Indian Institute of Information Technology, Allahabad

CVBL IIIT Allahabad

Image Categorization

By

Dr. Satish Kumar Singh & Dr. Shiv Ram Dubey
Computer Vision and Biometrics Lab
Department of Information Technology
Indian Institute of Information Technology, Allahabad

TEAM

Computer Vision and Biometrics Lab (CVBL)

Department of Information Technology

Indian Institute of Information Technology Allahabad

Course Instructors

Dr. Satish Kumar Singh, Associate Professor, IIIT Allahabad (Email: sk.singh@iiita.ac.in)

Dr. Shiv Ram Dubey, Assistant Professor, IIIT Allahabad (Email: srdubey@iiita.ac.in)

DISCLAIMER

The content (text, image, and graphics) used in this slide are adopted from many sources for Academic purposes. Broadly, the sources have been given due credit appropriately. However, there is a chance of missing out some original primary sources. The authors of this material do not claim any copyright of such material.

Visual Recognition and Learning

- Image Categorization
- Image Features
- 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

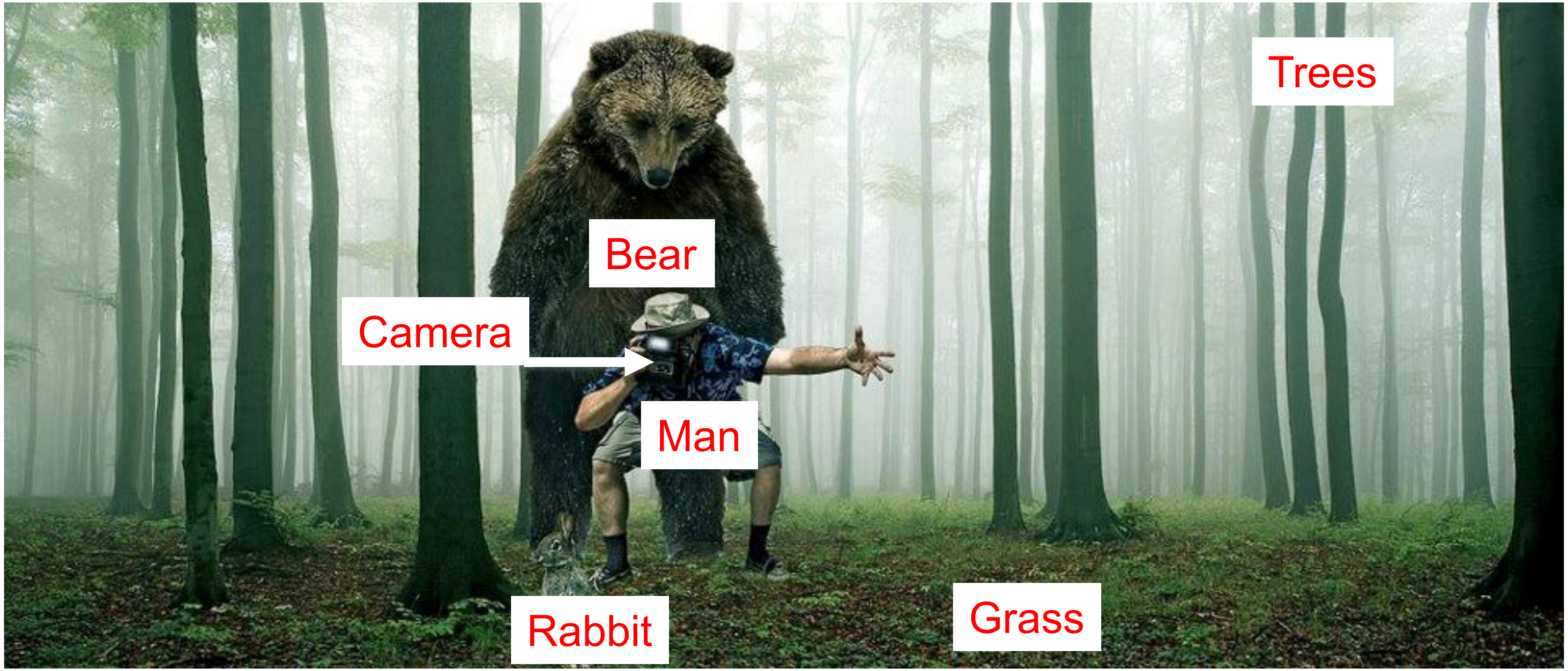
- Image and region categorization

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

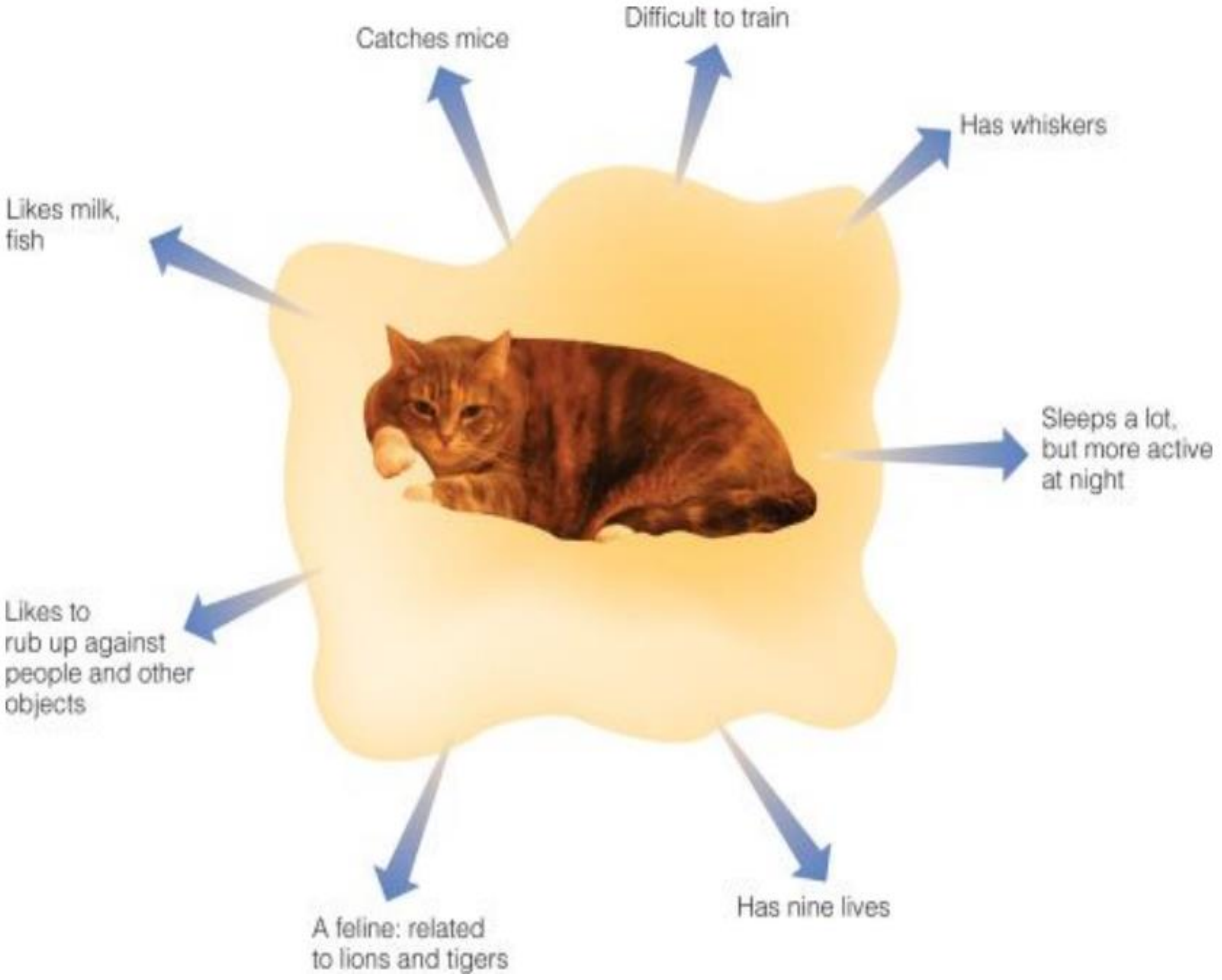
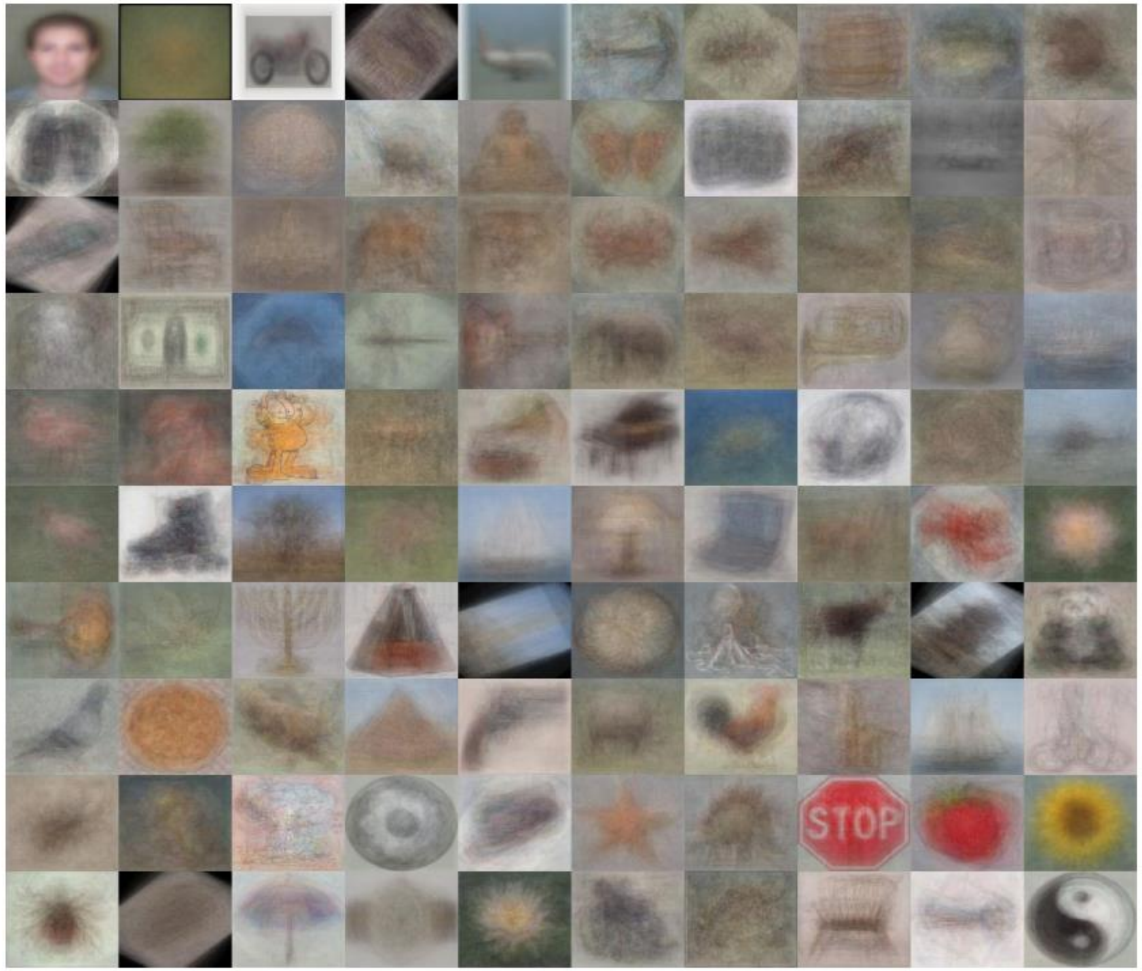


