Computer Vision

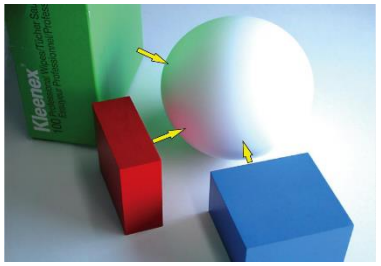
Image Categorization

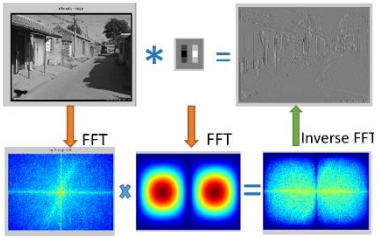
Dr. Mrinmoy Ghorai

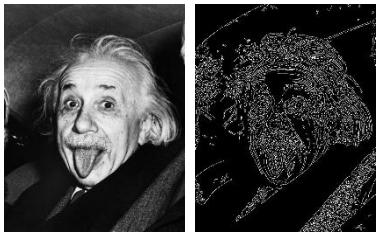
Indian Institute of Information Technology
Sri City, Chittoor

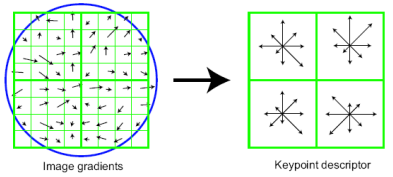
Topics Covered

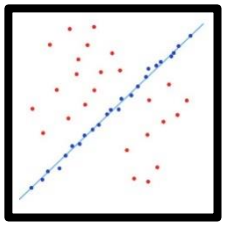
- **Light and color**
  - What an image records
- **Filtering in spatial and frequency domain**
  - Filtering, Smoothing, sharpening, measuring texture
  - Denoising, sampling
- **Image pyramid (Gaussian and Laplacian)**
  - Multi-scale analysis
- **Edge detection**
  - Canny edge, Finding straight lines
- **Interest points**
  - Find *distinct* and *repeatable* points in images
  - Harris-> corners, DoG -> blobs
  - SIFT -> feature descriptor
- **Feature tracking and optical flow**
  - Find motion of a keypoint/pixel over time
  - Lucas-Kanade, Handle large motion
- **Fitting and alignment**
  - find the transformation parameters that best align matched points

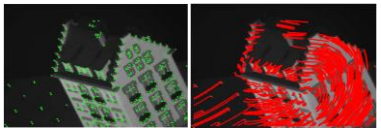






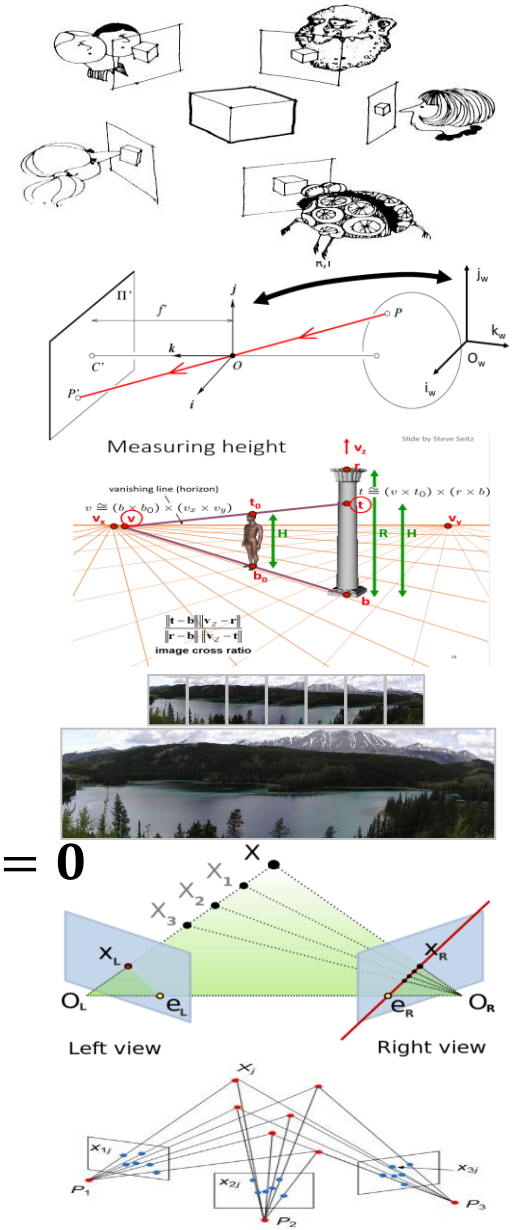






Topics Covered

- **Camera Models and Projective Geometry**
  - Perspective projection
  - Vanishing points/lines
- **Projection Matrix and Calibration**
  - $\mathrm{x} = \mathbf{K}[\mathbf{R}\ \ \mathbf{t}]\mathbf{X}$
  - Calibration using known 3D object or vanishing points
- **Single-view Metrology and Camera Properties**
  - Measuring size using perspective cues
  - Focal length, Field of View, etc.
- **Photo stitching**
  - Homography relates rotating cameras $\mathbf{x}' = \mathbf{H}\mathbf{x}$
  - Recover homography using RANSAC
- **Epipolar Geometry and Stereo Vision**
  - Fundamental/essential matrix relates two cameras $\mathbf{x}'\mathbf{F}\mathbf{x} = \mathbf{0}$
  - Recover $\mathbf{F}$ using RANSAC + normalized 8-point algorithm
  - Enforce rank 2 using SVD
- **Structure from motion**
  - How can we recover 3D points from multiple images?



Recognition and Learning

- Image Features and Categorization
- Classifiers
- Neural Networks
- Convolutional Neural Networks
- Object Detection
- Segmentation
- Image Generation
- Etc.

Today: Image features and categorization

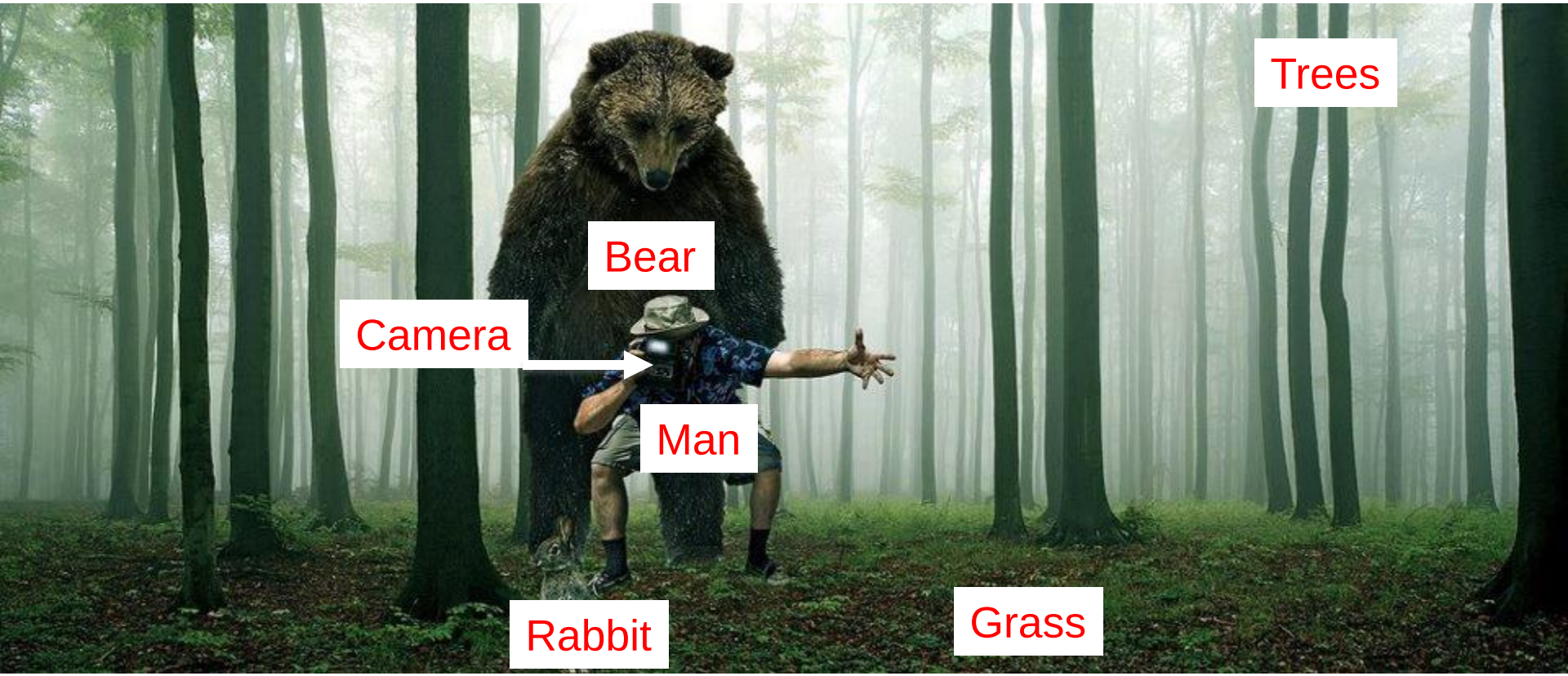
- General concepts of categorization
  - Why? What? How?
- Image features
  - Color, texture, gradient, shape, interest points
  - Histograms, SIFT, LBP, HoG
  - Bag of Visual Words
  - CNN as feature
- Image and region categorization

What do you see in this image?

What do you see in this image?

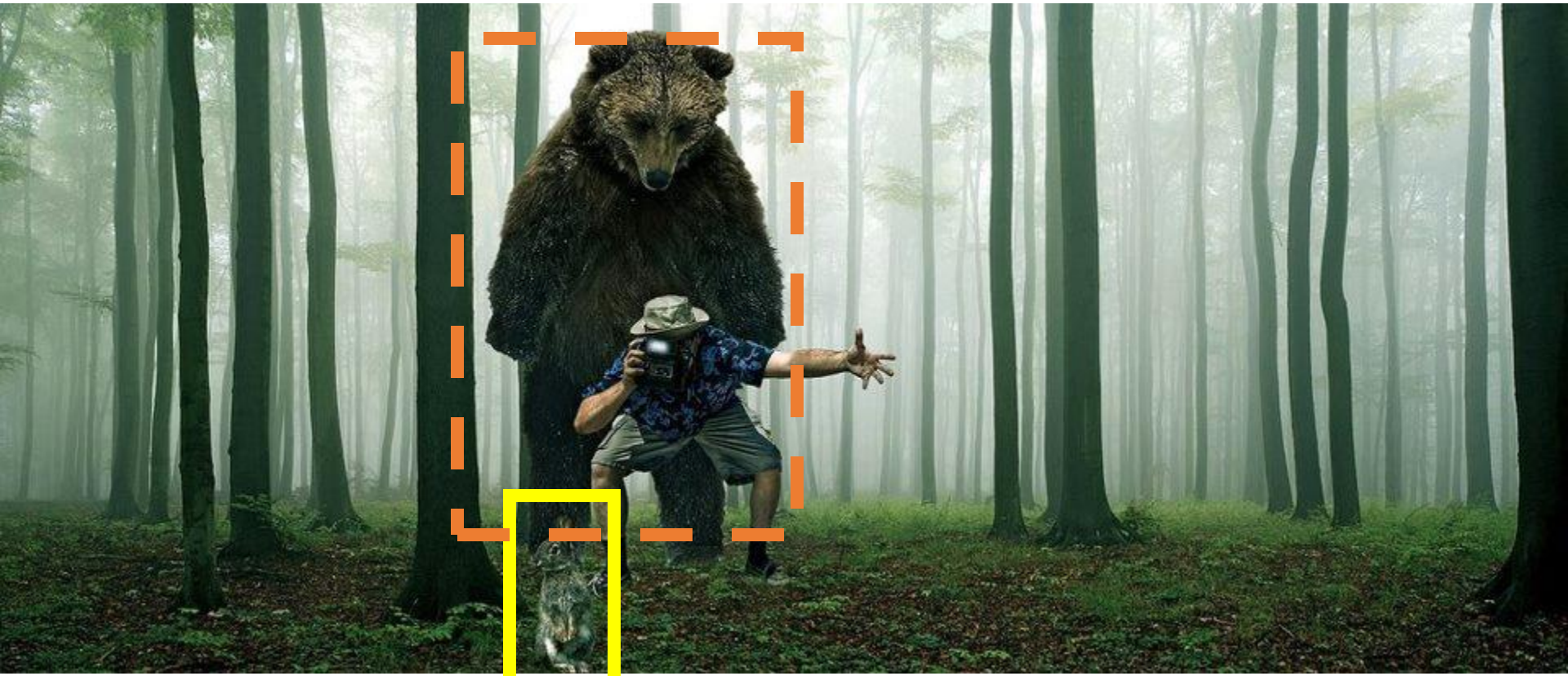
Forest

What do you see in this image?



Forest

Describe, predict, or interact with the object based on visual cues



Is it **dangerous**?

Is it **alive**?

How **fast** does it run?

Is it **soft**?

Does it have a **tail**?

Can I **poke with it**?

Why do we care about categories?

- From an object's category, we can make predictions about its behavior in the future, beyond of what is immediately perceived.
- Pointers to knowledge
  - Help to understand individual cases not previously encountered

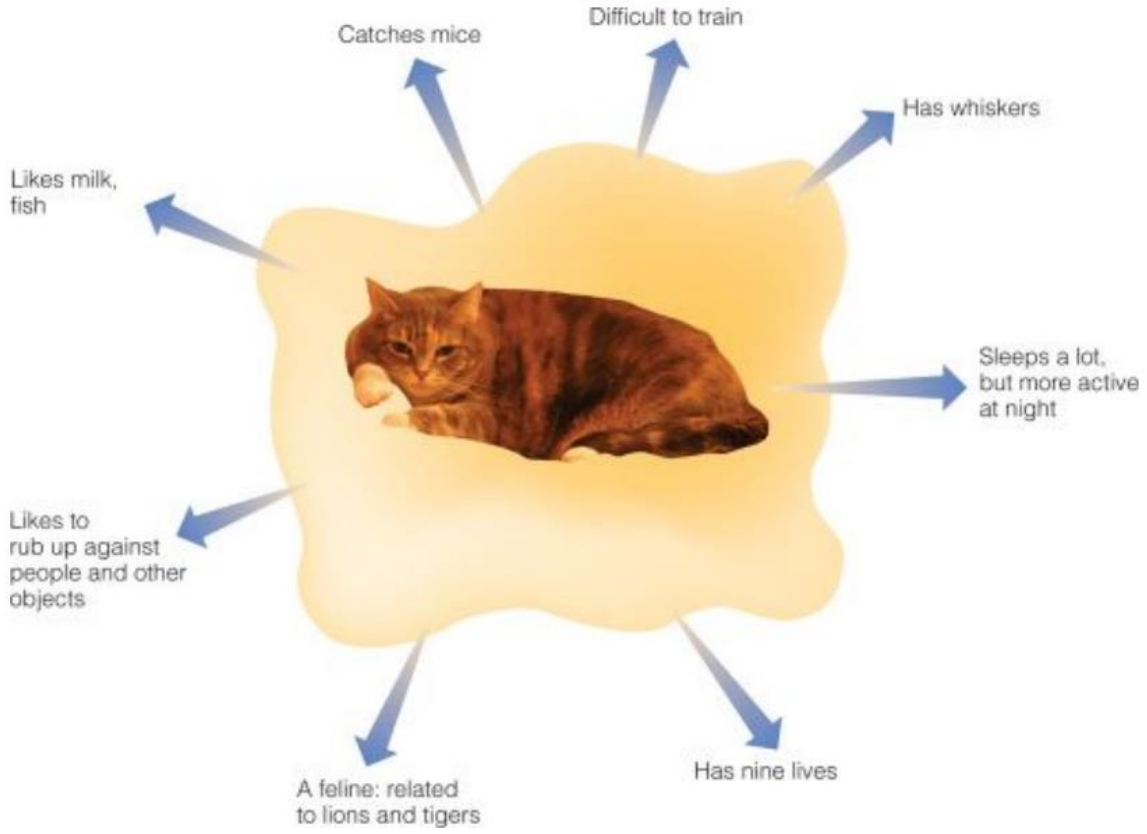
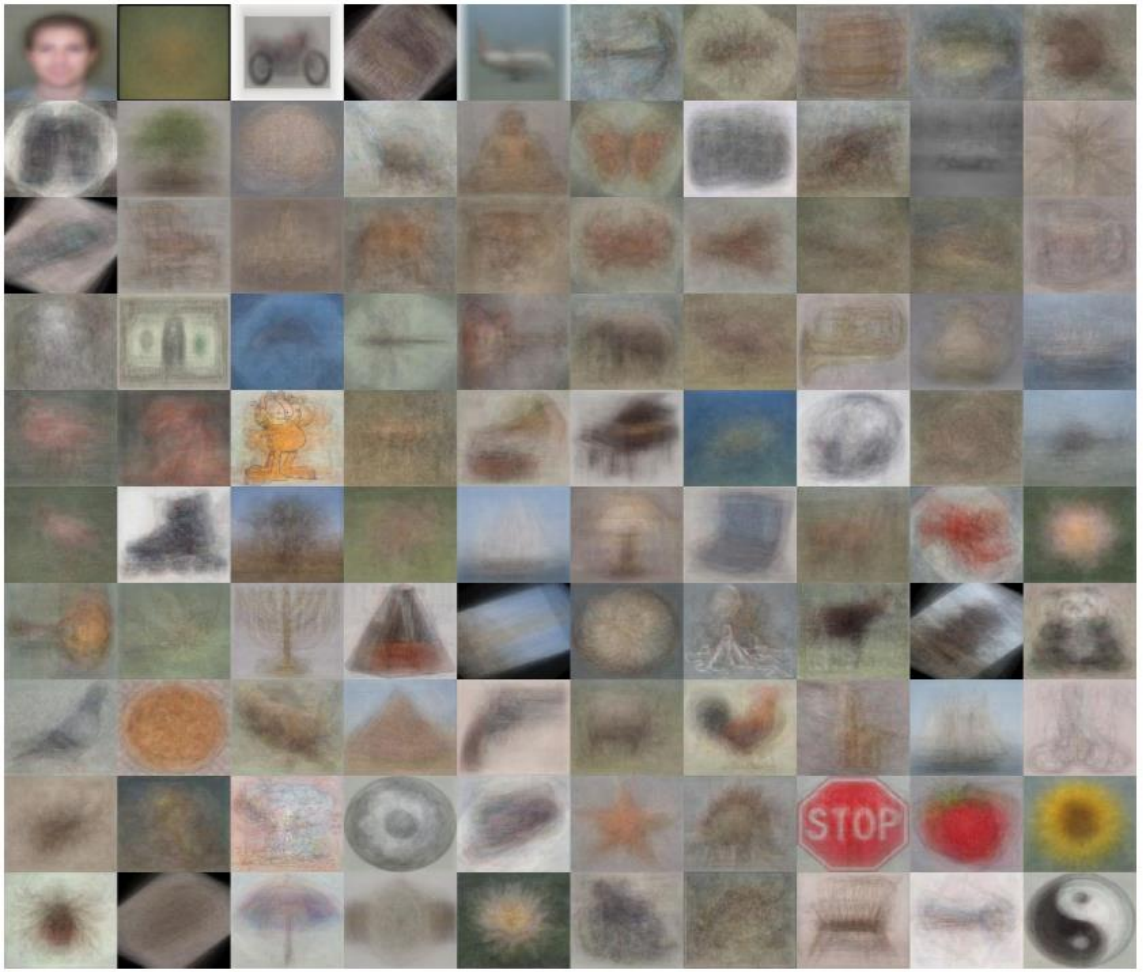


