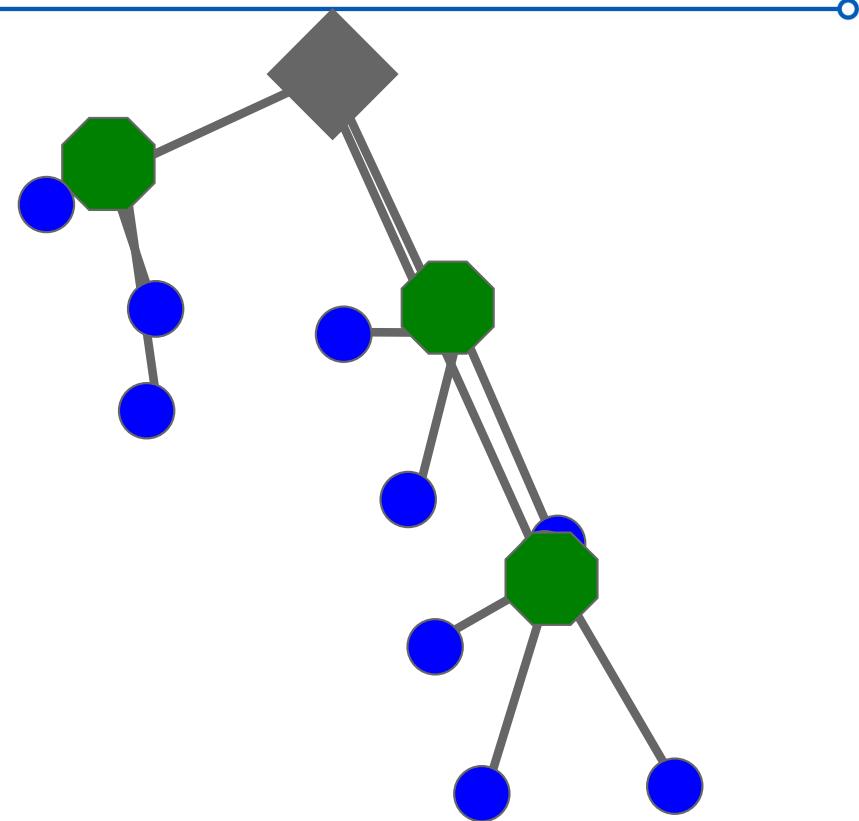
Recognition with K-tree



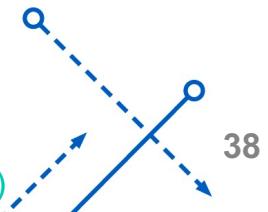
Following slides by David Nister (CVPR 2006)



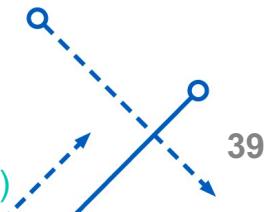
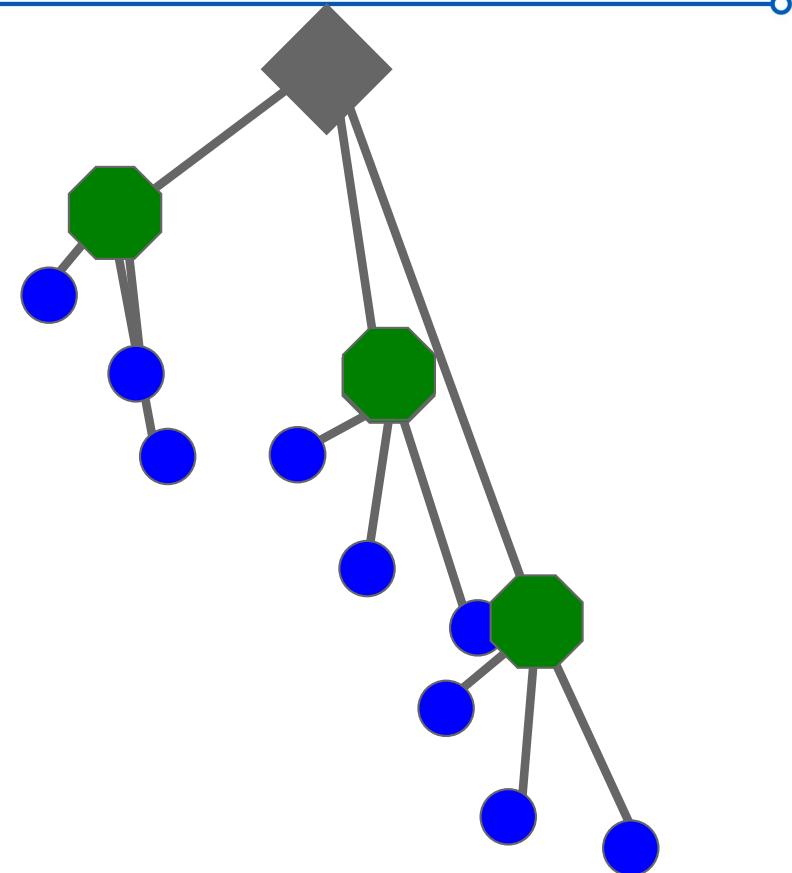
Recognition with K-tree



Following slides by David Nister (CVPR 2006)

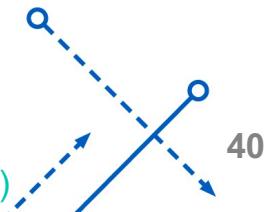
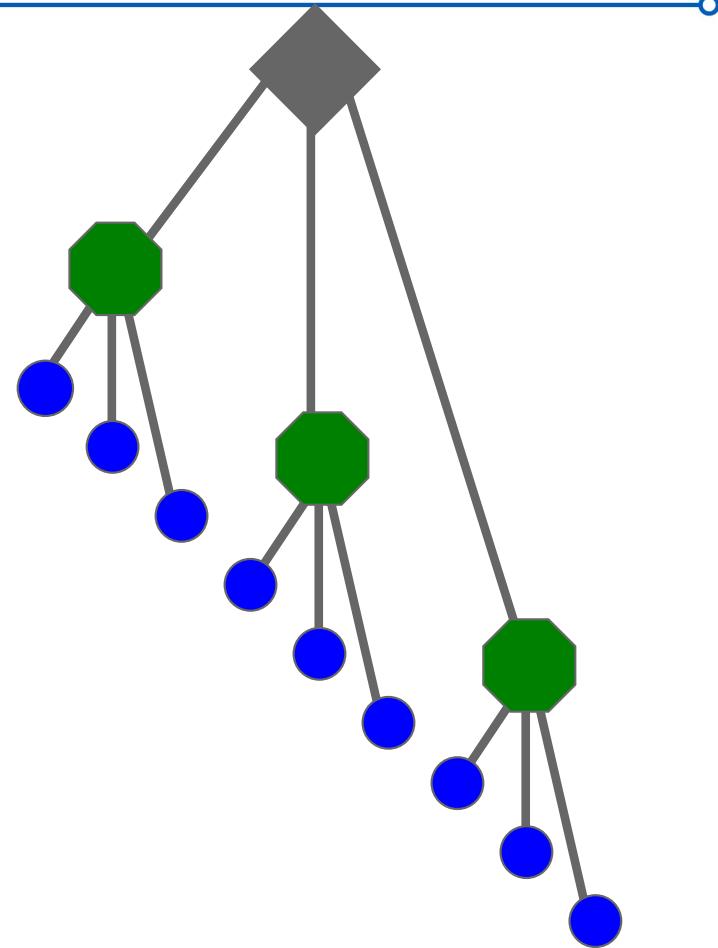


Recognition with K-tree



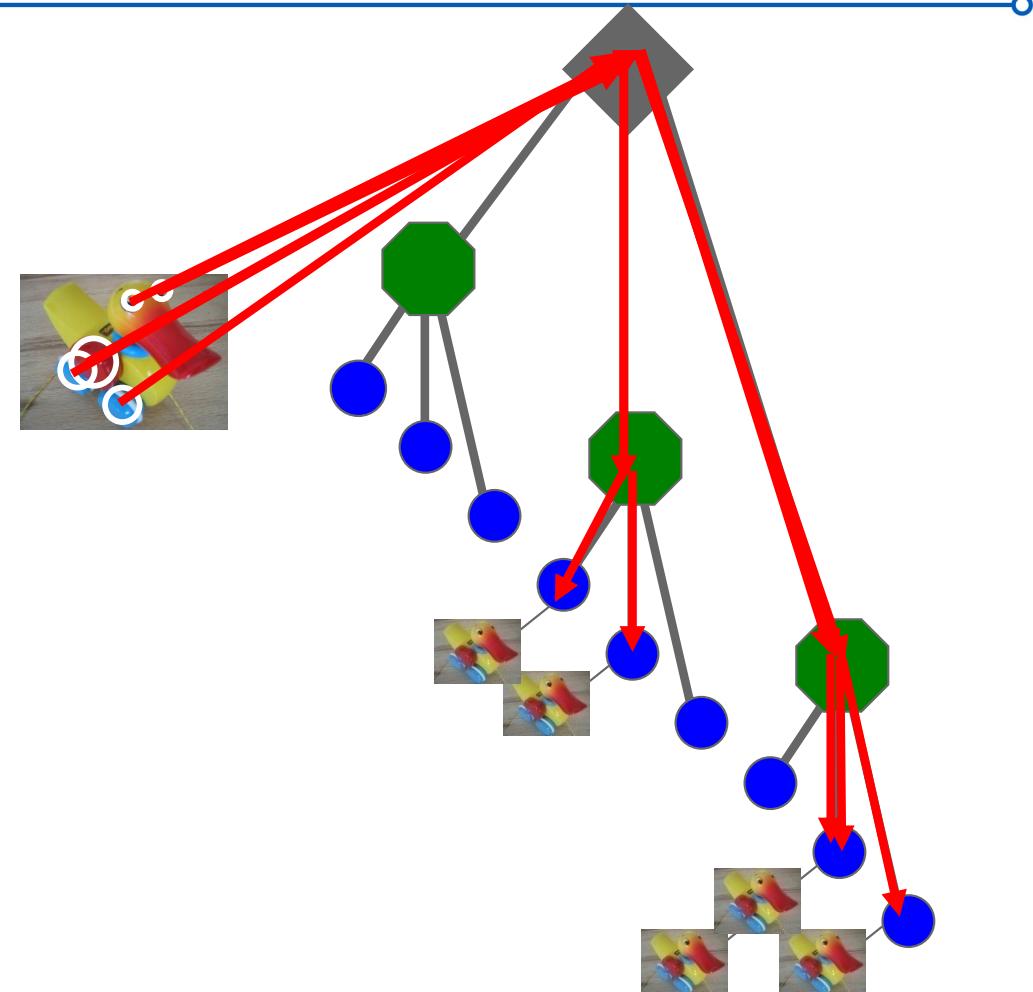
Following slides by David Nister (CVPR 2006)

Recognition with K-tree



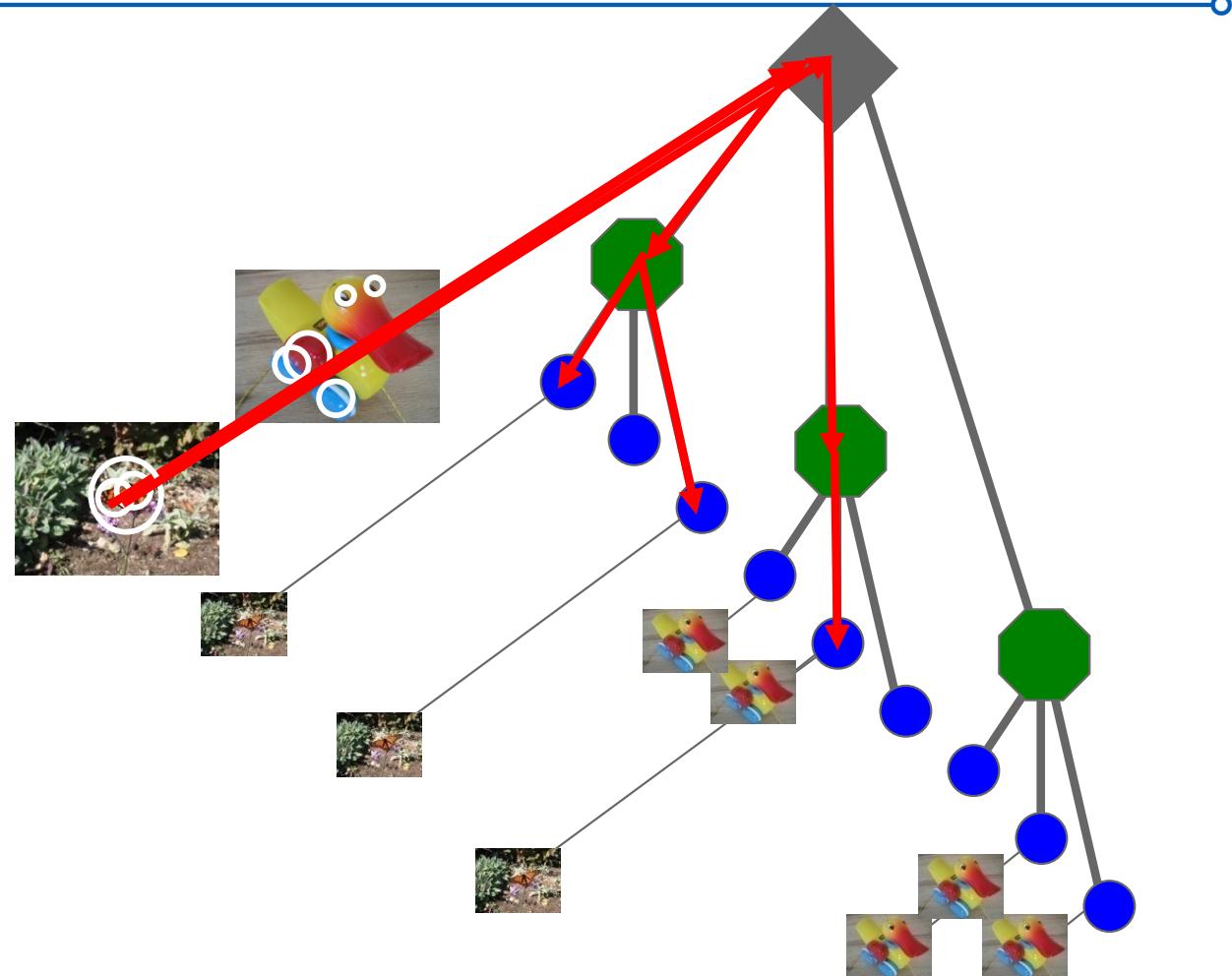
Following slides by David Nister (CVPR 2006)

Recognition with K-tree



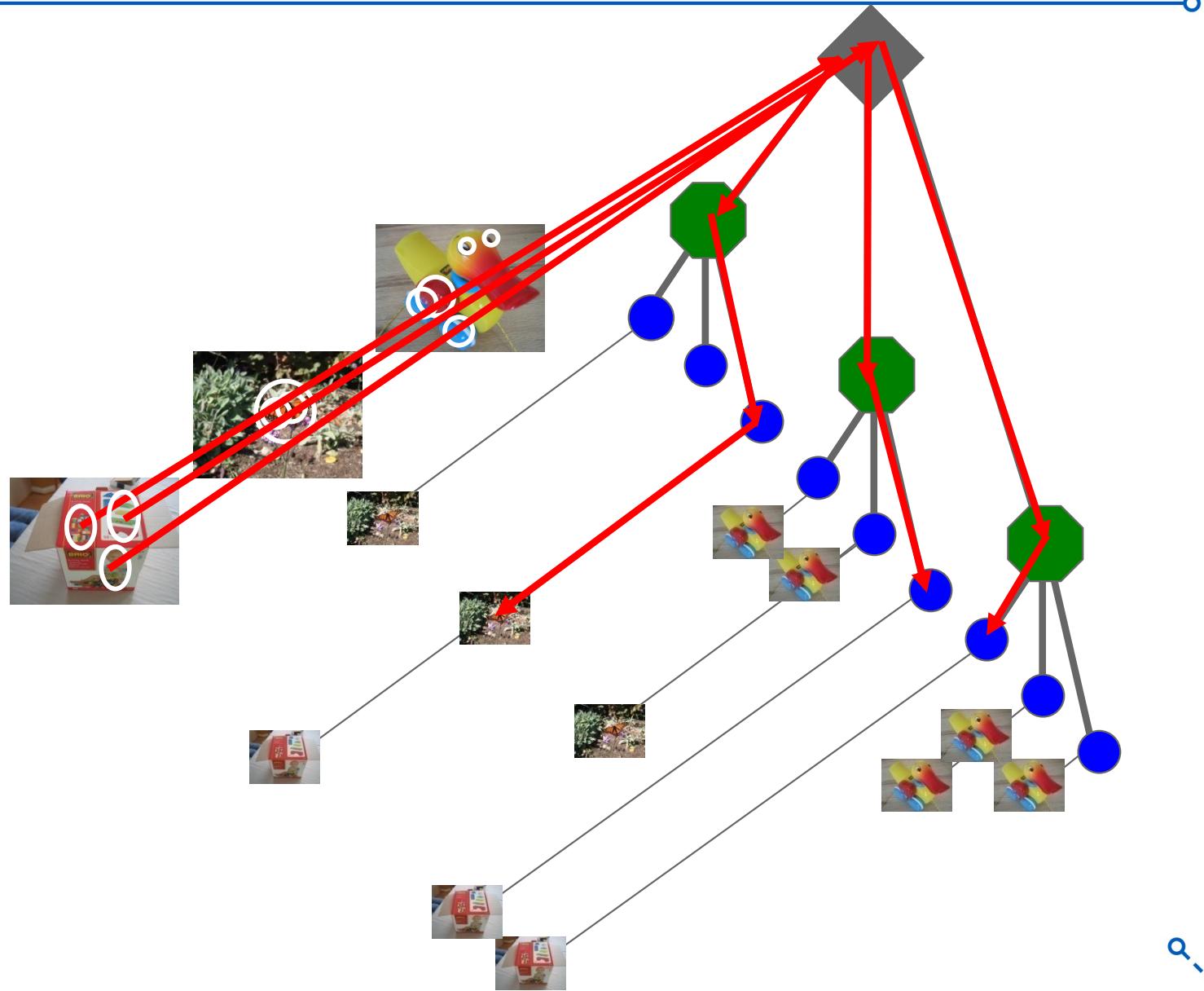
Following slides by David Nister (CVPR 2006)

