Image Classification and Object Recognition

Selim Aksoy
Department of Computer Engineering
Bilkent University
saksoy@cs.bilkent.edu.tr

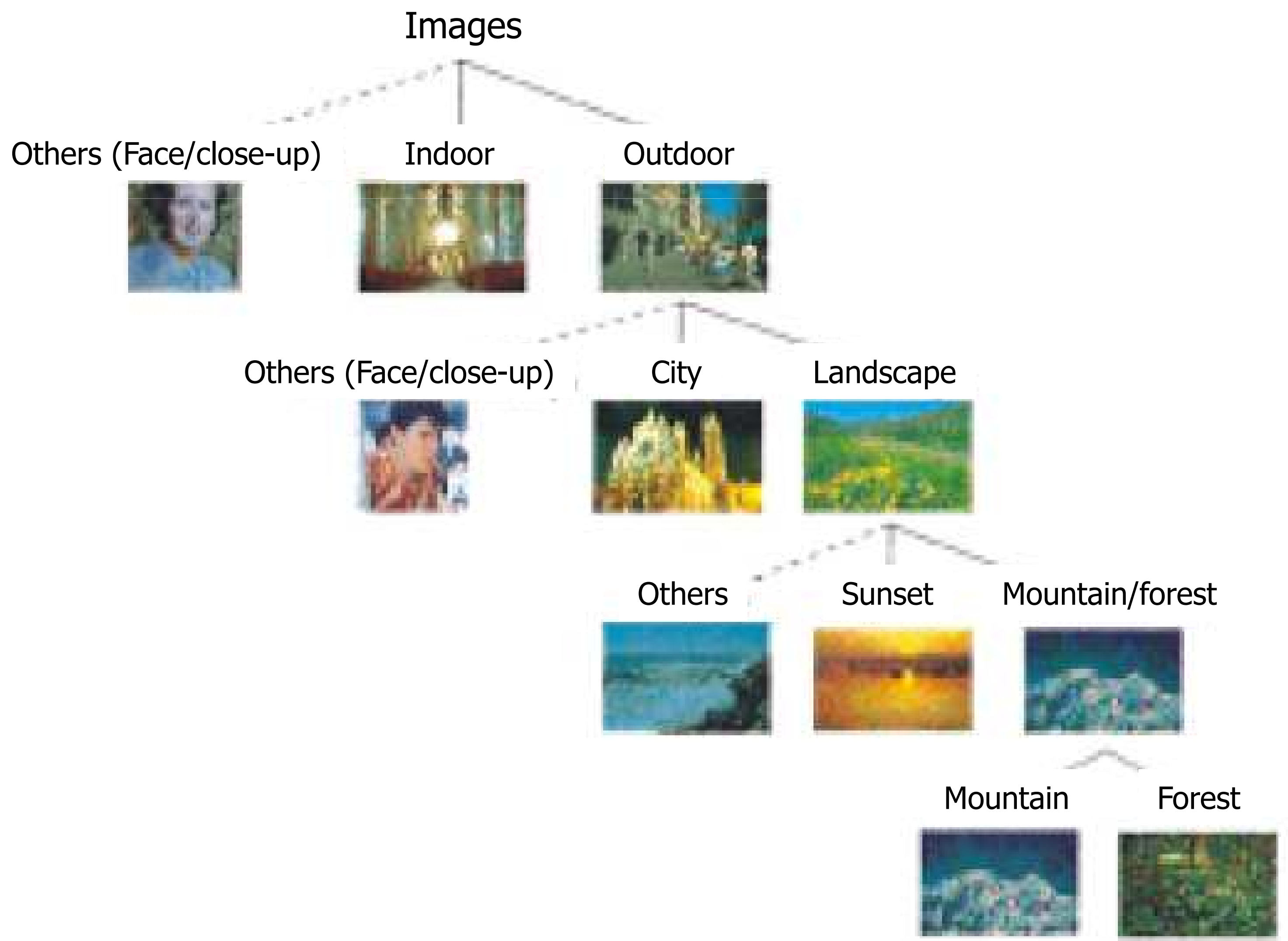
Image classification

- Image (scene) classification is a fundamental problem in image understanding.
- Automatic techniques for associating scenes with semantic labels have a high potential for improving the performance of other computer vision applications such as
  - browsing (natural grouping of images instead of clusters based only on low-level features),
  - retrieval (filtering images in archives based on content), and
  - object recognition (the probability of an unknown object/region that exhibits several local features of a ship actually being a ship can be increased if the scene context is known to be a coast with high confidence but can be decreased if no water related context is dominant in that scene).

Image classification

- The image classification problem has two critical components: representing images and learning models for semantic categories using these representations.
- Early work used low-level global features extracted from the whole image or from a fixed spatial layout.
- More recent approaches exploit local statistics in images using patches extracted by interest point detectors.
- Other configurations that use regions and their spatial relationships are also proposed.

Hierarchical image classification



Hierarchy of 11 scene categories (Vailaya et al., "Image classification for content-based indexing," IEEE Trans. Image Processing, 2001).

Hierarchical image classification

- Image representation:
  - Mean and std. dev. of LUV values in 10x10 blocks for indoor/outdoor classification.
  - Edge direction histograms for city/landscape classification.
  - Histograms of HSV and LUV values for sunset/mountain/forest classification.
- Classification:
  - Class-conditional density estimation using vector quantization.
  - Bayesian classification.

Hierarchical image classification

TABLE III
ACCURACIES (IN PERCENT) FOR INDOOR/OUTDOOR CLASSIFICATION USING COLOR MOMENTS; TEST SET 1 AND TEST SET 2 ARE INDEPENDENT TEST SETS

Test Data	Database Size	Accuracy (%)
Training Set	2,541	94.2
Test Set 1	2,540	88.2
Test Set 2	1,850	88.7
Entire Database	6,931	90.5

TABLE IV
CLASSIFICATION ACCURACIES (IN PERCENT) FOR CITY/LANDSCAPE CLASSIFICATION; THE FEATURES ARE ABBREVIATED AS FOLLOWS: EDGE DIRECTION HISTOGRAM (EDH), EDGE DIRECTION COHERENCE VECTOR (EDCV), COLOR HISTOGRAM (CH), AND COLOR COHERENCE VECTOR (CCV)

Test Data	EDH	EDCV	CH	CCV	EDH & CH	EDH & CCV	EDCV & CH	EDCV & CCV
Training Set	94.7	97.0	83.7	83.5	94.8	95.4	96.4	96.9
Test Set	92.0	92.9	75.4	76.0	92.5	92.8	93.4	93.8
Entire Database	93.4	95.0	79.6	79.8	93.7	94.1	94.9	95.3

Hierarchical image classification

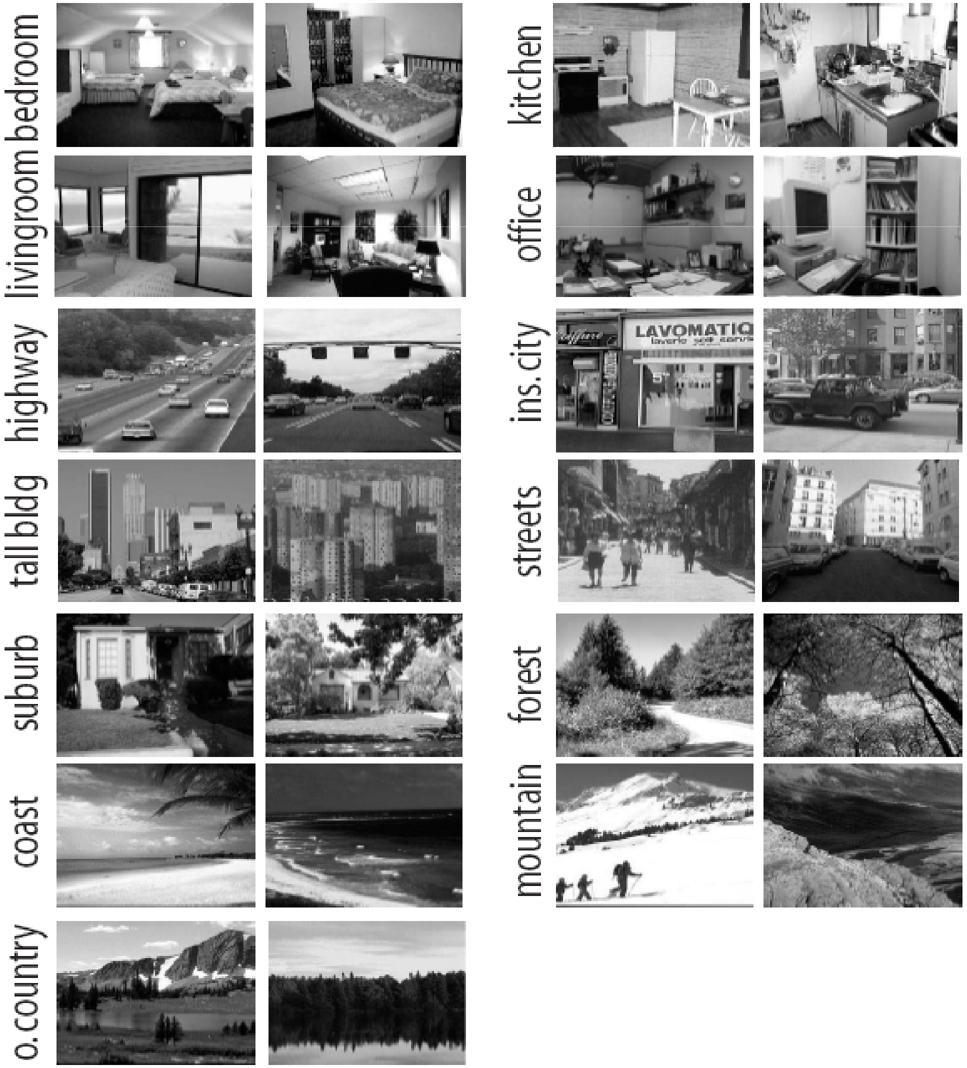
TABLE V
CLASSIFICATION ACCURACIES (IN PERCENT) FOR SUNSET/FOREST/MOUNTAIN CLASSIFICATION; *SPM* STANDS FOR "SPATIAL COLOR MOMENTS"

Test Data	EDH	EDCV	CH	CCV	SPM	EDH & CH	EDH & CCV	EDCV & CH	EDCV & CCV
Training Set	88.3	88.3	96.2	99.2	98.9	95.9	96.6	95.5	97.0
Test Set	86.3	89.0	89.7	93.9	93.9	90.1	95.4	90.5	95.1
Entire Database	87.4	88.7	93.0	96.6	96.4	93.0	96.0	93.0	96.1

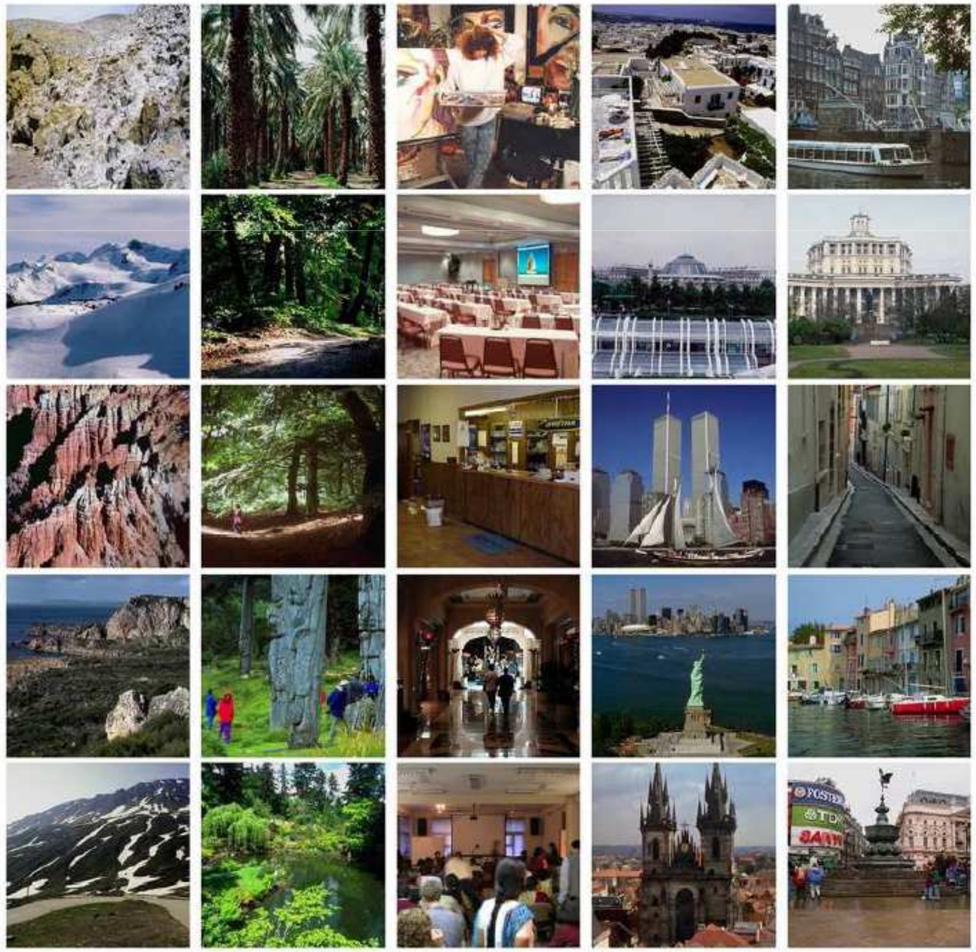
TABLE VI
CLASSIFICATION ACCURACIES (IN PERCENT) FOR FOREST/MOUNTAIN CLASSIFICATION

Test Data	EDH	EDCV	CH	CCV	SPM	EDH & CH	EDH & CCV	EDCV & CH	EDCV & CCV
Training Set	83.4	78.1	92.0	98.9	98.4	94.1	98.4	93.6	98.4
Test Set	87.1	77.2	91.4	91.9	93.6	93.0	92.5	93.5	91.9
Entire Database	85.3	77.7	91.7	95.5	96.0	93.6	95.5	93.6	95.2