Image categorization

- Cat vs Dog

Image categorization

- Object recognition



Caltech 101 Average Object Images

Image categorization

- Place recognition

Places Database [Zhou et al. NIPS 2014]

Image categorization

- Image style recognition


HDR


Macro


Baroque


Roccoco


Vintage


Noir


Northern Renaissance


Cubism


Minimal


Hazy


Impressionism


Post-Impressionism


Long Exposure


Romantic


Abs. Expressionism


Color Field Painting

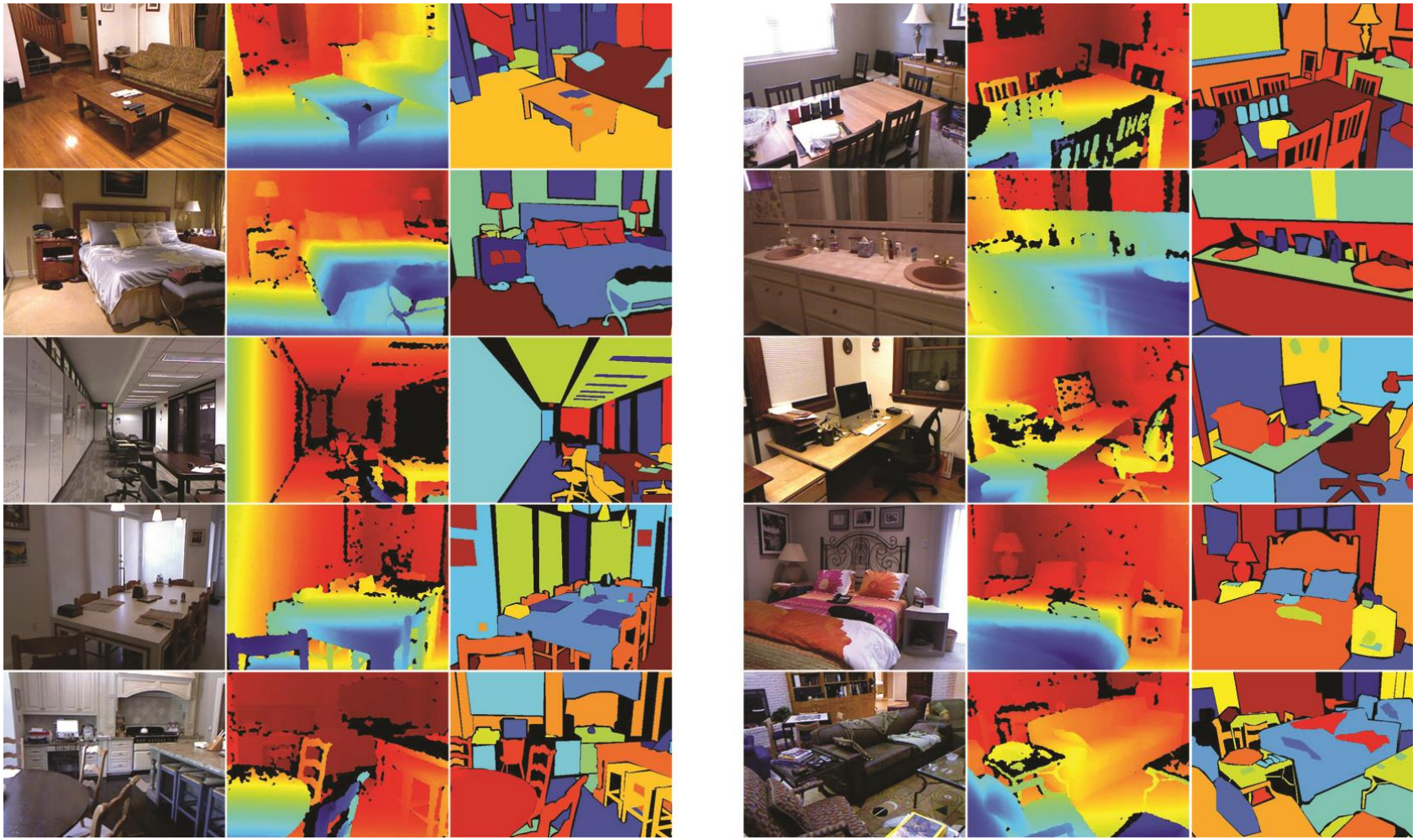
Flickr Style: 80K images covering 20 styles.

Wikipaintings: 85K images for 25 art genres.

[Karayev et al. BMVC 2014]

Region categorization

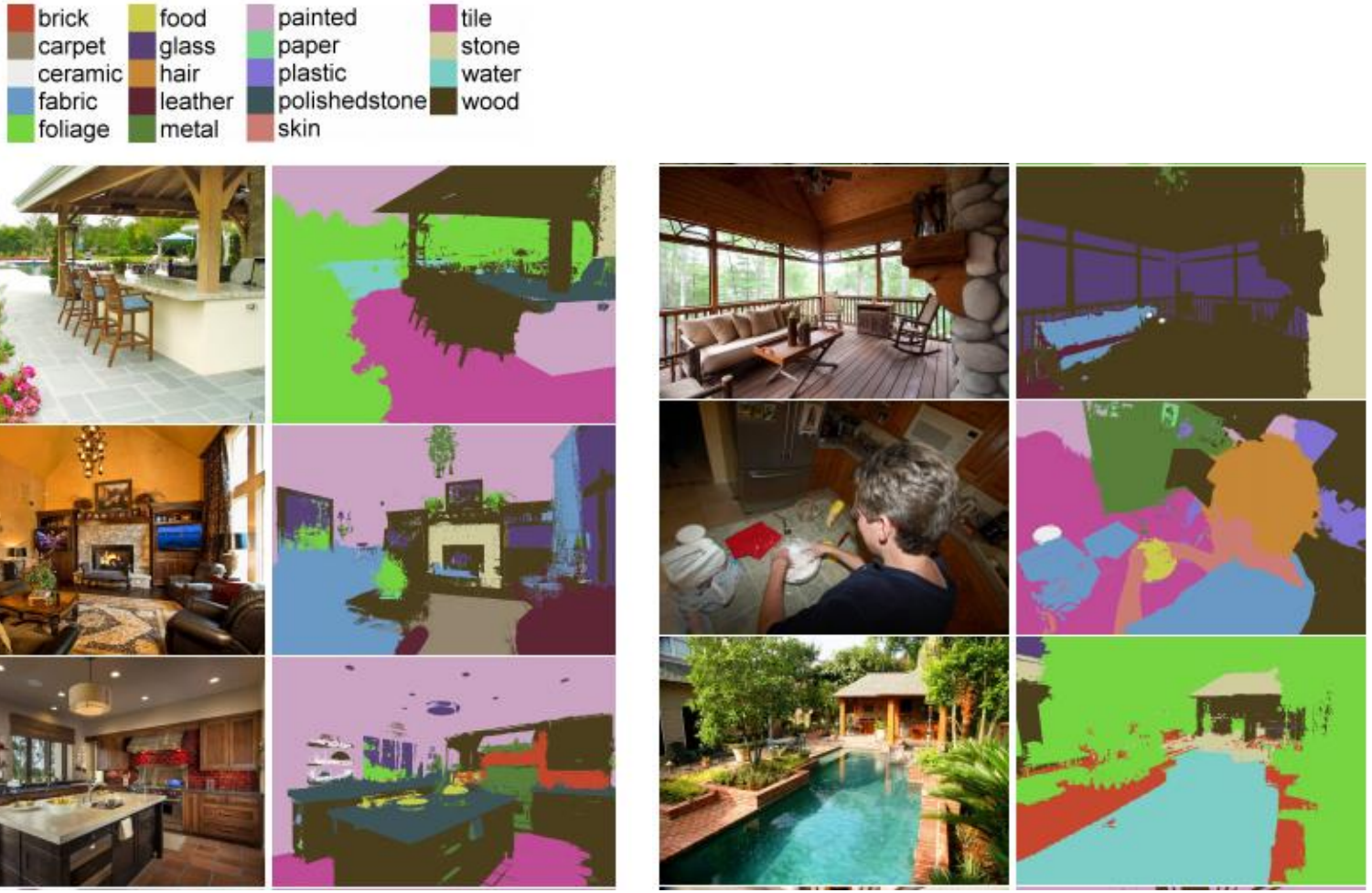
- Semantic segmentation from RGBD images



[Silberman et al. ECCV 2012]

Region categorization

- Material recognition



[Bell et al. CVPR 2015]

Image categorization

- Primary Tumor vs Metastasis

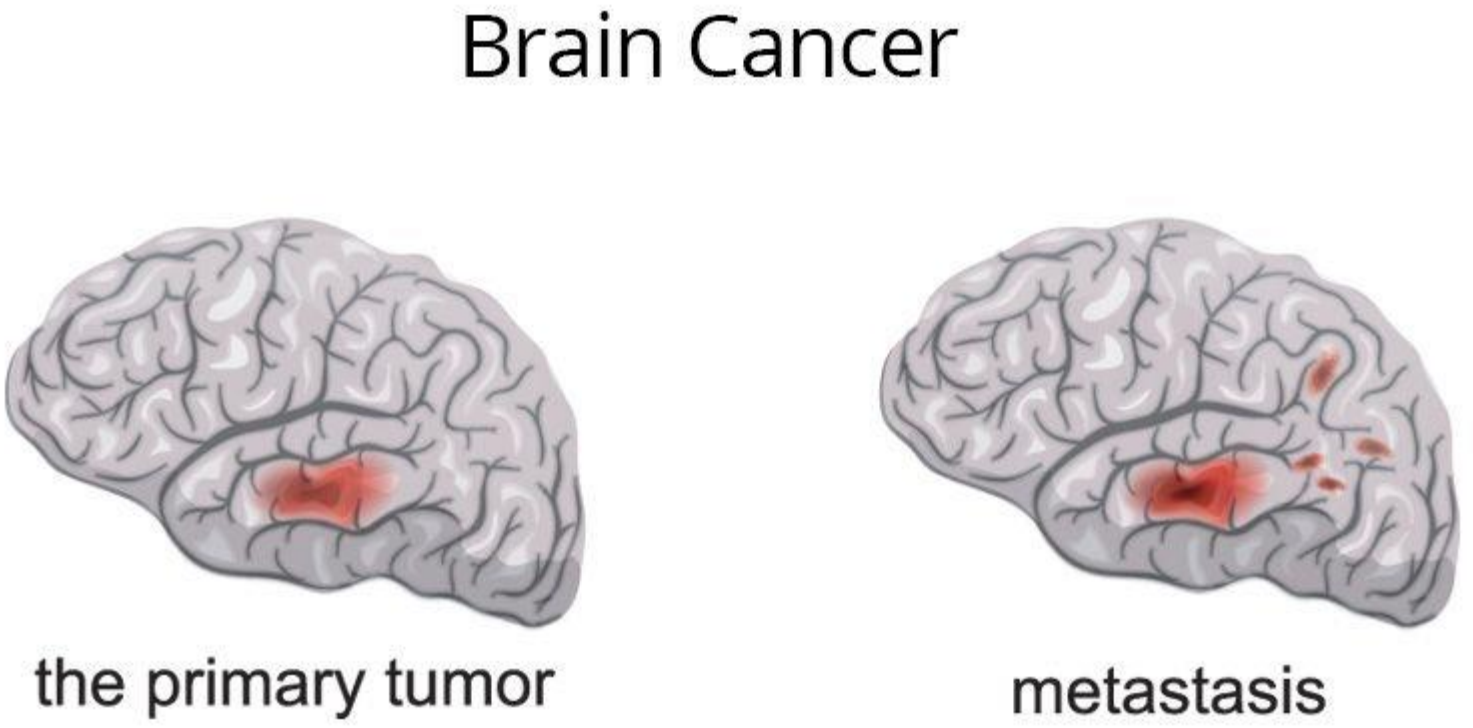
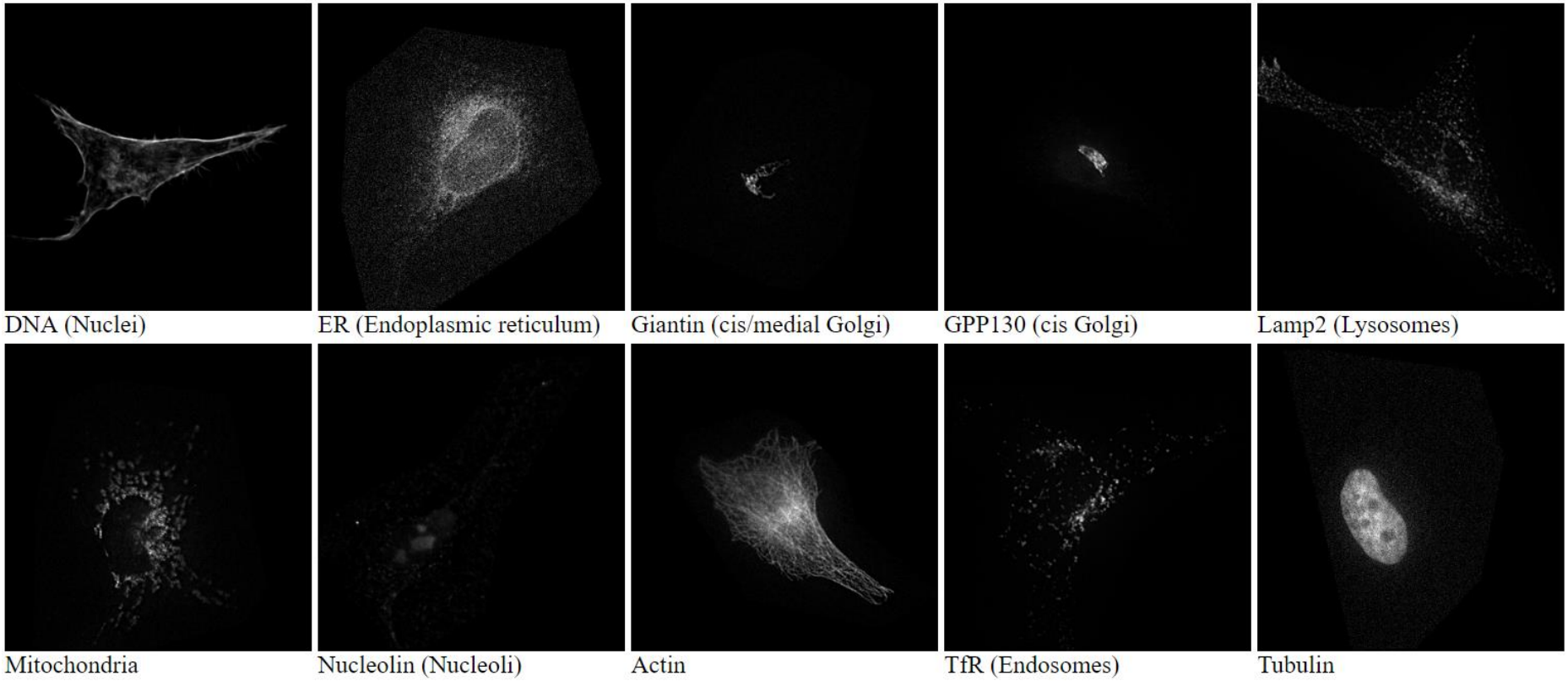


Image categorization

- Identification of sub-cellular organelles

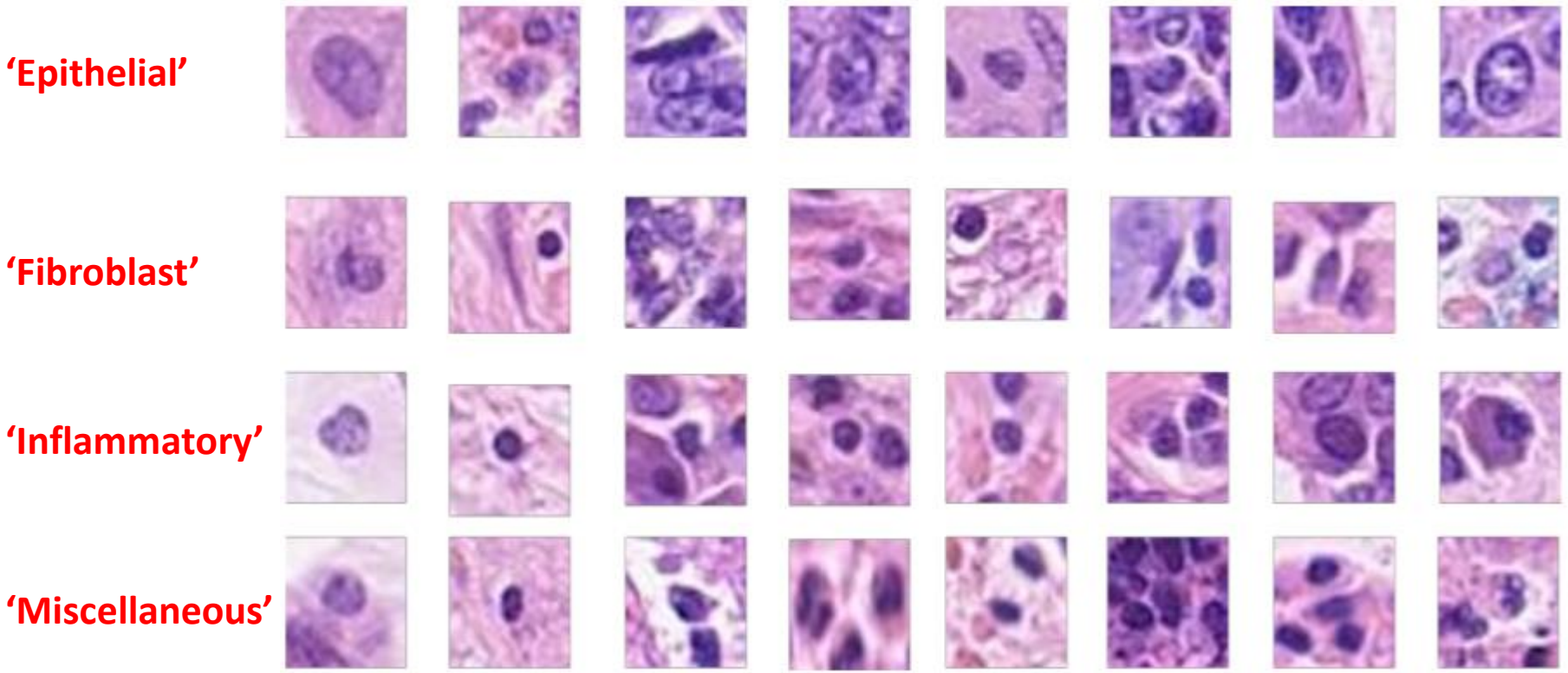


Fluorescence microscopy images of HeLa cells

https://ome.grc.nia.nih.gov/iicbu2008/hela/index.html

Image categorization

- Colon Cancer Nuclei Classification



"CRCHistoPhenotypes" dataset images

https://warwick.ac.uk/fac/sci/dcs/research/tia/data/crchistolabelednucleihe/

Image categorization/ classification

The statistical learning framework

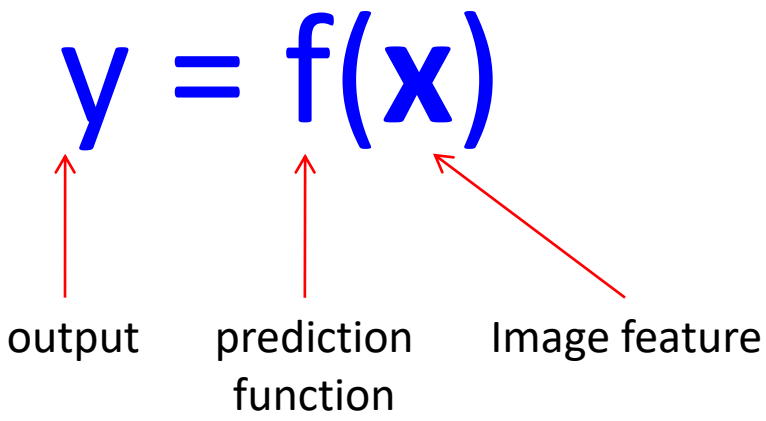
- Apply a prediction function to a feature representation of the image to get the desired output:

f() = “apple”

f() = “tomato”

f() = “cow”