Recognition with K-tree



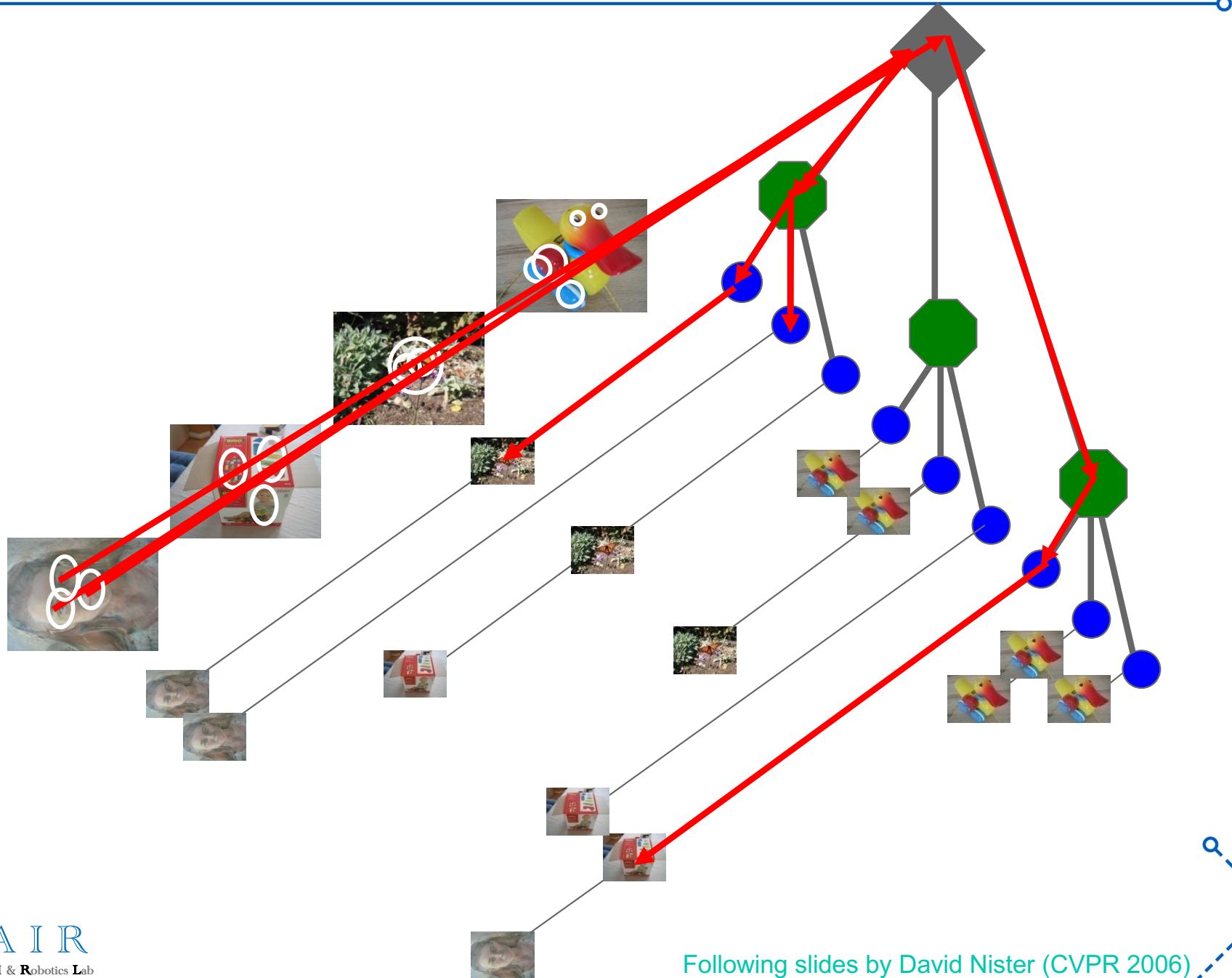
Following slides by David Nister (CVPR 2006)

Recognition with K-tree



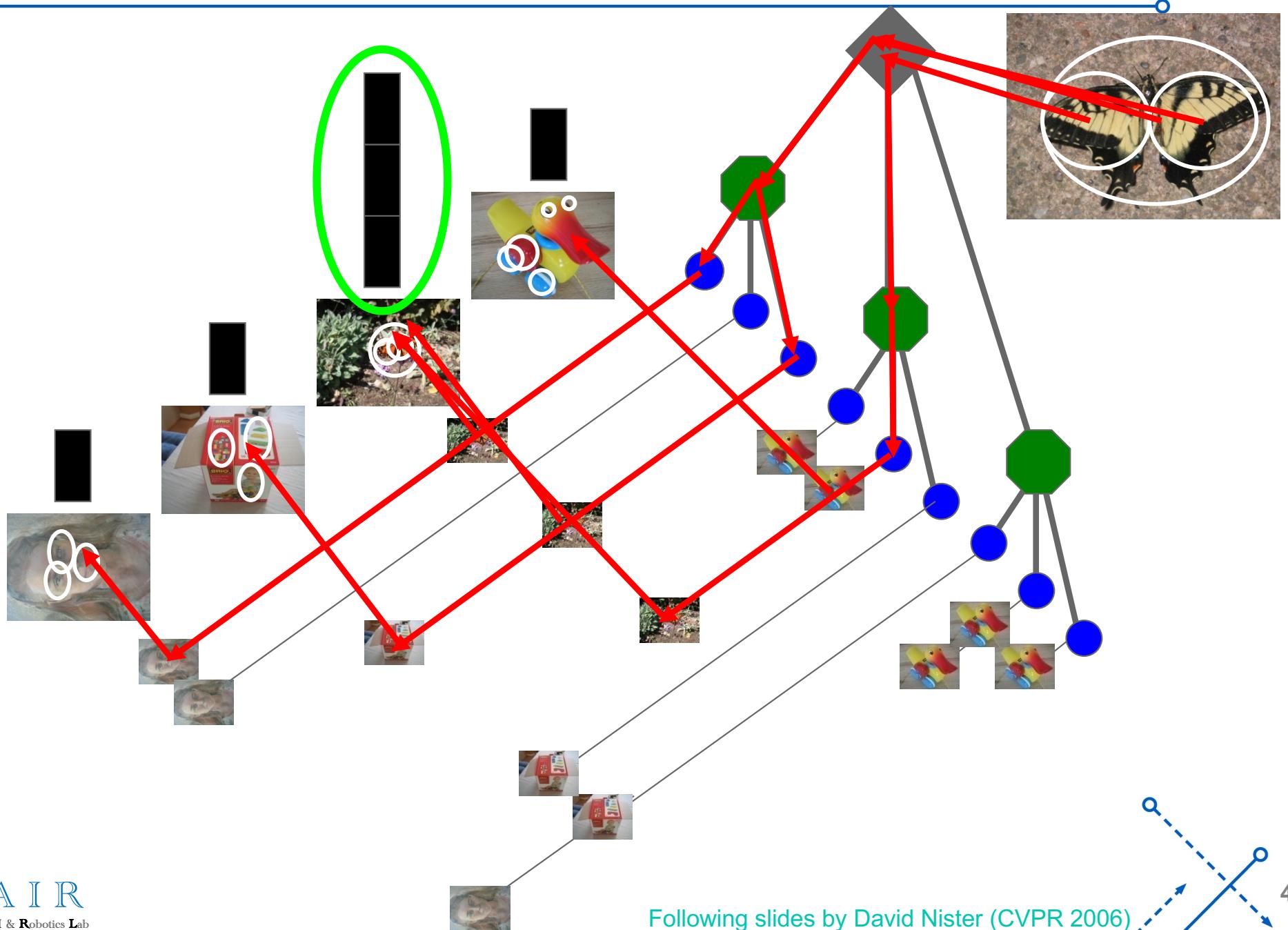
Following slides by David Nister (CVPR 2006)

Recognition with K-tree



Following slides by David Nister (CVPR 2006)

Recognition with K-tree



Vocabulary Tree: Performance

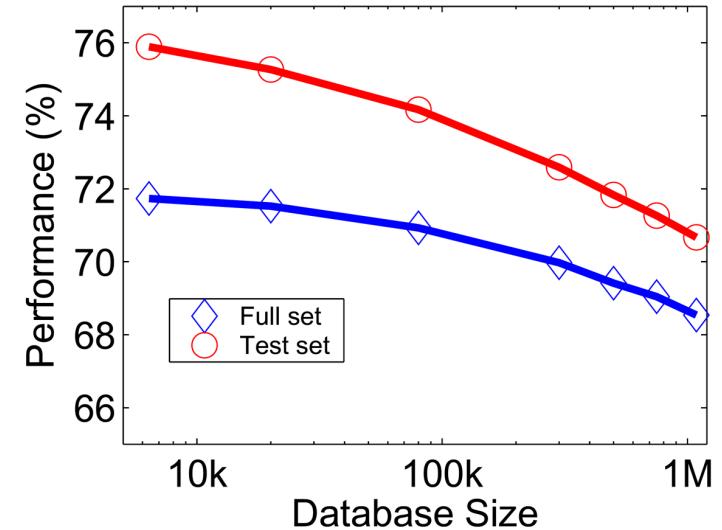
Evaluated on large databases

- Indexing with up to 1M images

Online recognition for database
of 50,000 CD covers

- Retrieval in ~1s

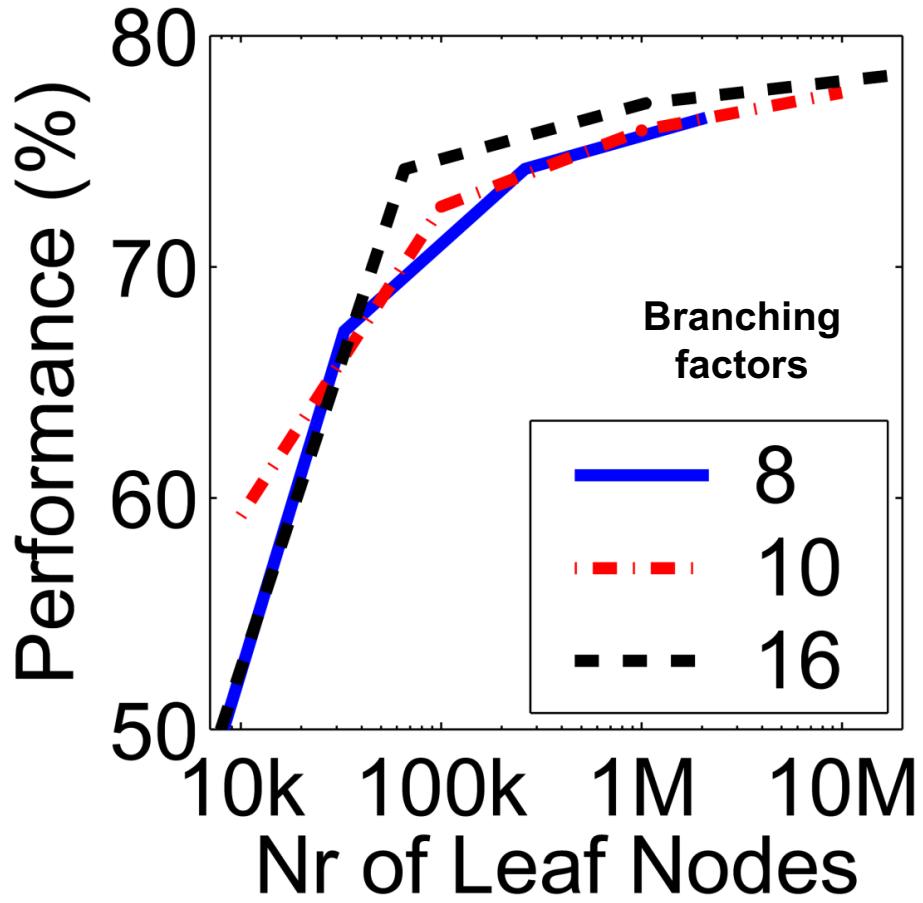
Find experimentally that large
vocabularies can be beneficial for
recognition



Instance recognition: remaining issues

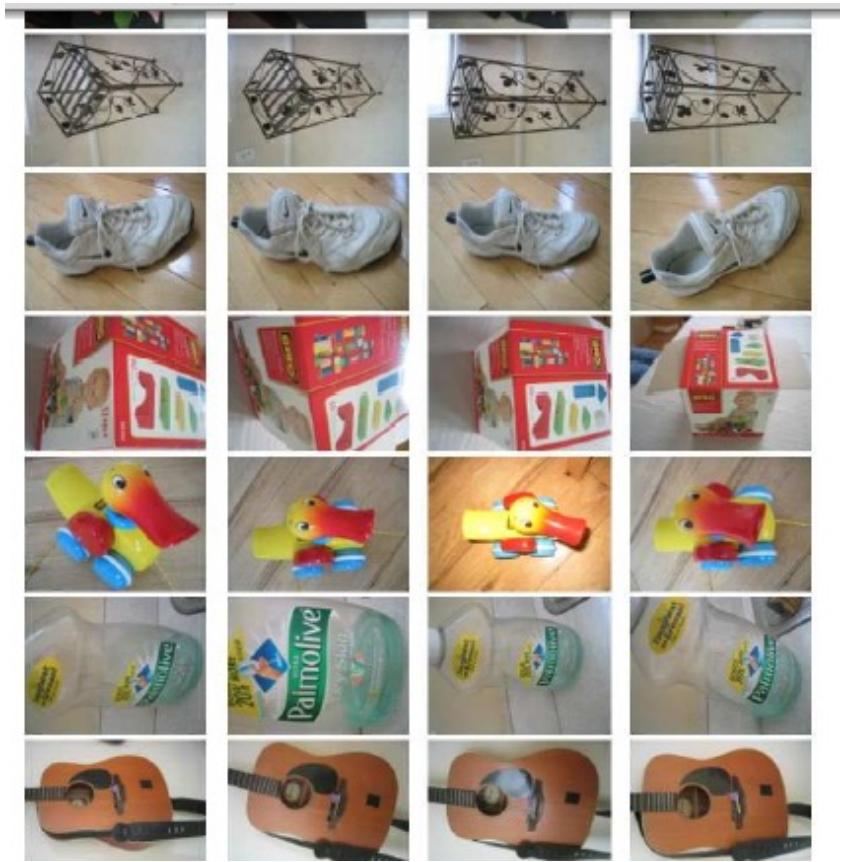
- How to summarize the content of an entire image?
And gauge overall similarity?
- How large should the vocabulary be? How to perform quantization efficiently?
- Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?
- How to score the retrieval results?

Vocabulary size



Influence on performance, sparsity

Results for recognition task with 6347 images



Nister & Stewenius, CVPR 2006

Kristen Grauman
48

Visual words/bags of words

Pro

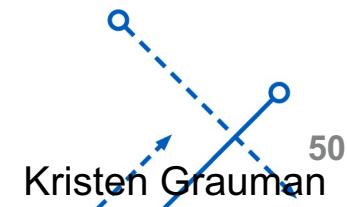
- + flexible to geometry / deformations / viewpoint
- + compact summary of image content
- + provides fixed dimensional vector representation for sets
- + good results in practice

Cons

- background and foreground mixed when bag covers whole image
- optimal vocabulary formation remains unclear
- basic model ignores geometry – must verify afterwards, or encode via features

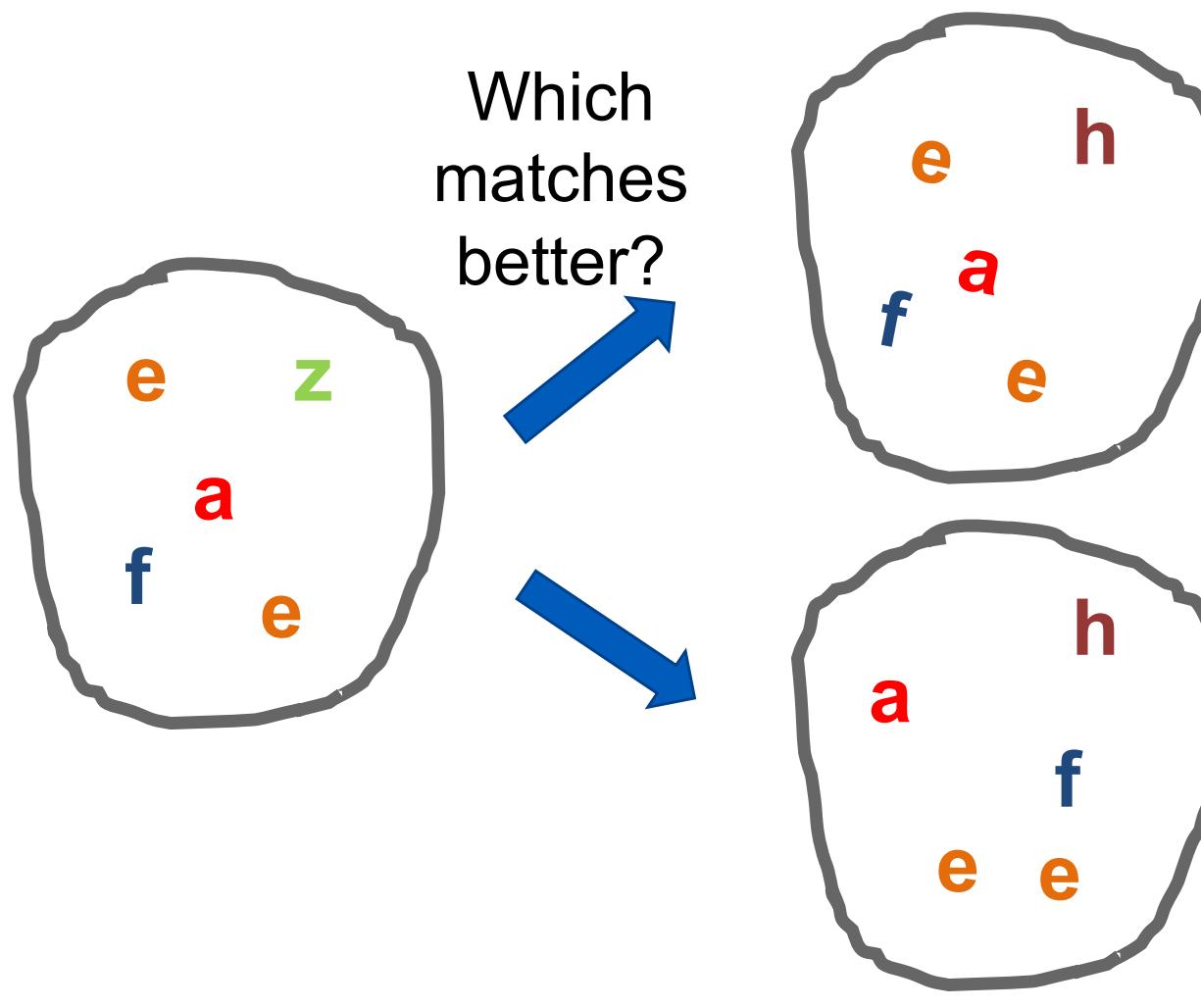
Instance recognition: remaining issues

- How to summarize the content of an entire image?
And gauge overall similarity?
- How large should the vocabulary be? How to perform quantization efficiently?
- Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?
- How to score the retrieval results?



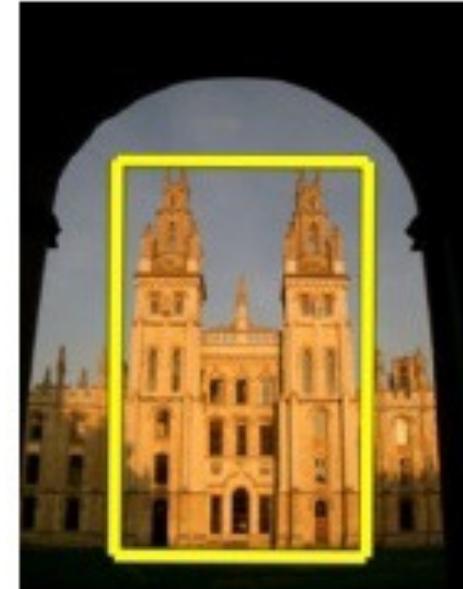
Can we be more accurate?