Image classification using bag-of-words

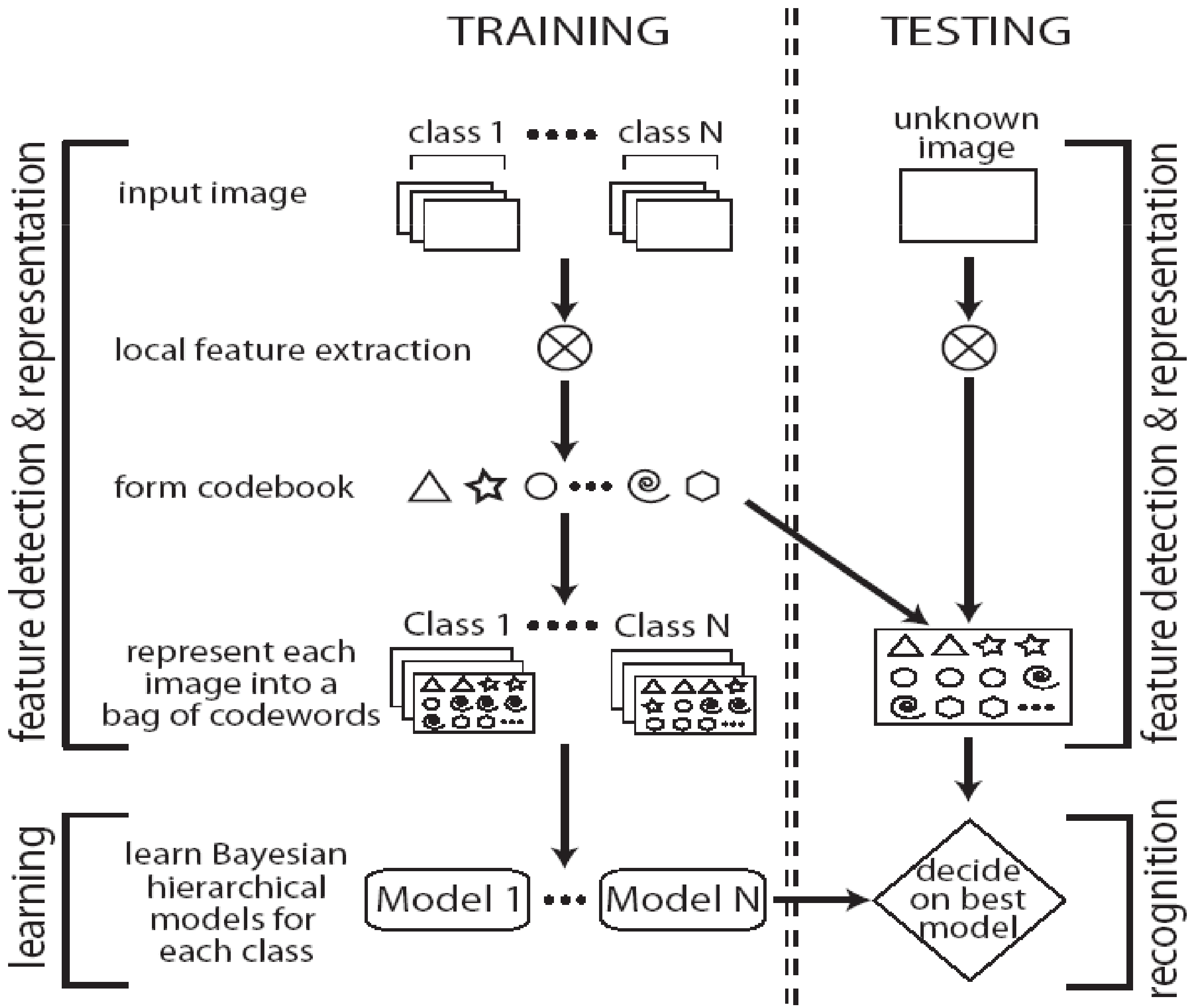


Caltech data set: 13 natural scene categories.



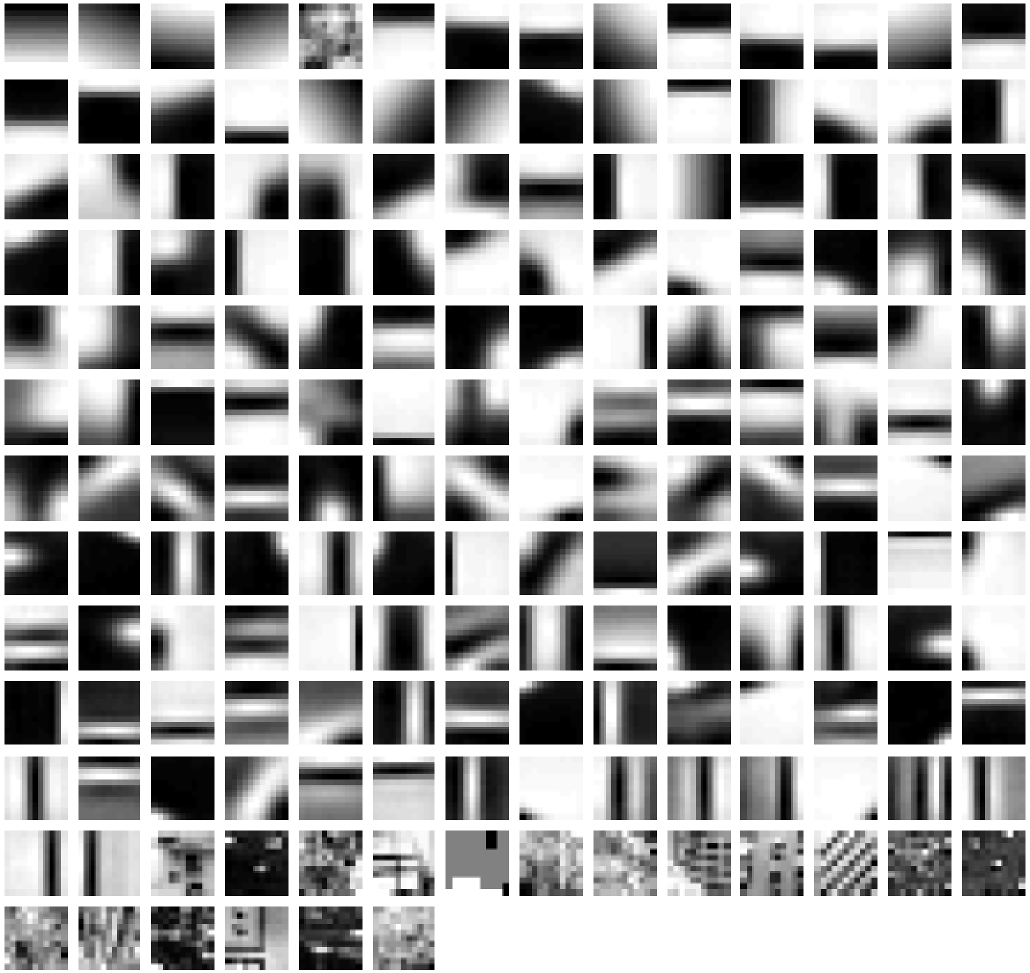
IDIAP data set (left to right): mountain, forest, indoor, city-panorama, city-street.

Image classification using bag-of-words



Flowchart from Fei-Fei Li, Pietro Perona, "A Bayesian hierarchical model for learning natural scene categories," IEEE CVPR, 2005.

Image classification using bag-of-words



A codebook obtained from 650 training examples from 13 categories.
Image patches are detected by a sliding grid and random sampling of scales.

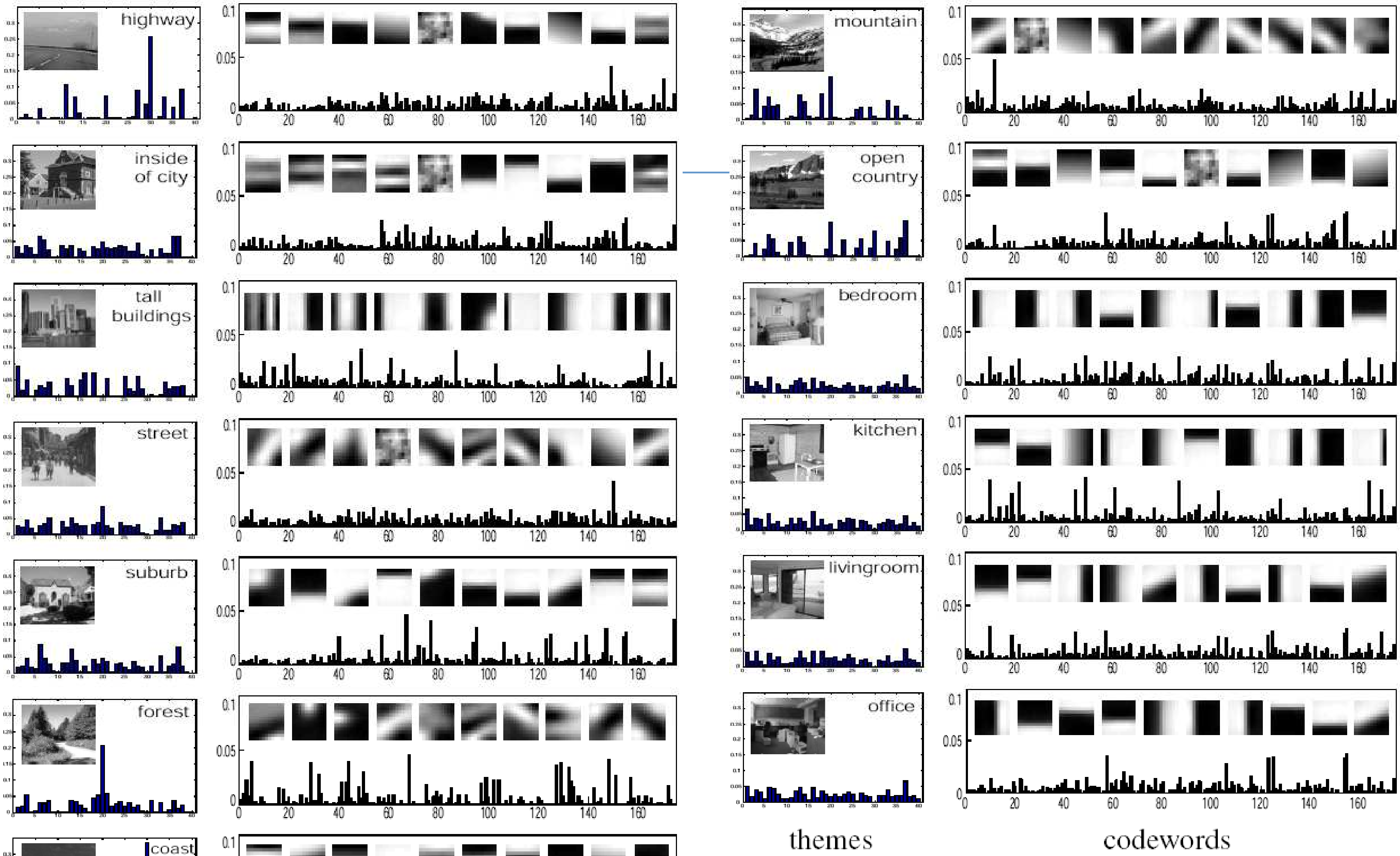


Figure 5. Internal structure of the models learnt for each category. Each row represents one category. The left panel shows the distribution of the 40 intermediate themes. The right panel shows the distribution of codewords as well as the appearance of 10 codewords selected from the top 20 most likely codewords for this category model.

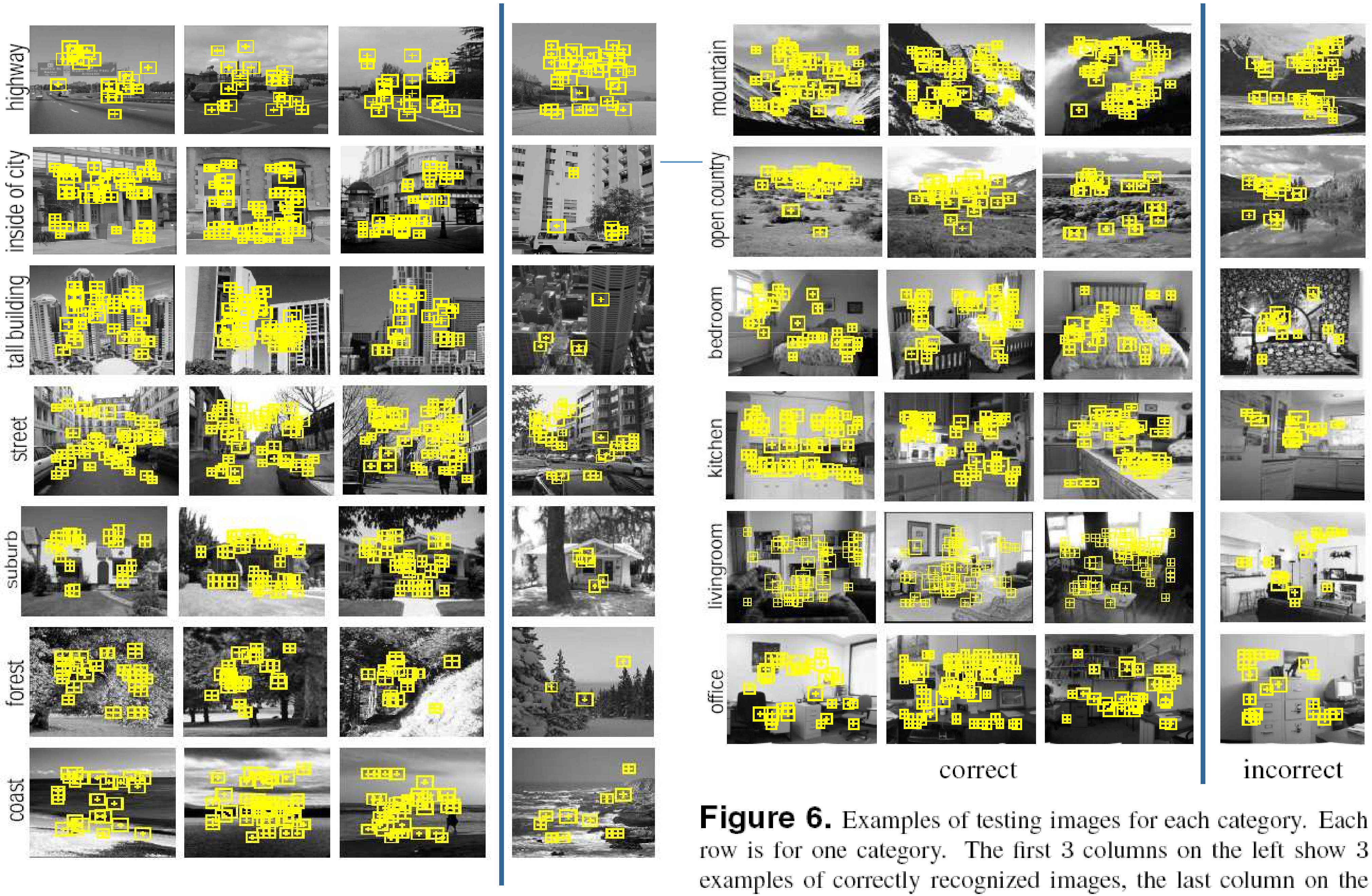


Figure 6. Examples of testing images for each category. Each row is for one category. The first 3 columns on the left show 3 examples of correctly recognized images, the last column on the right shows an example of incorrectly recognized image. Superimposed on each image, we show samples of patches that belong to the most significant set of codewords given the category model.