The statistical learning framework



The statistical learning framework

$$y = f(\mathbf{x})$$

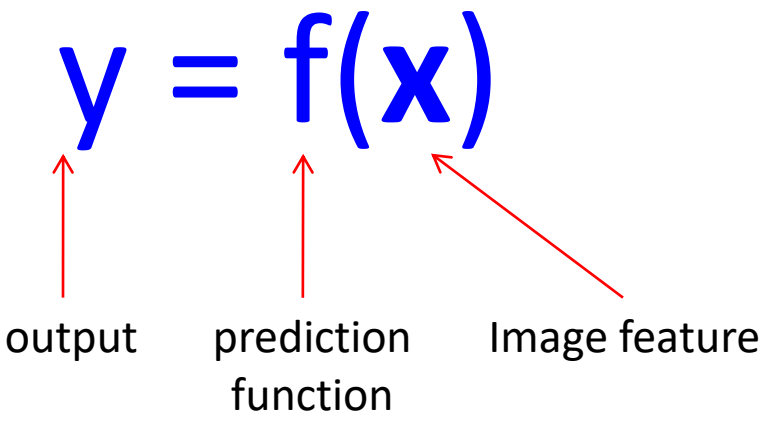
output

prediction function

Image feature

- **Training:** given a *training set* of labeled examples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set

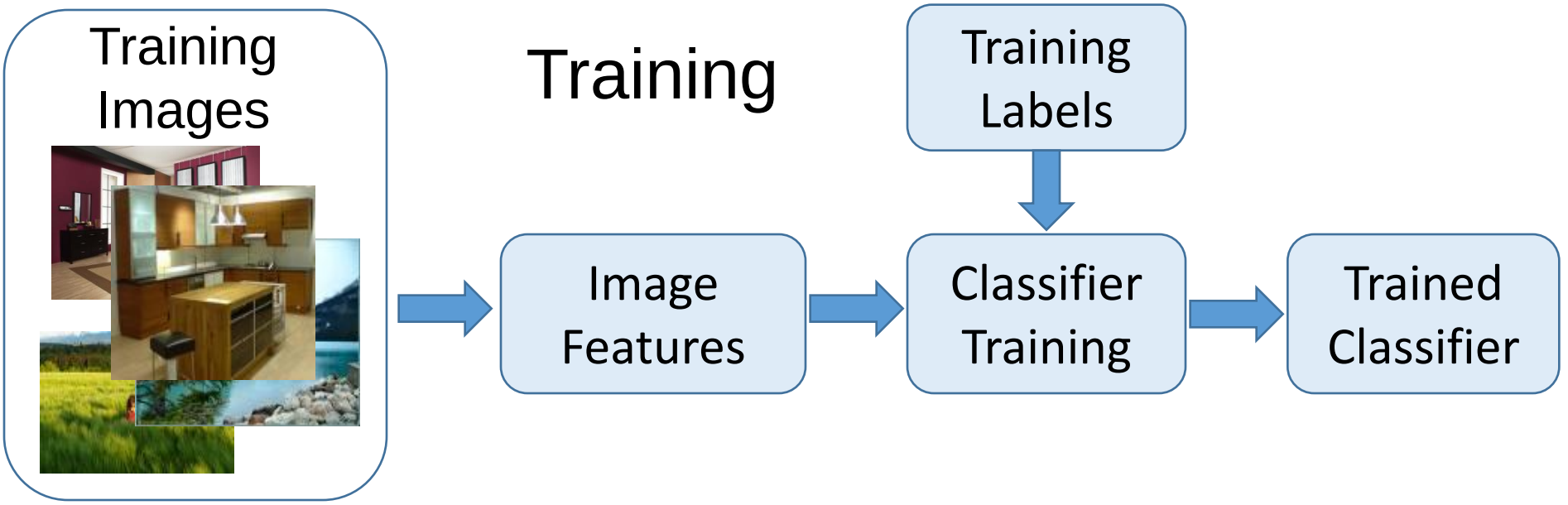
The statistical learning framework



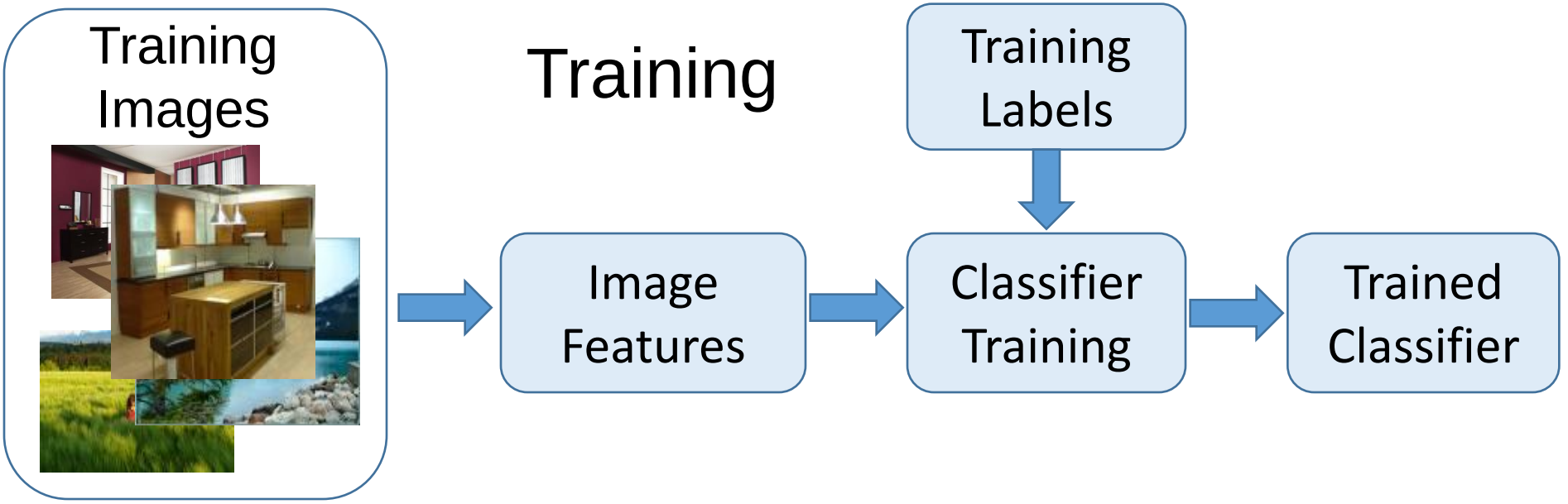
- **Training:** given a *training set* of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set

- **Testing:** apply f to a never before seen *test example* **x** and output the predicted value $y = f(\mathbf{x})$

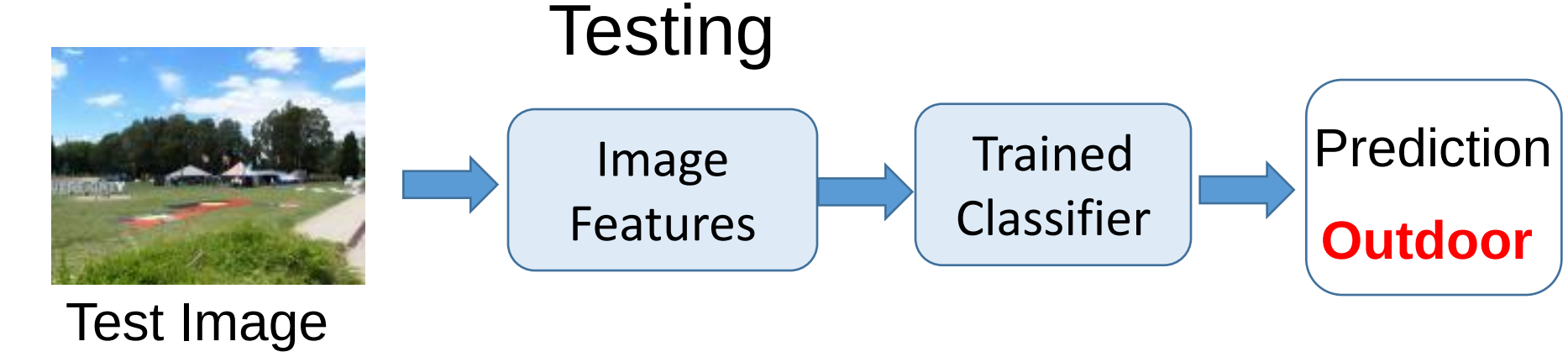
Training phase



Training phase

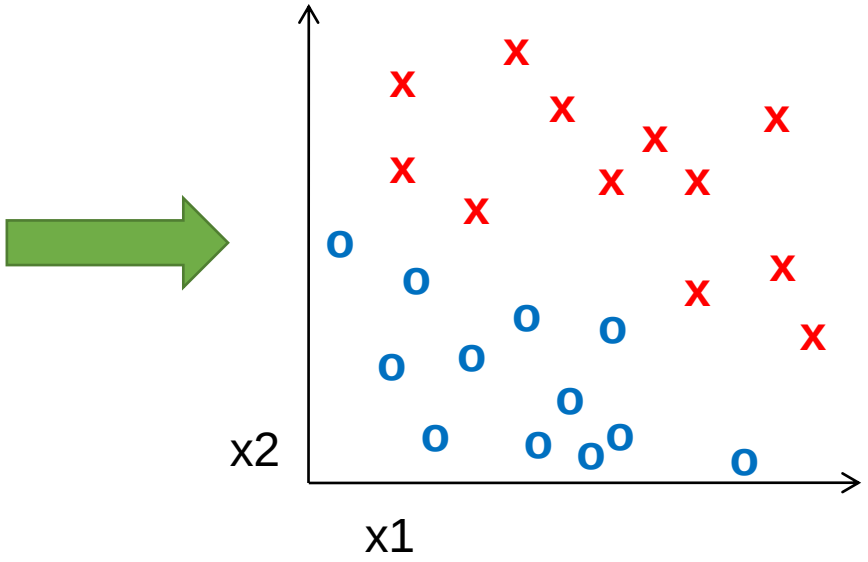


Testing phase

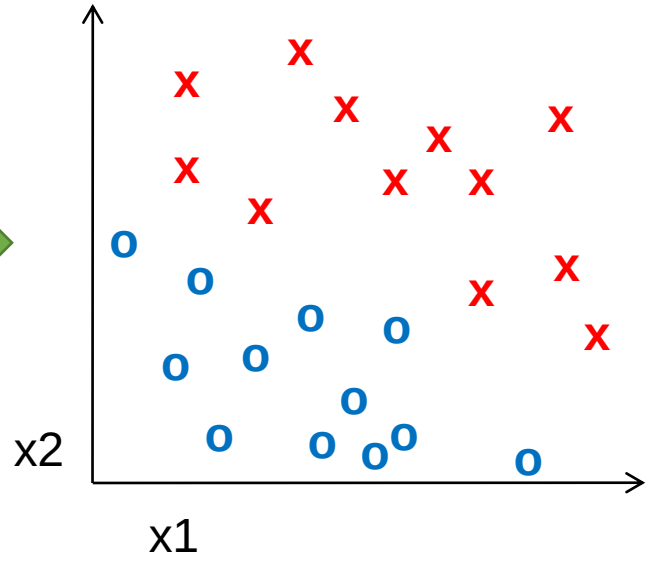


- **Image features**: map images to feature space

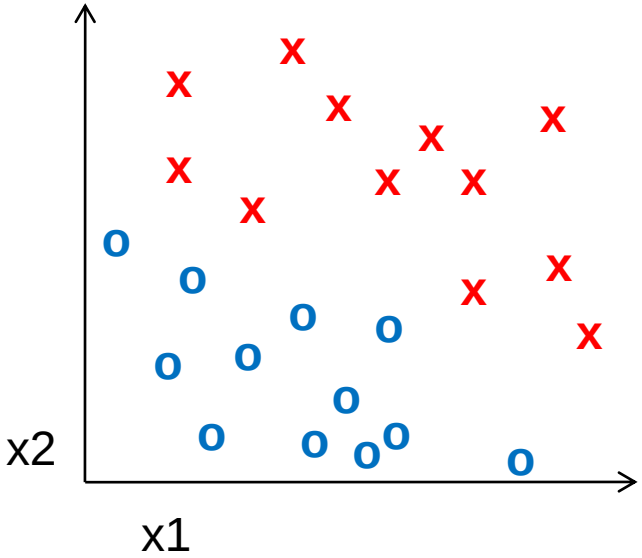
- **Image features**: map images to feature space



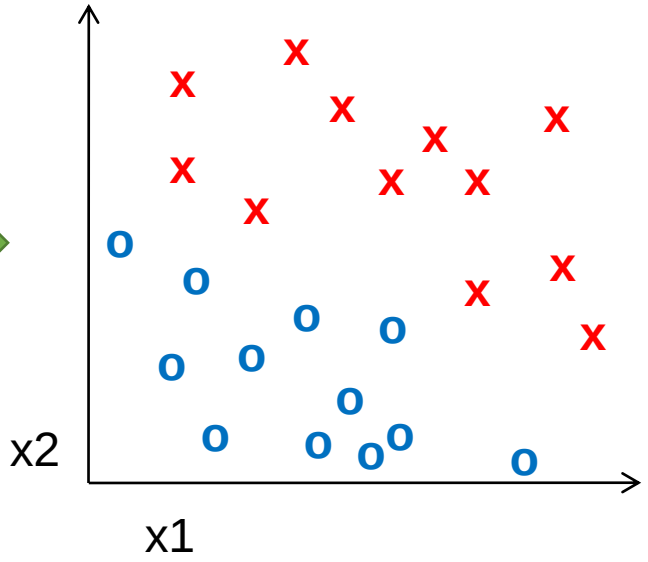
- **Image features**: map images to feature space



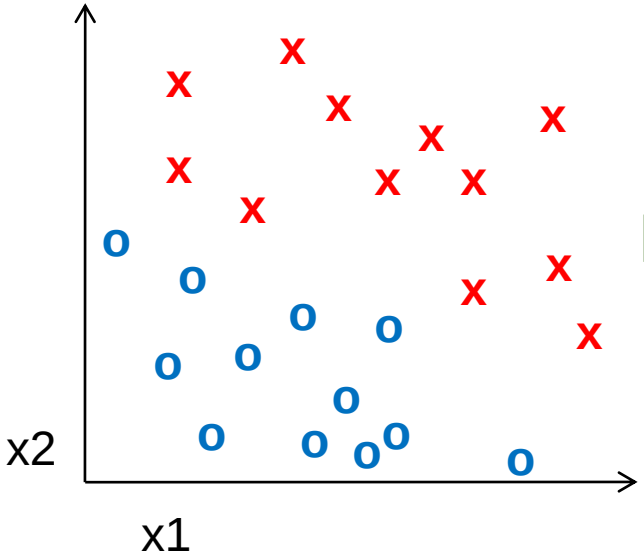
- **Classifiers**: map feature space to label space