So far, we treat each image as containing a “bag of words”, with no spatial information

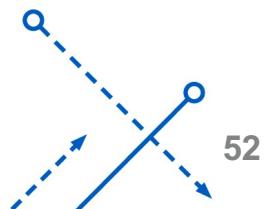


Can we be more accurate?

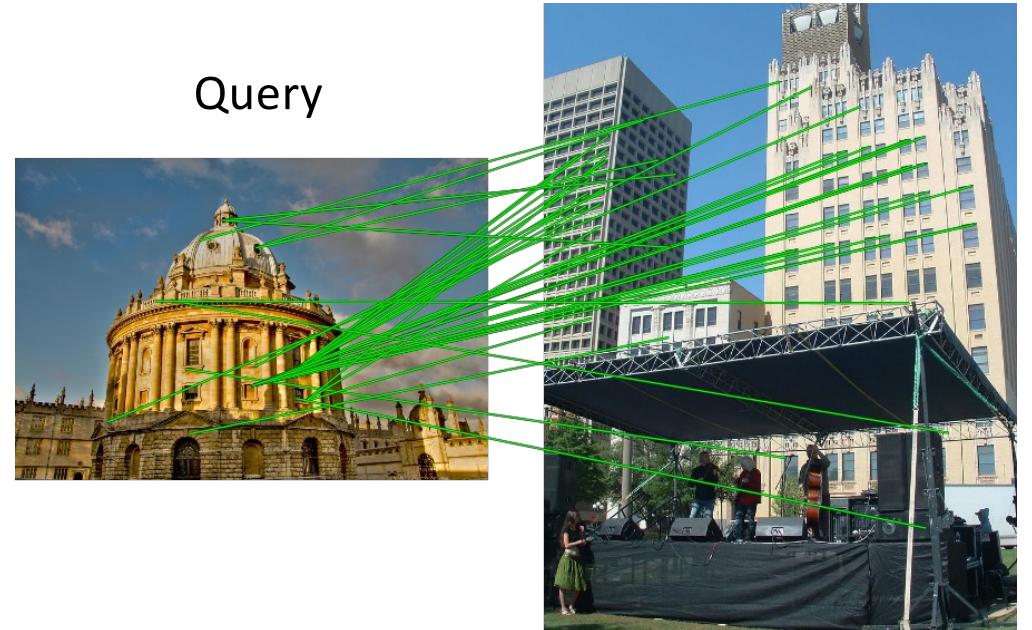
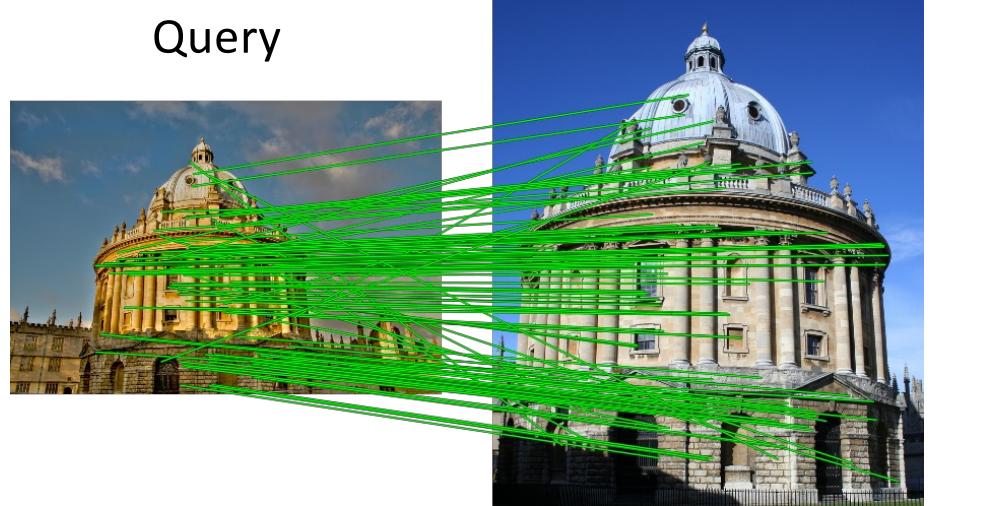
So far, we treat each image as containing a “bag of words”, with no spatial information



Real objects have consistent geometry

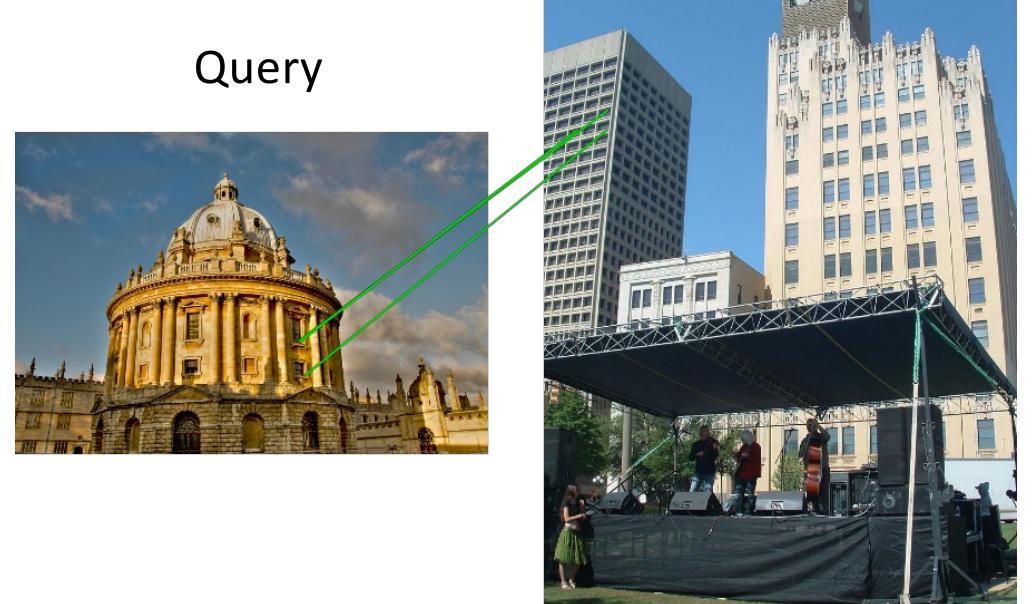
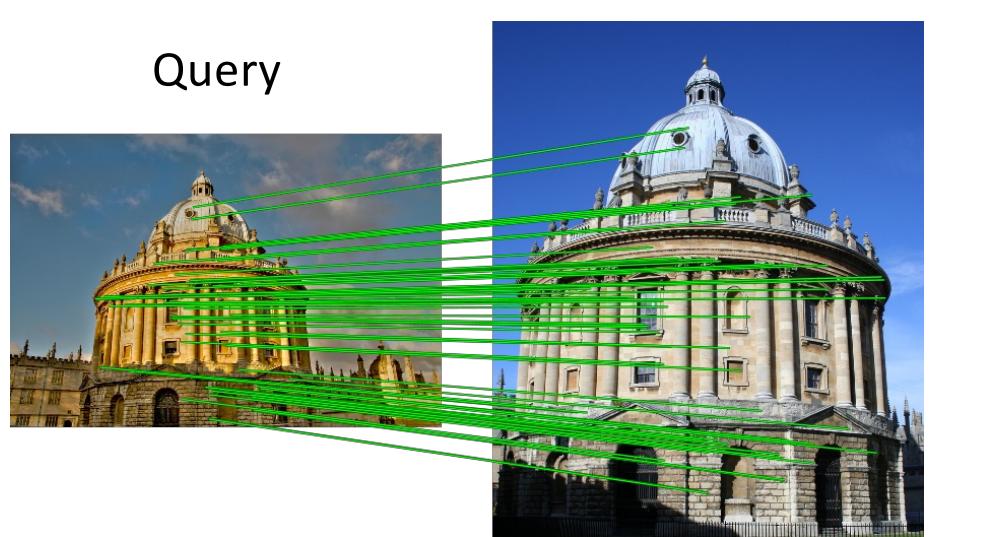


Spatial Verification



Both image pairs have many visual words in common.

Spatial Verification



Only some of the matches are mutually consistent

Spatial Verification: three basic strategies

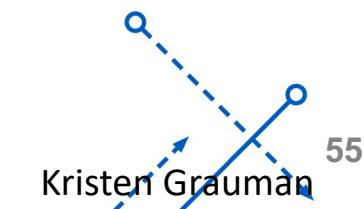
- RANSAC

- Typically sort by BoW similarity as initial filter
- Verify by checking support (inliers) for possible transformations
 - e.g., “success” if find a transformation with $> N$ inlier correspondences

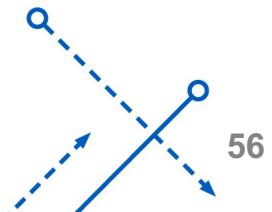
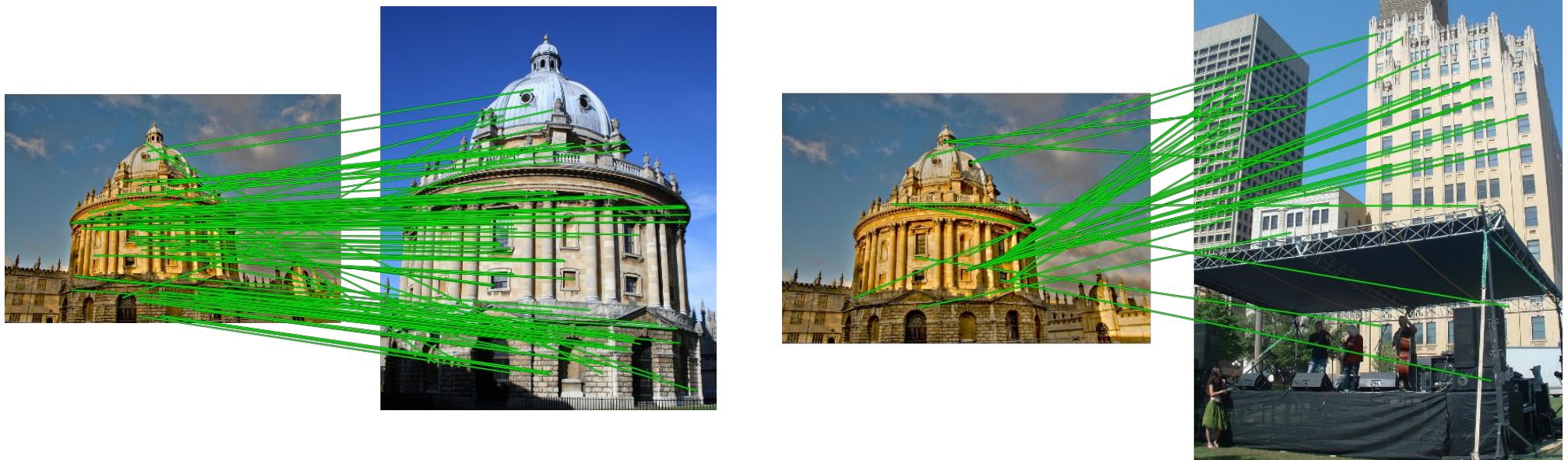
- Generalized Hough Transform

- Let each matched feature cast a vote on location, scale, orientation of the model object
- Verify parameters with enough votes

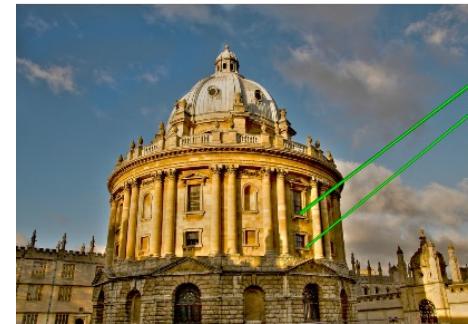
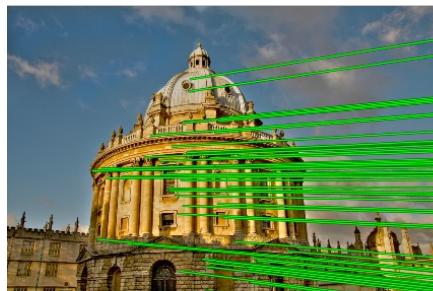
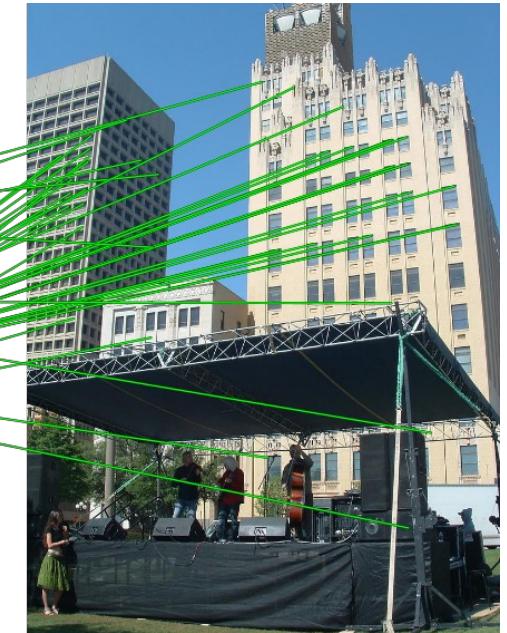
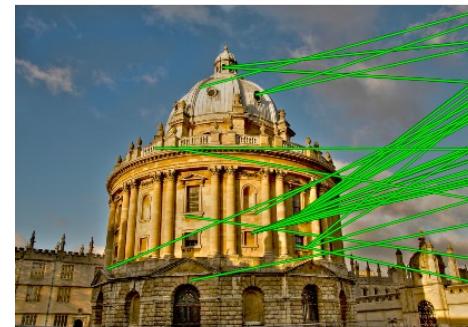
- Triplet Verification



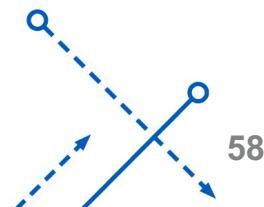
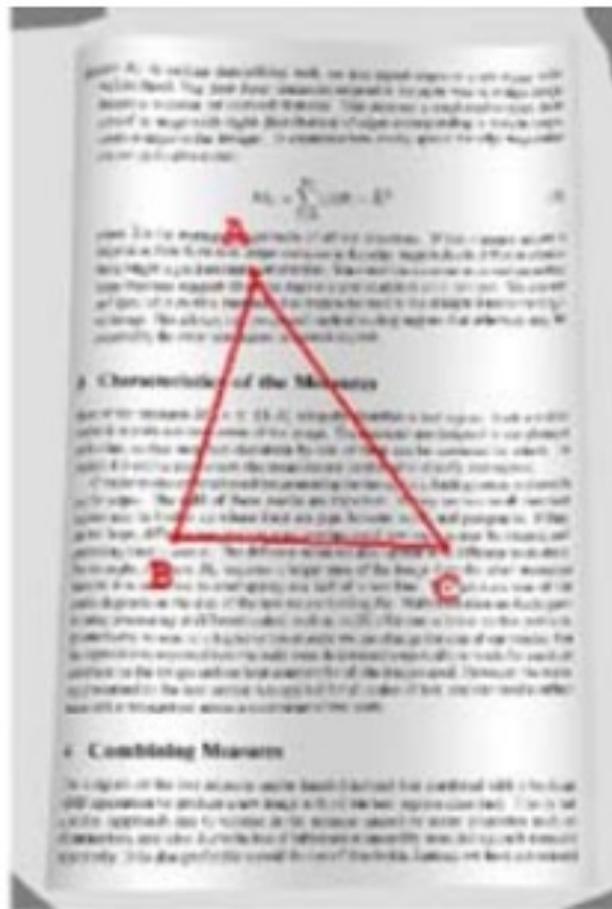
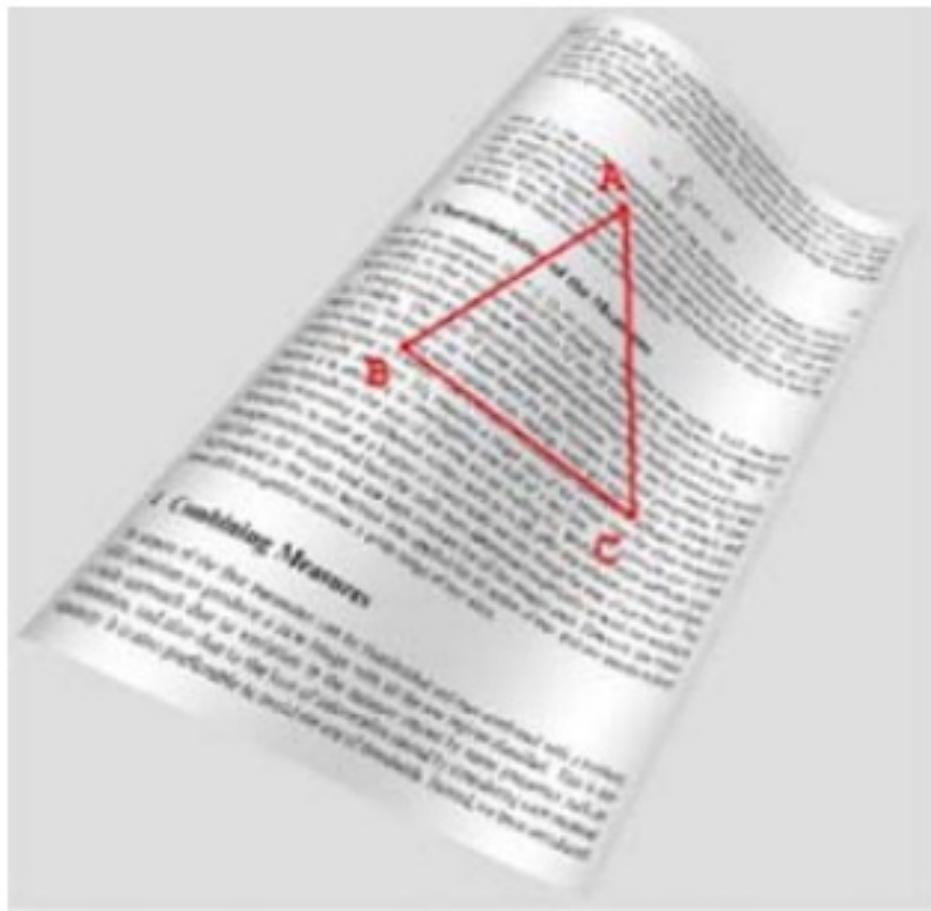
RANSAC verification



RANSAC

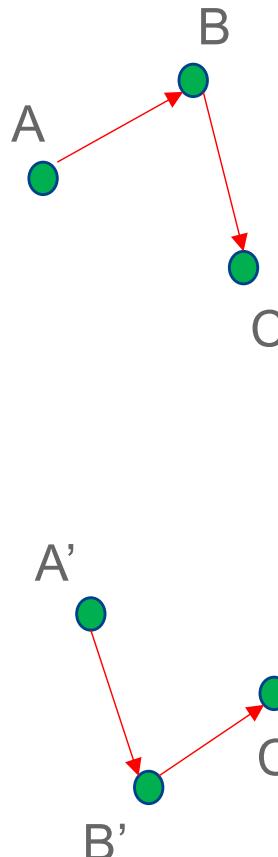


Triplet Verification



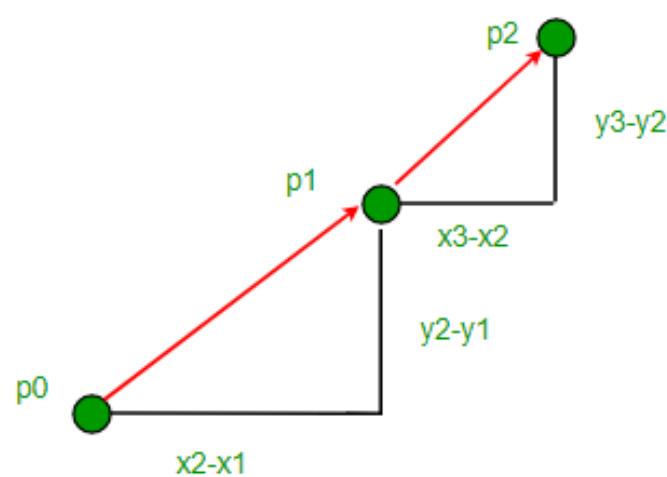
Using Slope to Determine Orientation

Consider the slopes....