Image classification using bag-of-words

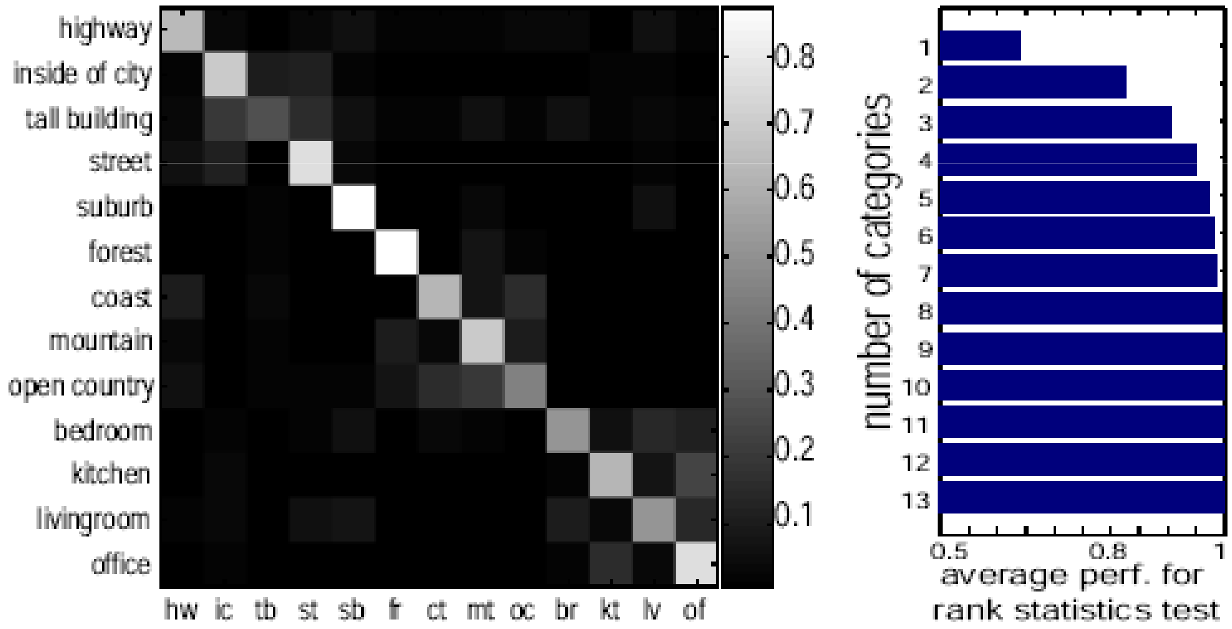
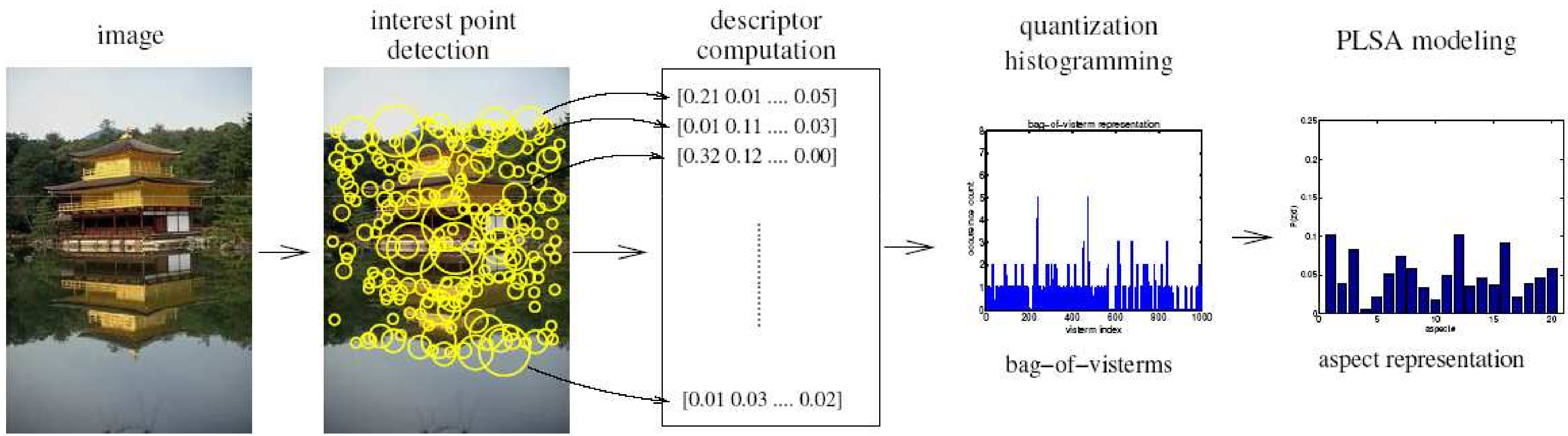


Figure 7. Left Panel. Confusion table of Theme Model 1 using 100 training and 50 test examples from each category, the grid detector and patch based representation. The average performance is 64.0%. **Right Panel.** Rank statistics of the confusion table, which shows the probability of a test scene correctly belong to one of the top N most probable categories. N ranges from 1 to 13.

Image classification using bag-of-words



Flowchart from Quelhas et al., “A thousand words in a scene,” IEEE Trans. PAMI, 2007.

- Probabilistic Latent Semantic Analysis (PLSA) is used to learn aspect models to capture co-occurrences of visterms (visual terms).
- Bag-of-visterms representation or the aspect parameters are given as input to Support Vector Machines for classification.

Image classification using bag-of-words

Total class. error				11.1 (0.8)	
Gr. Truth	Classification (%)			Class. Error (%)	# of images
	indoor	city	land.		
indoor	89.7	9.0	1.3	10.3	2777
city	14.5	74.8	10.7	25.2	2505
landscape	1.2	2.0	96.8	3.1	4175

TABLE III

CONFUSION MATRIX FOR THE THREE-CLASS CLASSIFICATION PROBLEM, USING VOCABULARY V_{1000}.

Total class. error				11.9(1.0)	
	indoor	city	land.	class error(%)	# images
indoor	86.6	11.8	1.6	13.4	2777
city	14.8	75.4	9.8	24.5	2505
land.	1.3	1.9	96.8	3.1	4175

TABLE VIII

CLASSIFICATION ERROR AND CONFUSION MATRIX FOR THE THREE-CLASS PROBLEM USING PLSA, WITH V_{1000} AND 60 ASPECTS.

Total class. error rate: 20.8 (2.1) (Baseline: 30.1 (1.1))							
	m.	f.	i.	c.-p.	c.-s.	error (%)	# of images
mount.	85.8	8.6	2.5	0.5	2.6	14.2	590
forest	8.9	80.3	1.6	2.4	6.7	19.7	492
indoor	0.4	0	91.1	0.4	8.1	8.9	2777
city-pan.	3.5	1.8	8.0	46.9	39.8	53.1	549
city-str.	2.0	2.2	20.8	6.0	68.9	31.1	1957

TABLE V

CLASSIFICATION RATE AND CONFUSION MATRIX FOR THE FIVE-CLASS, USING BOV AND VOCABULARY V_{1000}.

Total error rate (BOV: 20.8 (2.1), Baseline: 30.1 (1.1))						
	m.	f.	i.	c.-p.	c.-s.	error (%)
mountain	85.5	12.2	0.8	0.3	1.2	14.5
forest	12.8	78.3	0.8	0.4	7.7	21.7
indoor	0.3	0.1	88.9	0.2	10.5	11.1
city-pan.	3.6	4.9	8.8	12.6	70.1	87.4
city-str.	1.6	1.4	20.4	1.7	74.9	25.1

TABLE X

CLASSIFICATION ERROR AND CONFUSION MATRIX FOR THE FIVE-CLASS PROBLEM USING PLSA-O WITH 60 ASPECTS.

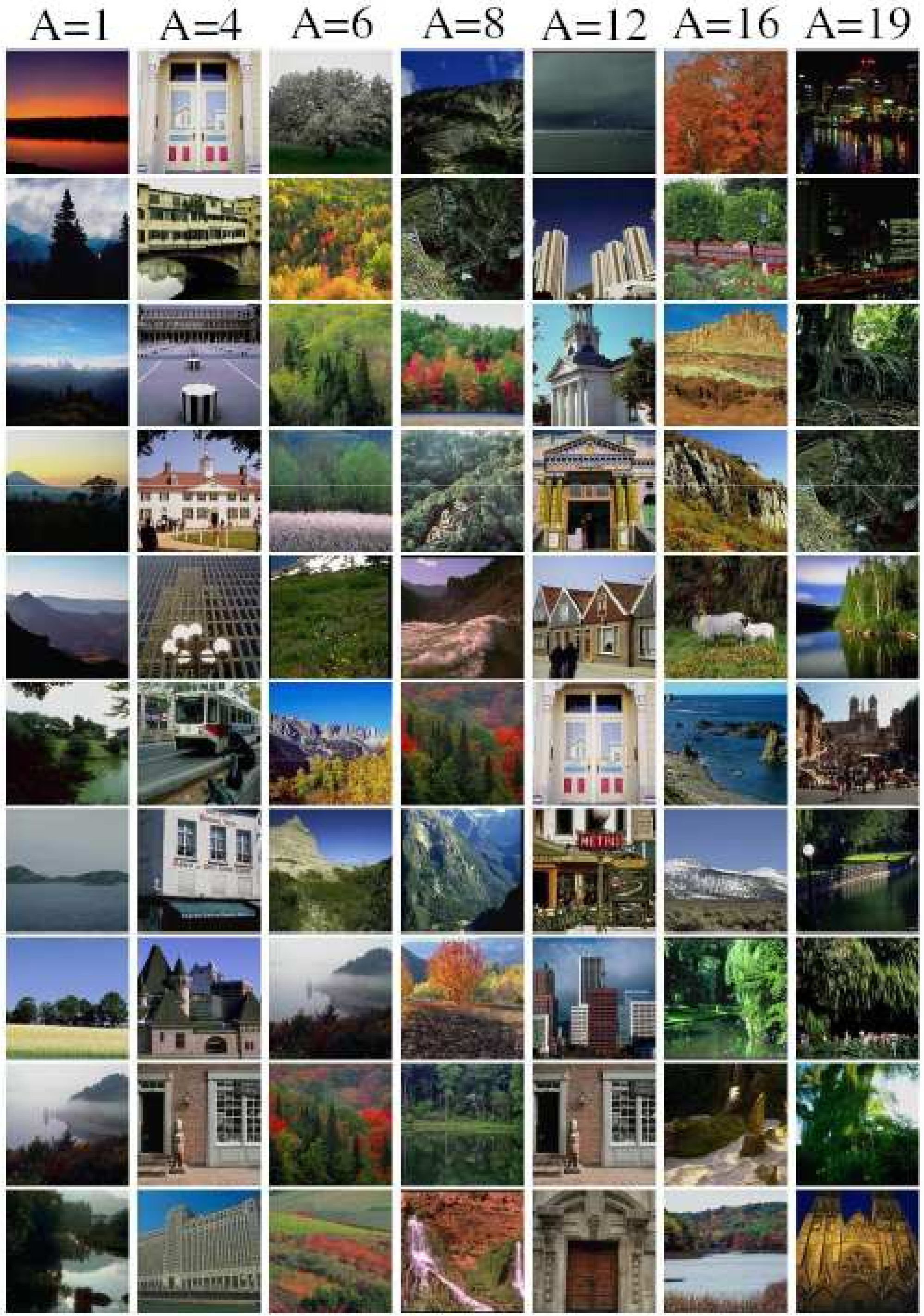


Fig. 7. The 10 most probable images from the **D1** data set for seven aspects (out of 20) learned on the **D3** data set.

Image classification using bag-of-regions

- D. Gökalp, S. Aksoy, "Scene classification using bag-of-regions representations," IEEE CVPR, Beyond Patches Workshop, 2007.
  - Region segmentation
  - Region clustering → region codebook
  - Above-below spatial relationships → region pairs
  - Statistical region selection: identify region types that
    - are frequently found in a particular class of scenes but rarely exist in other classes, and
    - consistently occur together in the same class of scenes.
  - Bayesian scene classification using
    - bag of individual regions,
    - bag of region pairs.

Image classification using bag-of-regions

Examples for region clusters.
Each row represents a different cluster.

Image classification using bag-of-regions

Table 3. Confusion matrix for the bag of individual regions representation after region selection.

		Assigned							Total	% Agree
		coast	forest	highway	insidecity	mountain	opencountry	street		
True	coast	38	2	2	1	3	4	0	50	76.00
	forest	4	36	0	0	7	2	1	50	72.00
	highway	2	2	32	6	0	2	6	50	64.00
	insidecity	3	1	12	22	2	0	10	50	44.00
	mountain	2	3	5	0	32	6	2	50	64.00
	opencountry	9	8	3	1	14	14	1	50	28.00
	street	0	0	9	6	2	6	27	50	54.00
Total		58	52	63	36	60	34	47	350	57.43

Table 4. Confusion matrix for the bag of region pairs representation after region selection.