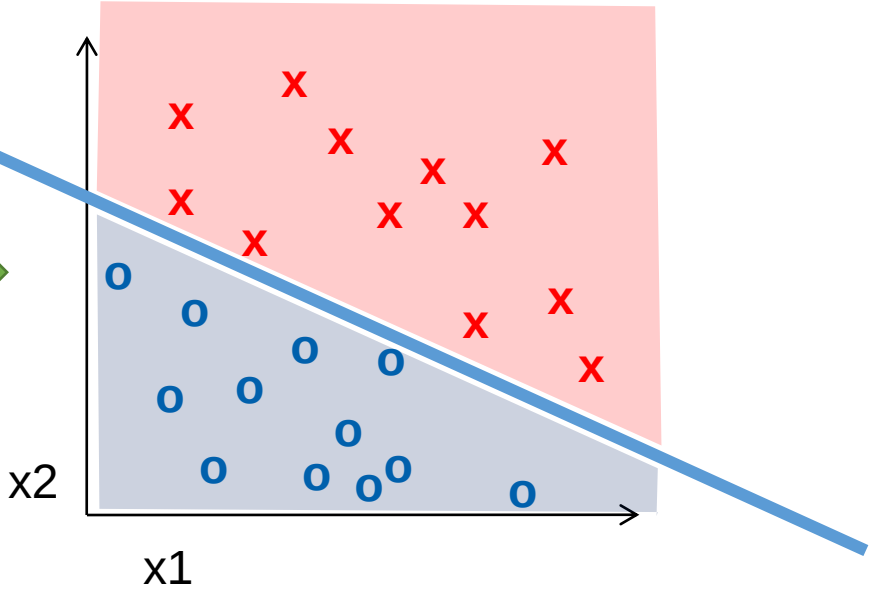


- **Image features**: map images to feature space

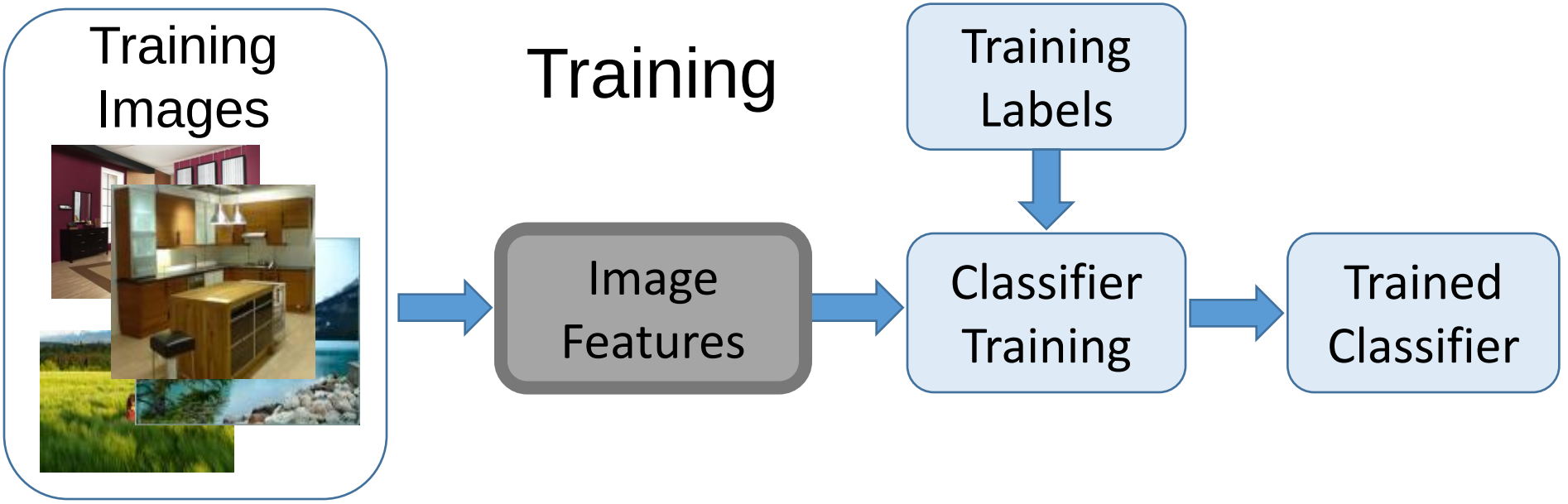


- **Classifiers**: map feature space to label space

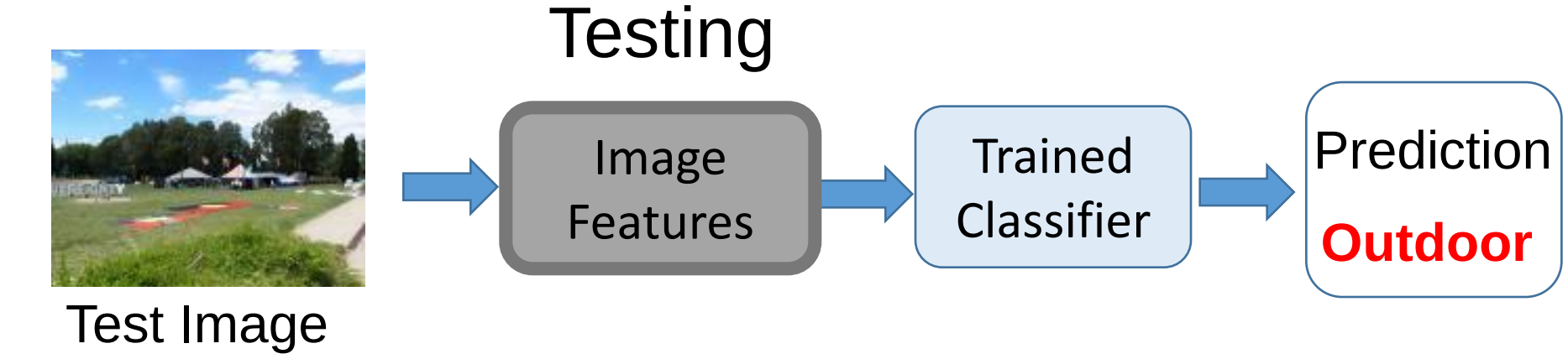




Training phase



Testing phase



Q: What are good features for…

- recognizing a beach?






Q: What are good features for…

- recognizing cloth fabric?






Q: What are good features for…

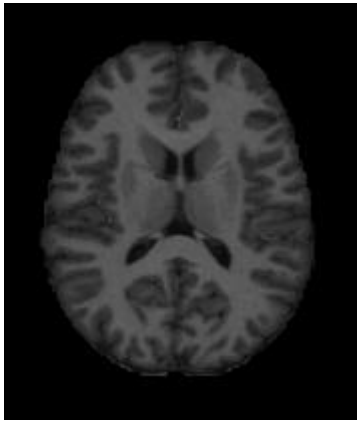
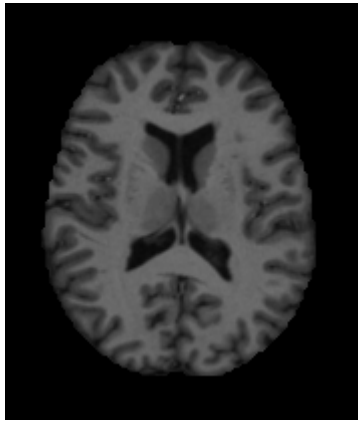
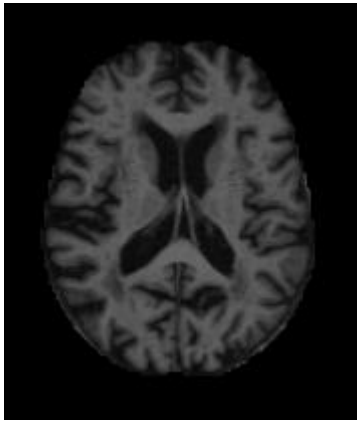
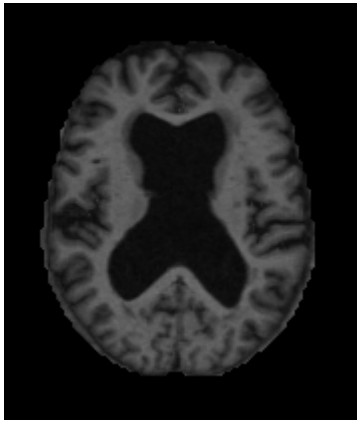
- recognizing a mug?





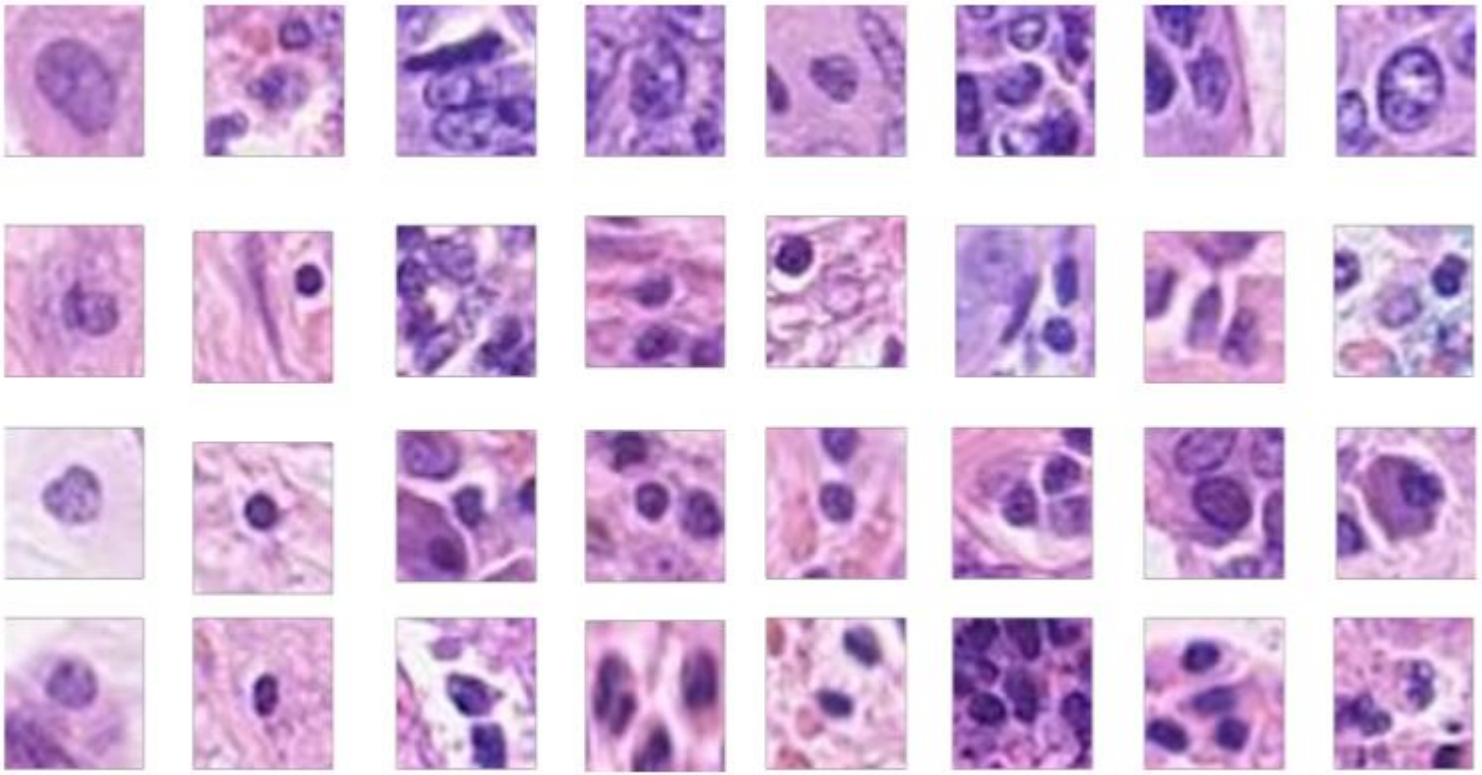

Q: What are good features for…

- recognizing the nodule in MRI data?

Q: What are good features for…

- recognizing the type of colon cancer?



"CRCHistoPhenotypes" dataset images

https://warwick.ac.uk/fac/sci/dcs/research/tia/data/crchistolabelednucleihe/

What are the right features?

What are the right features?

Depend on what you want to know!

What are the right features?

Depend on what you want to know!

- Object: shape
  - Local shape info, shading, shadows, texture

What are the right features?

Depend on what you want to know!

- Object: shape
  - Local shape info, shading, shadows, texture
- Scene: geometric layout
  - linear perspective, gradients, line segments

What are the right features?

Depend on what you want to know!

- Object: shape
  - Local shape info, shading, shadows, texture
- Scene: geometric layout
  - linear perspective, gradients, line segments
- Material properties: albedo, feel, hardness
  - Color, texture

What are the right features?

Depend on what you want to know!

- Object: shape
  - Local shape info, shading, shadows, texture
- Scene: geometric layout
  - linear perspective, gradients, line segments
- Material properties: albedo, feel, hardness
  - Color, texture
- Action: motion
  - Optical flow, tracked points

Image representations

- Templates
  - Intensity, gradients, etc.

Image Intensity

Gradient template

- Histograms
  - Color, texture, SIFT descriptors, etc.

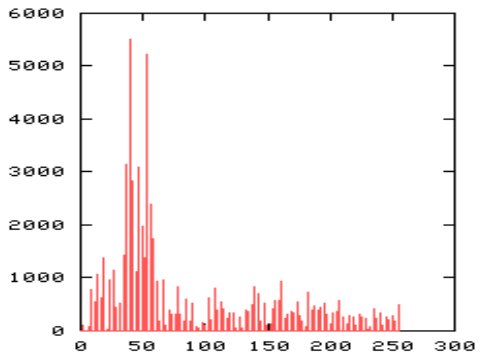
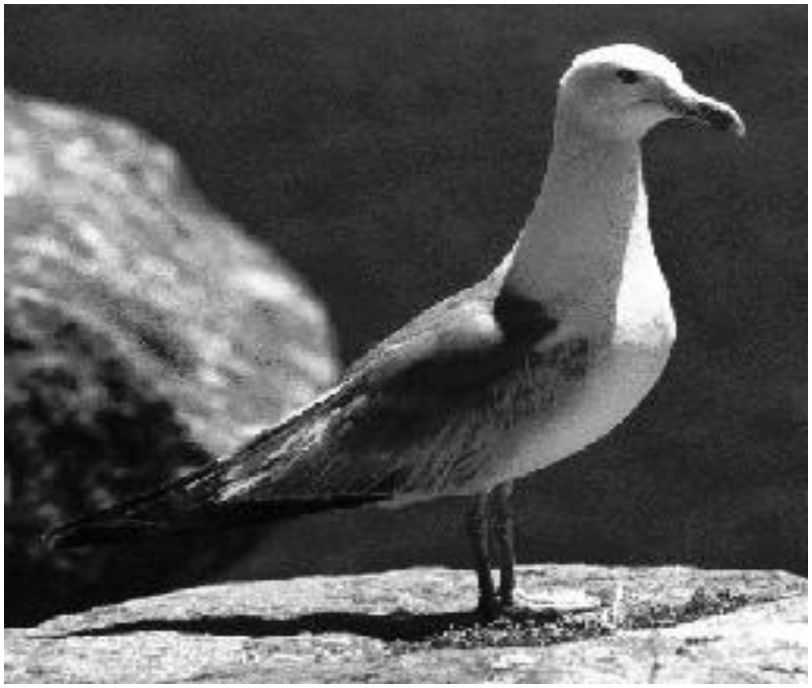
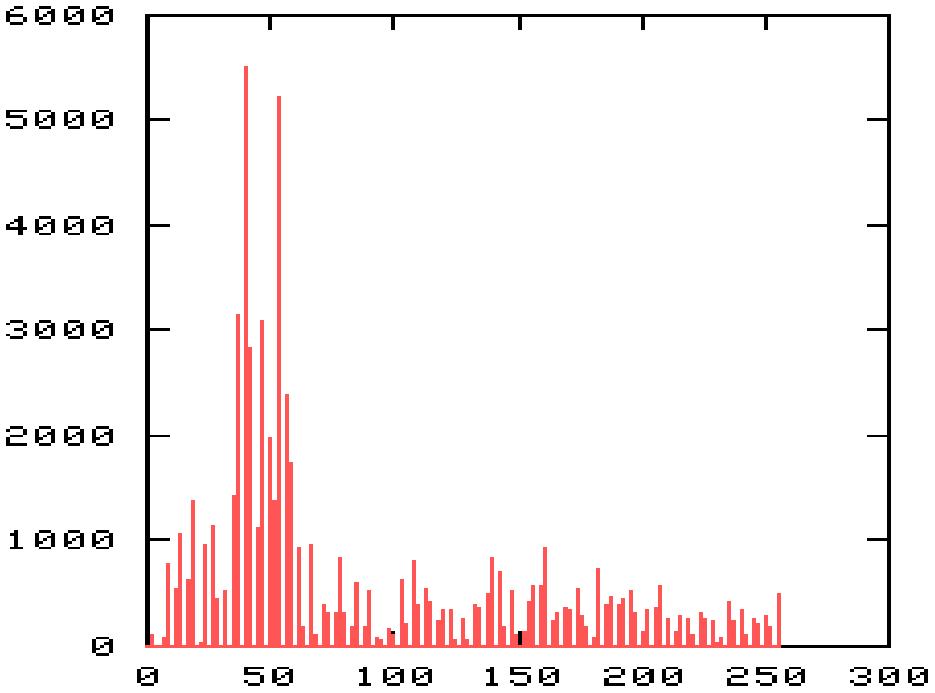


Image representations: histograms





Global histogram

- Represent distribution of features
  - Color, texture, depth, …

Images from Dave Kauchak

Computing histogram distance

Computing histogram distance

- Histogram intersection

$$\text{histint}(h_i, h_j) = 1 - \sum_{m=1}^{K} \min\left(h_i(m), h_j(m)\right)$$

- Chi-squared Histogram matching distance

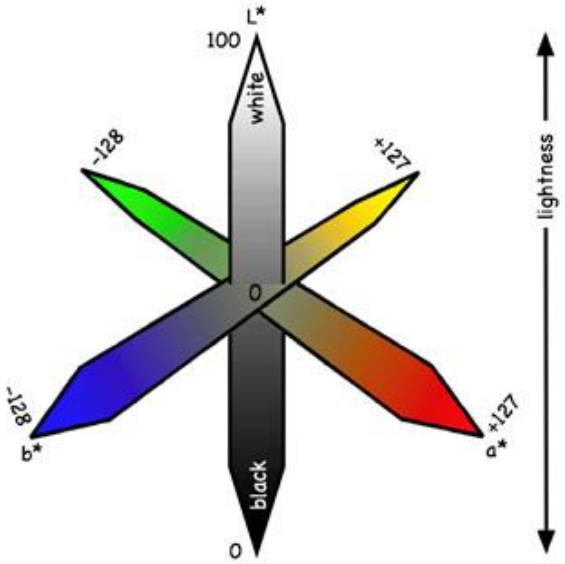
$$\chi^2(h_i, h_j) = \frac{1}{2} \sum_{m=1}^{K} \frac{[h_i(m) - h_j(m)]^2}{h_i(m) + h_j(m)}$$

- Earth mover's distance
  (Cross-bin similarity measure)
  - minimal cost paid to transform one distribution into the other

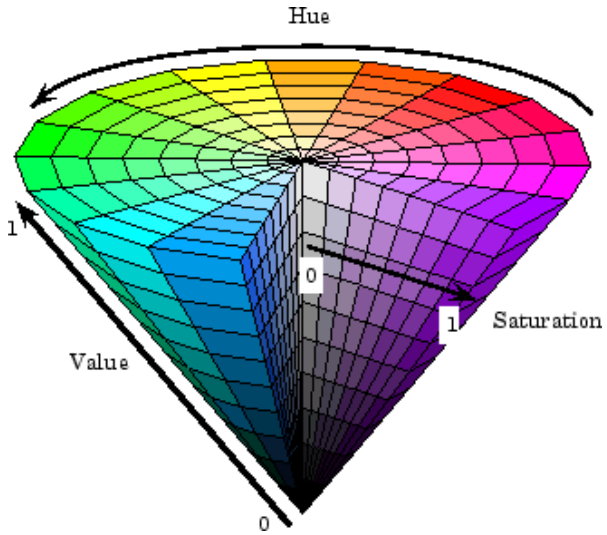
[Rubner et al. The Earth Mover's Distance as a Metric for Image Retrieval, IJCV 2000]

What kind of things do we compute histograms of?