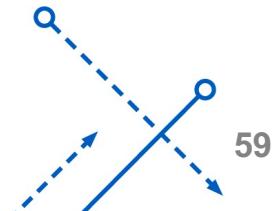
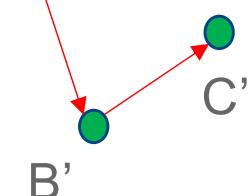
Slope(BC) – Slope(AB)?

< 0



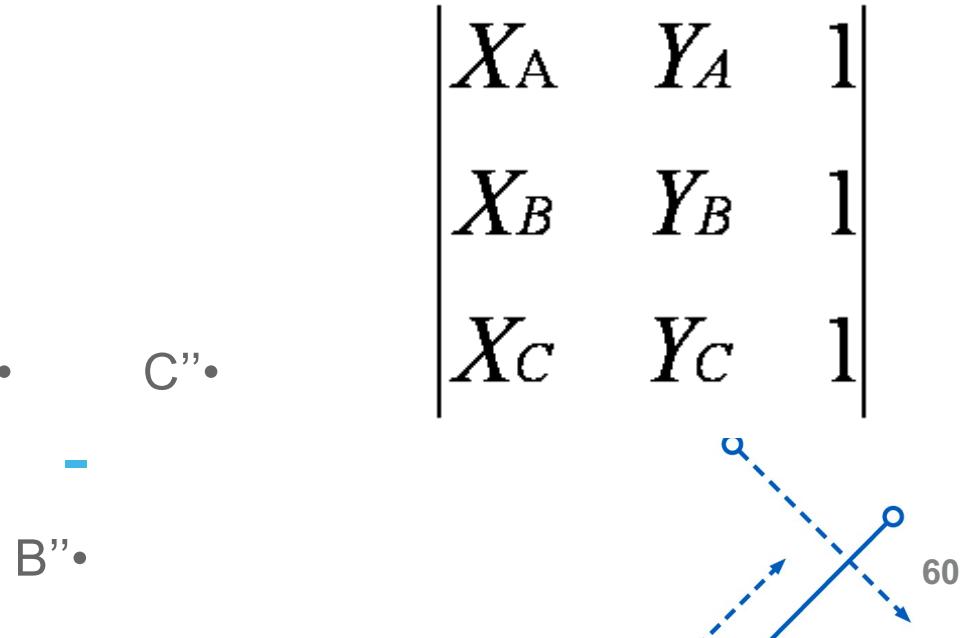
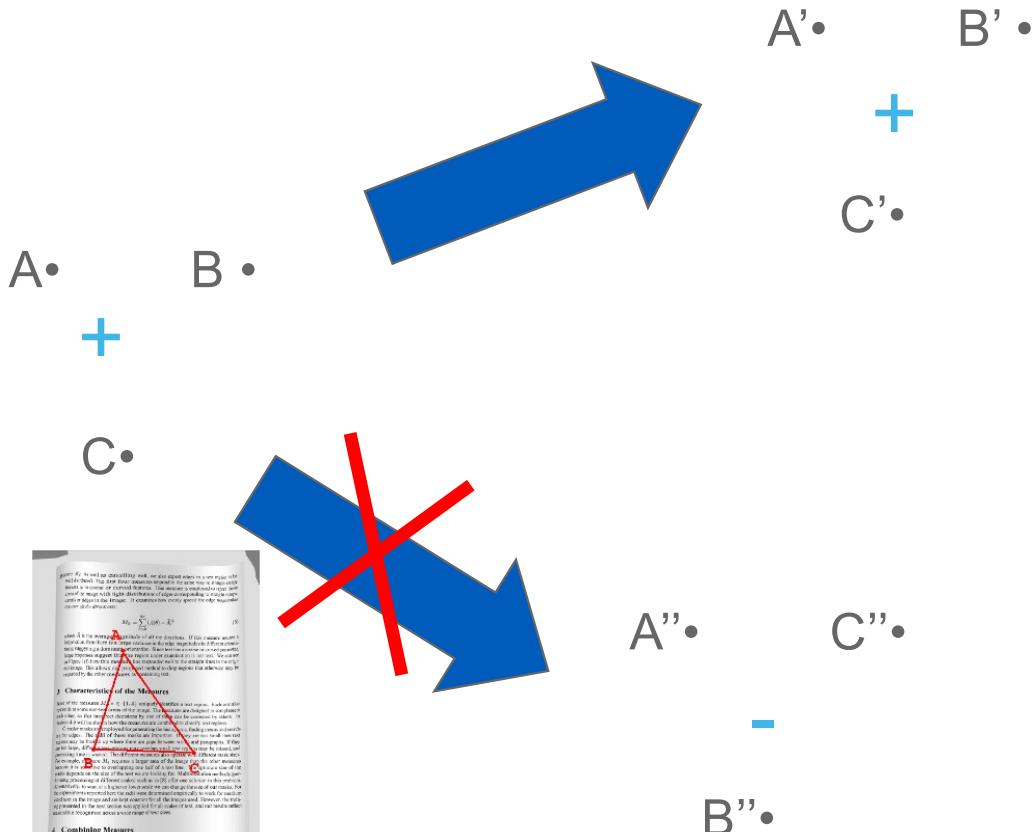
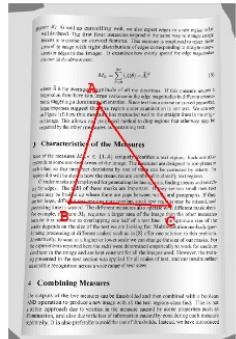
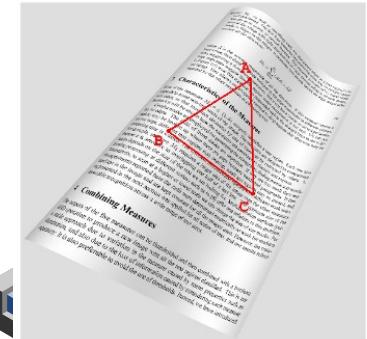
Slope(BC) – Slope(AB)?

> 0



Triplet Verification

- Use the spatial relationship between “visual words”.
- Orientation of triplet is invariant under rotation, translation, scaling, warping and even crinkling



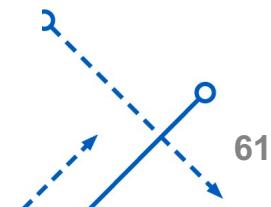
Triplet Verification: Scoring

- When viewed from another angle

$$Sign\left(\begin{vmatrix} X_A & Y_A & 1 \\ X_B & Y_B & 1 \\ X_C & Y_C & 1 \end{vmatrix}\right) \times Sign\left(\begin{vmatrix} X'_A & Y'_A & 1 \\ X'_B & Y'_B & 1 \\ X'_C & Y'_C & 1 \end{vmatrix}\right) = 1$$

- Score is simply

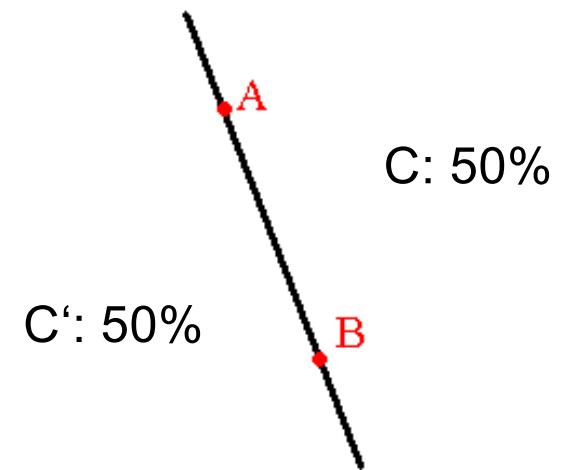
$$\sum_{A,B,C \in S} (Sign\left(\begin{vmatrix} X_A & Y_A & 1 \\ X_B & Y_B & 1 \\ X_C & Y_C & 1 \end{vmatrix}\right) \times Sign\left(\begin{vmatrix} X'_A & Y'_A & 1 \\ X'_B & Y'_B & 1 \\ X'_C & Y'_C & 1 \end{vmatrix}\right))$$



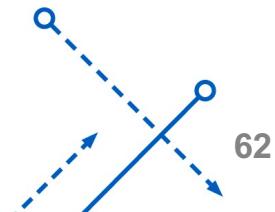
Triplet Verification: Failure rate

- Assume one triplet:
 - $P(A, B, C \text{ clockwise}) = 0.5$
 - $P(A, B, C \text{ counterclockwise}) = 0.5$
- Triplets are independent.
- Then possibility that M triplets out of N triplets accidentally satisfies orientation verification is

$$Q(N, M) = \left(\frac{1}{2}\right)^N \binom{N}{M}$$



N	10	30	50	60
M	5	20	40	50
Q(N,M)	0.25	0.027	0.0001	6.5×10^{-8}

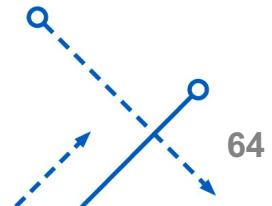


Instance recognition: remaining issues

- How to summarize the content of an entire image?
And gauge overall similarity?
- How large should the vocabulary be? How to perform quantization efficiently?
- Is having the same set of visual words enough to identify the object/scene? How to verify spatial agreement?
- How to score the retrieval results?

Precision and Recall

- Precision – How precise are the answers you gave?
 - $(\# \text{ Relevant}) / (\# \text{ Total Returned})$
- Recall – How many did you find of the ones that could be found?
 - $(\# \text{ Relevant}) / (\# \text{ Total Relevant})$
- F Measure – Harmonic Mean of Precision and Recall
 - $(2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$



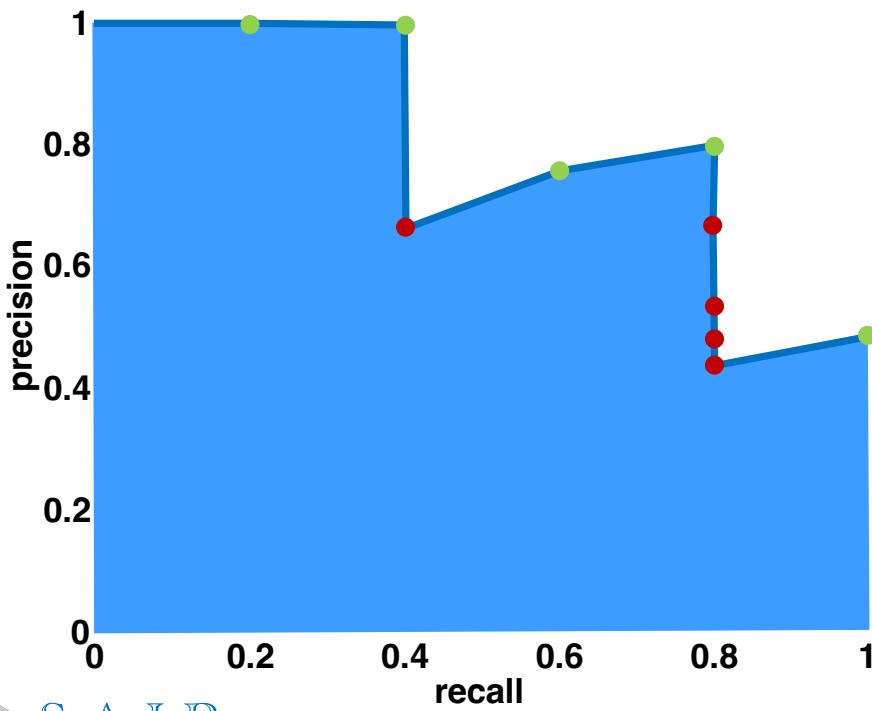
Precision and Recall Curve (PR curve)

Database size: 10 images

Relevant (total): 5 images

precision = #relevant / #returned

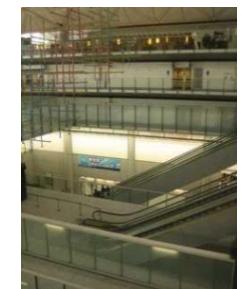
recall = #relevant / #total relevant



Query



Results (ordered):



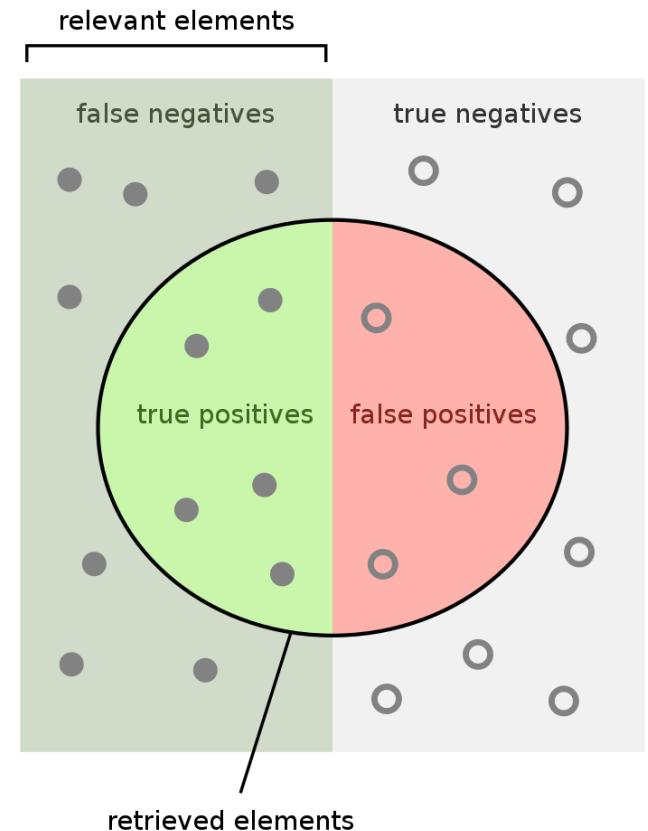
Slide credit: Ondrej Chum 65

Precision, Recall, F1 (Recap)

$$\text{Precision} = \frac{\text{True Positive}}{\text{Predicted Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Actual Positive}}$$

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



How many retrieved items are relevant?

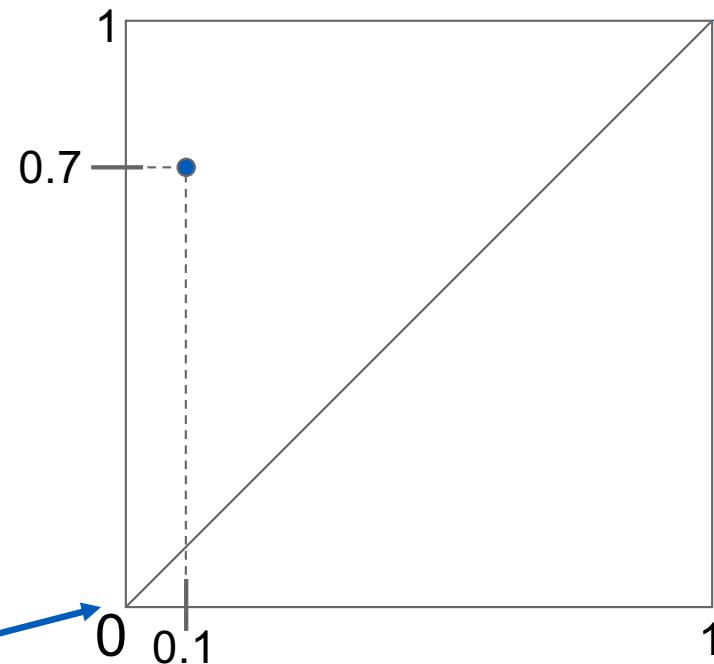
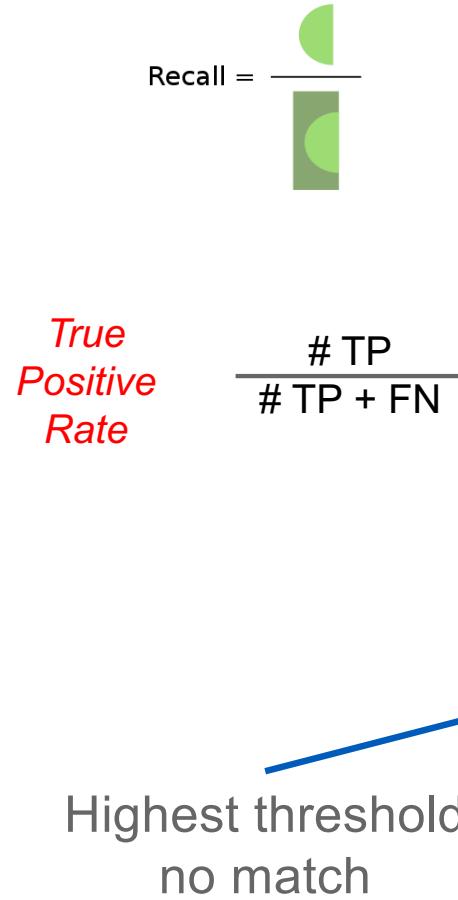
$$\text{Precision} = \frac{\text{green}}{\text{green} + \text{red}}$$

How many relevant items are retrieved?

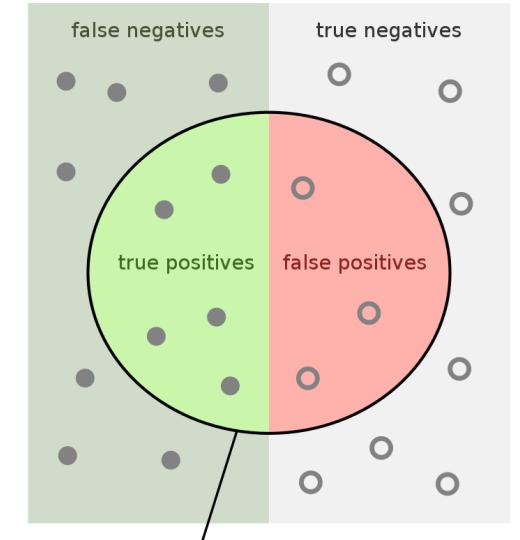
$$\text{Recall} = \frac{\text{green}}{\text{green} + \text{grey}}$$

ROC curve (Recap)

- How can we measure the performance of a feature matcher?