		Assigned							Total	% Agree
		coast	forest	highway	insidecity	mountain	opencountry	street		
True	coast	42	0	0	1	3	4	0	50	84.00
	forest	1	38	0	2	4	4	1	50	76.00
	highway	1	1	31	4	2	2	9	50	62.00
	insidecity	3	4	12	19	1	1	10	50	38.00
	mountain	1	5	0	0	40	3	1	50	80.00
	opencountry	8	5	1	2	9	25	0	50	50.00
	street	2	1	8	12	2	3	22	50	44.00
Total		58	54	52	40	61	42	43	350	62.00

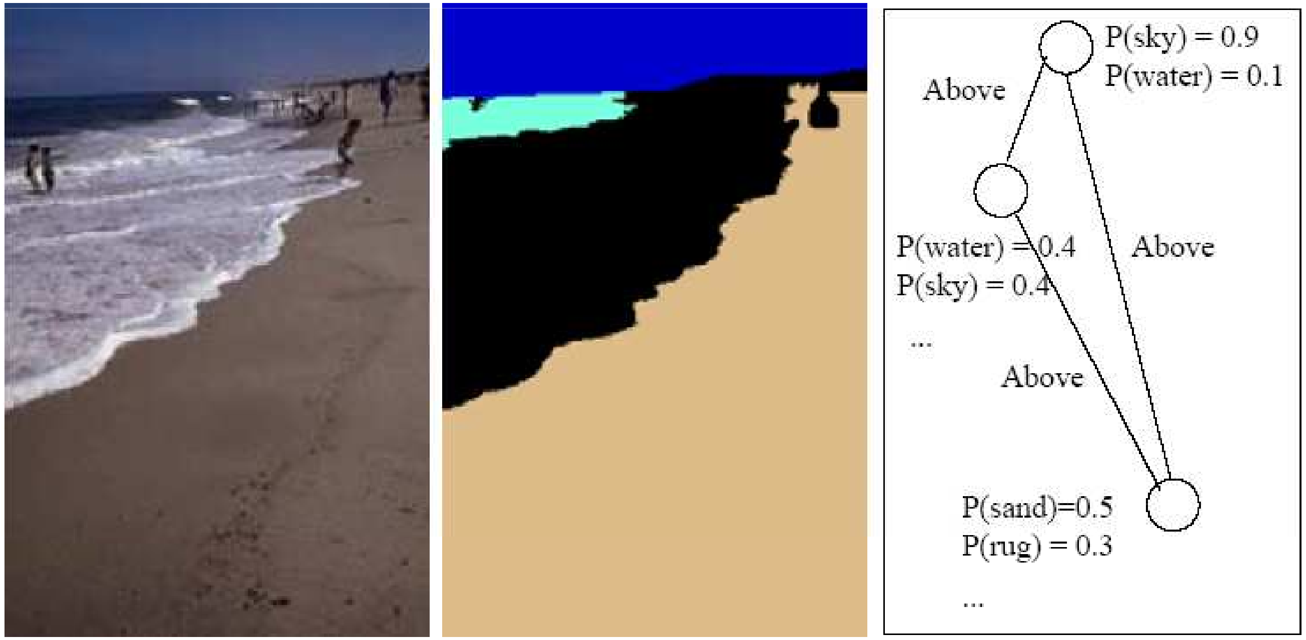
Image classification using bag-of-regions

Examples for correctly classified scenes.

Examples for wrongly classified scenes.

Image classification using factor graphs

- Boutell et al., “Scene Parsing Using Region-Based Generative Models,” IEEE Trans. Multimedia, 2007.



(a) (b) (c)

Figure 1. (a) A beach scene. (b) Its manually-labeled materials. The true configuration includes *sky above water*, *water above sand*, and *sky above sand*. (c) The underlying graph showing detector results and spatial relations.

Recognizing and Learning Object Categories

Li Fei-Fei, UIUC

Rob Fergus, MIT

Antonio Torralba, MIT

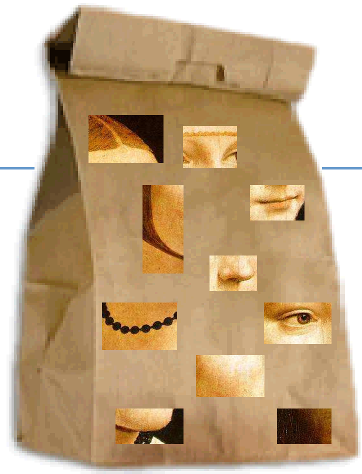
ILLINOIS

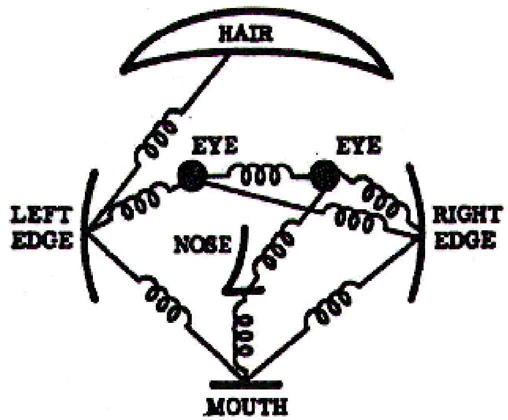
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

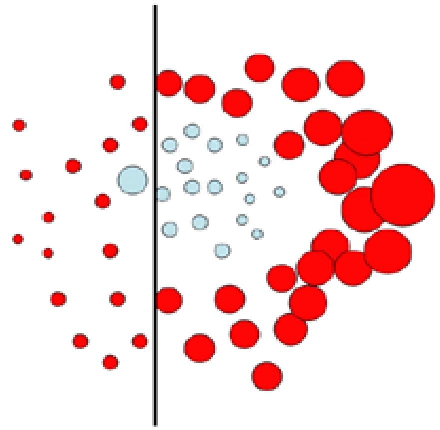
TM

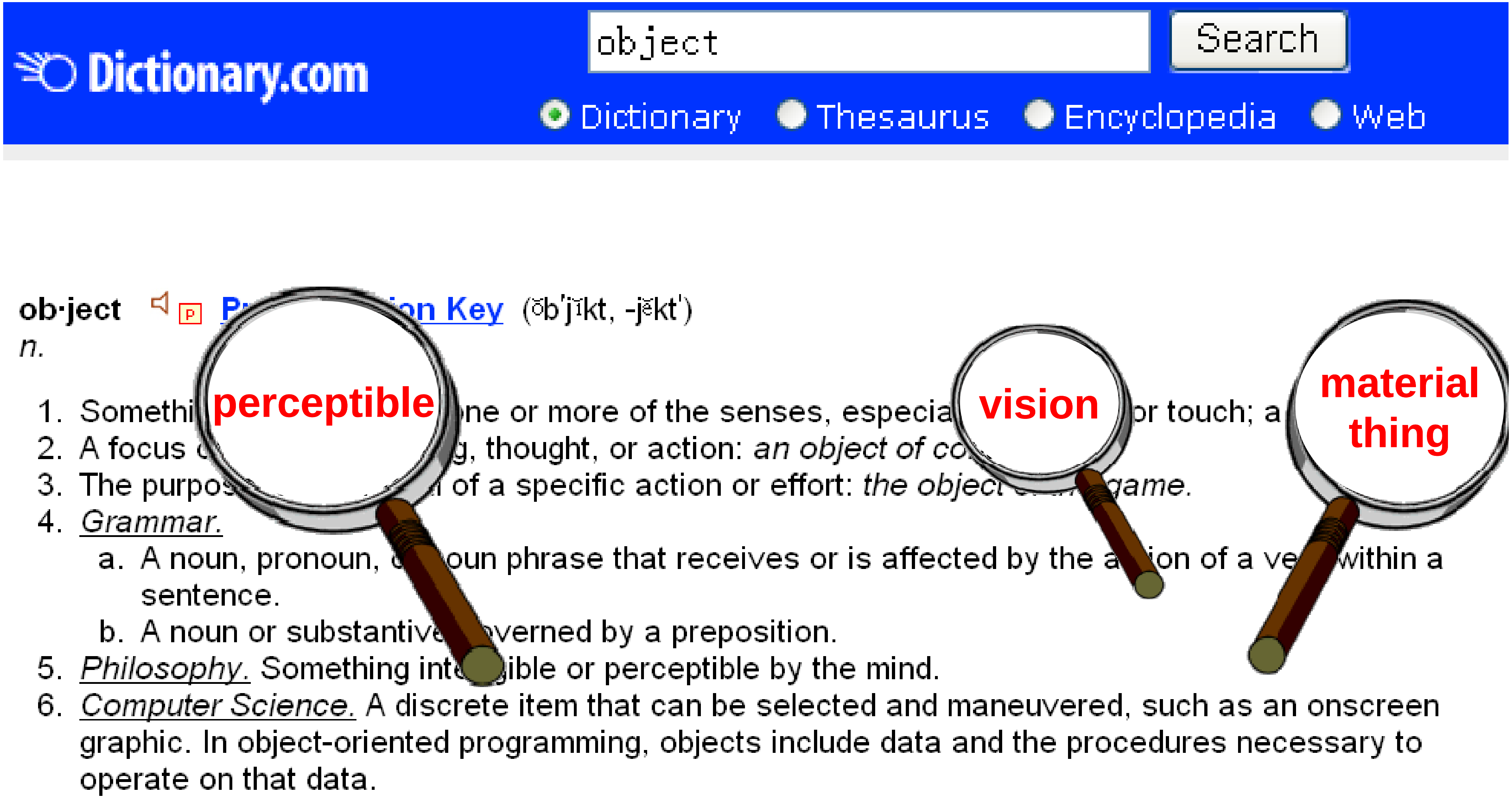
Agenda

- Introduction
- Bag of words models
- Part-based models
- Discriminative methods
- Segmentation and recognition
- Conclusions

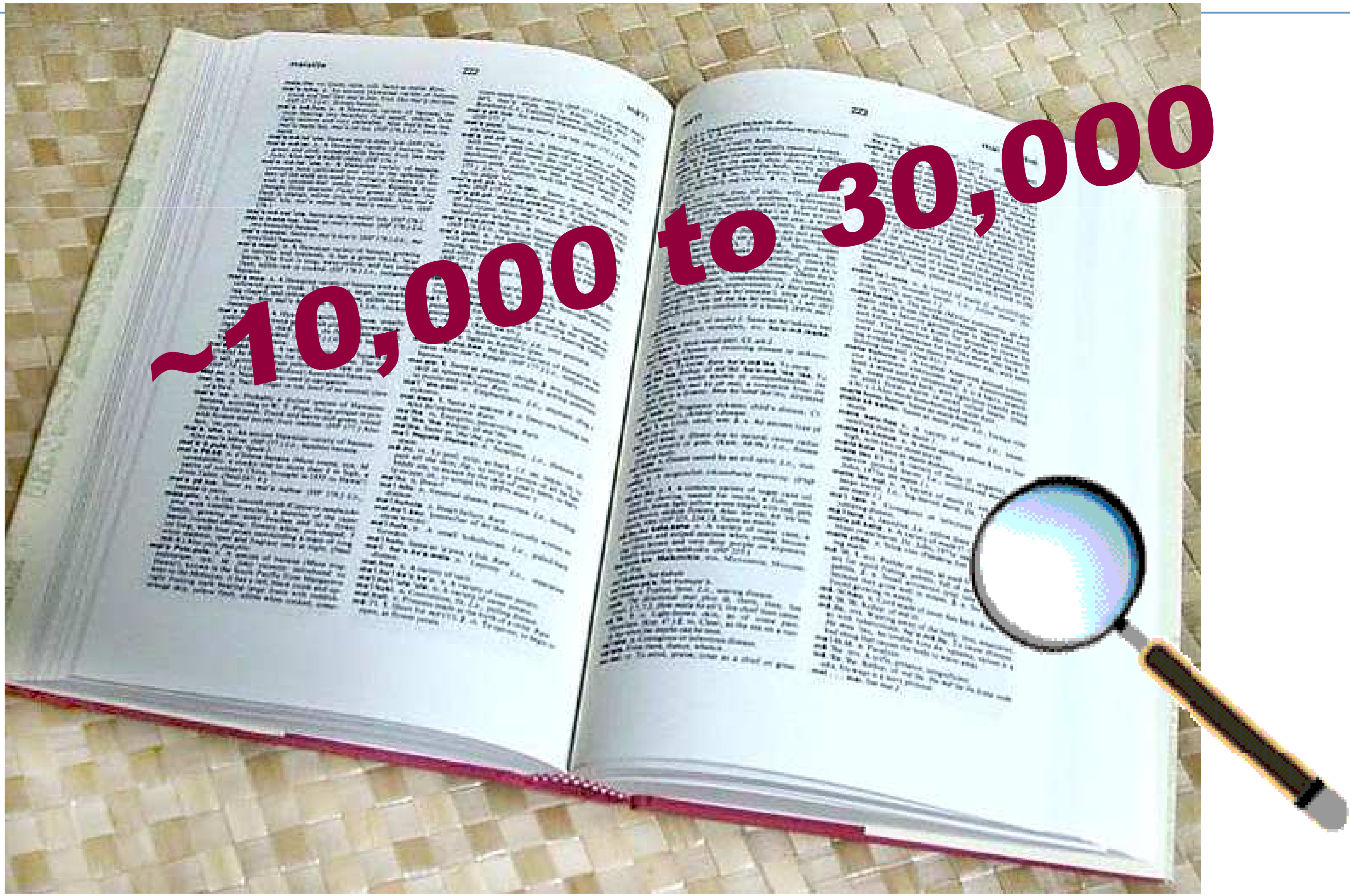







Dictionary.com
object
Search
Dictionary
Thesaurus
Encyclopedia
Web
ob·ject
P
on Key (ŏb'jĭkt, -jĕkt')
n.
1. Somethi
ne or more of the senses, especia
or touch; a
2. A focus
g, thought, or action: an object of co
3. The purpo
of a specific action or effort: the object
game.
4. Grammar.
a. A noun, pronoun,
oun phrase that receives or is affected by the a
ion of a ve
within a sentence.
b. A noun or substantive
verned by a preposition.
5. Philosophy. Something int
ible or perceptible by the mind.
6. Computer Science. A discrete item that can be selected and maneuvered, such as an onscreen graphic. In object-oriented programming, objects include data and the procedures necessary to operate on that data.
perceptible
vision
material thing

How many object categories are there?



Biederman 1987

So what does object recognition involve?

Verification: is that a bus?