- Color

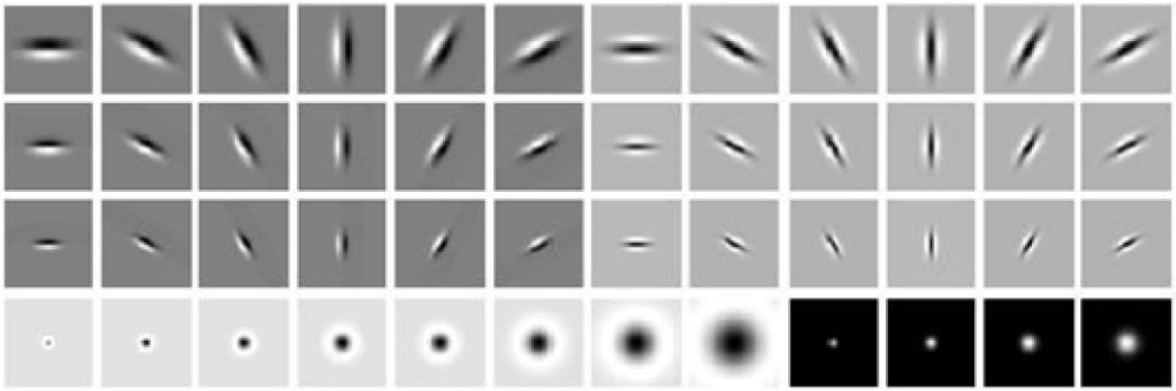


L*a*b* color space



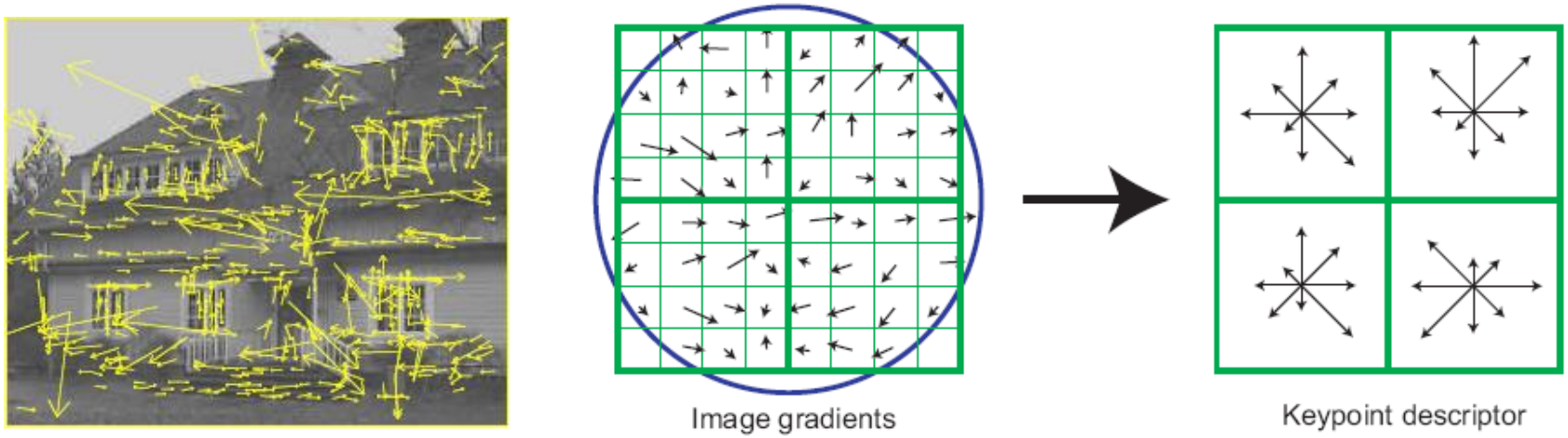
HSV color space

- Texture (filter banks or HOG over regions)



What kind of things do we compute histograms of?

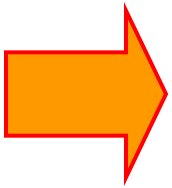
Histograms of descriptors



SIFT – [Lowe IJCV 2004]

What kind of things do we compute histograms of?

Bags of visual words



Analogy to documents

Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach the brain from our eyes. For a long time it was thought that the retinal image was transmitted point by point to visual centers in the brain; the cerebral cortex was a movie screen, so to speak, upon which the image in the eye was projected. Through the discoveries of Hubel and Wiesel we now know that behind the origin of the visual perception in the brain there is a considerably more complicated course of events. By following the visual impulses along their path to the various cell layers of the optical cortex, Hubel and Wiesel have been able to demonstrate that the *message about the image falling on the retina undergoes a step-wise analysis in a system of nerve cells stored in columns. In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.*

China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be created by a predicted 30% jump in exports to $750bn, compared with a 18% rise in imports to $660bn. The figures are likely to further annoy the US, which has long argued that China's exports are unfairly helped by a deliberately undervalued yuan. Beijing agrees the surplus is too high, but says the yuan is only one factor. Bank of China governor Zhou Xiaochuan said the country also needed to do more to boost domestic demand so more goods stayed within the country. China increased the value of the yuan against the dollar by 2.1% in July and permitted it to trade within a narrow band, but the US wants the yuan to be allowed to trade freely. However, Beijing has made it clear that it will take its time and tread carefully before allowing the yuan to rise further in value.

Analogy to documents

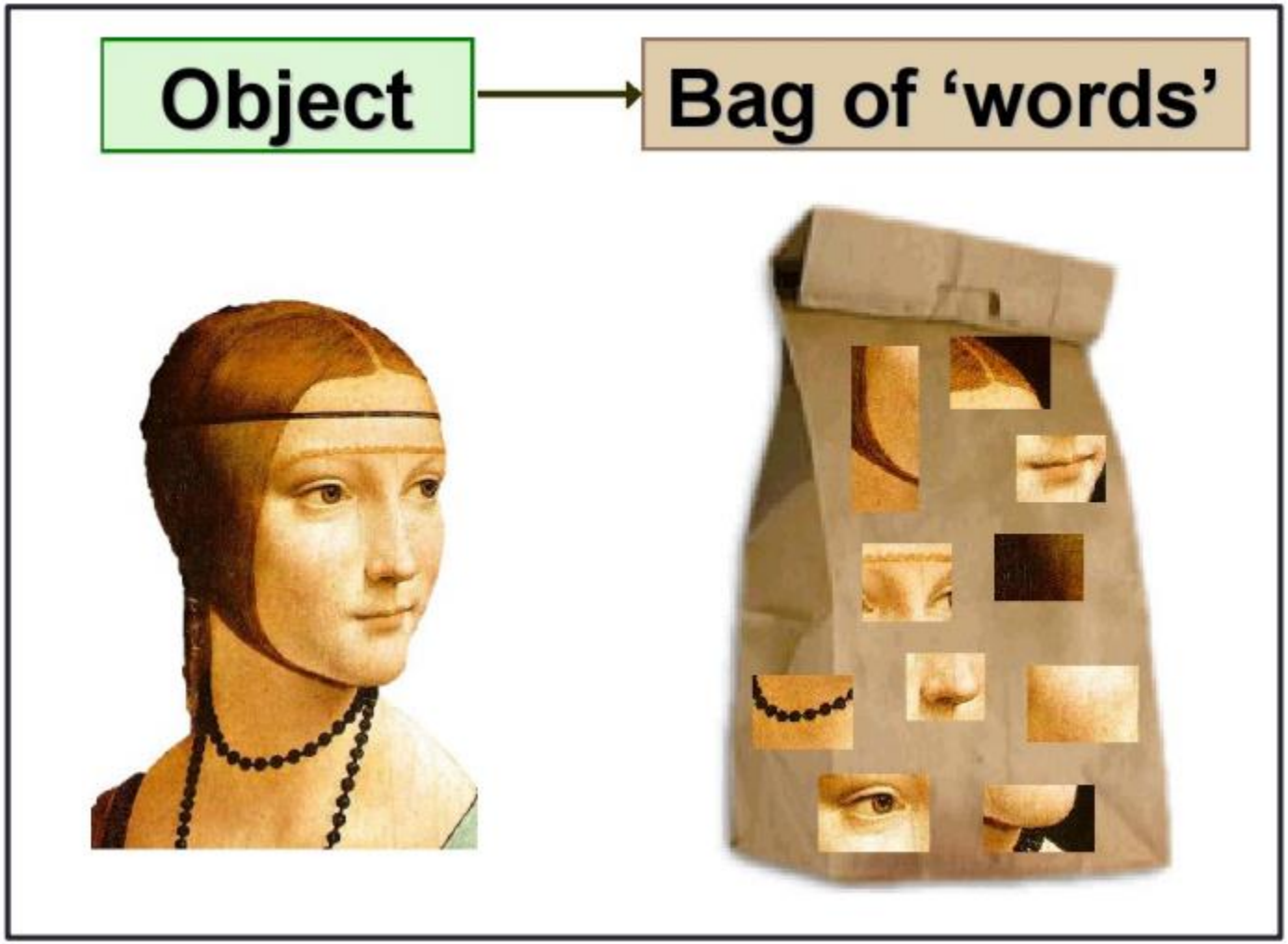
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that r… our eyes. For a long ti… retinal image wa… isual centers i… s a movie s… image … discove… know th… perceptio… more com… following the … ath to the various c… ortex, Hubel and Wiesel ha… demonstrate that the *message abou… image falling on the retina undergoes … wise analysis in a system of nerve cell… stored in columns. In this system each c… has its specific function and is responsible… a specific detail in the pattern of the retinal image.*

sensory, brain, visual, perception, retinal, cerebral cortex, eye, cell, optical nerve, image Hubel, Wiesel

China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be created by a predicted 30% … $750bn, compared wi… $660bn. T… annoy th… China's … deliber… agrees … yuan is … governo… also need… demand so … country. China … e yuan against the dol… nd permitted it to trade within a narrow … but the US wants the yuan to be allowed … de freely. However, Beijing has made it cl… t it will take its time and tread carefully be… allowing the yuan to rise further in value.

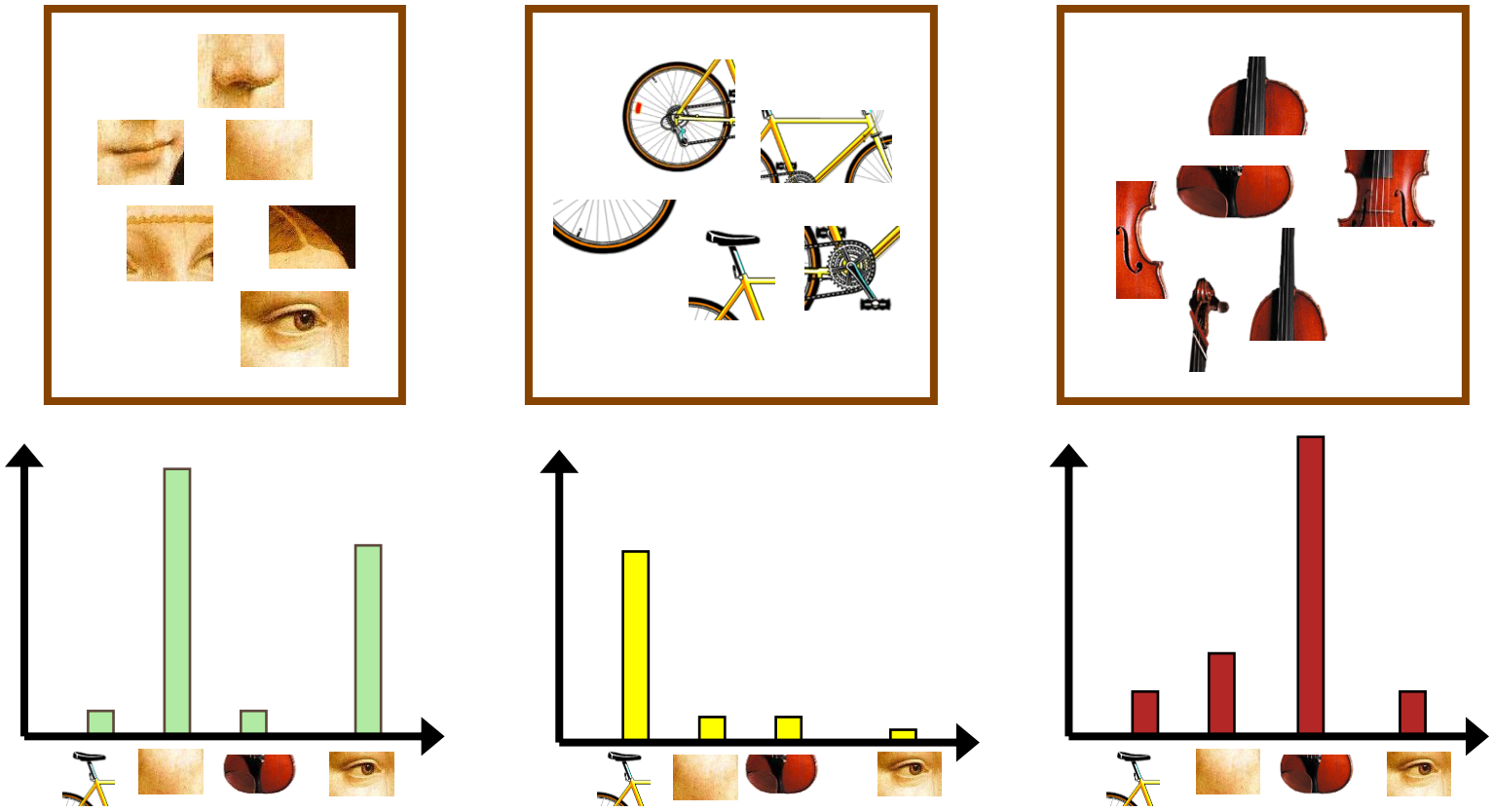
China, trade, surplus, commerce, exports, imports, US, yuan, bank, domestic, foreign, increase, trade, value

Bags of visual words: Motivation



Bags-of-visual-words

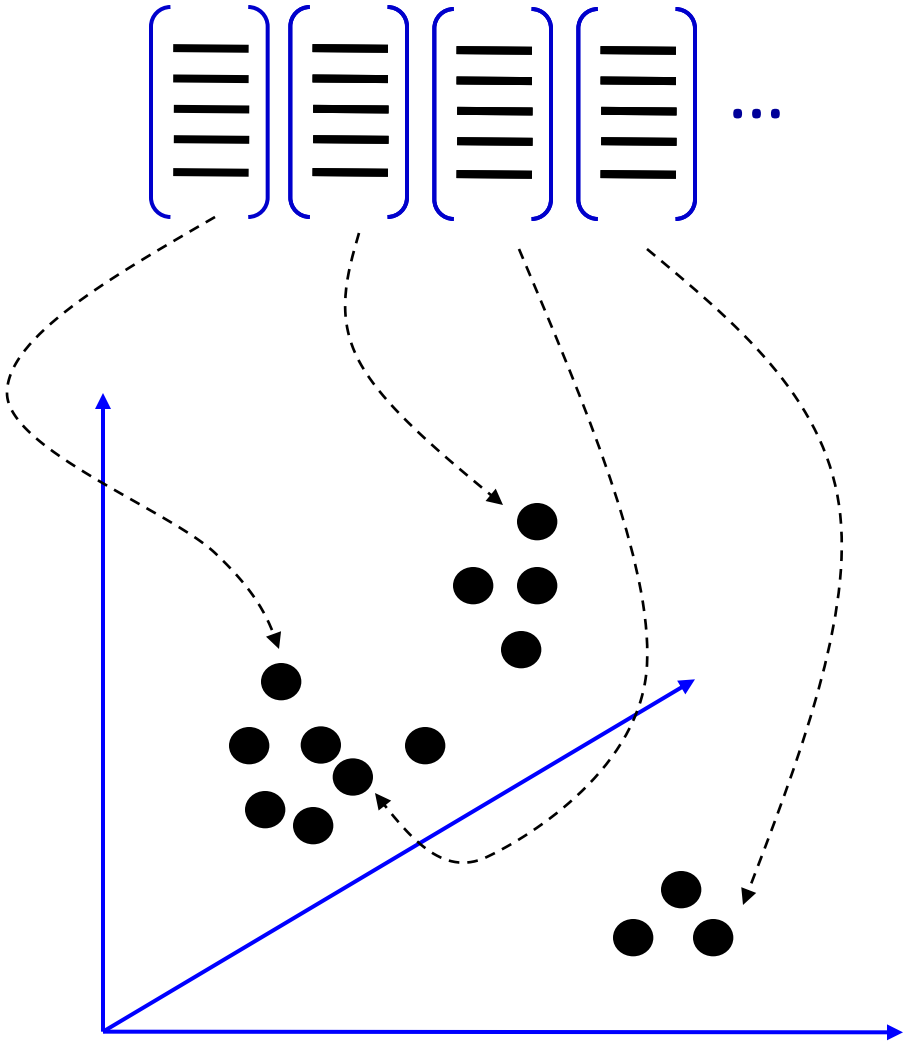
1. Extract local features
2. Learn “visual vocabulary”
3. Quantize local features using visual vocabulary
4. Represent images by frequencies of “visual words”