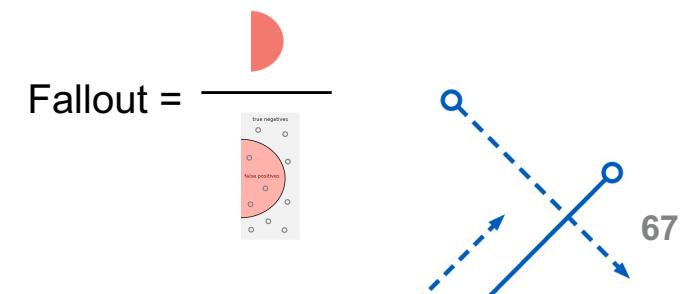


Lowest threshold
all pairs as match



False Positive Rate

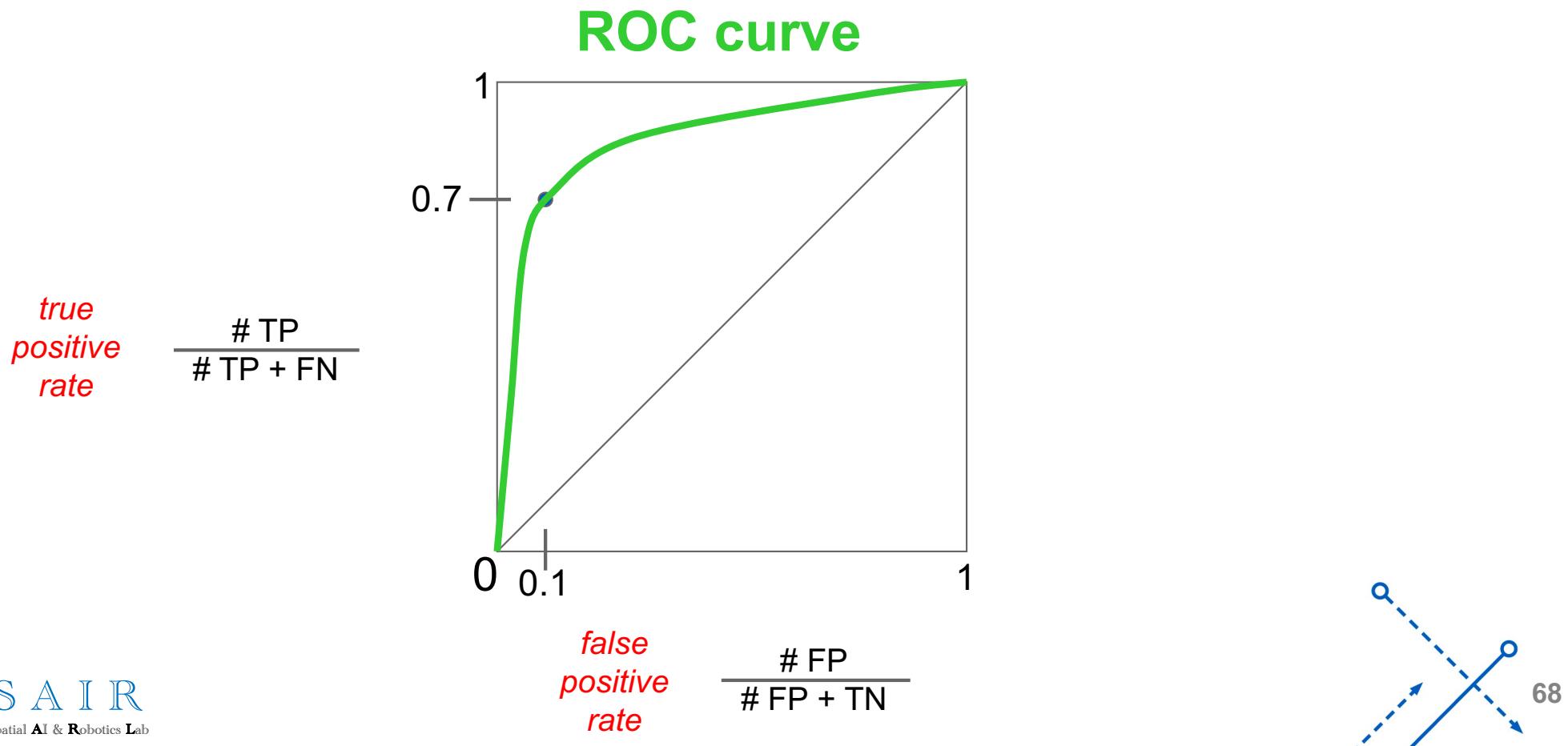
$$\frac{\# \text{FP}}{\# \text{FP} + \text{TN}}$$



ROC curve (Recap)

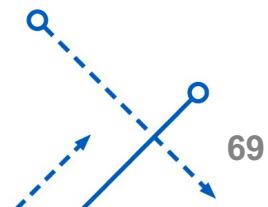
ROC Curves (Receiver Operator Characteristic)

- Generated by counting # correct/incorrect matches, for different thresholds.
- Want to maximize area under the curve (AUC)
- Useful for comparing different feature matching methods.

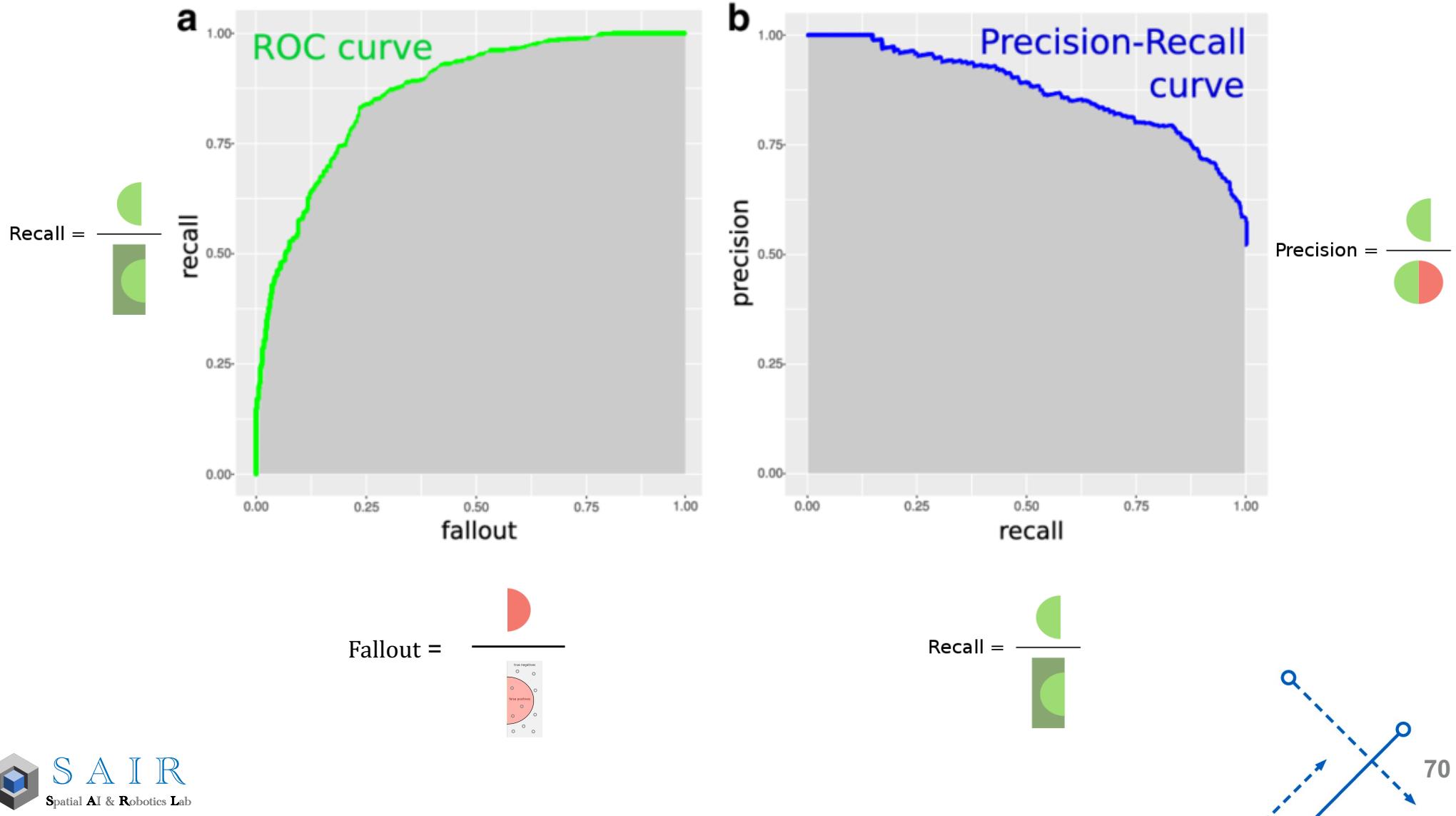


ROC curve VS. PR curve

- ROC Curves (Lecture 4)
 - summarize the trade-off between the **true positive rate** and **false positive rate** for a predictive model using different probability thresholds.
- PR curves
 - summarize the trade-off between the **true positive rate** and the **positive predictive value** for a predictive model using different probability thresholds.



ROC curve VS. PR curve



ROC curve VS. PR curve

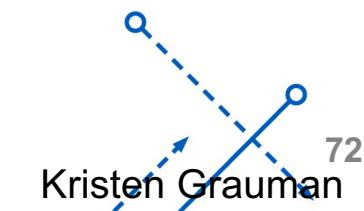
- ROC: appropriate for balanced datasets/classes.
- PR curves: appropriate for imbalanced datasets.
- If the proportion of positive to negative instances changes in a test set, the ROC curves will not change, so do not depend on class distributions.
- ROC “weights” equally true positives and true negatives (not expected sometimes, e.g., place recognition).



[More explanation](#)

Summary

- **Matching local invariant features**
 - Useful not only to provide matches for multi-view geometry, but also to find objects and scenes.
- **Bag of words** representation: quantize feature space to make discrete set of visual words
 - Summarize image by distribution of words
 - Index individual words
- **Inverted index**: pre-compute index to enable faster search at query time
- **Recognition of instances via alignment**: matching local features followed by spatial verification
 - Robust fitting : RANSAC, GHT

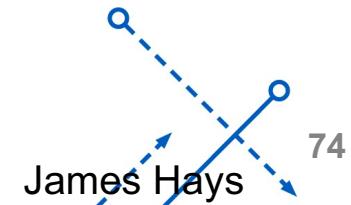


Lessons from a Decade Later

- For Category recognition
 - Bag of Feature models remained the state of the art until Deep Learning.
 - Spatial layout either isn't that important or its too difficult to encode.
 - Quantization error is, in fact, the bigger problem. Advanced feature encoding methods address this.
 - Bag of feature models are nearly obsolete. At best they seem to be inspiring tweaks to deep models, e.g., [NetVLAD](#) (deep learning model for place recognition).

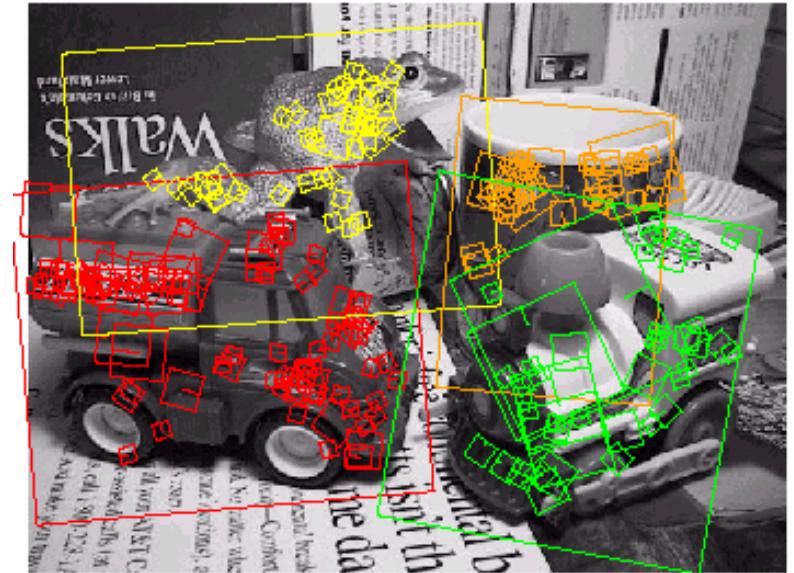
Lessons from a Decade Later

- For *instance* retrieval (this lecture)
 - Deep learning has taking over except in the field of SLAM.
 - Learn better local features (replace SIFT), e.g., MatchNet
 - or learn better image embeddings (replace the histograms of visual features), e.g., Vo and Hays 2016.
 - or learn to do spatial verification, e.g., DeTone, Malisiewicz, and Rabinovich 2016.
 - or learn a monolithic deep network to recognition all locations, e.g., Google's PlaNet 2016.

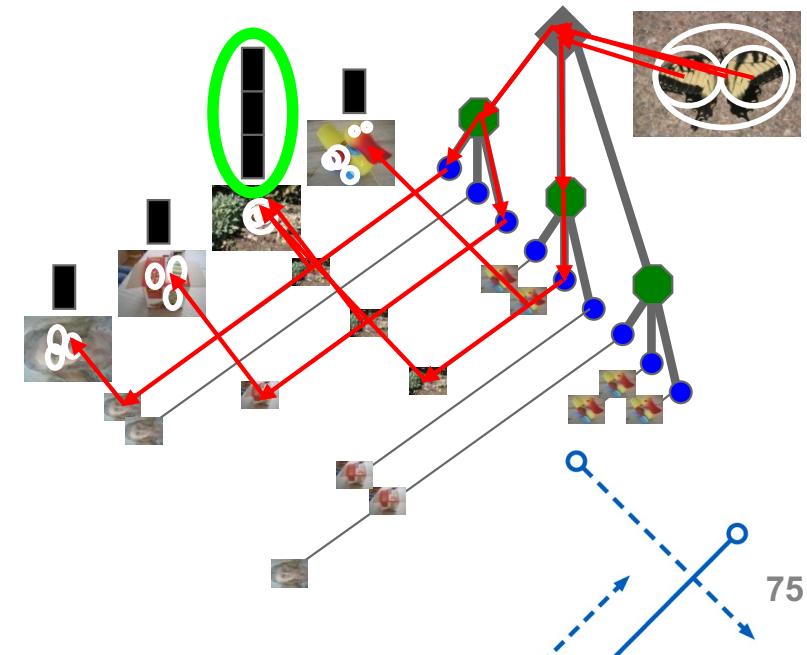


Things to remember

- Object instance recognition
 - Find keypoints, compute descriptors
 - Match descriptors
 - Vote for / fit affine parameters
 - Return object if # inliers > T



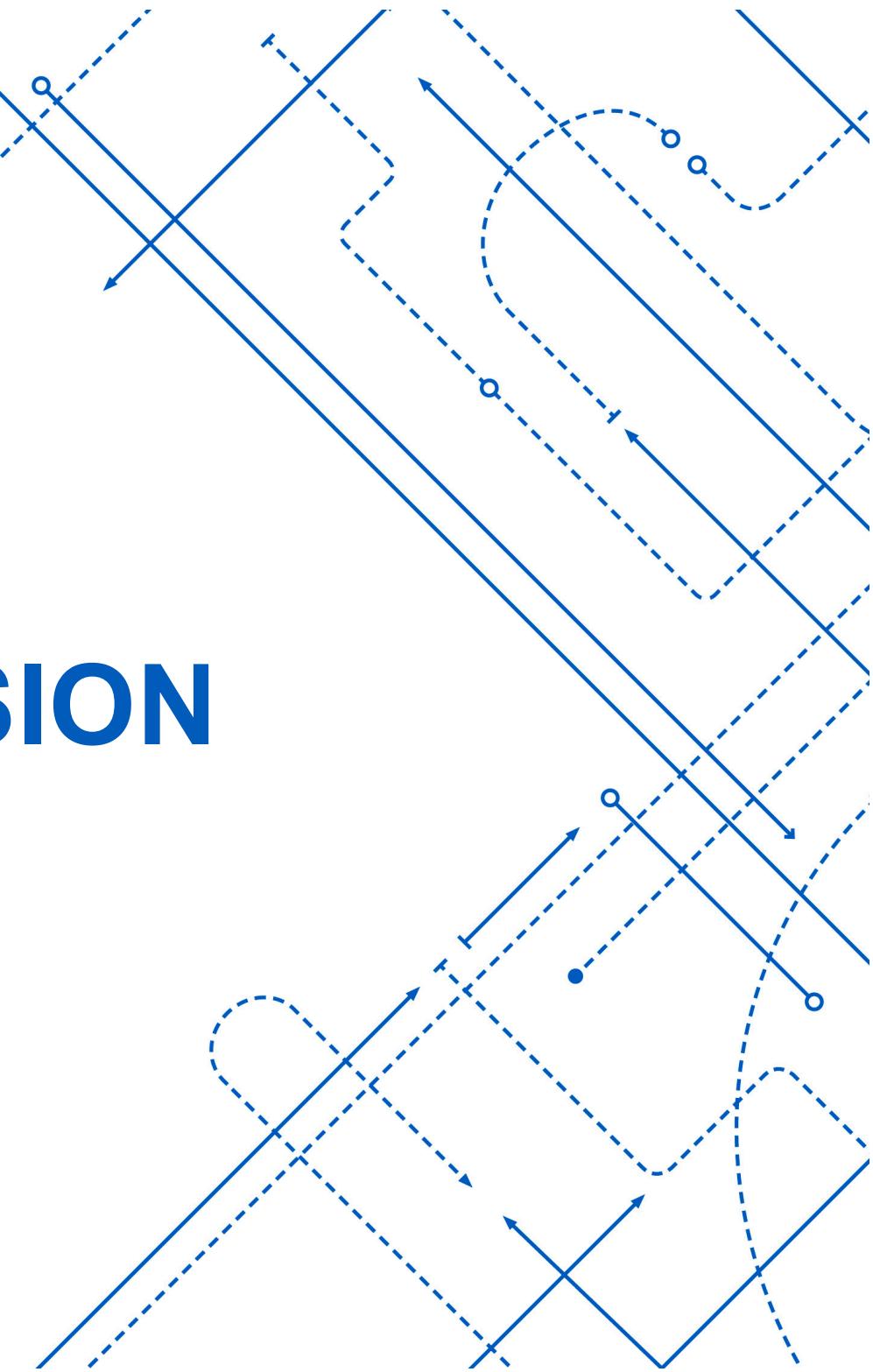
- Keys to efficiency
 - Visual words
 - Used for many applications
 - Inverse document file
 - Used for web-scale search



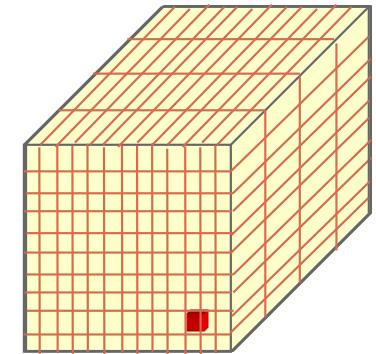


COMPUTER VISION

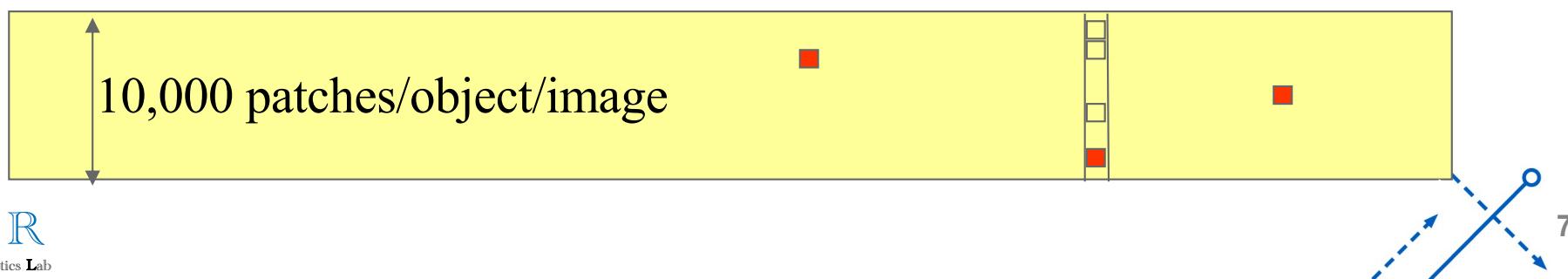
Object Detection



Why is detection hard?