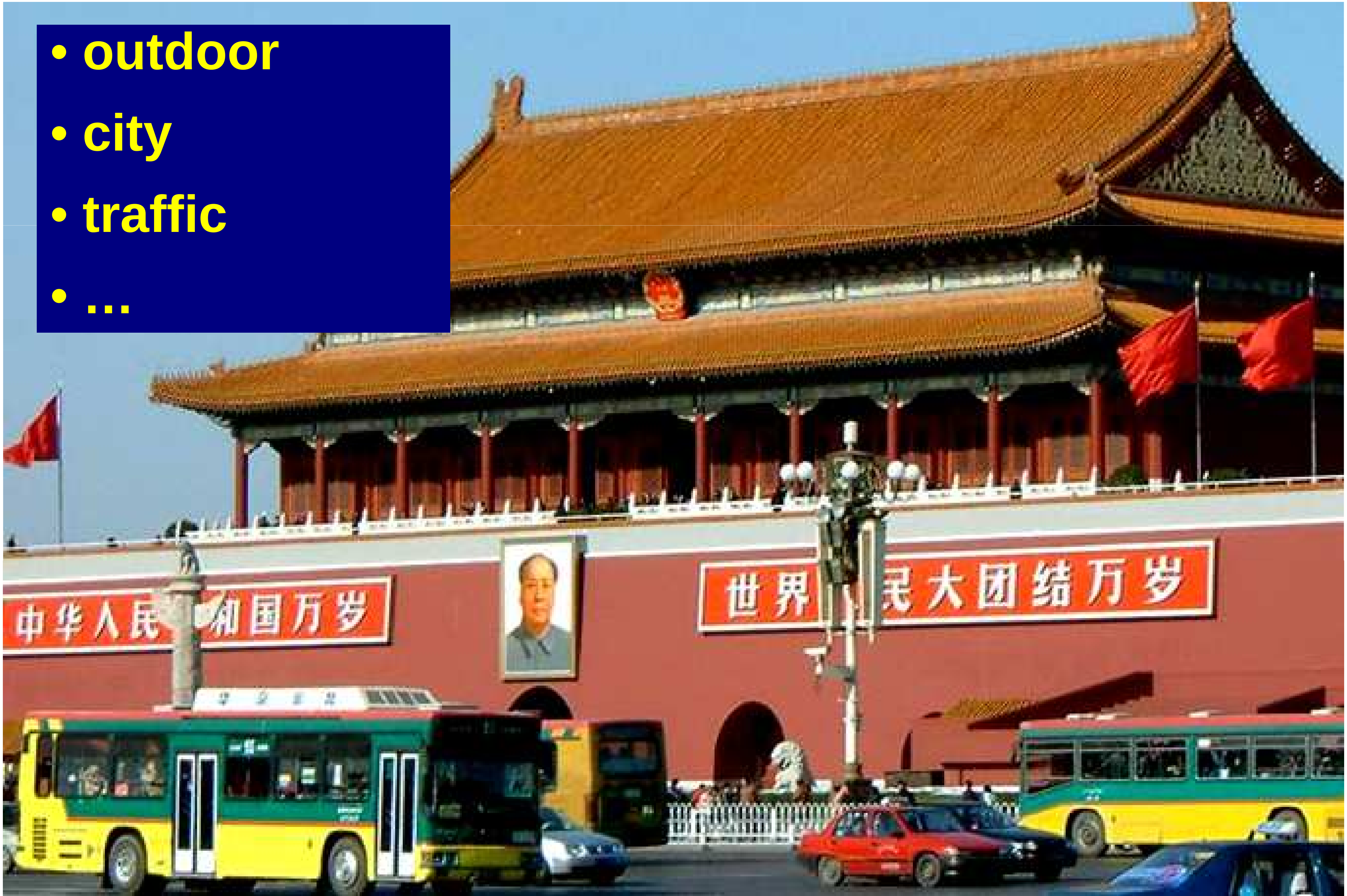
Detection: are there cars?

Identification: is that a picture of Mao?