1. Local feature extraction

- Sample patches and extract descriptors

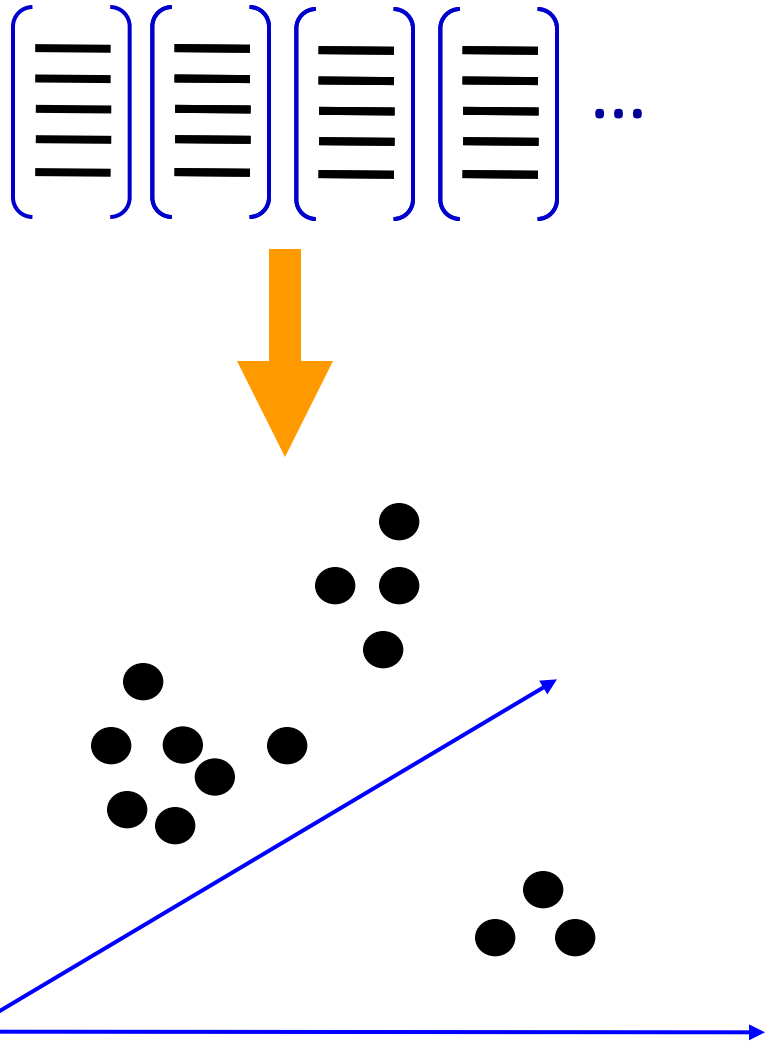
2. Learning the visual vocabulary



Extracted descriptors from the training set

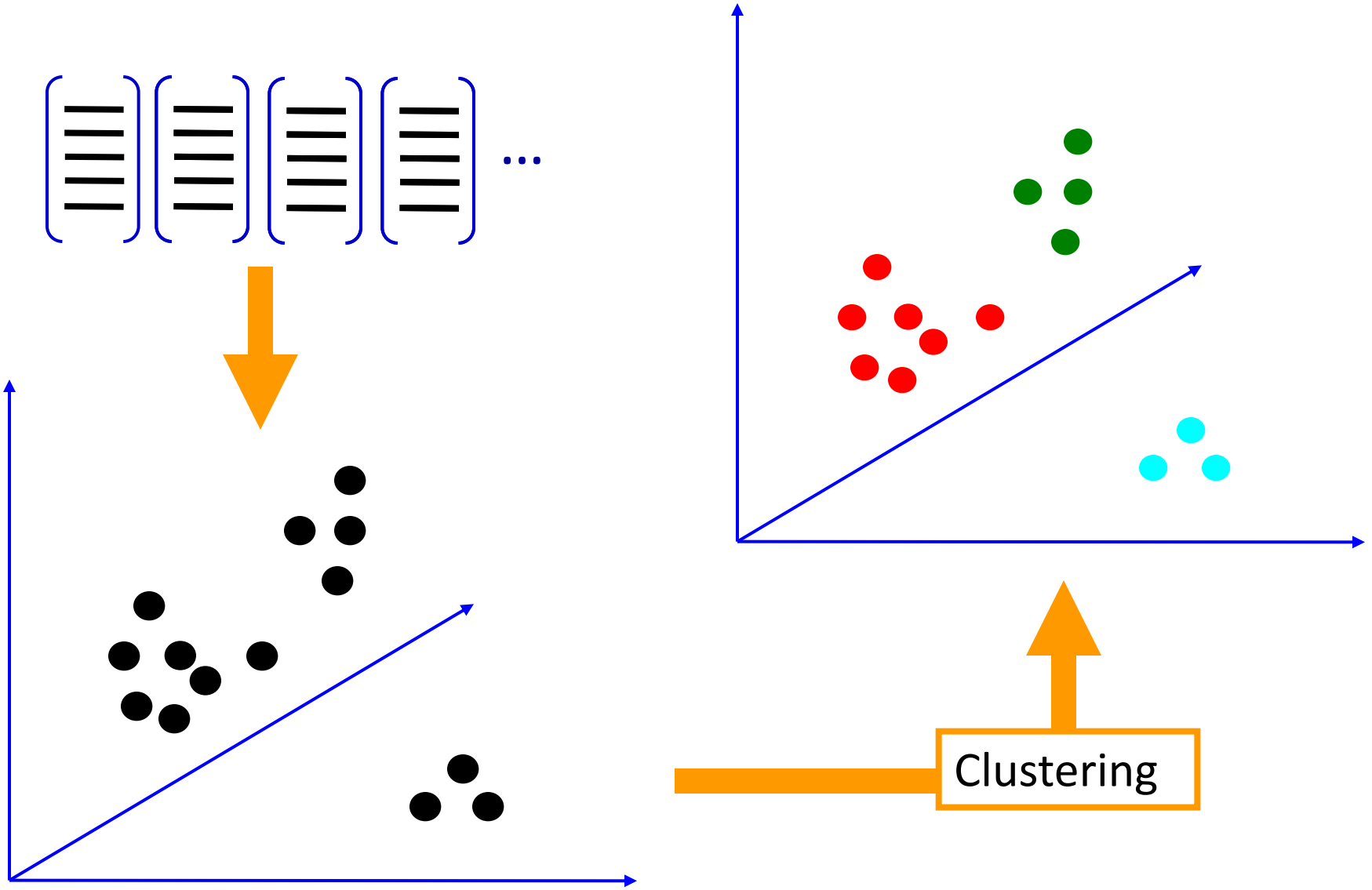
Slide credit: Josef Sivic

2. Learning the visual vocabulary



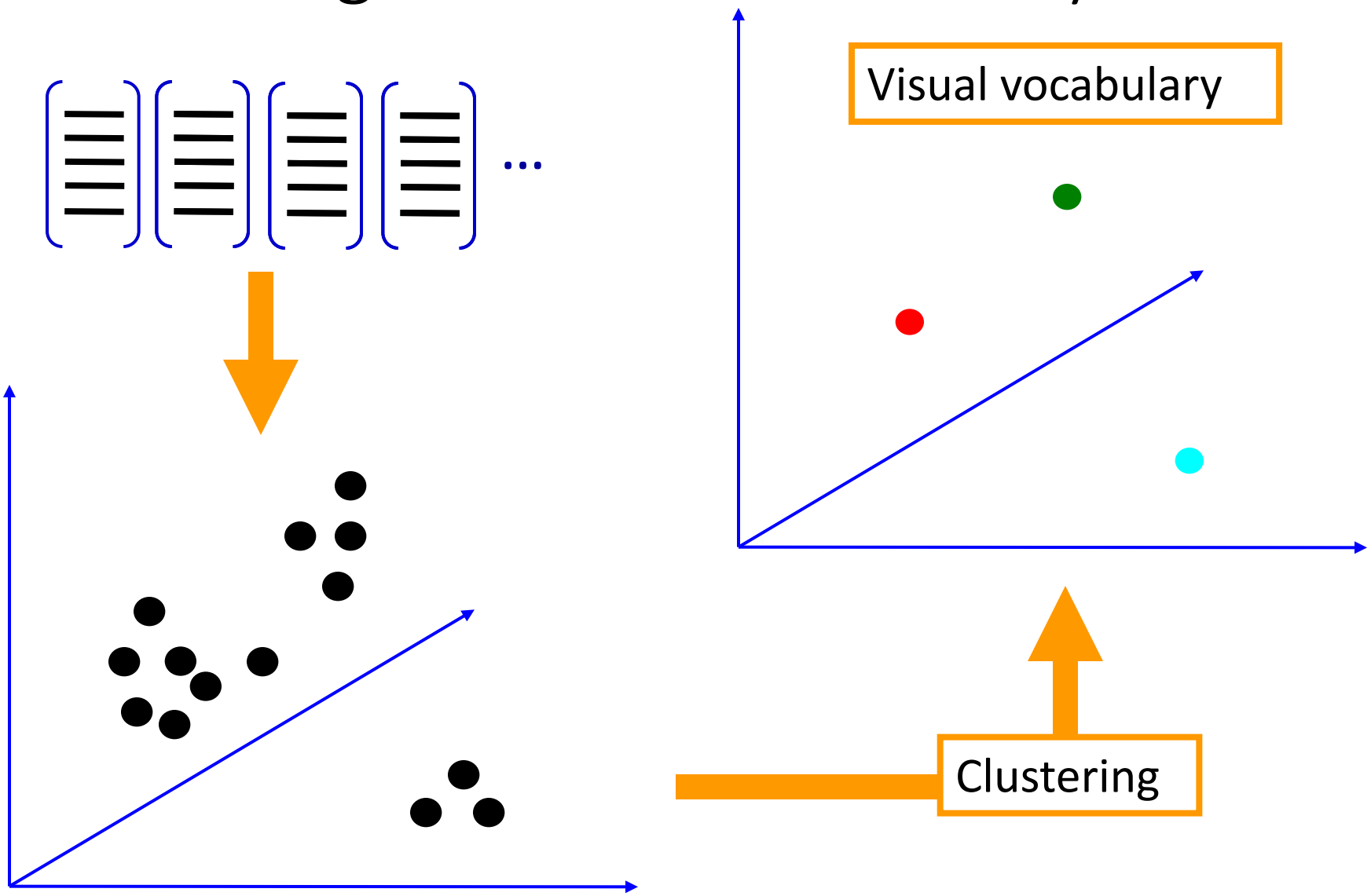
Slide credit: Josef Sivic

2. Learning the visual vocabulary



Slide credit: Josef Sivic

2. Learning the visual vocabulary



Slide credit: Josef Sivic

Review: K-means clustering

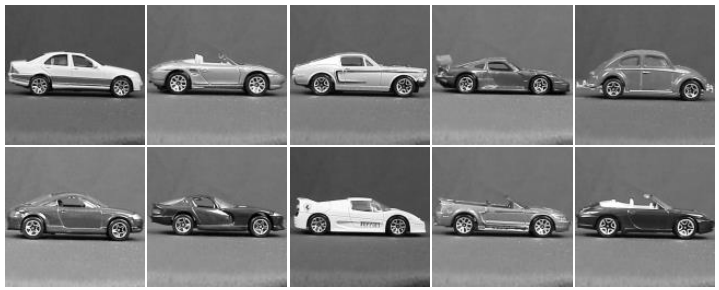
- Want to minimize sum of squared Euclidean distances between features $\mathbf{x}_i$ and their nearest cluster centers $\mathbf{m}_k$

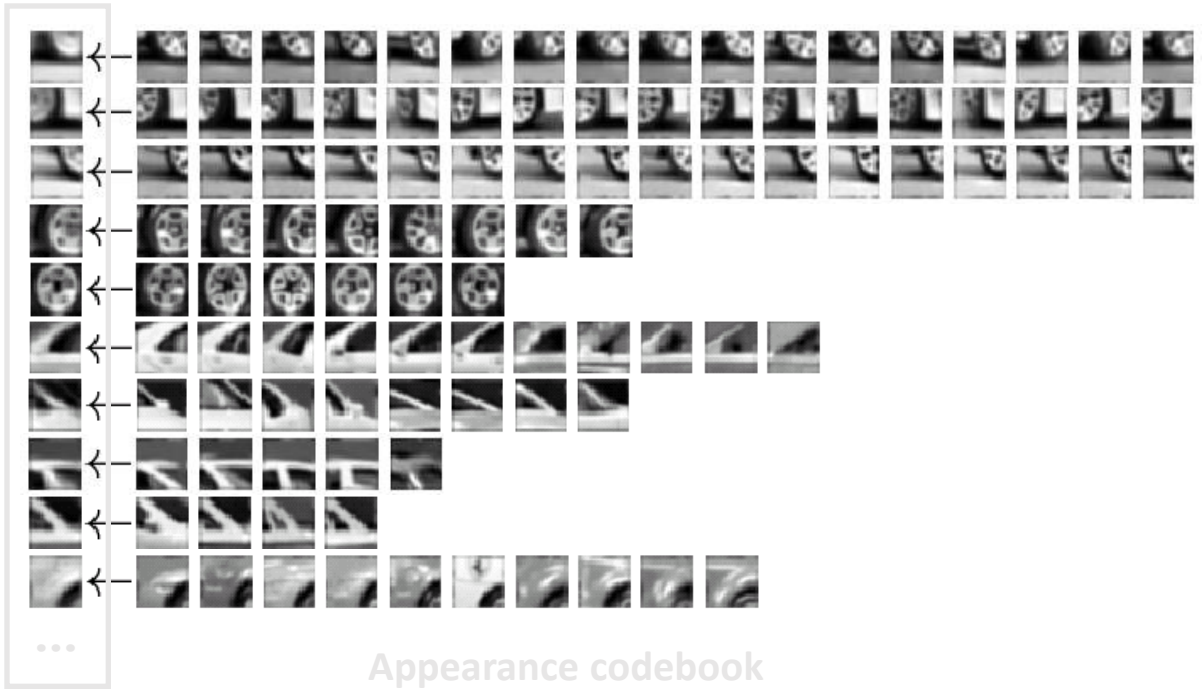
$$D(X, M) = \sum_{\text{cluster } k} \sum_{\substack{\text{point } i \text{ in} \\ \text{cluster } k}} (\mathbf{x}_i - \mathbf{m}_k)^2$$

Algorithm:

- Randomly initialize K cluster centers
- Iterate until convergence:
  - Assign each feature to the nearest center
  - Recompute each cluster center as the mean of all features assigned to it

Example visual vocabulary

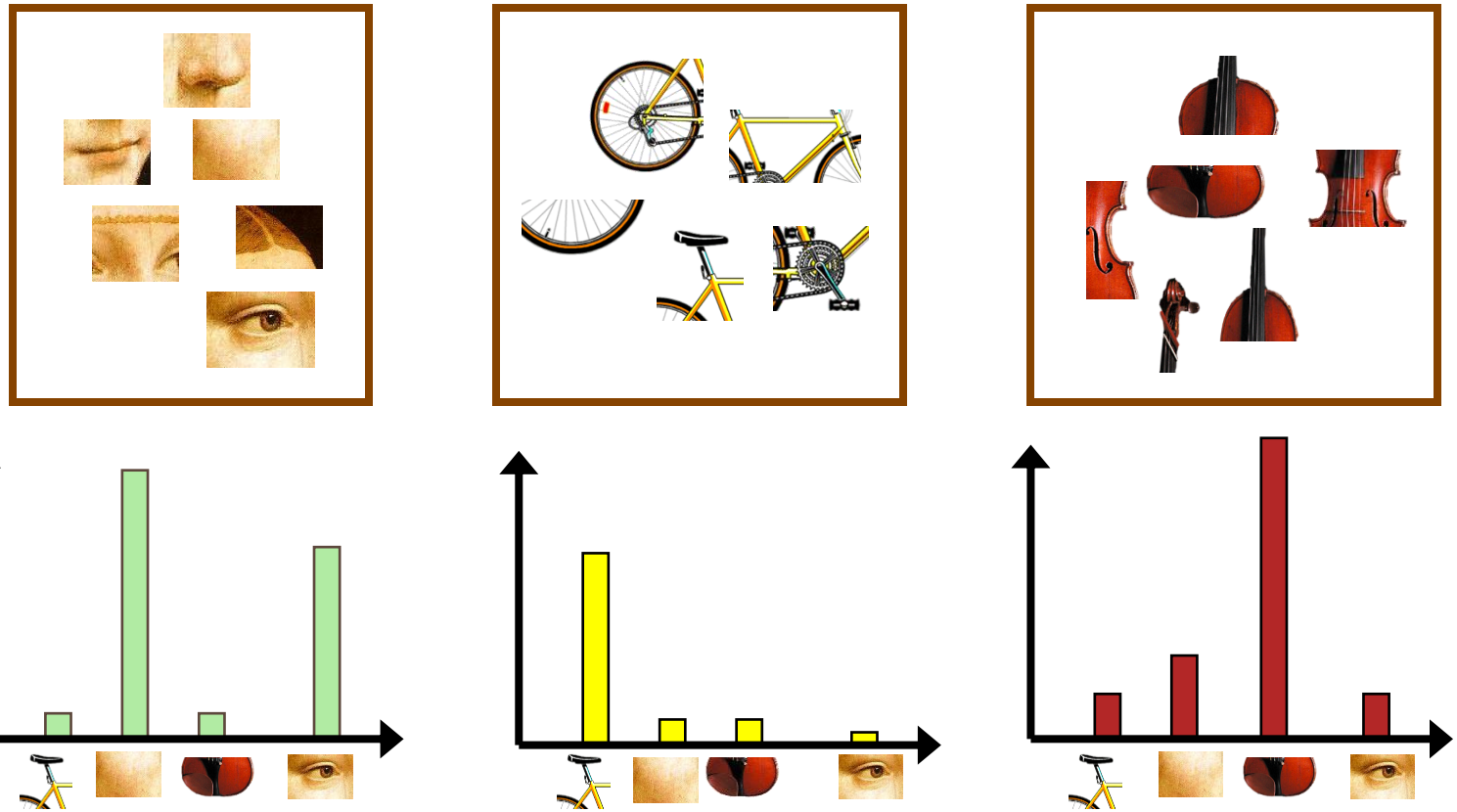




Source: B. Leibe

Bag-of-features steps

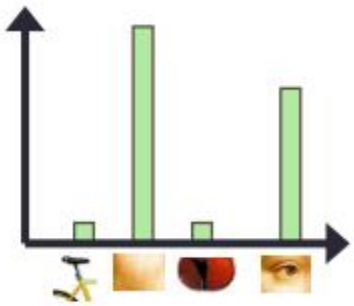
1. Extract local features
2. Learn “visual vocabulary”
3. **Quantize local features using visual vocabulary**
4. **Represent images by frequencies of “visual words”**



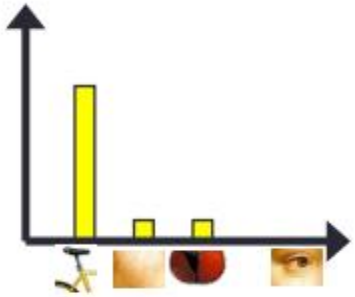
Comparing bags of words

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

[1 8 1 4]



[5 1 1 0]



$\vec{d}_j$

$\vec{q}$

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

$$= \frac{\sum_{i=1}^{V} d_j(i) * q(i)}{\sqrt{\sum_{i=1}^{V} d_j(i)^2} * \sqrt{\sum_{i=1}^{V} q(i)^2}}$$

for vocabulary of V words

Kristen Grauman

Image categorization with bag of words

Training

1. Extract keypoints and descriptors for all training images
2. Cluster descriptors
3. Quantize descriptors using cluster centers to get "visual words"
4. Represent each image by normalized counts of "visual words"
5. Train classifier on labeled examples using histogram values as features

Image categorization with bag of words

Training

1. Extract keypoints and descriptors for all training images
2. Cluster descriptors
3. Quantize descriptors using cluster centers to get “visual words”
4. Represent each image by normalized counts of “visual words”
5. Train classifier on labeled examples using histogram values as features

Testing

1. Extract keypoints/descriptors and quantize into visual words
2. Compute visual word histogram
3. Compute label or confidence using classifier

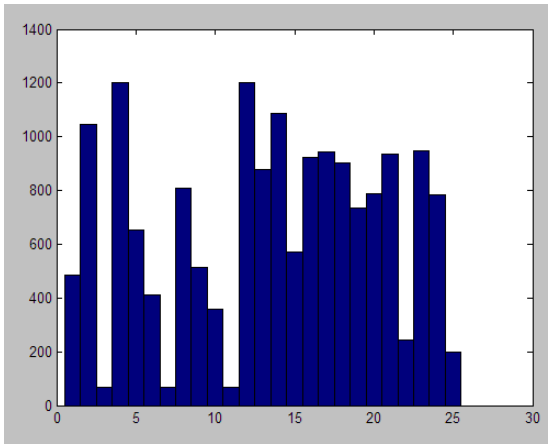
Object classification with bag of words

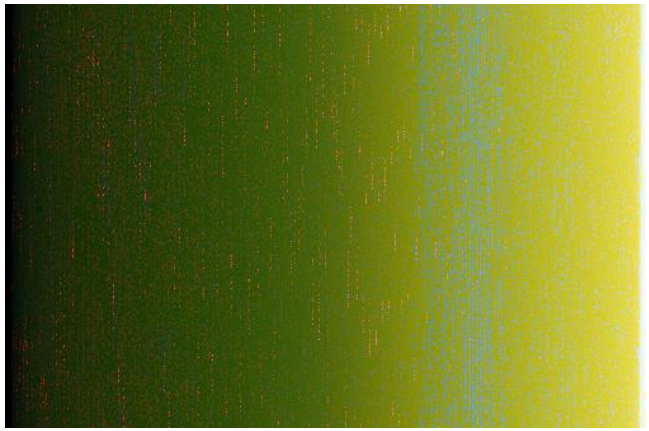
- Performance on Caltech 101 dataset with linear SVM on bag-of-word vectors:

True classes →	*faces (frontal)*	*airplanes (side)*	*cars (rear)*	*cars (side)*	*motorbikes (side)*
faces(frontal)	**94**	0.4	0.7	0	1.4
airplanes (side)	1.5	**96.3**	0.2	0.1	2.7
cars (rear)	1.9	0.5	**97.7**	0	0.9
cars(side)	1.7	1.9	0.5	**99.6**	2.3
motorbikes (side)	0.9	0.9	0.9	0.3	**92.7**

[Csurka et al., '04]