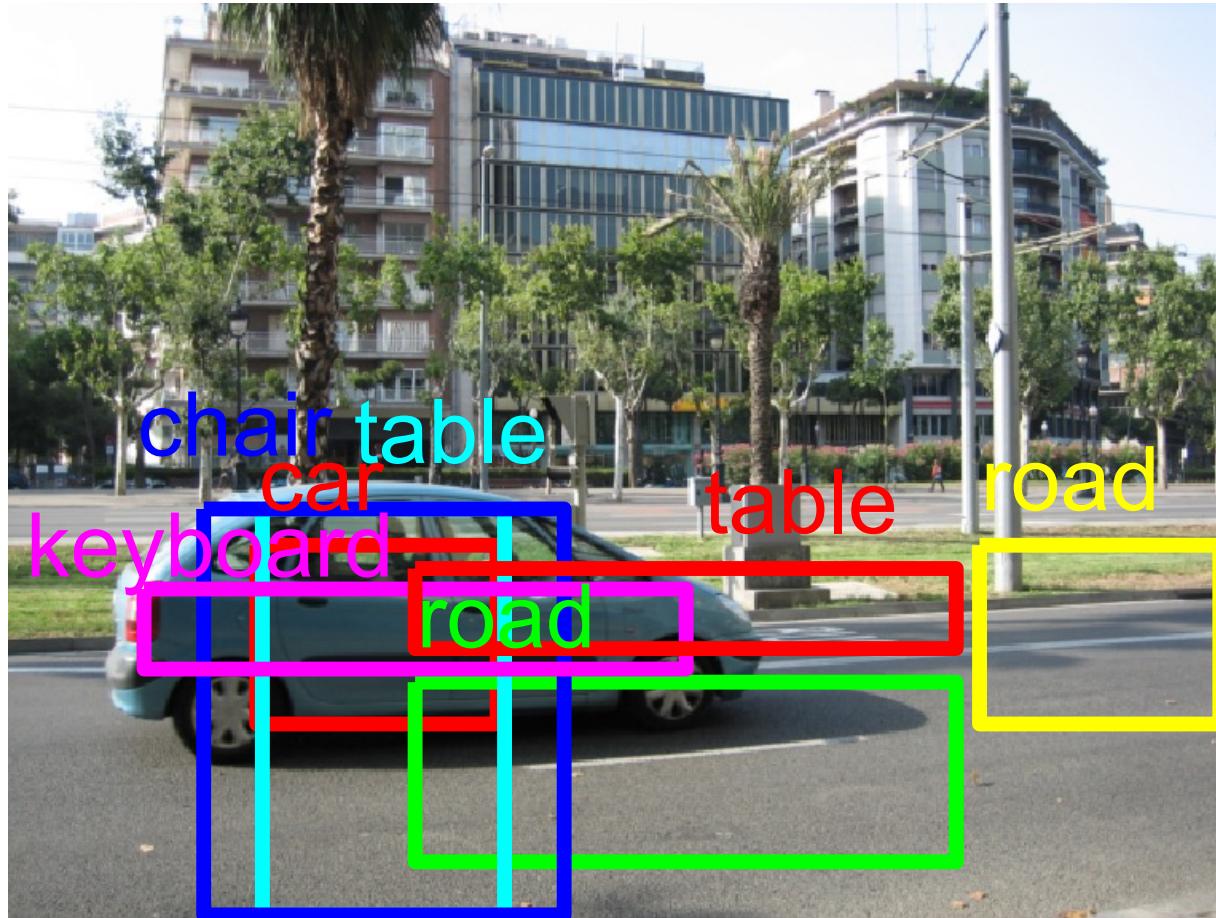


We want to do this
for ~ 1000 objects
1,000,000 images/day



Why is detection hard?

If we have 1000 categories (detectors), and each detector produces 1 false every 10 images, we will have 100 false alarms per image... pretty much garbage...



Local information is helpful



Slide credit: A. Torralba 79

Is local information enough?



Information

Local features

Contextual features

Distance

Slide credit: A. Torralba 80

Is local information even enough?

We know there is a keyboard present in this scene even if we cannot see it clearly.



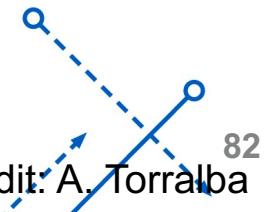
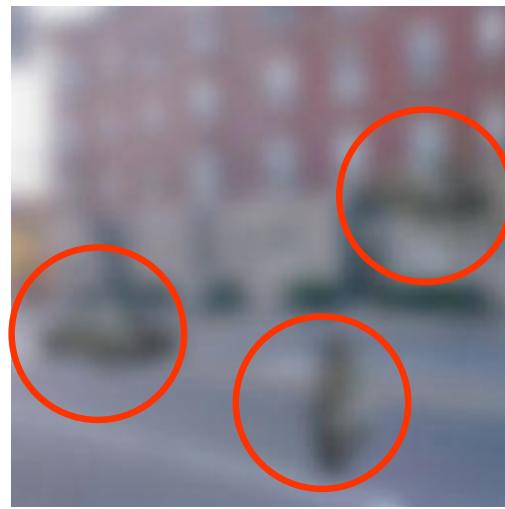
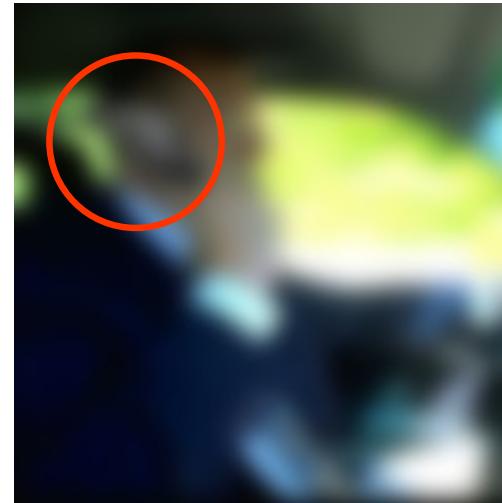
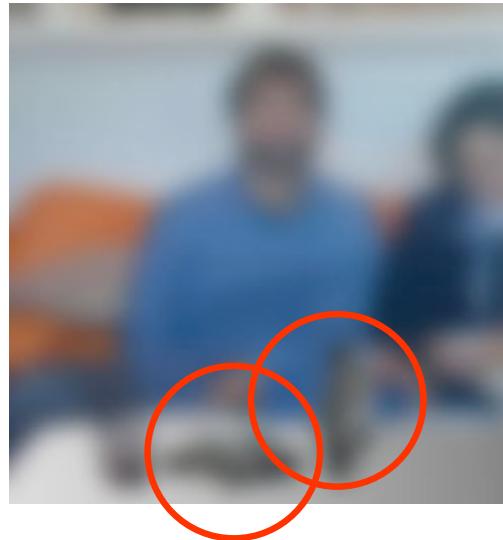
We know there is no keyboard present in this scene



... even if there is one indeed.

Slide credit: A. Torralba

The multiple personalities of a blob



Slide credit: A. Torralba 82

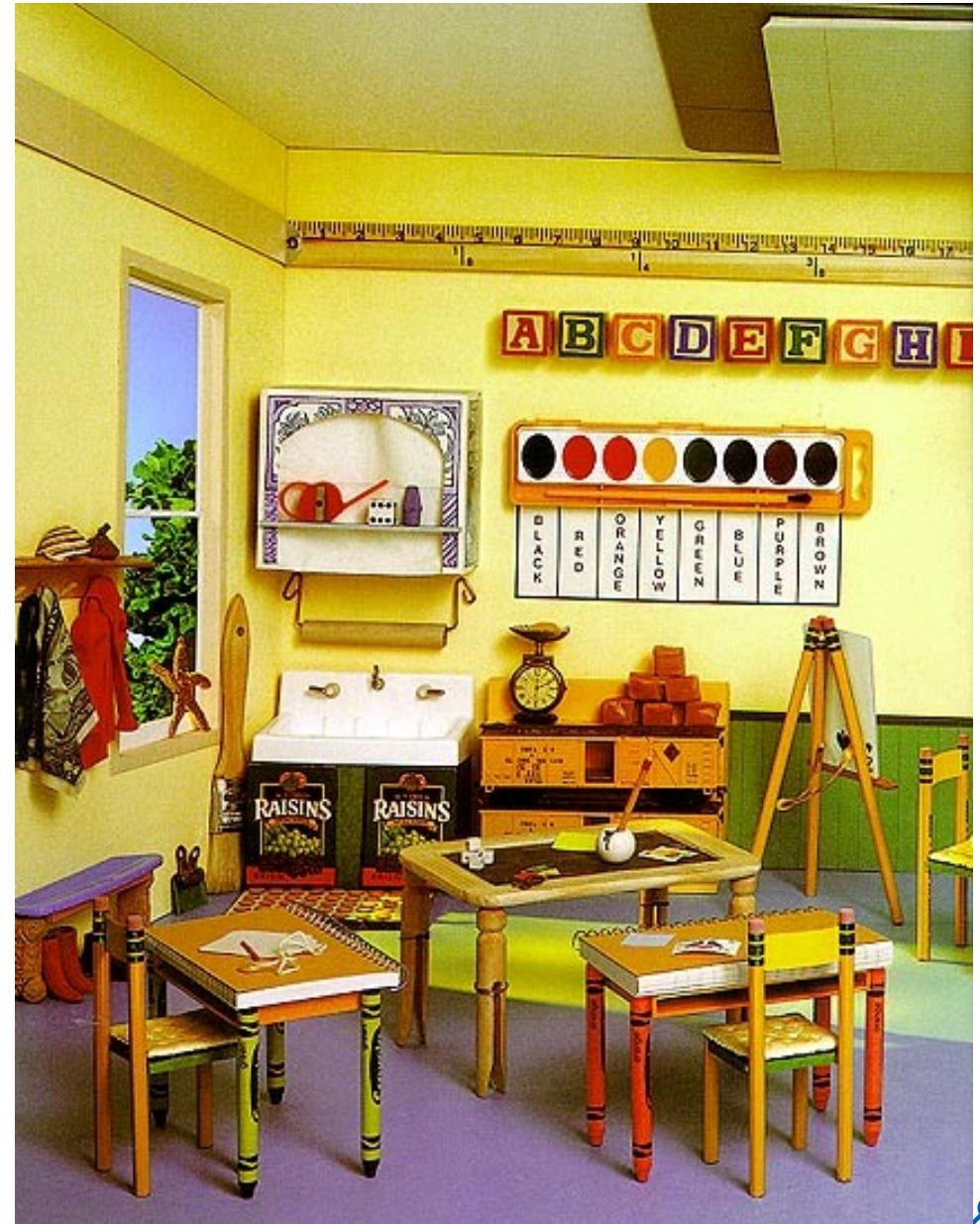
The multiple personalities of a blob

A B C

12
13
14

12
A B C
14

Look-Alikes by Joan Steiner



Look-Alikes: The More You Look, the More You See! Hardcover – Picture Book, October 17, 2003

Slide credit: A. Torralba

The context challenge

- What are the hidden objects?



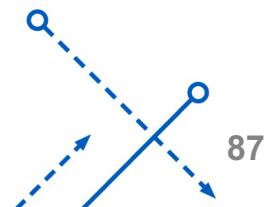
The context challenge

- What are the hidden objects?



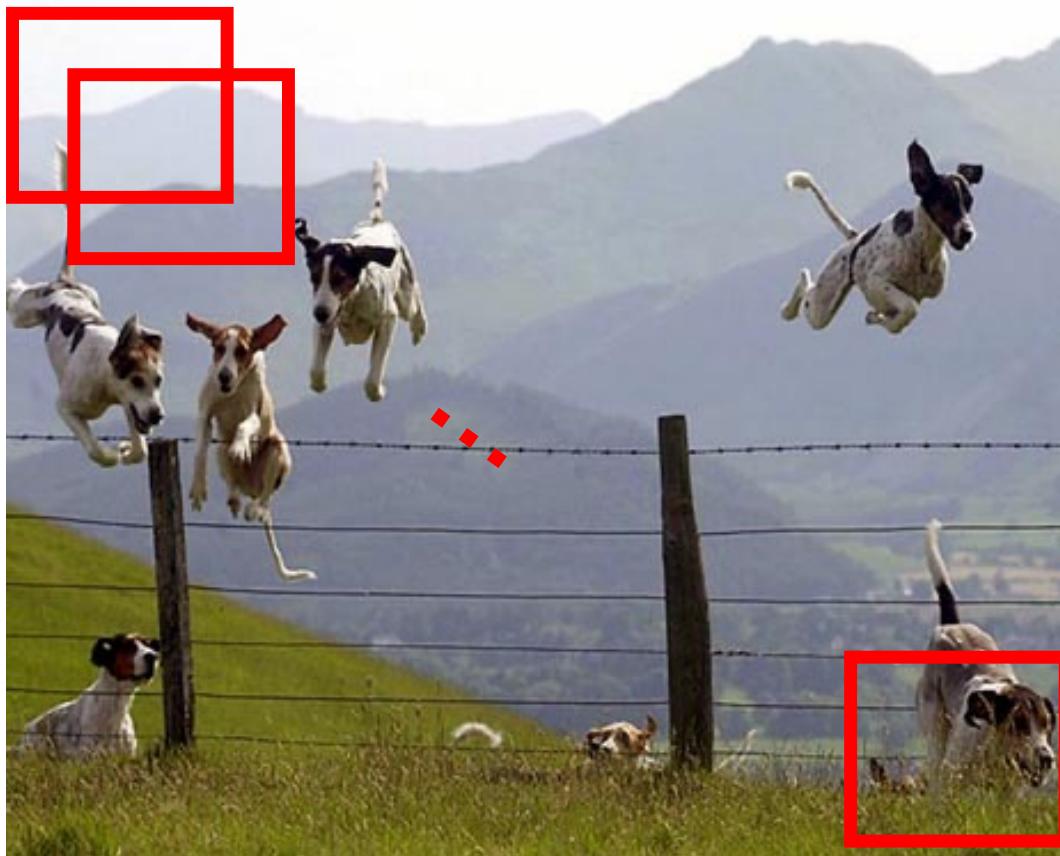
Object detection vs Scene Recognition

- Objects (even if deformable and articulated) probably have more **consistent shapes** than scenes.
- Scenes can be defined by distribution of “stuff” – materials and surfaces with **arbitrary shape**.
- Objects are “**things**” that own their **boundaries**
- Bag of words (BoW) were less popular for object detection because they **throw away shape** info.

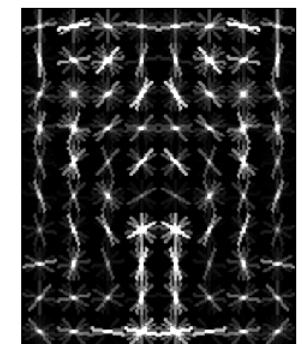


Object Category Detection

- Focus on object search: “Where is it?”
- Build templates that quickly differentiate object patch from background patch



Dog Model



**Object or
Non-Object?**



Challenges in modeling the object class



Illumination



Object pose



Clutter



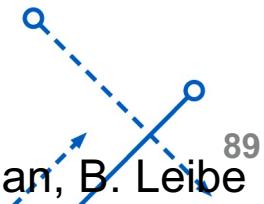
Occlusions



Intra-class
appearance



Viewpoint



Slide from K. Grauman, B. Leibe 89

Challenges in modeling the non-object class

True
Detections



Bad
Localization



Confused with
Similar Object



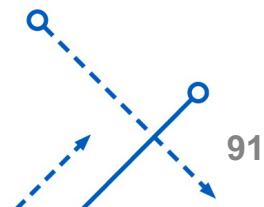
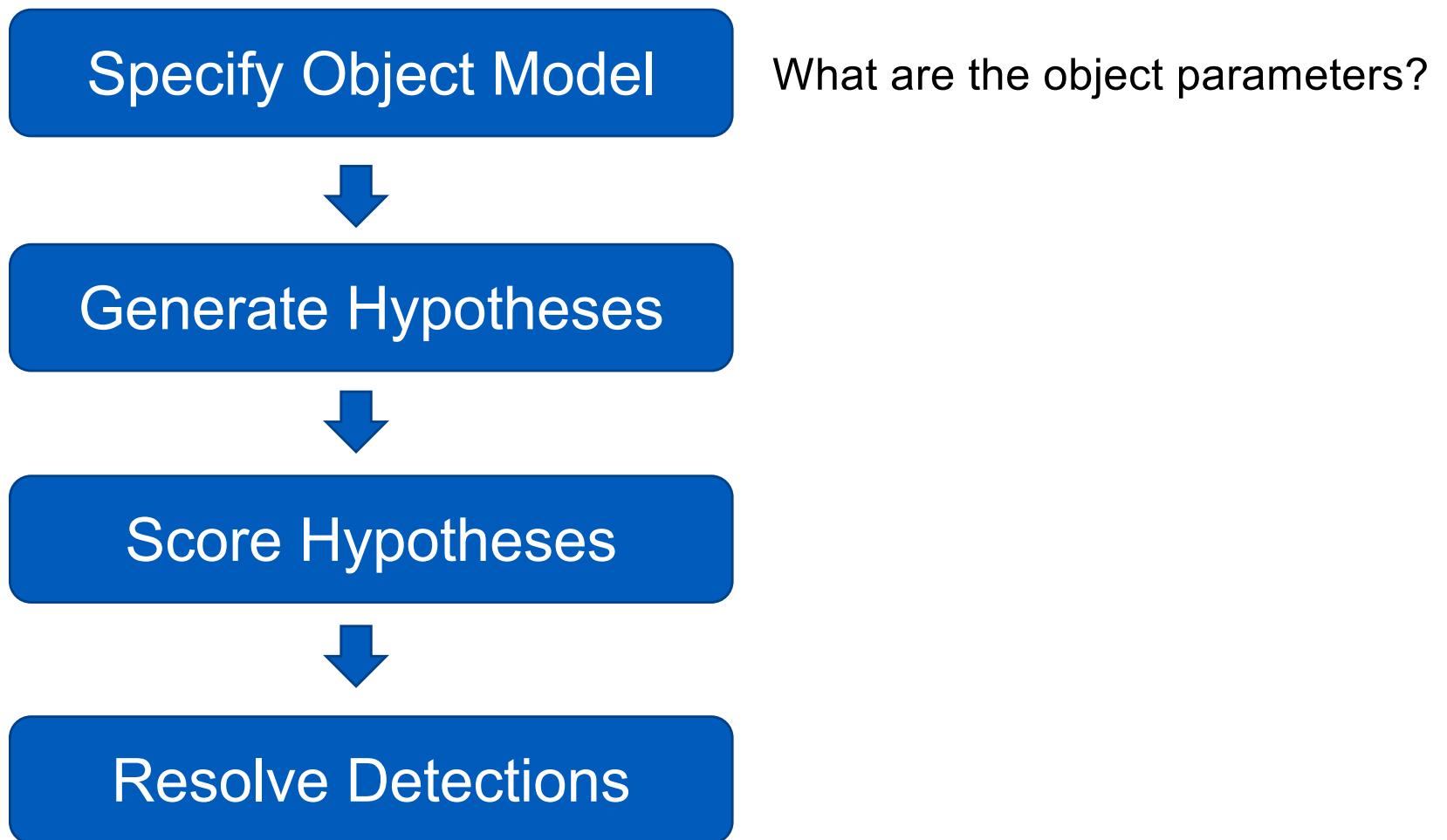
Misc. Background