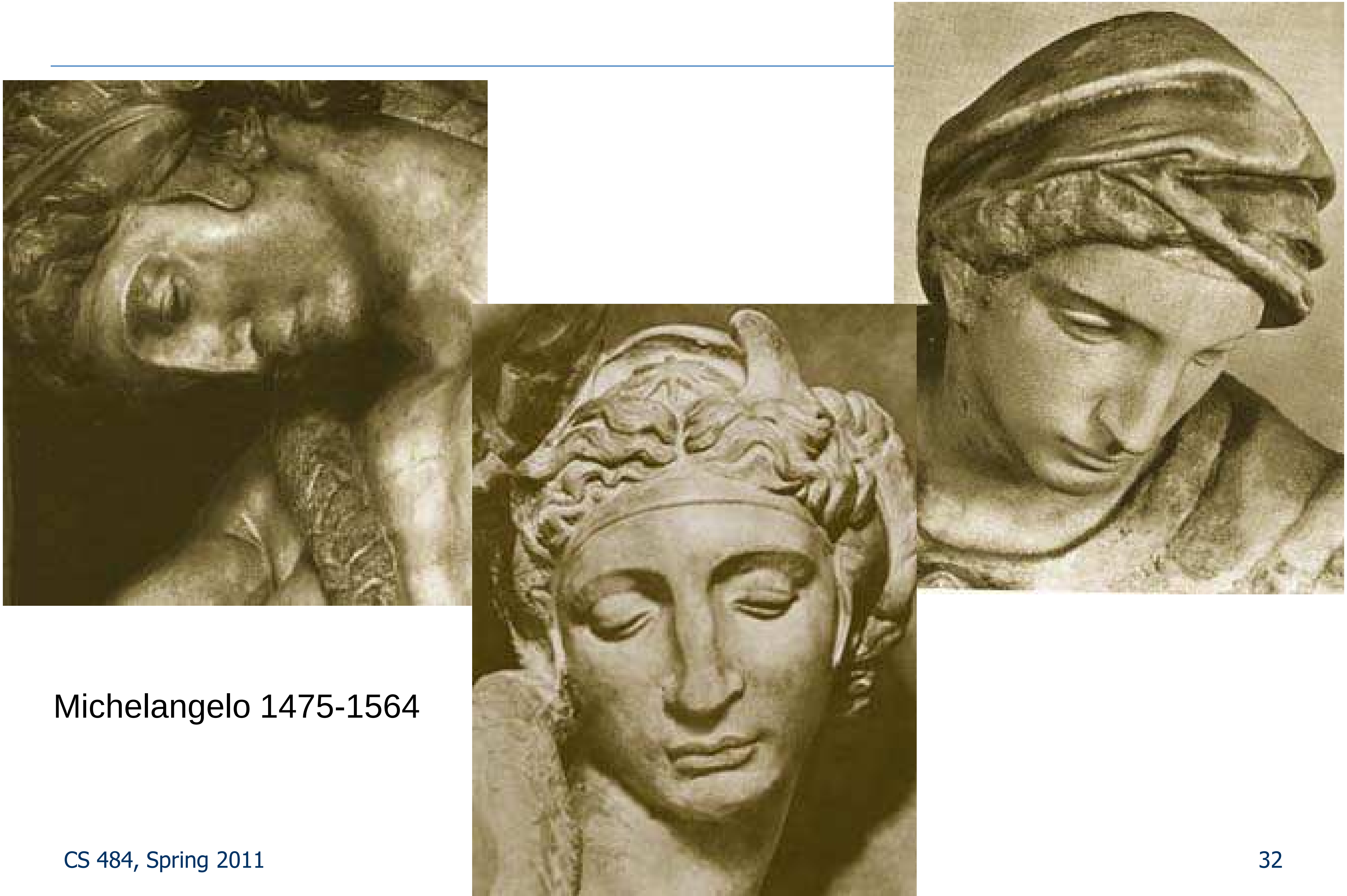
Object categorization

Scene and context categorization

- **outdoor**
- **city**
- **traffic**
- **...**



Challenges 1: view point variation

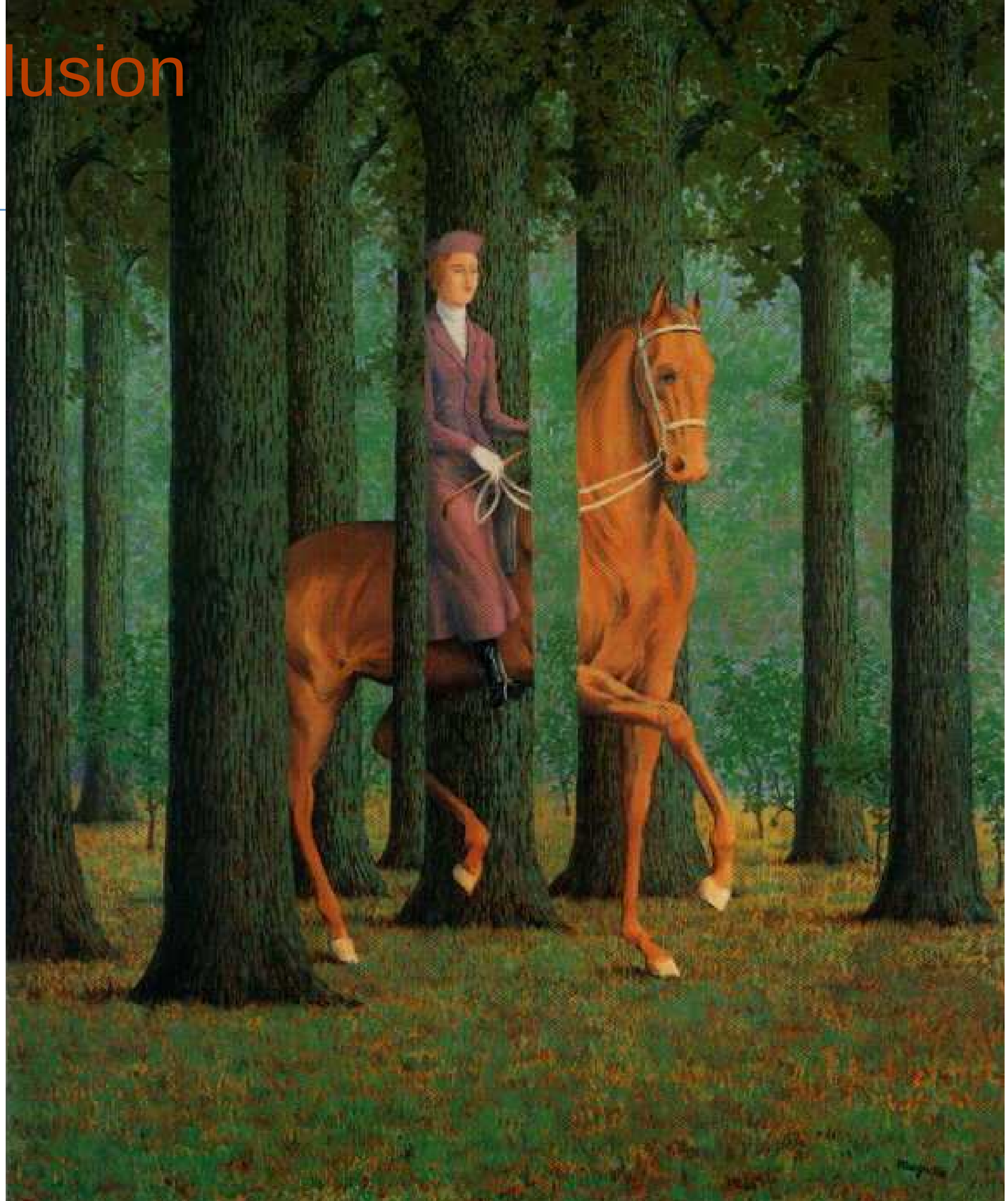


Michelangelo 1475-1564

Challenges 2: illumination

slide credit: S. Ullman

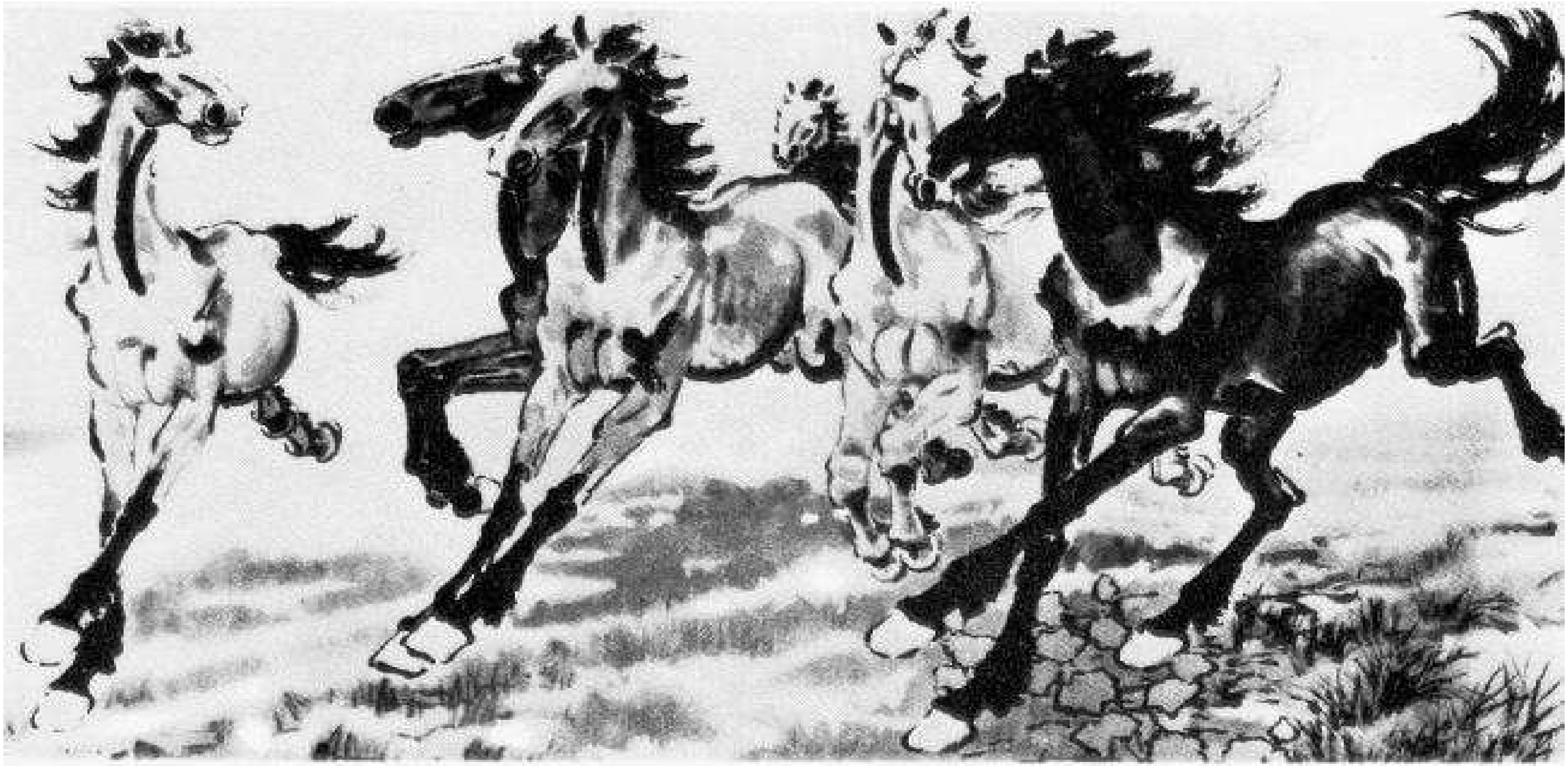
Challenges 3: occlusion



Magritte, 1957

Challenges 4: scale

Challenges 5: deformation



Xu, Beihong 1943

Challenges 6: background clutter

Klimt, 1913

History: single object recognition

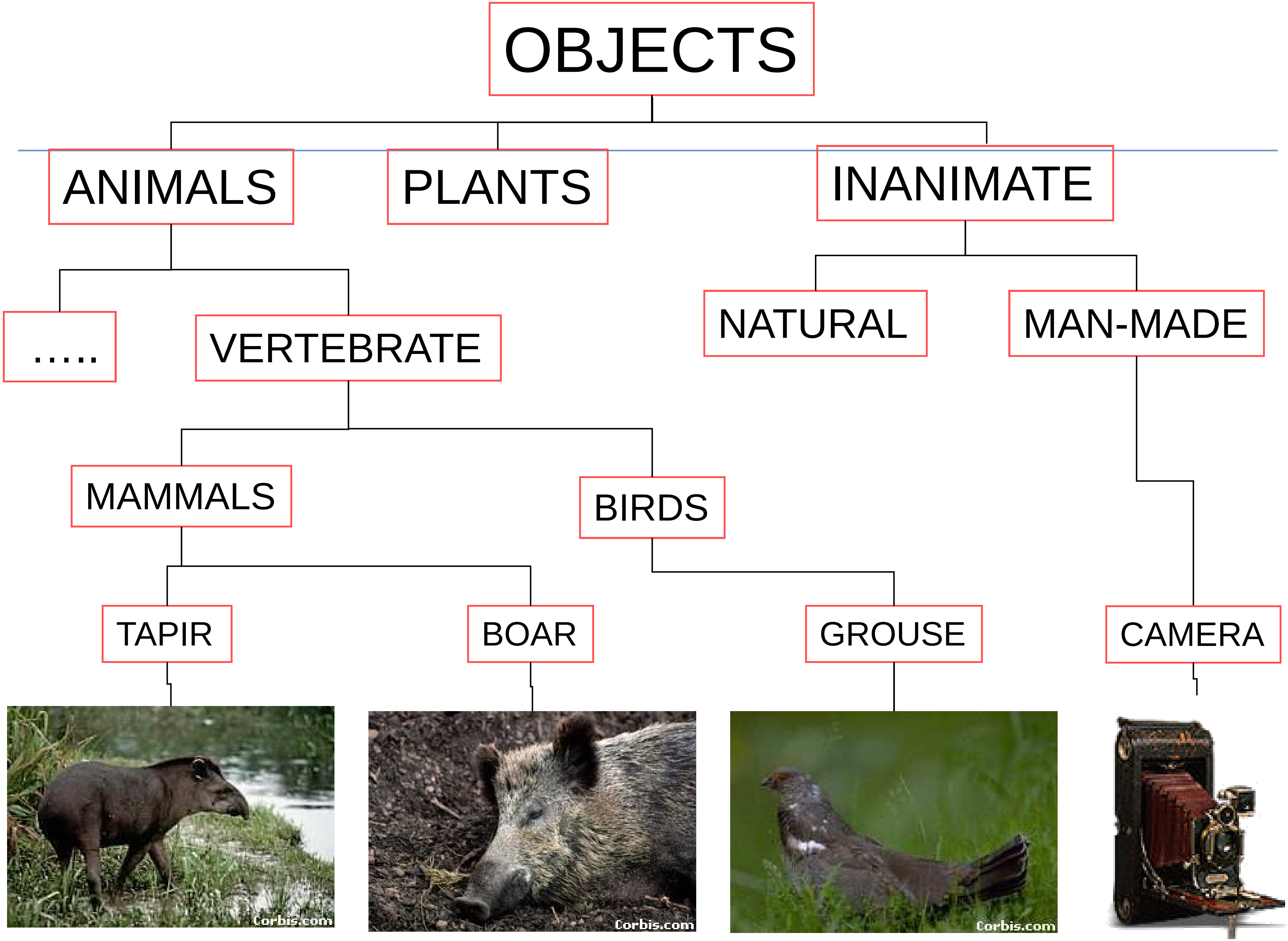
Challenges 7: intra-class variation

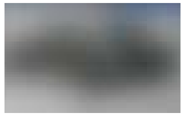




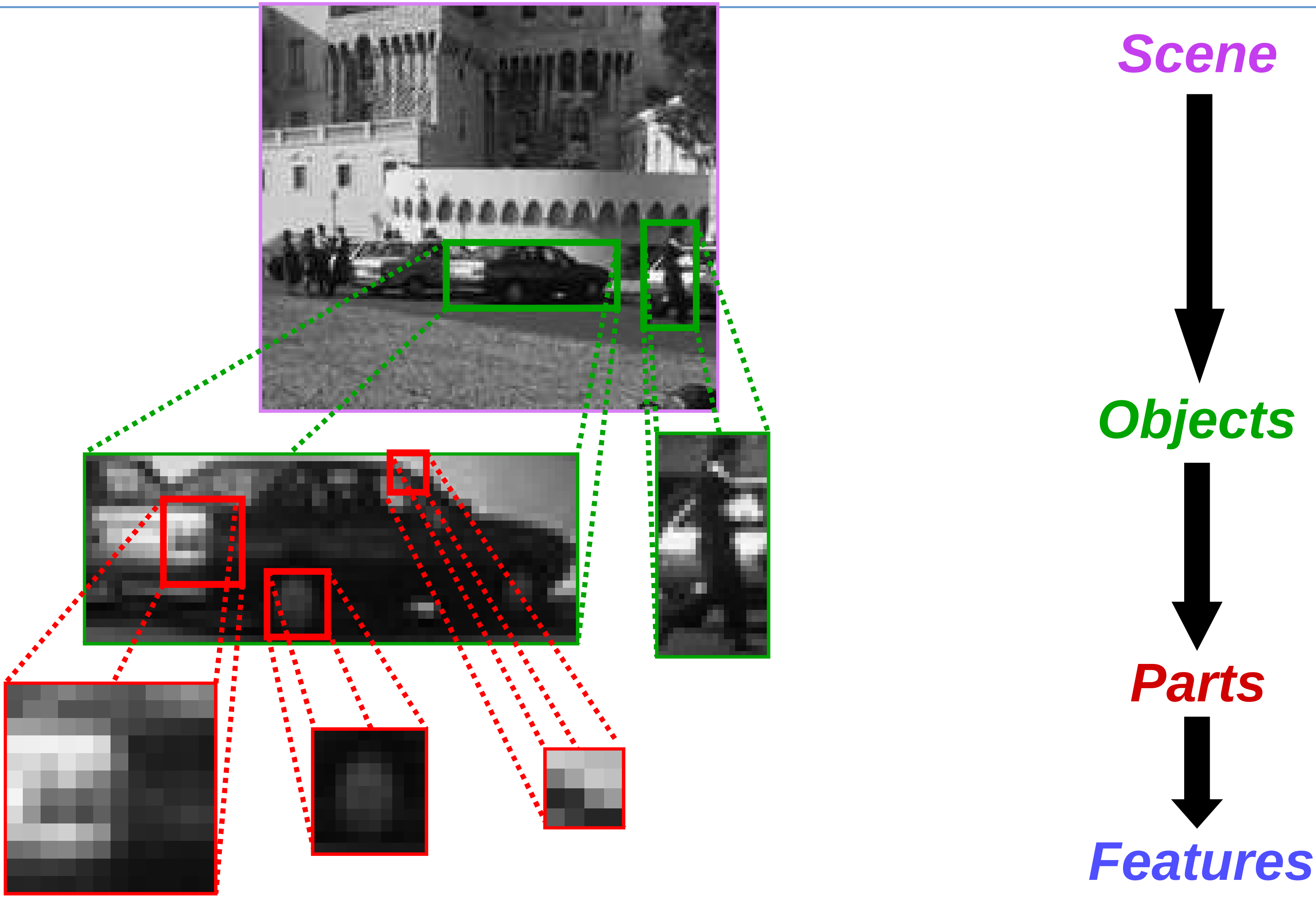

History: early object categorization


~10,000 to 30,000


OBJECTS
ANIMALS
PLANTS
INANIMATE
.....
VERTEBRATE
NATURAL
MAN-MADE
MAMMALS
BIRDS
TAPIR
BOAR
GROUSE
CAMERA
Corbis.com
Corbis.com
Corbis.com



Scenes, Objects, and Parts



E. Sudderth, A. Torralba, W. Freeman, A. Willsky. ICCV 2005.

Object categorization: the statistical viewpoint

$$p(zebra \mid image)$$

vs.

$$p(no\ zebra \mid image)$$

- Bayes rule:

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

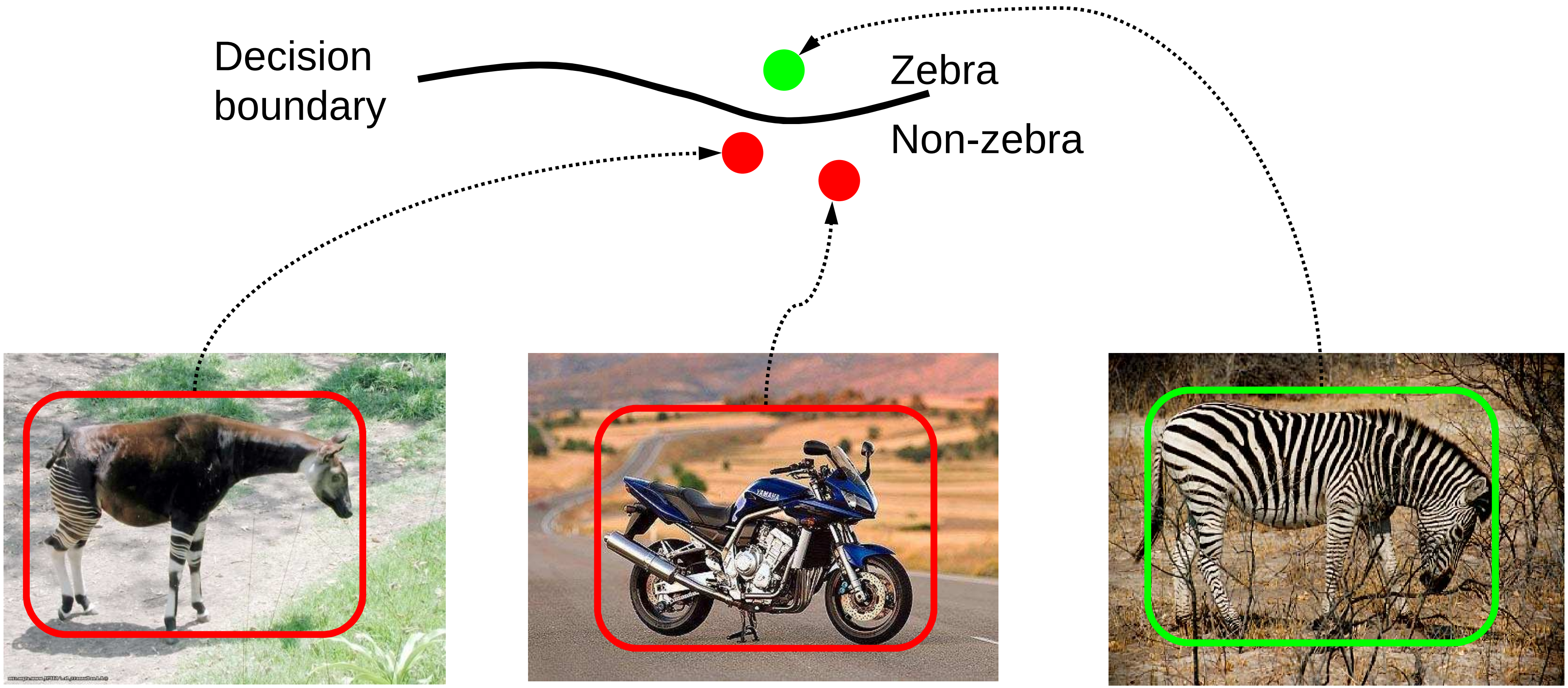
Object categorization: the statistical viewpoint

$$\underbrace{\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid zebra)}{p(image \mid no\ zebra)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(zebra)}{p(no\ zebra)}}_{\text{prior ratio}}$$

- **Discriminative methods model posterior**
- **Generative methods model likelihood and prior**

Discriminative

- Direct modeling of $$\frac{p(zebra \mid image)}{p(no\ zebra \mid image)}$$



Generative

- Model $p(image \mid zebra)$ and $p(image \mid no\ zebra)$

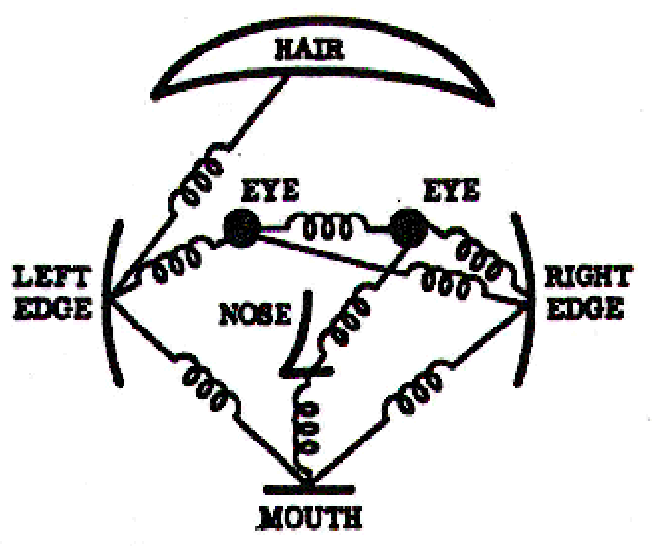
	$p(image \mid zebra)$	$p(image \mid no\ zebra)$
	Low	Middle
	High	Middle→Low