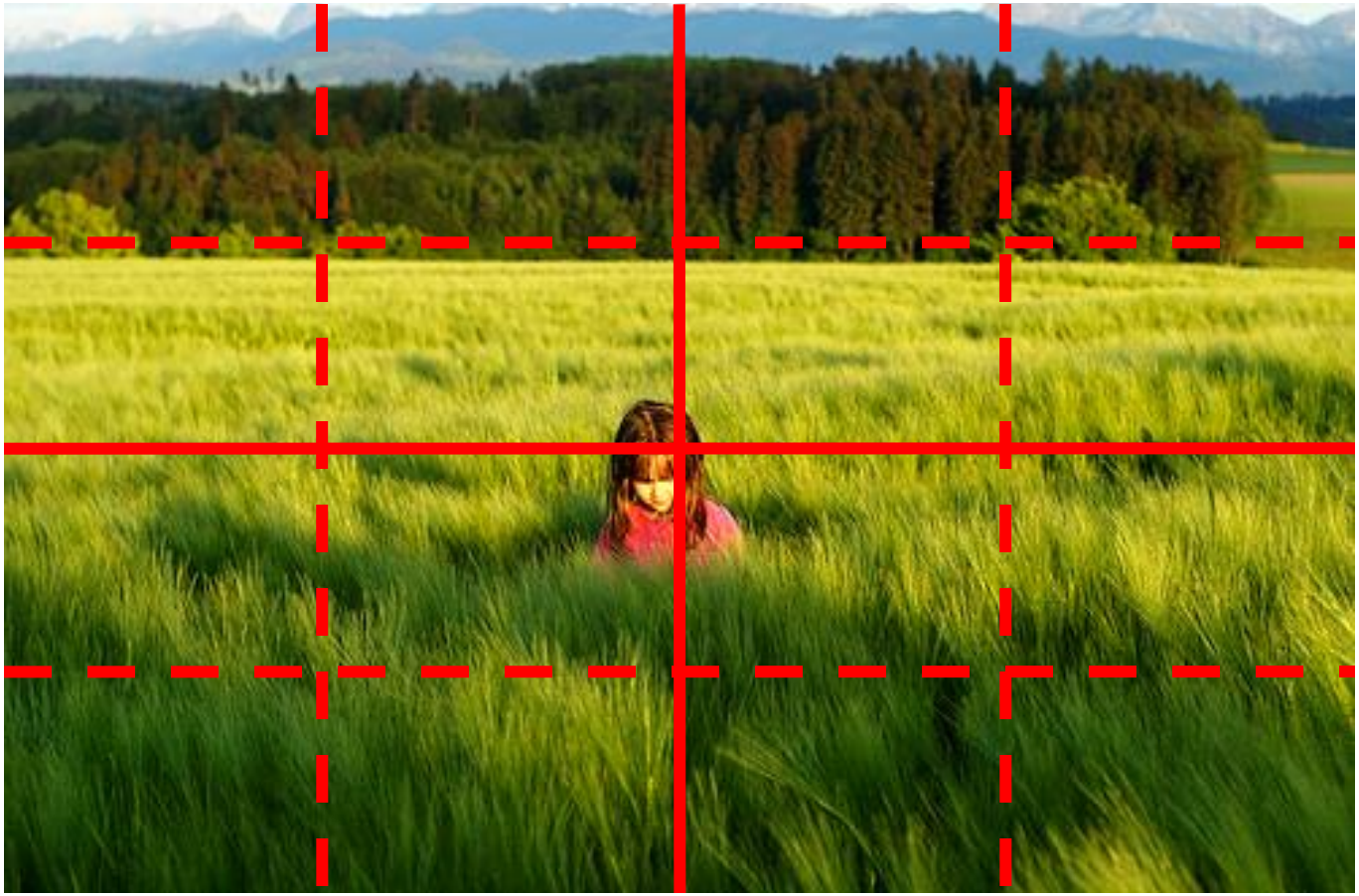
But what about spatial layout?





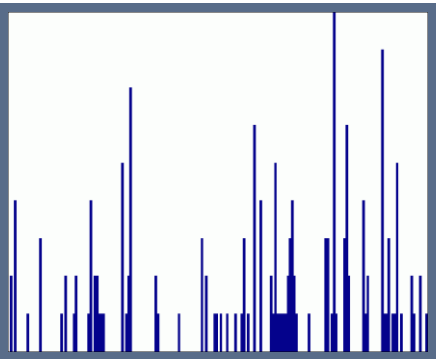
All of these images have the same color histogram

Spatial pyramid



Compute histogram in each spatial bin

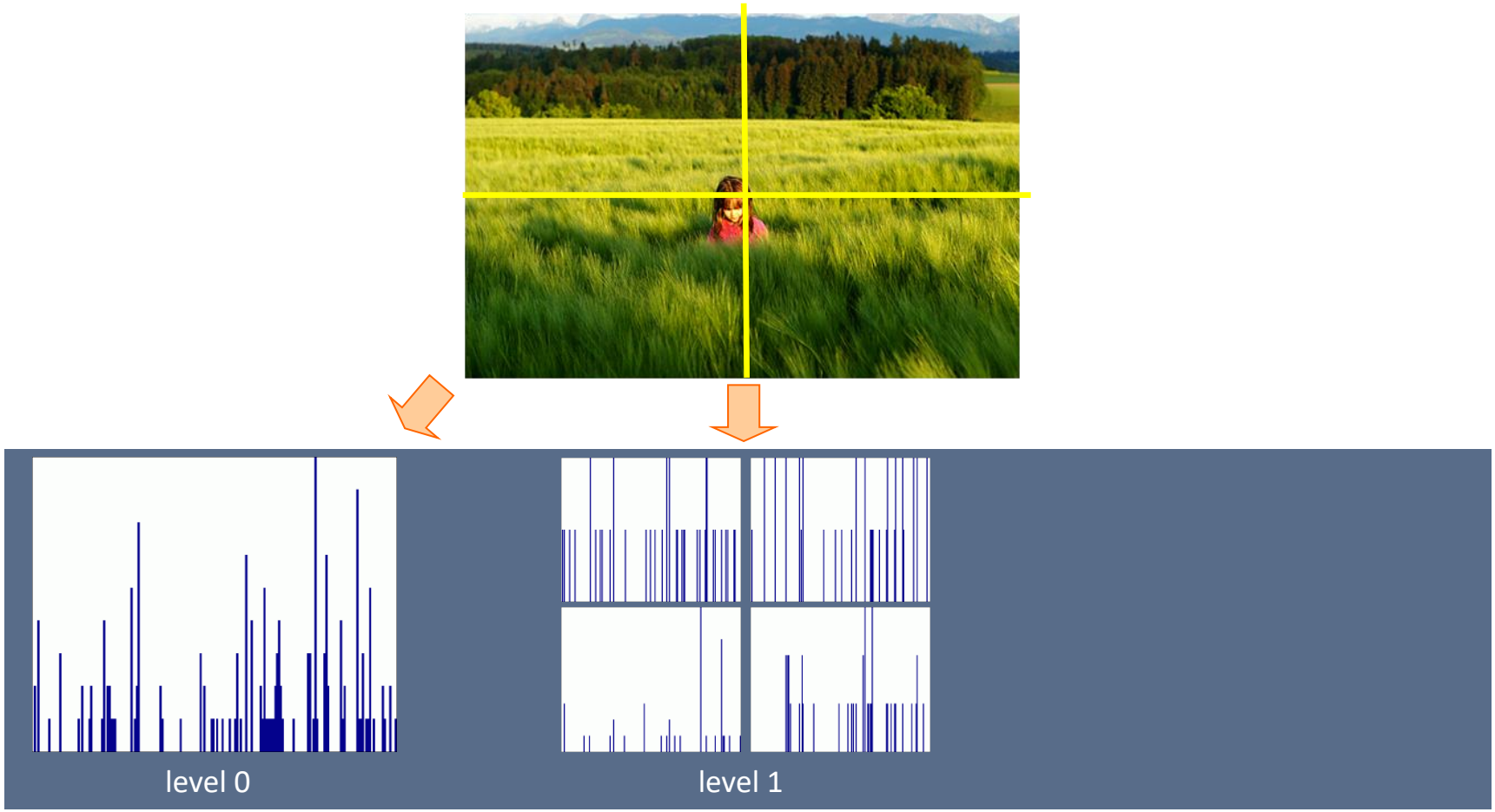
Spatial pyramids



level 0

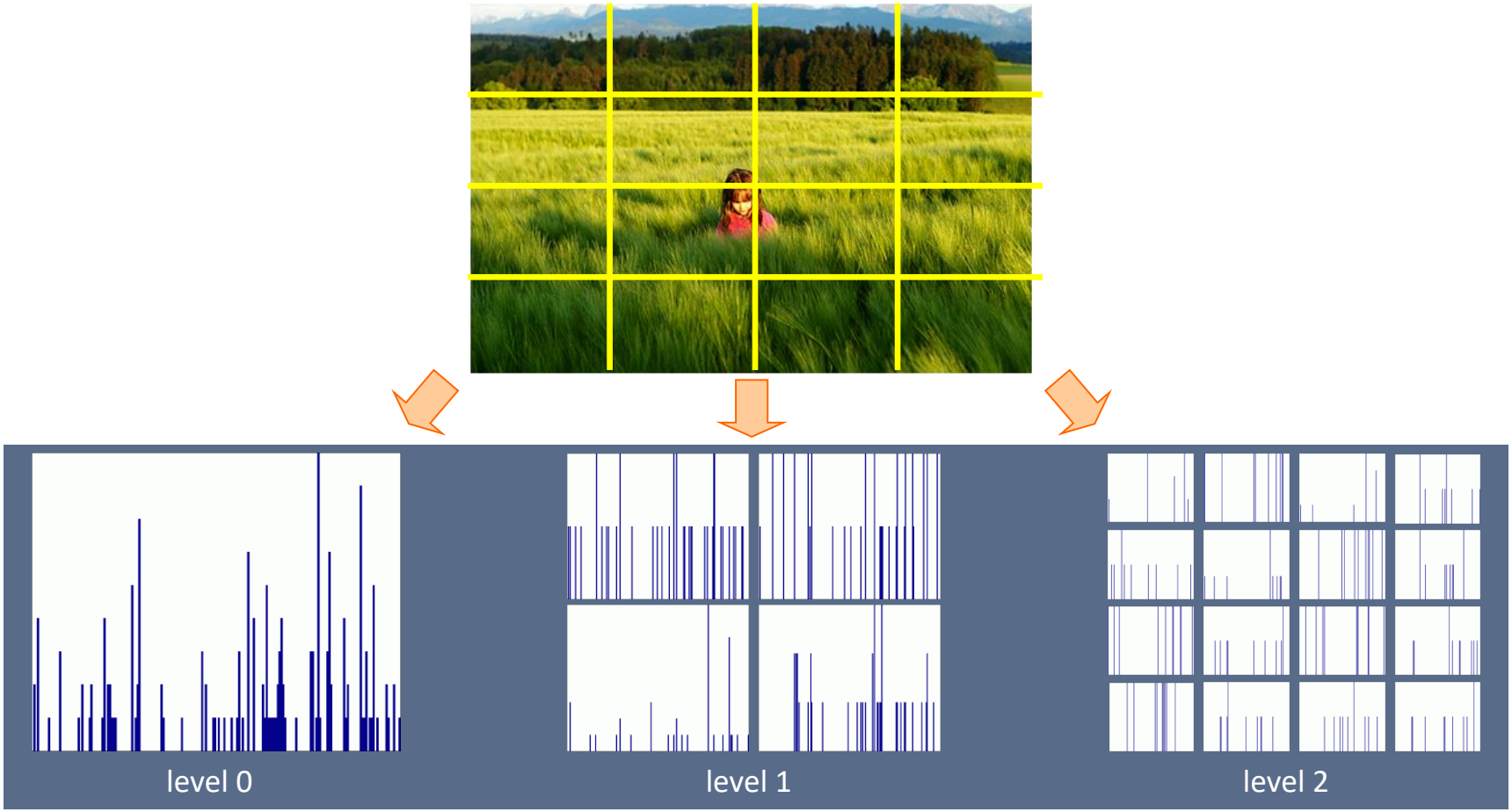
Lazebnik, Schmid & Ponce (CVPR 2006) –
Beyond bags of features: spatial pyramid matching for recognizing natural scene categories

Spatial pyramids



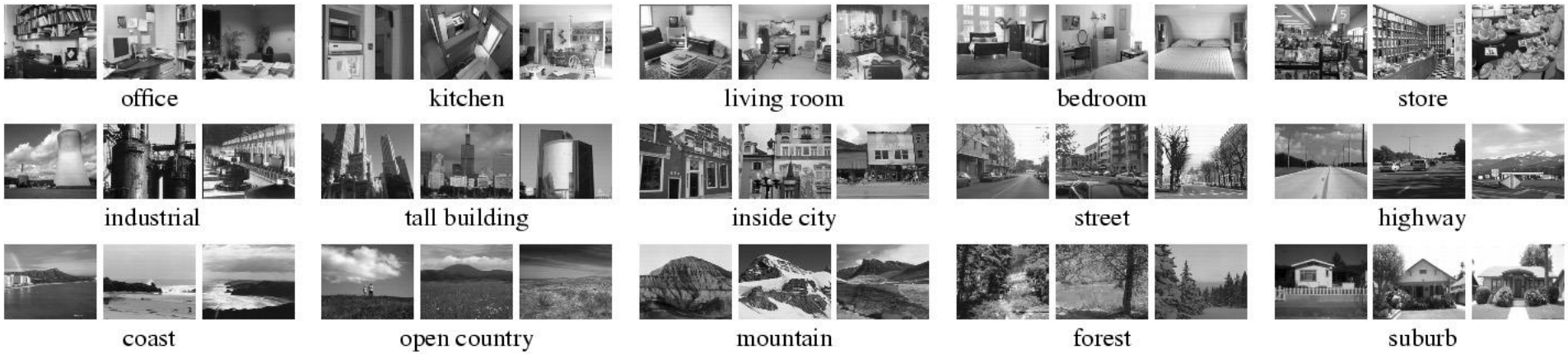
Lazebnik, Schmid & Ponce (CVPR 2006) –
Beyond bags of features: spatial pyramid matching for recognizing natural scene categories

Spatial pyramids



Lazebnik, Schmid & Ponce (CVPR 2006) –
Beyond bags of features: spatial pyramid matching for recognizing natural scene categories

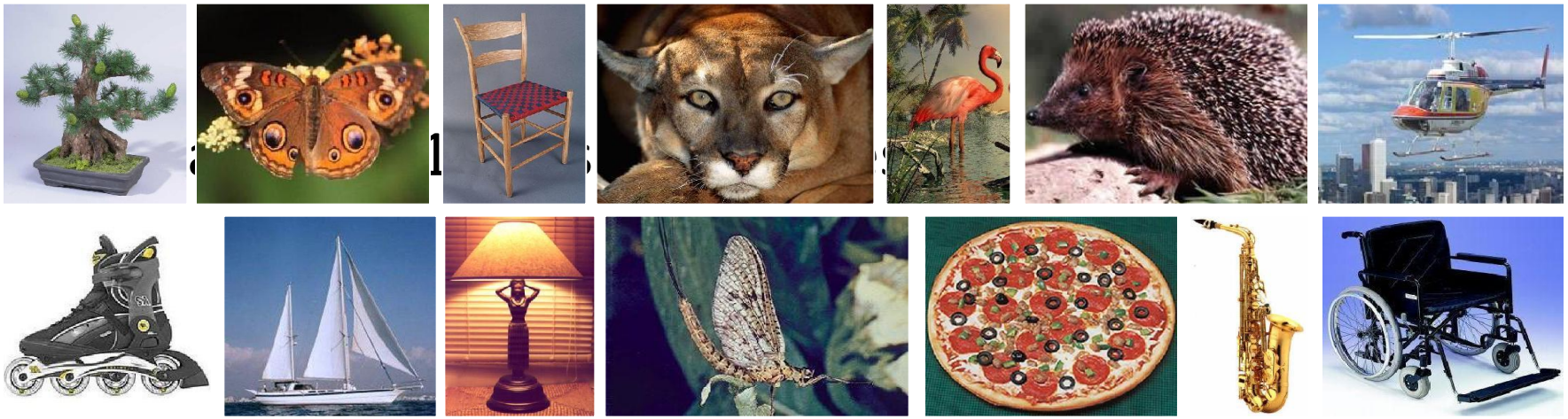
Spatial pyramids: Scene classification results



	Weak features (vocabulary size: 16)		Strong features (vocabulary size: 200)	
Level	Single-level	Pyramid	Single-level	Pyramid
0 (1×1)	45.3 ±0.5		72.2 ±0.6	
1 (2×2)	53.6 ±0.3	56.2 ±0.6	77.9 ±0.6	79.0 ±0.5
2 (4×4)	61.7 ±0.6	64.7 ±0.7	79.4 ±0.3	**81.1** ±0.3
3 (8×8)	63.3 ±0.8	**66.8** ±0.6	77.2 ±0.4	80.7 ±0.3

Lazebnik, Schmid & Ponce (CVPR 2006) –
Beyond bags of features: spatial pyramid matching for recognizing natural scene categories

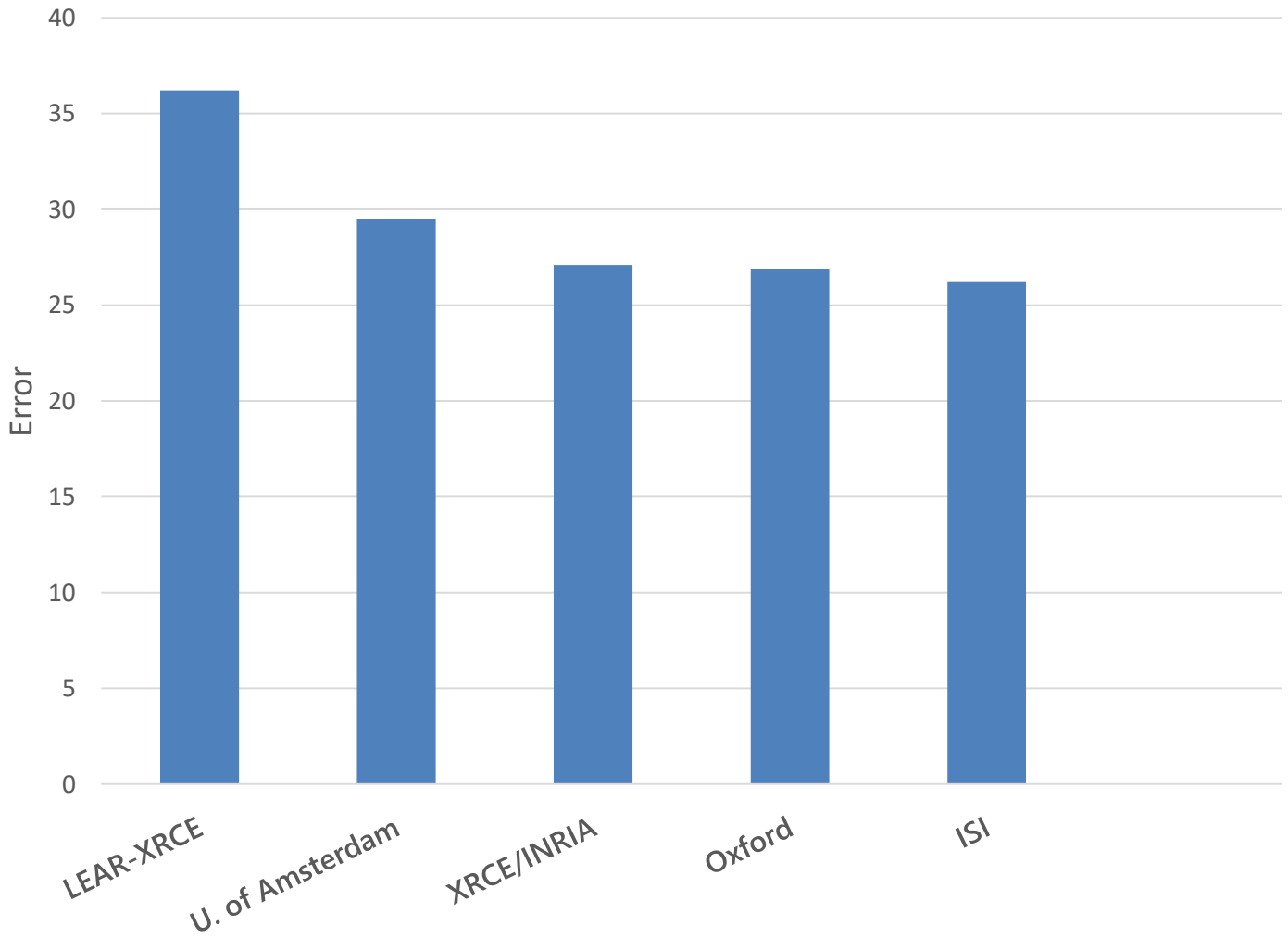
Spatial pyramids: Caltech101 classification results