Confused with
Dissimilar Objects



General Process of Object Recognition



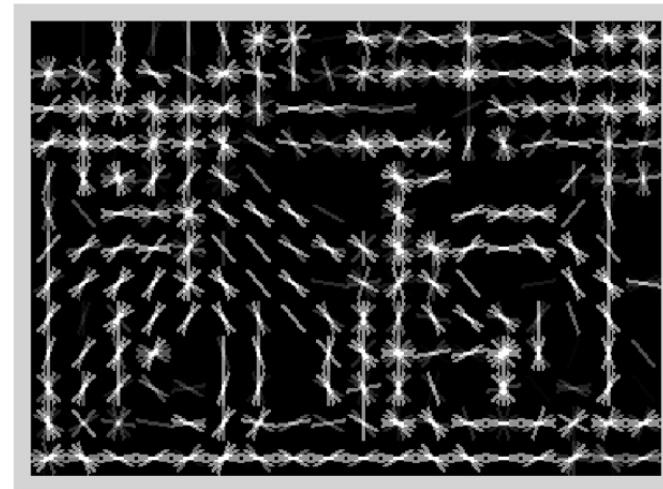
Specifying an object model

1. Statistical Template in Bounding Box

- Object is somewhere (x, y, w, h) in image
- Features defined wrt bounding box coordinates



Image

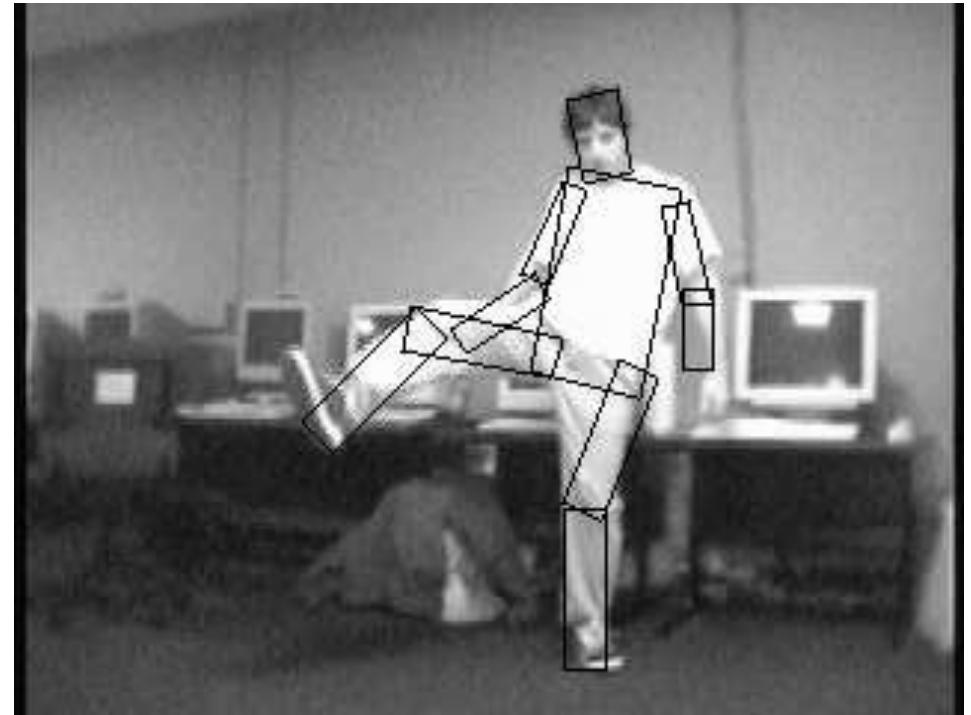
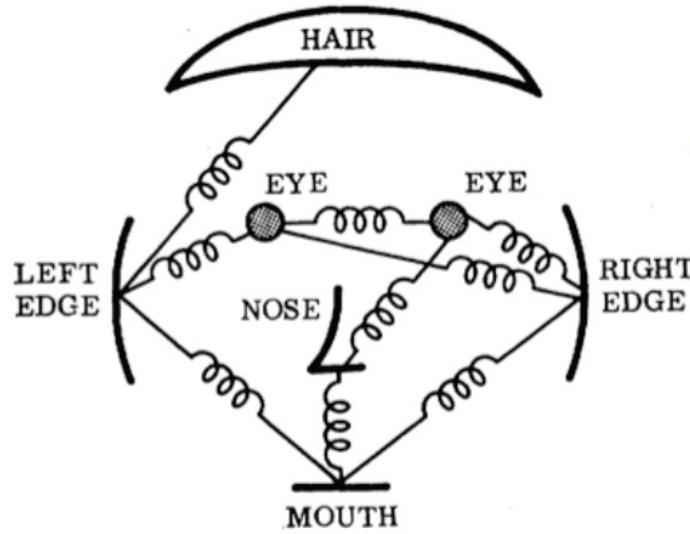


Template Visualization
(HoG features)

Specifying an object model

2. Articulated parts model

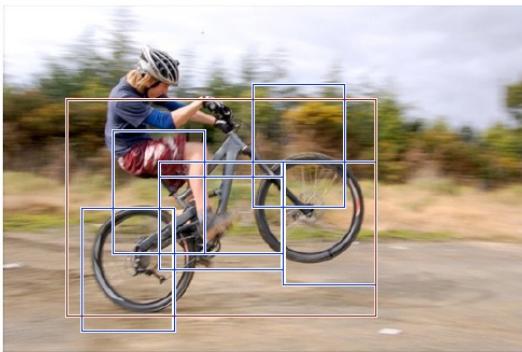
- Object is configuration of parts
- Each part is detectable



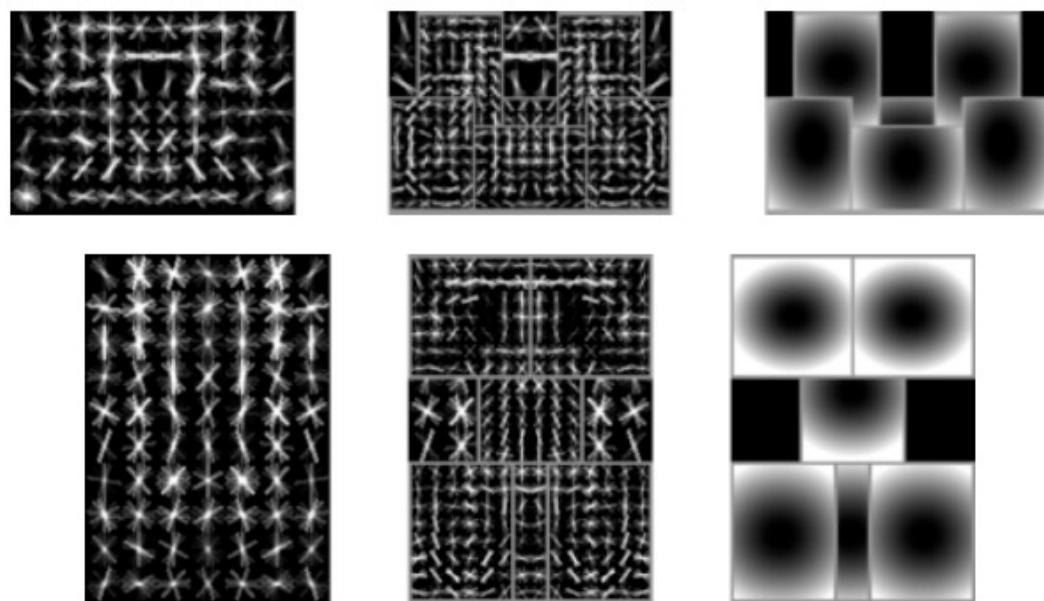
Specifying an object model

3. Hybrid template/parts model

Detections



Template Visualization
HoG features



root filters
coarse resolution

part filters
finer resolution

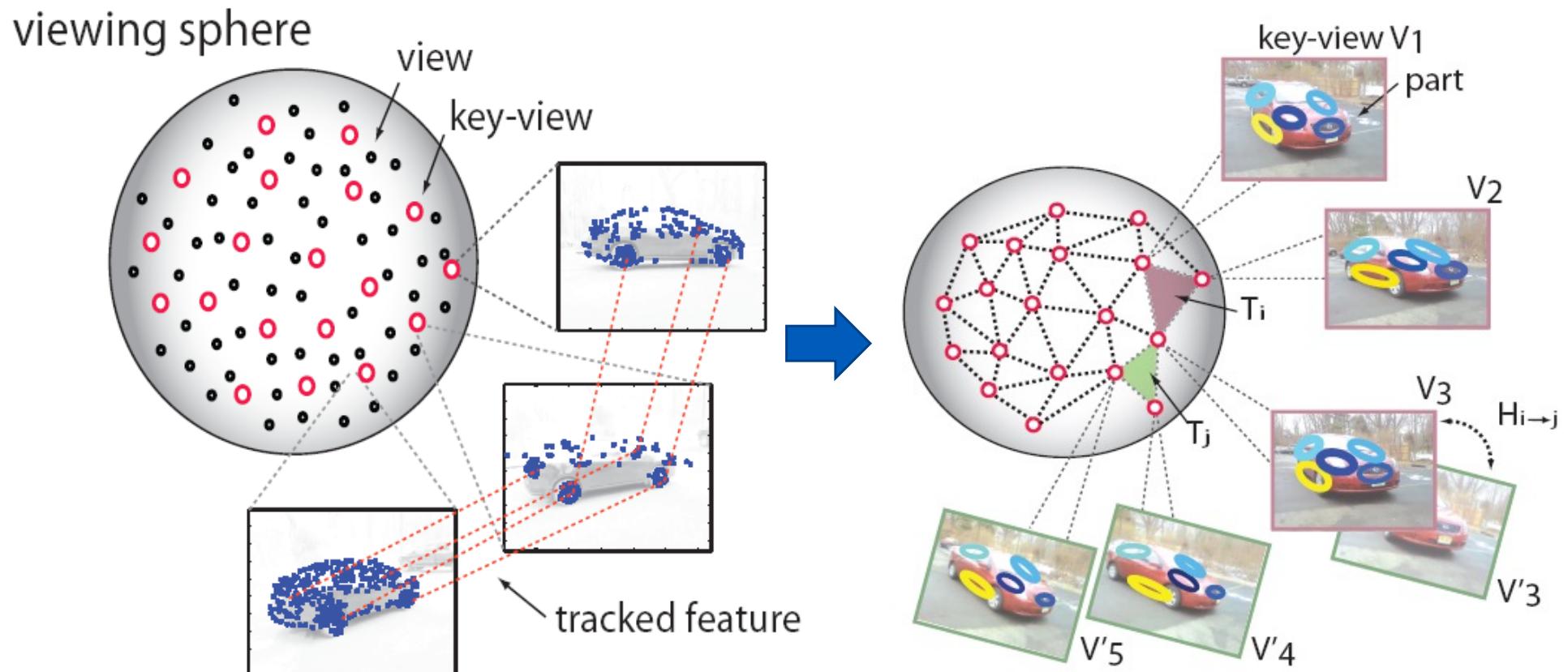
deformation
models

Felzenszwalb et al. 2008

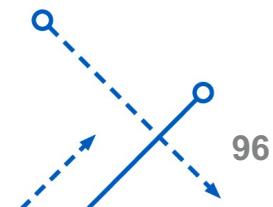
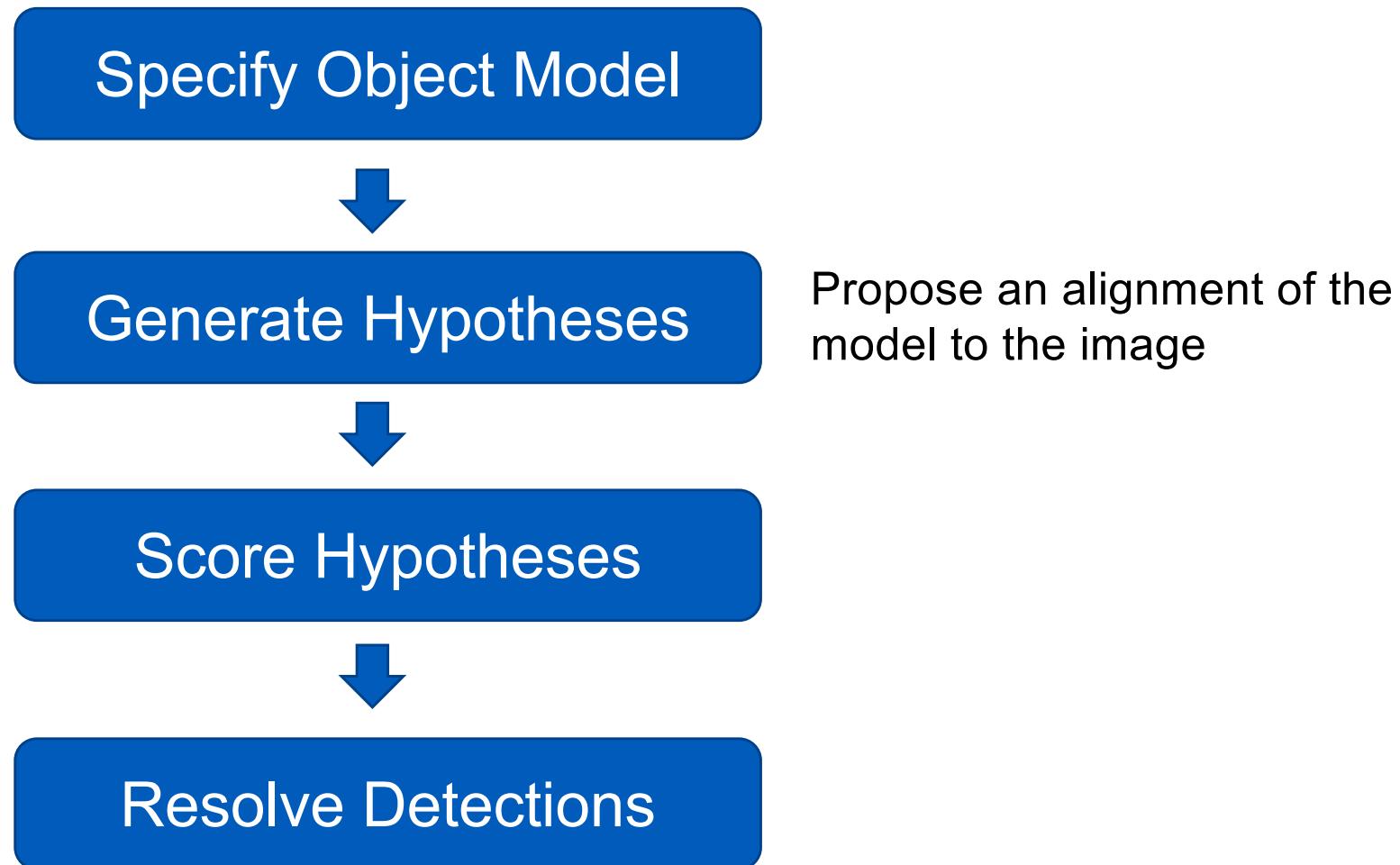
Specifying an object model

4. 3D-ish model

- Object is collection of 3D planar patches under affine transformation



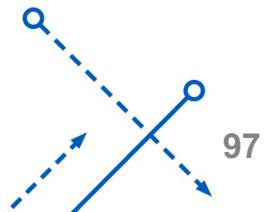
General Process of Object Recognition



Generating hypotheses

1. Sliding window

- Test patch at each **location** and **scale**



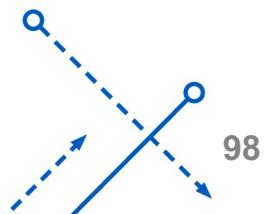
Generating hypotheses

1. Sliding window

- Test patch at each **location** and **scale**



Note – Template did not change size



Each window is separately classified



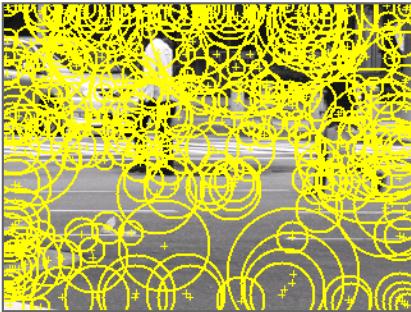
99
99

Generating hypotheses

2. Voting from patches/keypoints



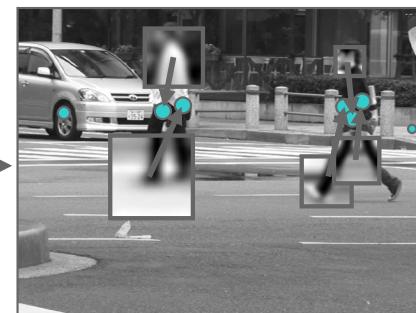
Interest Points



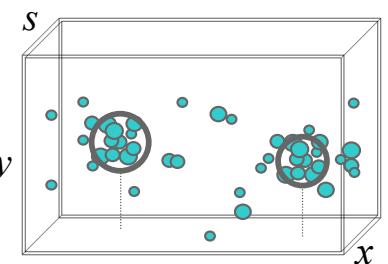
Matched Codebook
Entries



Probabilistic
Voting

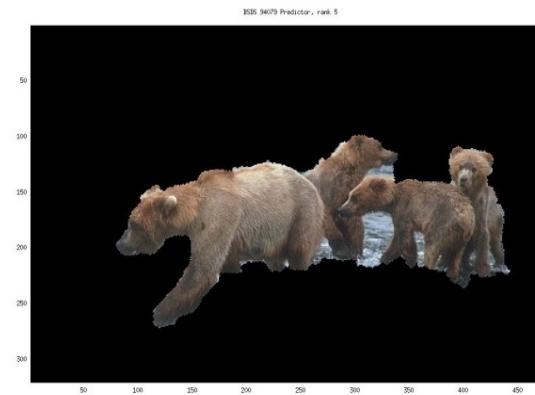
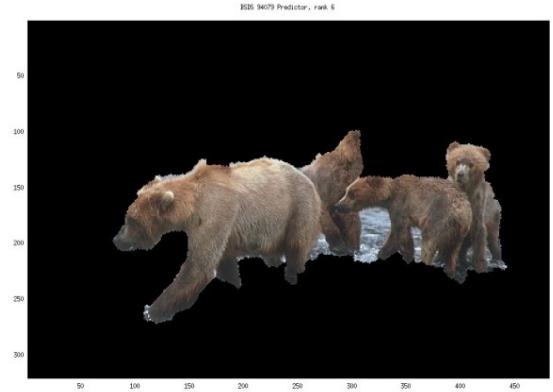