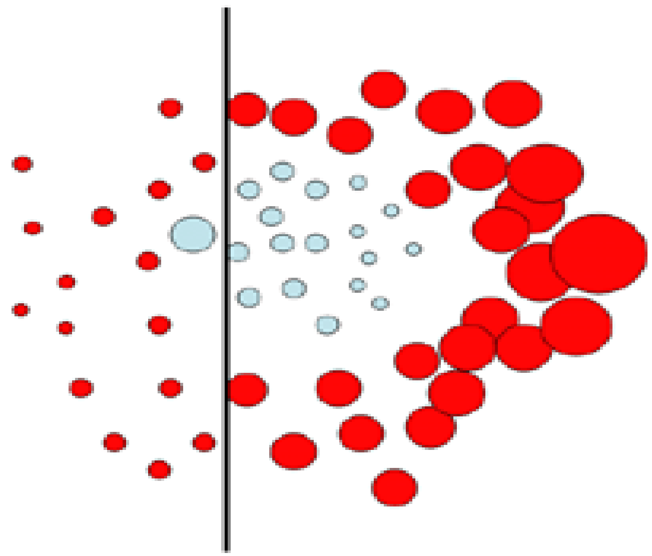
Three main issues

- Representation
  - How to represent an object category
- Learning
  - How to form the classifier, given training data
- Recognition
  - How the classifier is to be used on novel data

Representation

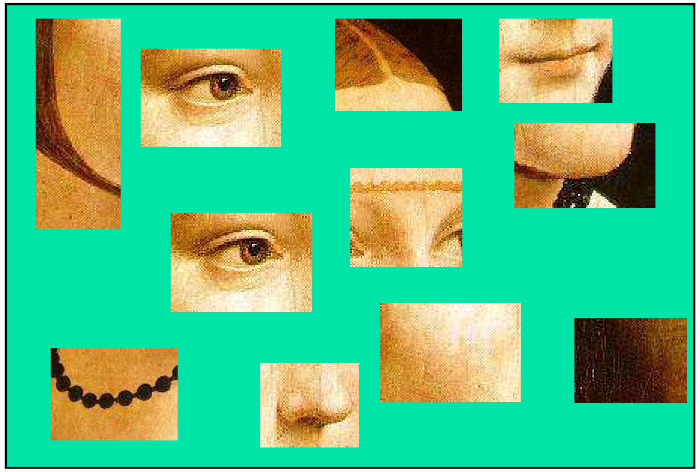
- Generative / discriminative / hybrid

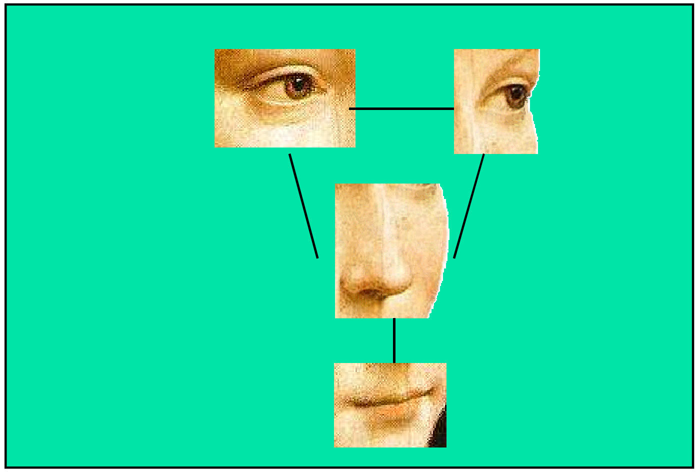




Representation

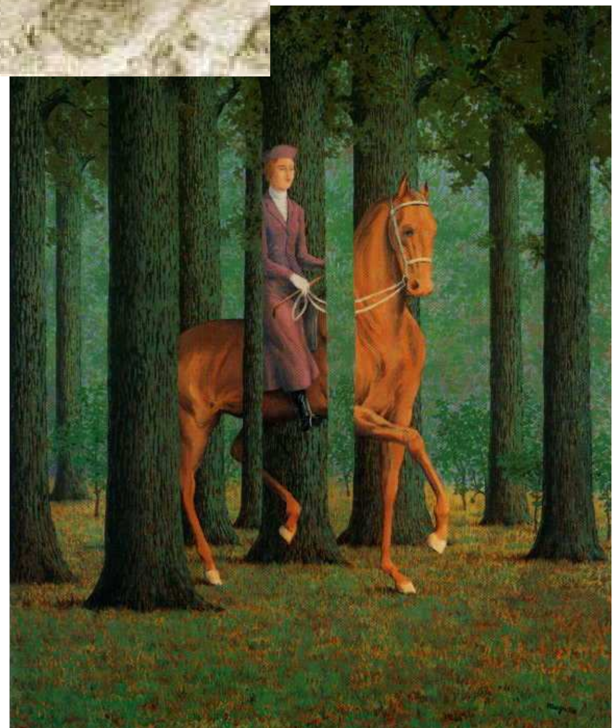
- Generative / discriminative / hybrid
- Appearance only or location and appearance





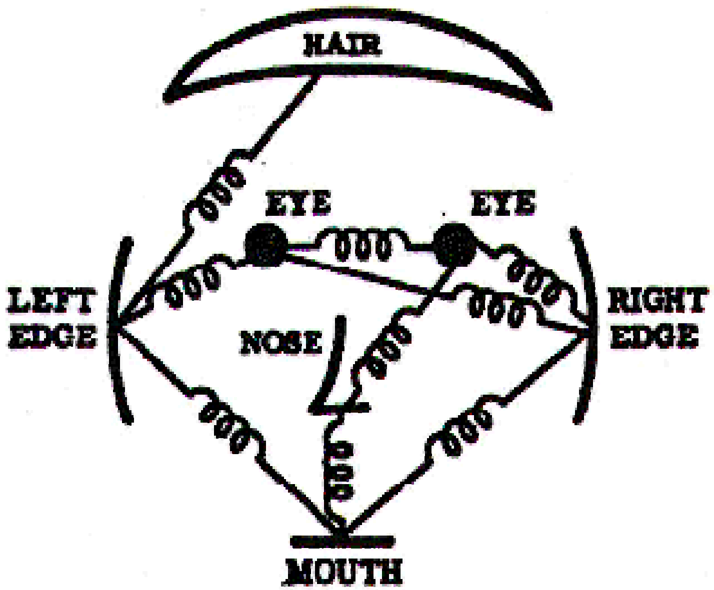
Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
- Invariances
  - View point
  - Illumination
  - Occlusion
  - Scale
  - Deformation
  - Clutter
  - etc.



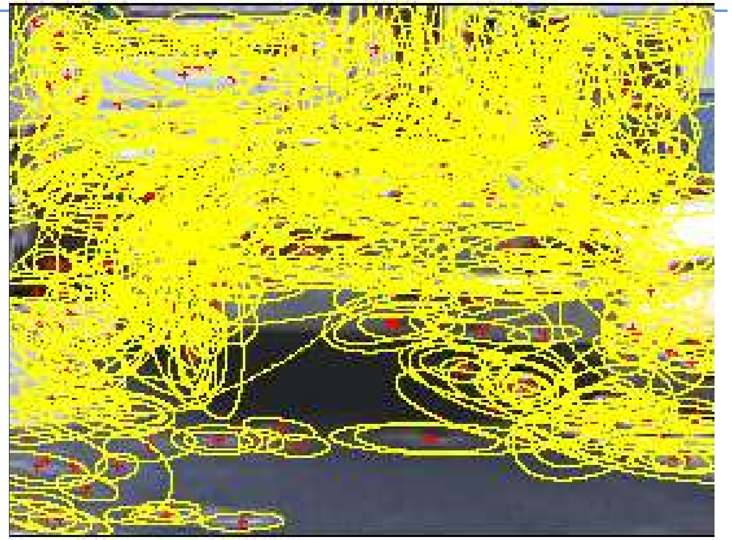
Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
- invariances
- Part-based or global w/sub-window



Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
- invariances
- Parts or global w/sub-window
- Use set of features or each pixel in image

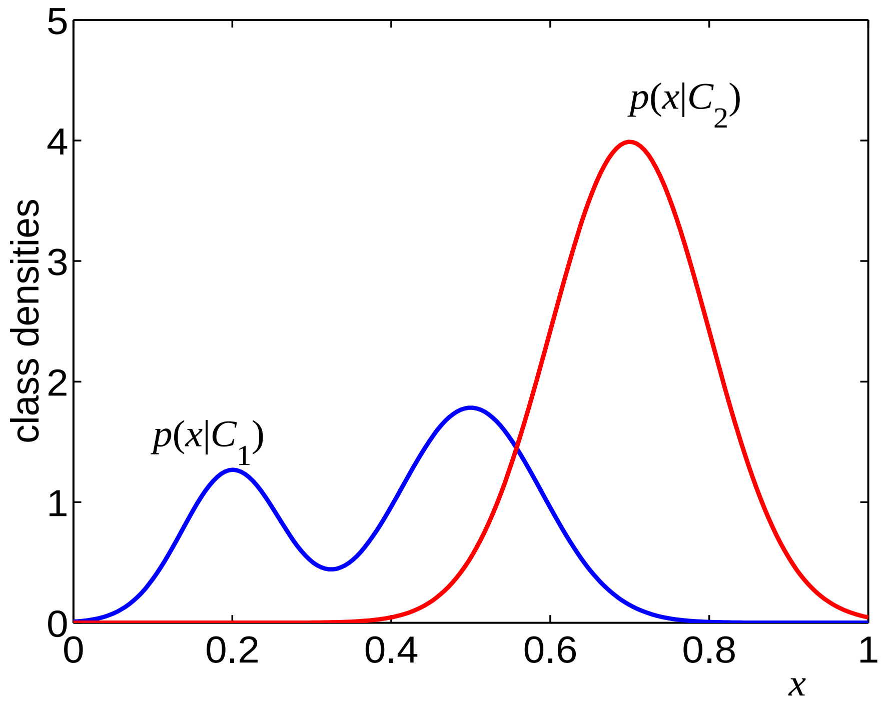


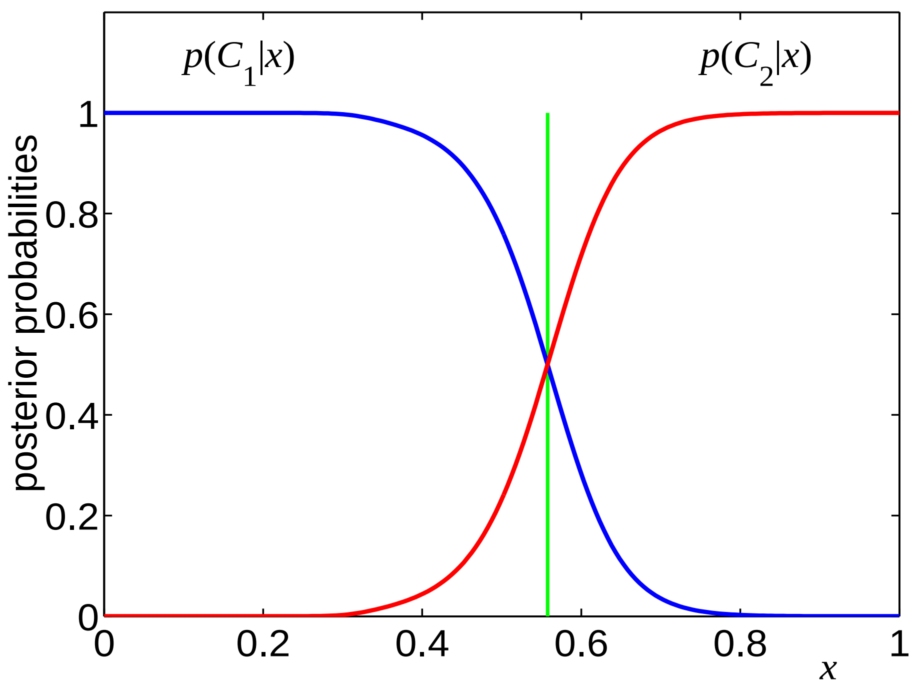
Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning

Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- Methods of training: generative vs. discriminative

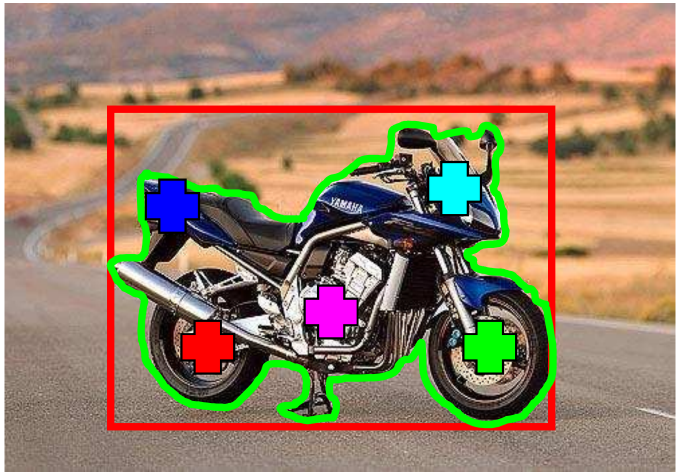




Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels

Contains a motorbike



Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels
- Batch/incremental (on category and image level; user-feedback)

Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels
- Batch/incremental (on category and image level; user-feedback)
- Training images:
  - Issue of overfitting
  - Negative images for discriminative methods

Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels
- Batch/incremental (on category and image level; user-feedback)
- Training images:
  - Issue of overfitting
  - Negative images for discriminative methods
- Priors