	Weak features (16)		Strong features (200)	
Level	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

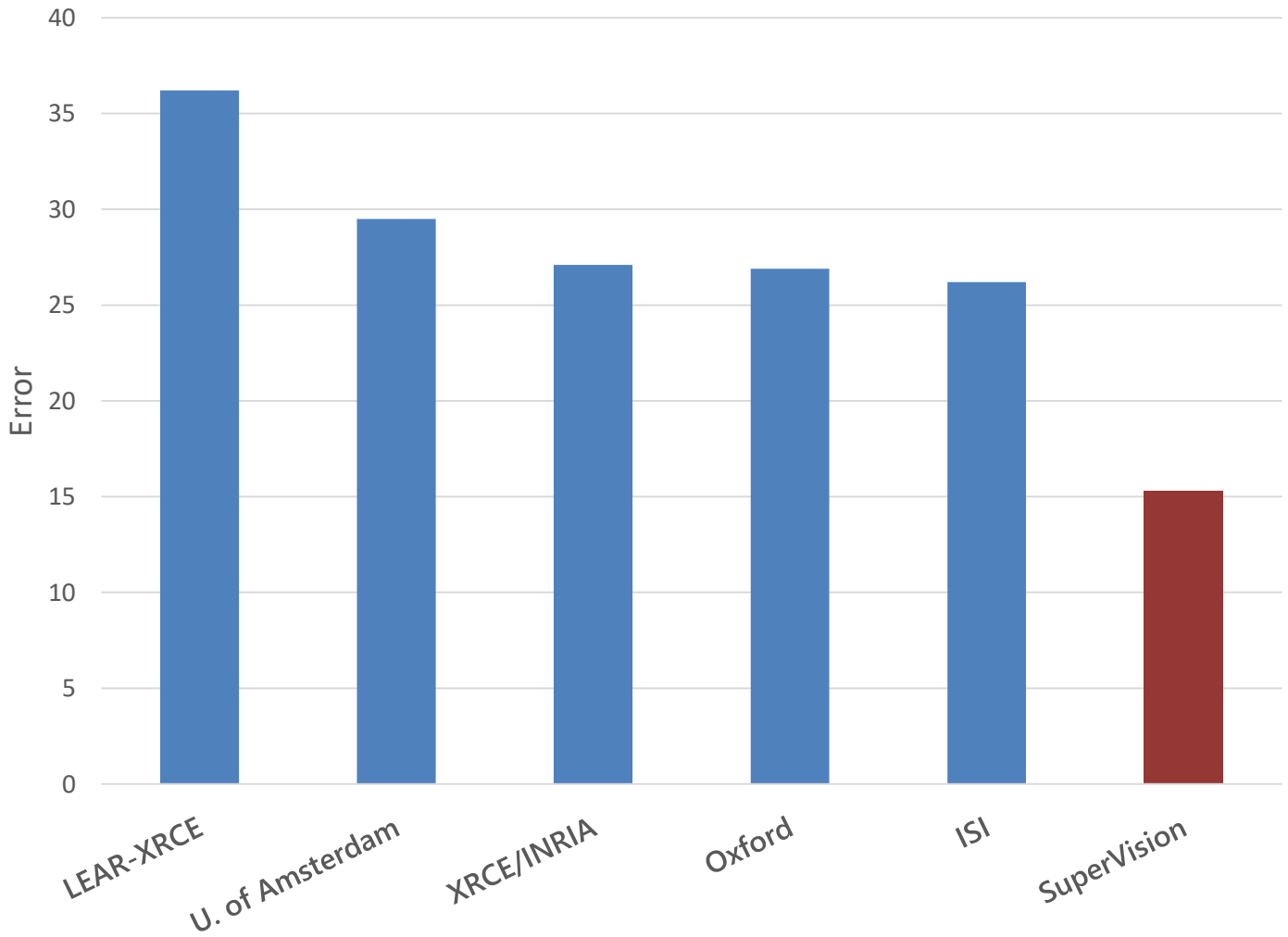
Lazebnik, Schmid & Ponce (CVPR 2006) –
Beyond bags of features: spatial pyramid matching for recognizing natural scene categories

ImageNet 1K


40
35
30
25
20
15
10
5
0
Error
LEAR-XRCE
U. of Amsterdam
XRCE/INRIA
Oxford
ISI

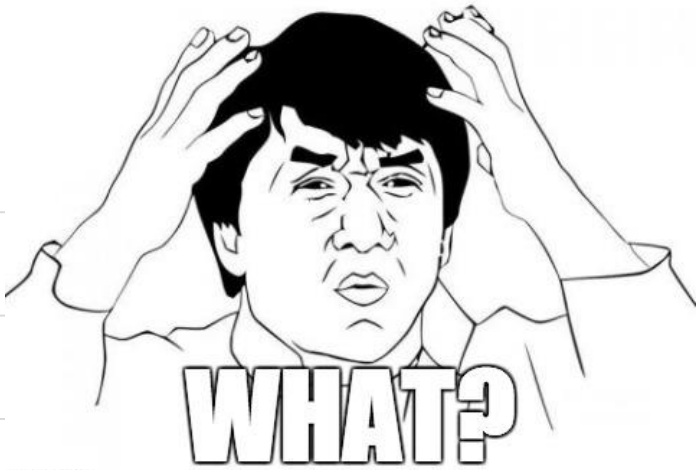
ImageNet 1K

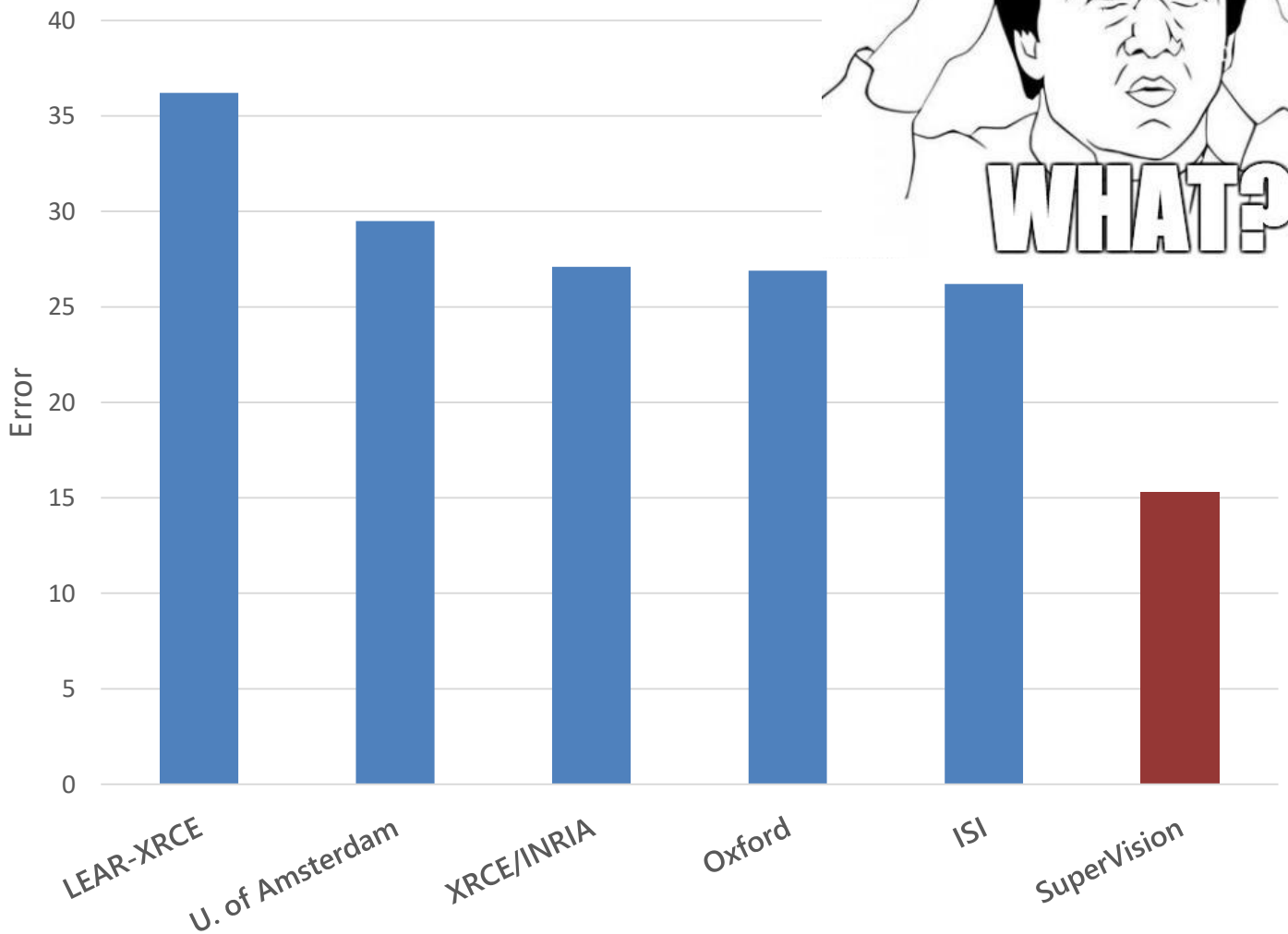
(Fall 2012)



ImageNet 1K

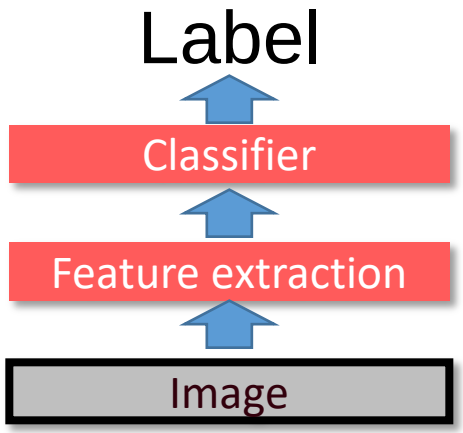
(Fall 2012)

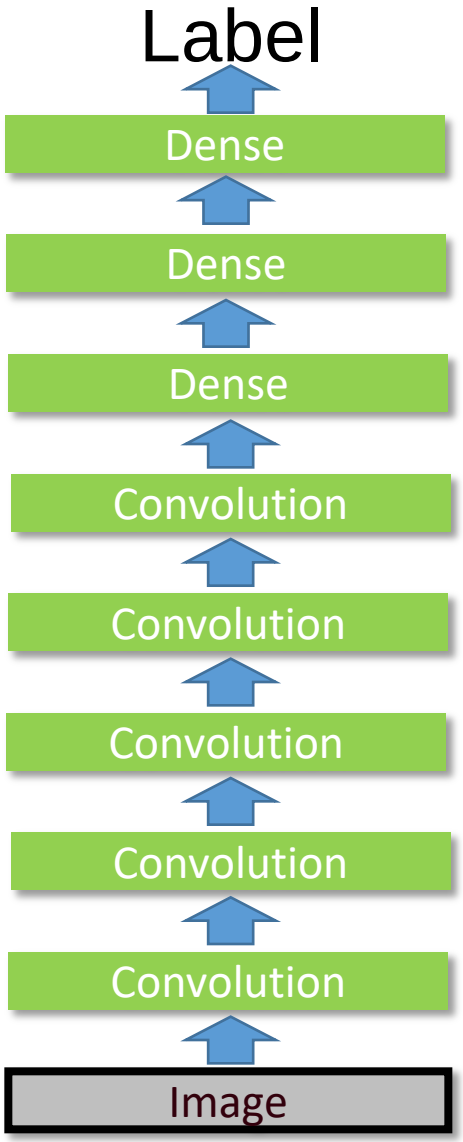




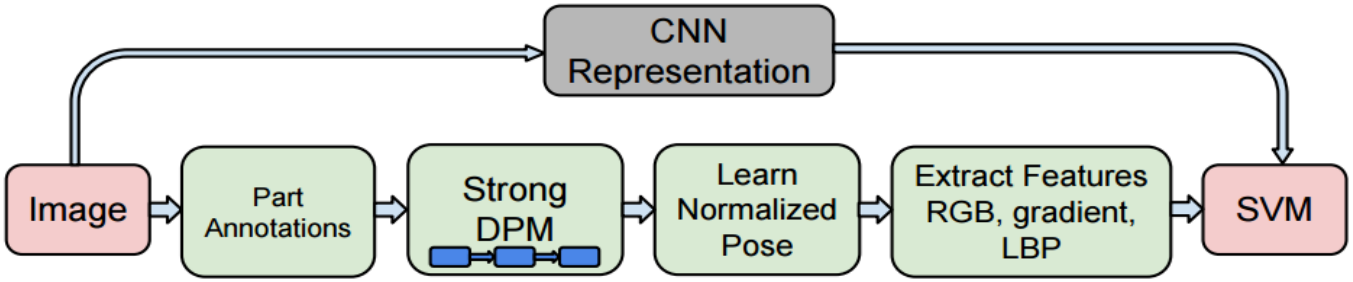
Shallow vs. deep learning

- Engineered vs. learned features





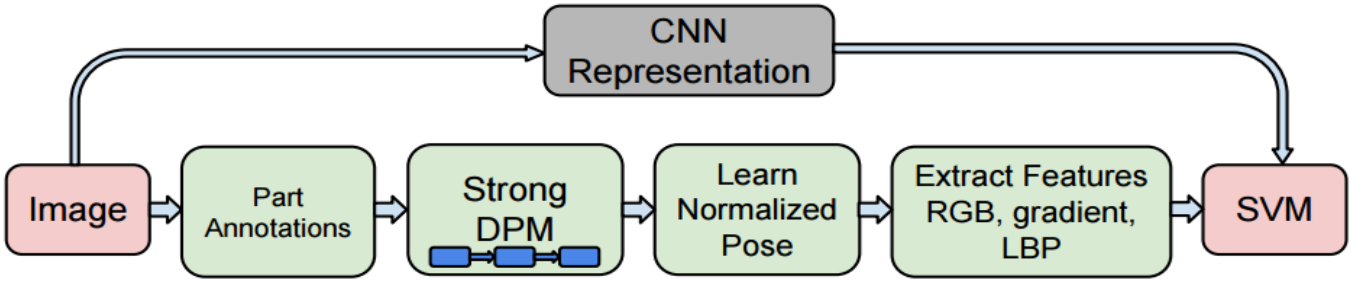
Convolutional activation features

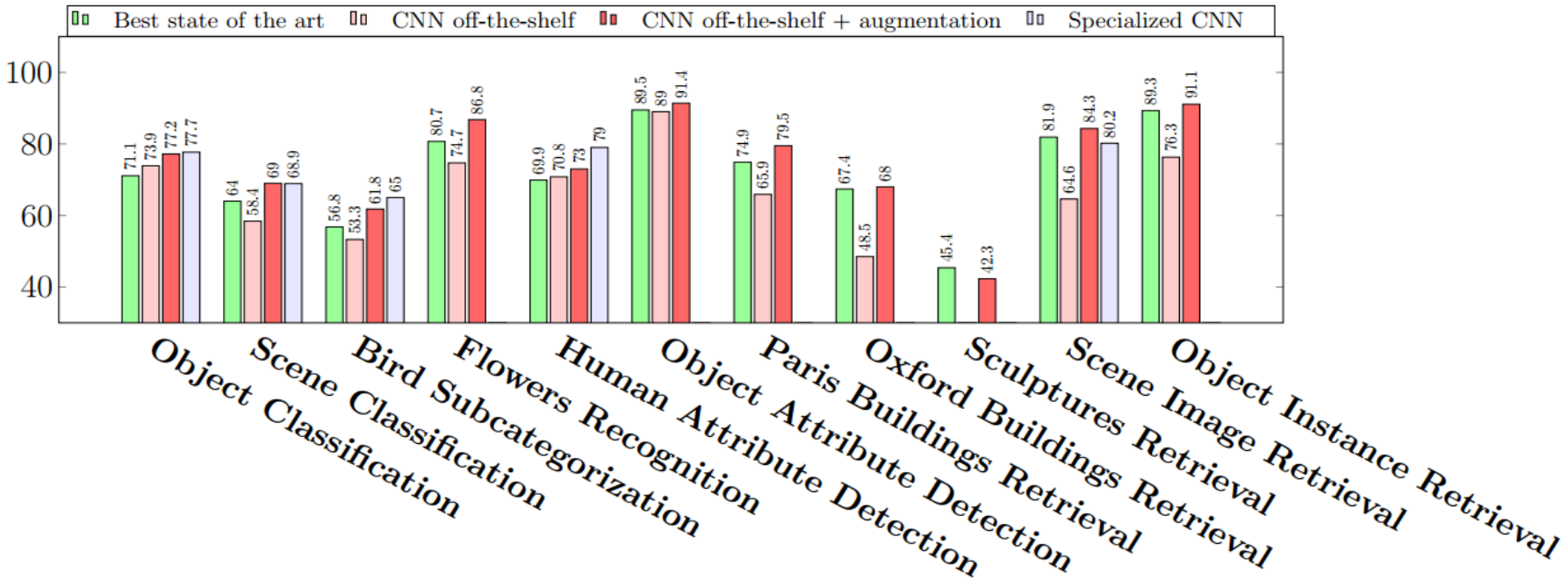


CNN Features off-the-shelf: an Astounding Baseline for Recognition

[Razavian et al. 2014]

Convolutional activation features





CNN Features off-the-shelf: an Astounding Baseline for Recognition

[Razavian et al. 2014]

Things to remember

- Visual categorization help transfer knowledge
- Image features
  - Color, gradients, textures, motion
  - Histogram, SIFT, Descriptors
  - Bag-of-visual-words
  - CNN Feature
- Image/region categorization

Acknowledgements

- Thanks to the following researchers for making their teaching/research material online
  - Forsyth
  - Steve Seitz
  - Noah Snavely
  - J.B. Huang
  - Derek Hoiem
  - D. Lowe
  - A. Bobick
  - S. Lazebnik
  - K. Grauman
  - R. Zaleski
  - Leibe
  - And many more …….

Next Lecture - Classifiers