3D Voting Space
(continuous)

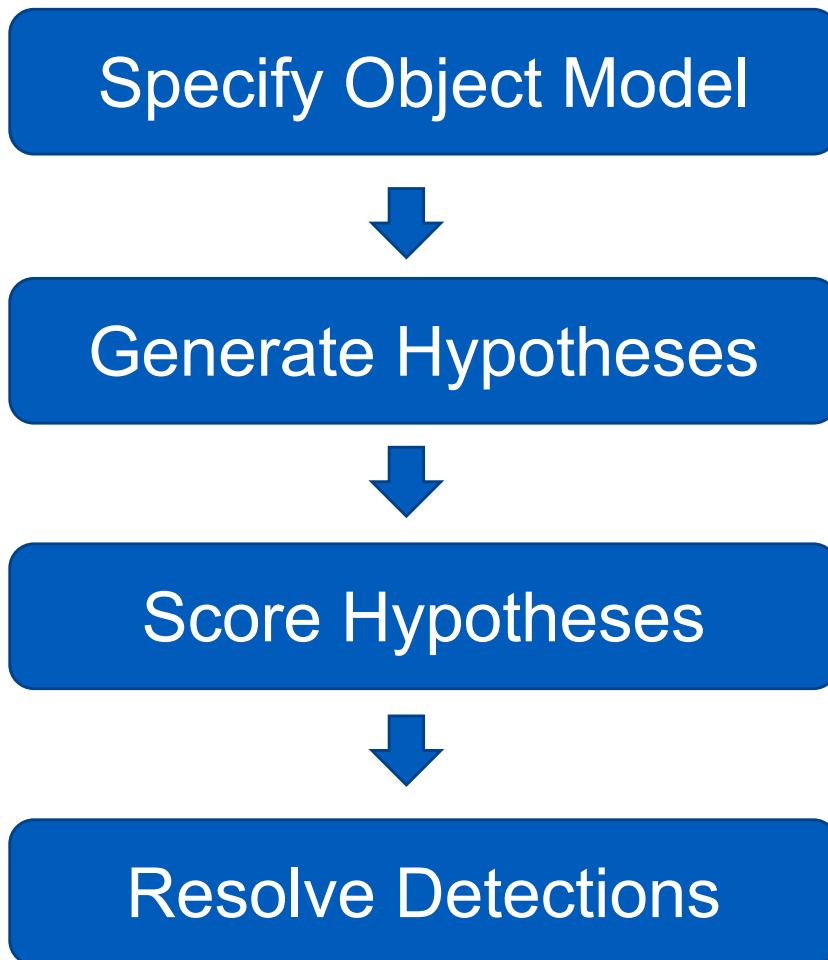


Generating hypotheses

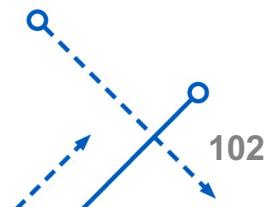
3. Region-based proposal



General Process of Object Recognition



Mainly-gradient based features,
usually based on summary
representation, many classifiers.



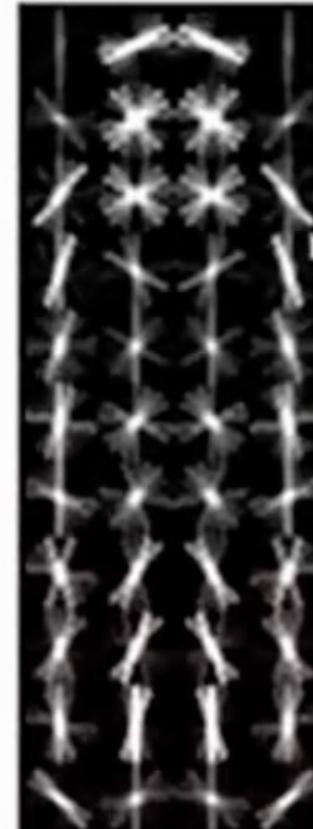
HOG + SVM for Object Detection



Image



HOG Features

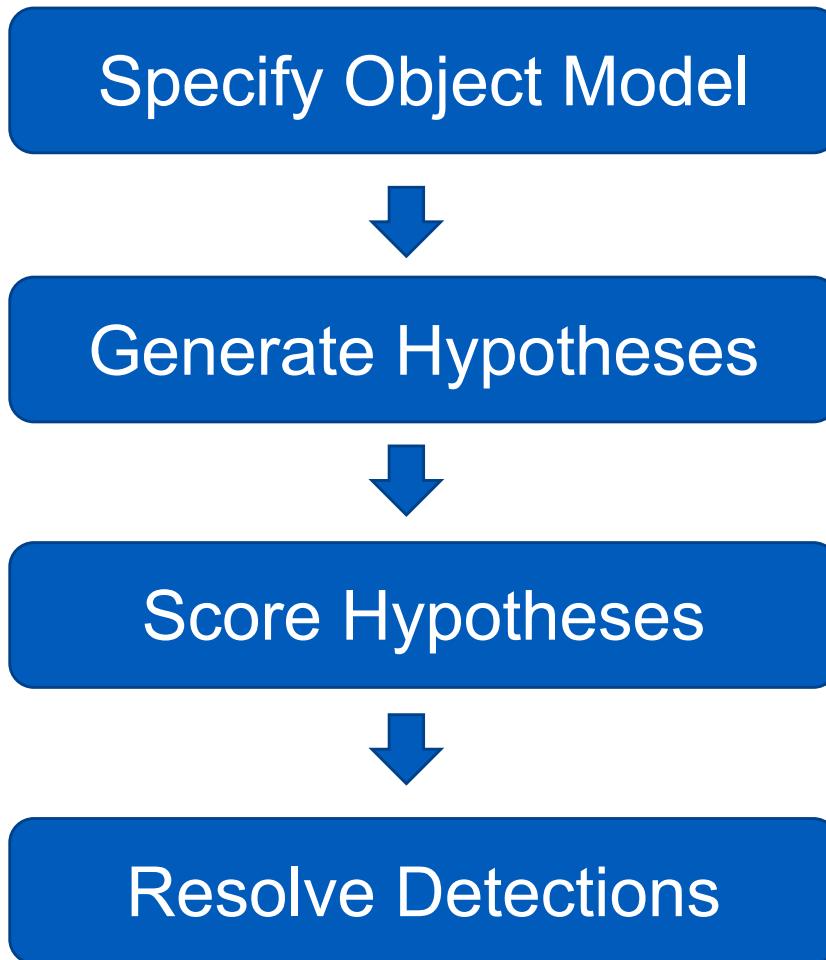


SVM Coefs
Human Model

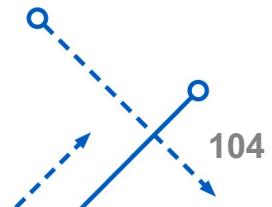


Human or Not

General Process of Object Recognition

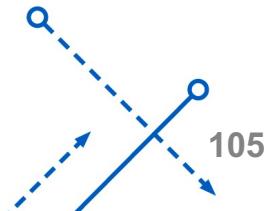
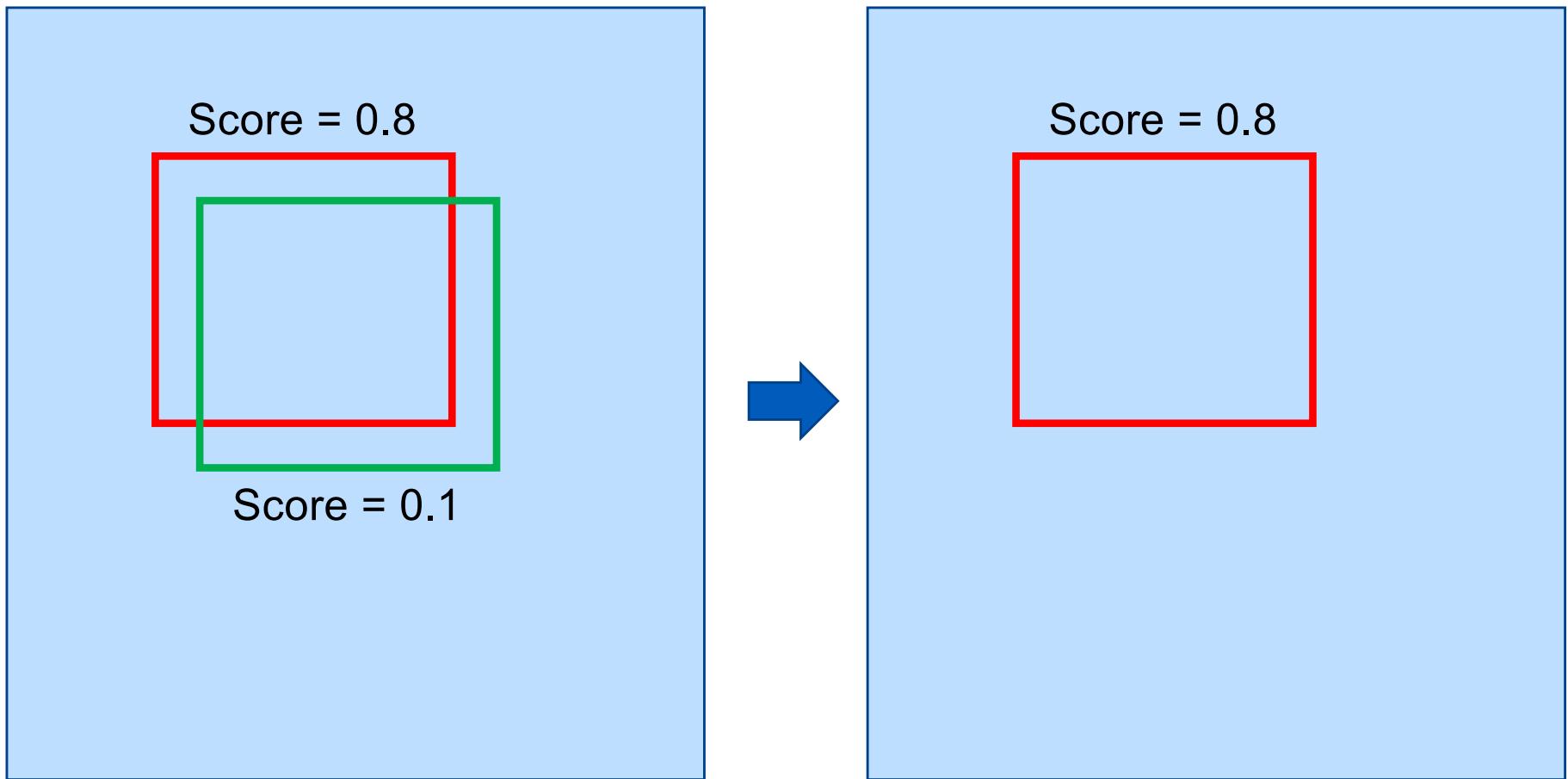


Rescore each proposed
object based on whole set



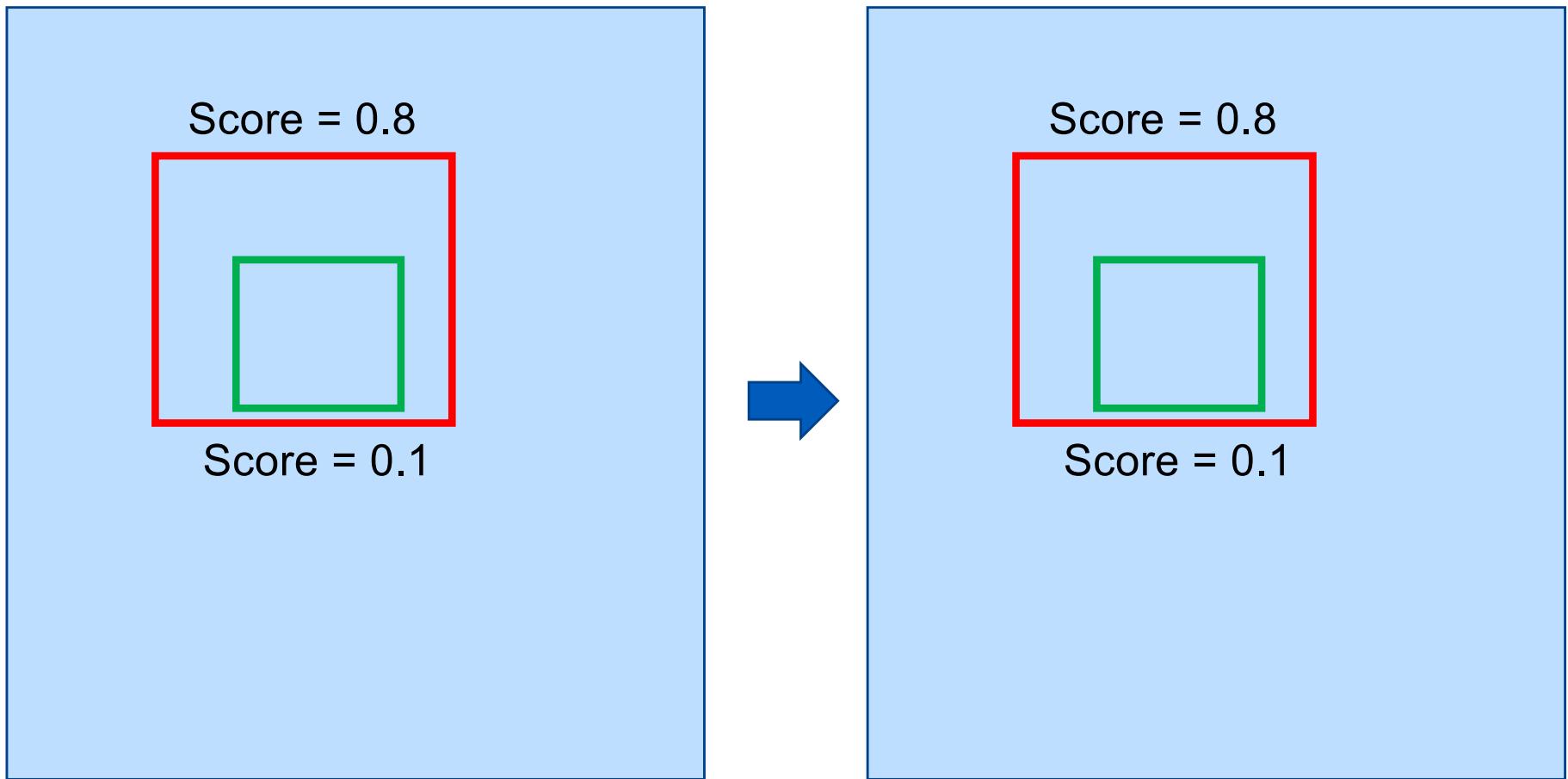
Resolving detection scores

1. Non-max suppression

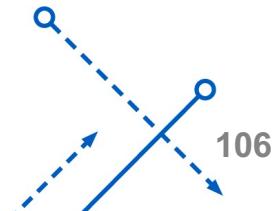


Resolving detection scores

1. Non-max suppression

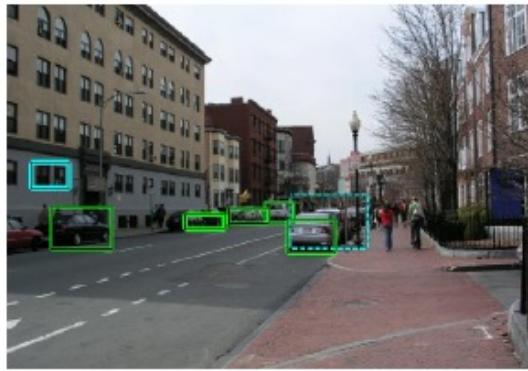


“Overlap” score is below some threshold



Resolving Detection Scores

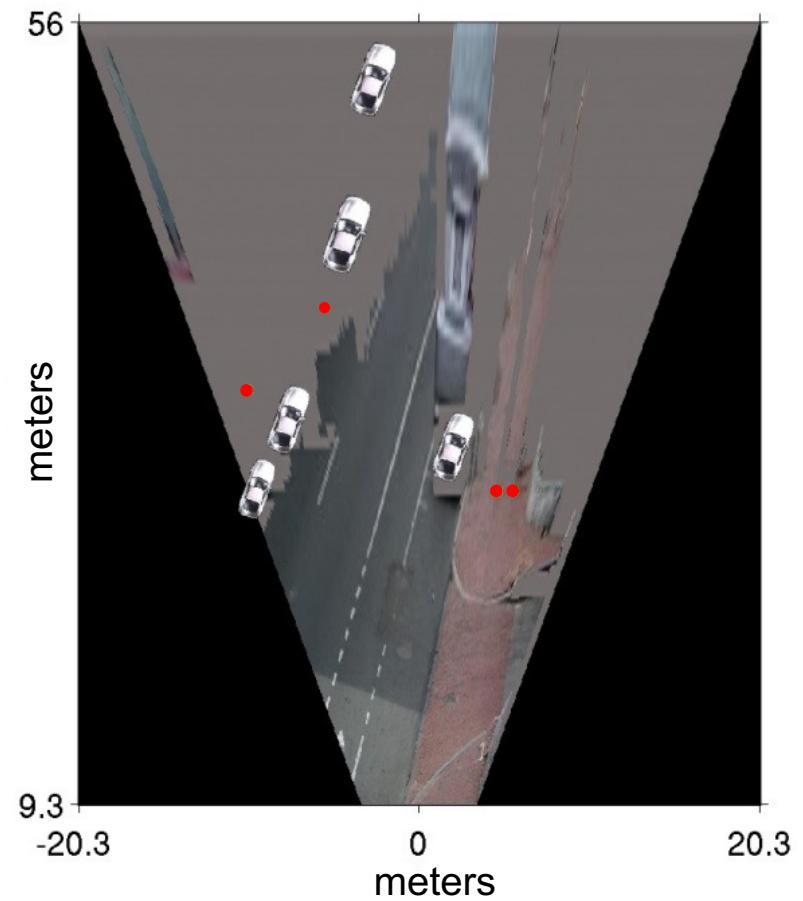
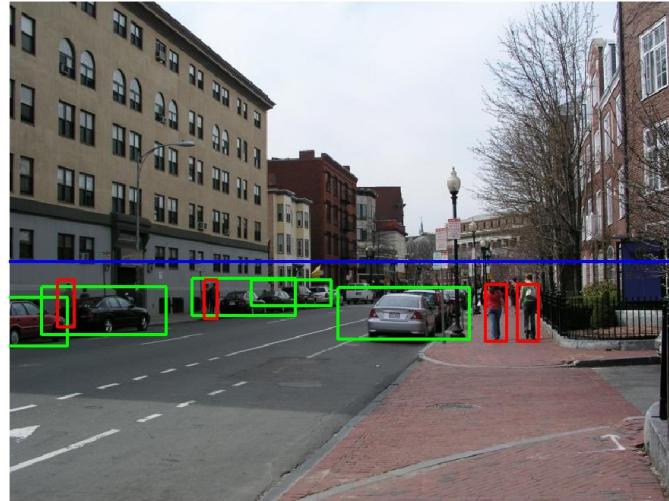
2. Context/reasoning



(g) Car Detections: Local



(h) Ped Detections: Local



Hoiem et al., 2006
107