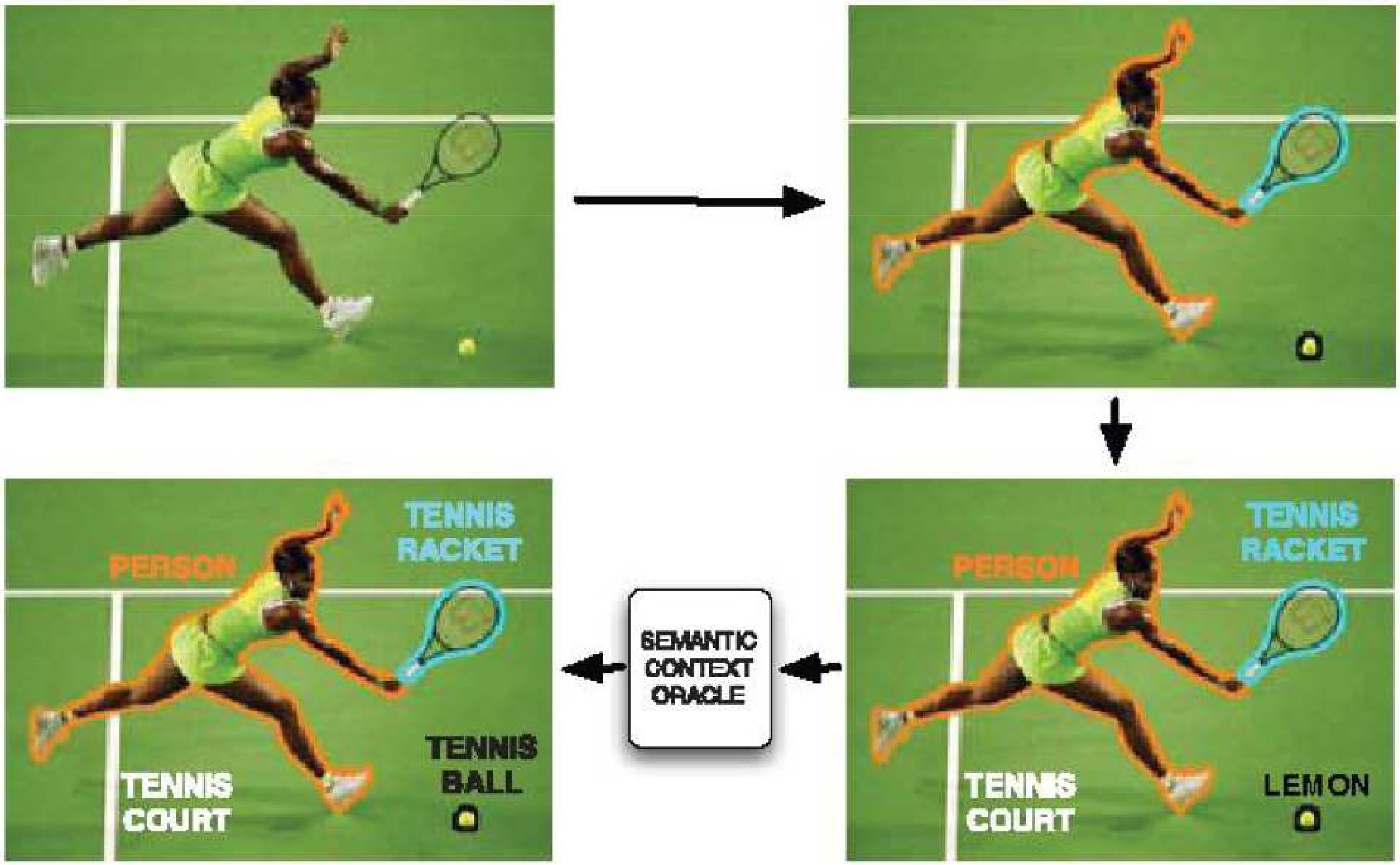
Recognition

- Scale / orientation range to search over
- Speed

Object recognition using context

- The relationships between objects and the scene setting can be characterized in five ways:
  - Interposition: objects interrupt their background
  - Support: objects tend to rest on surfaces
  - Probability: objects tend to be found in some contexts but not others
  - Position: given an object is probable in a scene, it is often found in some positions and not others
  - Familiar size: objects have a limited set of size relations with other objects

Object recognition using context



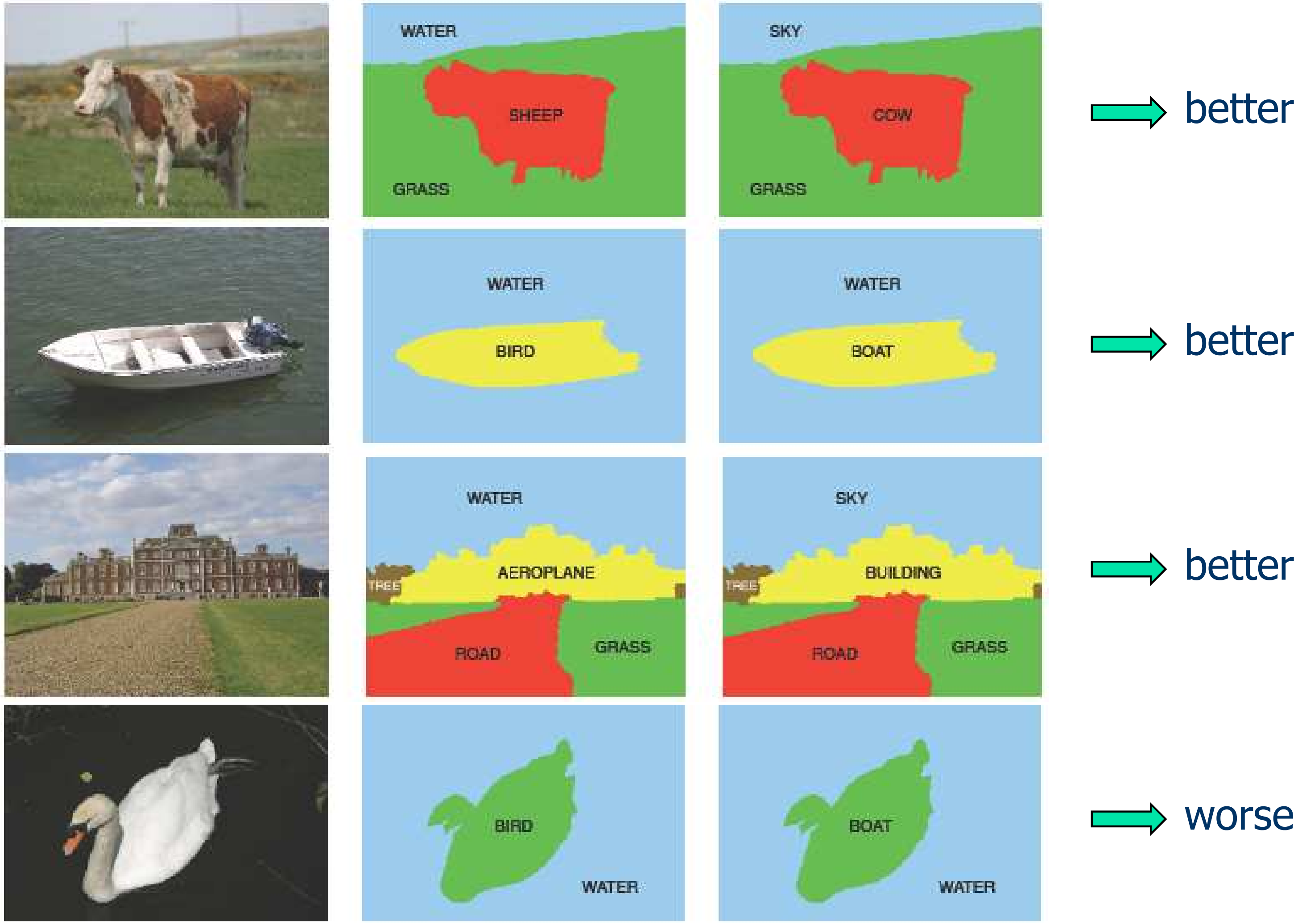
An ideal setting where the objects are perfectly segmented, the regions are classified, and the objects' labels are refined with respect to semantic context in the image. (Rabinovich et al. "Objects in Context," ICCV, 2007)

Object recognition using context

First column: segmentation results; second column: classification without contextual constraints; third column: classification with co-occurrence contextual constraints implemented using a conditional random field model. (Rabinovich et al. "Objects in Context," ICCV, 2007)

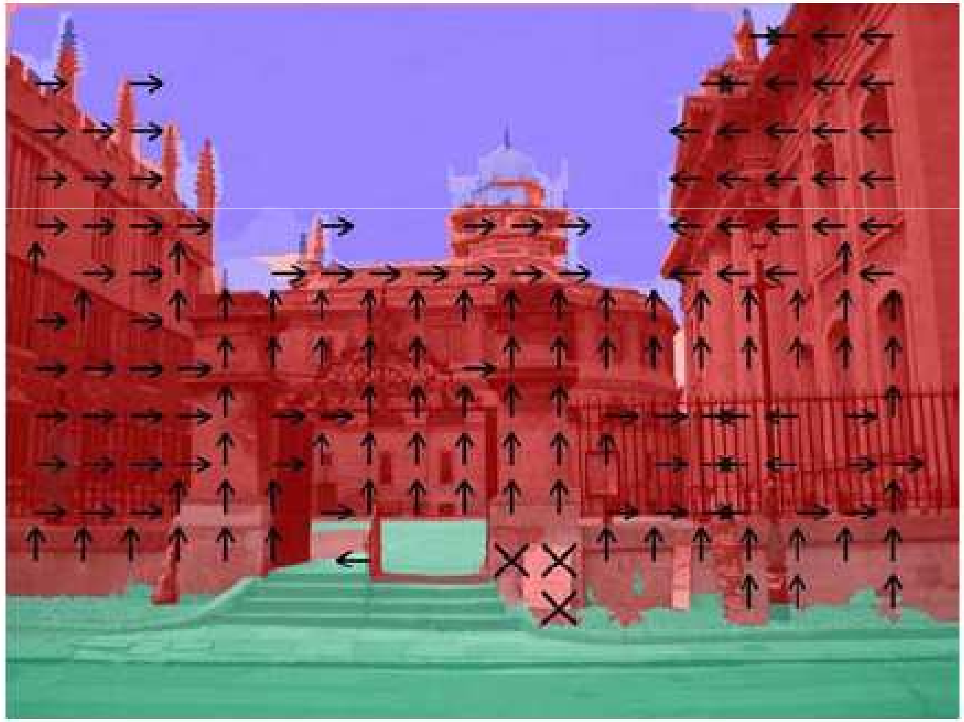
Object recognition using context

Using
•above
•below
•inside
•around
relationships as spatial constraints



First column: input image; second column: classification with co-occurrence contextual constraints; third column: classification with spatial and co-occurrence contextual constraints. (Galleguillos et al. "Object Categorization using Co-Occurrence, Location and Appearance," CVPR, 2008)

Object recognition using context



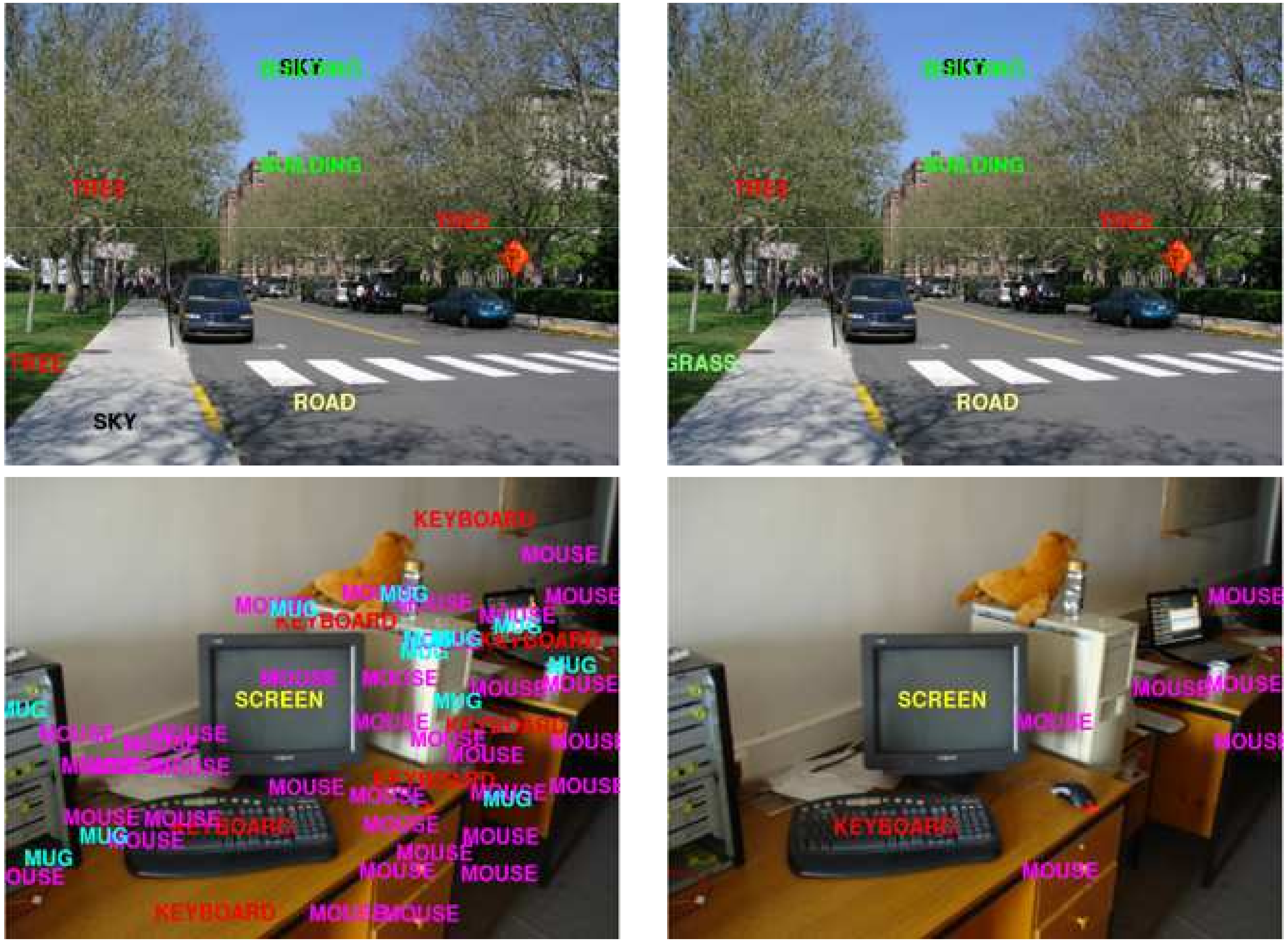
Geometric context: learn labeling of the image into geometric classes such as support/ground (green), vertical planar (green) and sky (blue). The arrows show the direction that the vertical planar surface is facing.
(Hoiem et al., “Geometric Context from a Single Image,” ICCV, 2005)

Object recognition using context

(a) (b) (c) (d)

Putting objects in perspective: Car (a) and pedestrian (b) detection by using only local information. Car (c) and pedestrian (d) detection by using context.
(Hoiem et al. "Putting Objects in Perspective," CVPR, 2006)

Object recognition using context



Contextual recognition by maximizing a scene probability function that incorporates outputs of individual object detectors (confidence values for assigned object labels) and pairwise interactions between objects (likelihood of pairwise spatial relationships). Left column: without using context; right column: with using context. (Firat Kalaycilar, "An object recognition framework using contextual interactions among objects," M.S. Thesis, Bilkent University, 2009)