IMAGE CATEGORIZATION

- Cat vs Dog

IMAGE CATEGORIZATION

- Object recognition



Caltech 101 Average Object Images

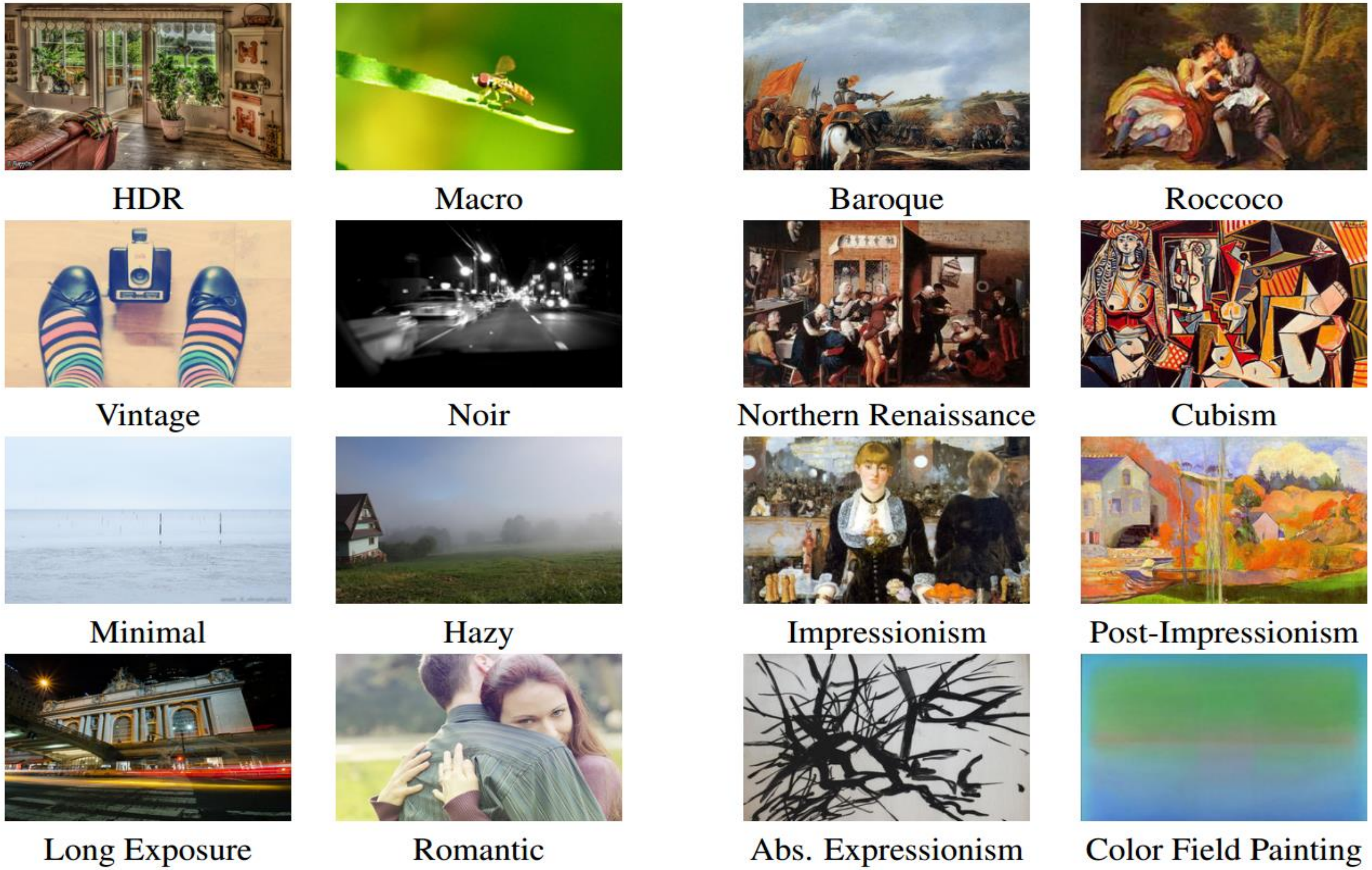
IMAGE CATEGORIZATION

- Place recognition

Places Database [Zhou et al. NIPS 2014]

IMAGE CATEGORIZATION

- Image style recognition



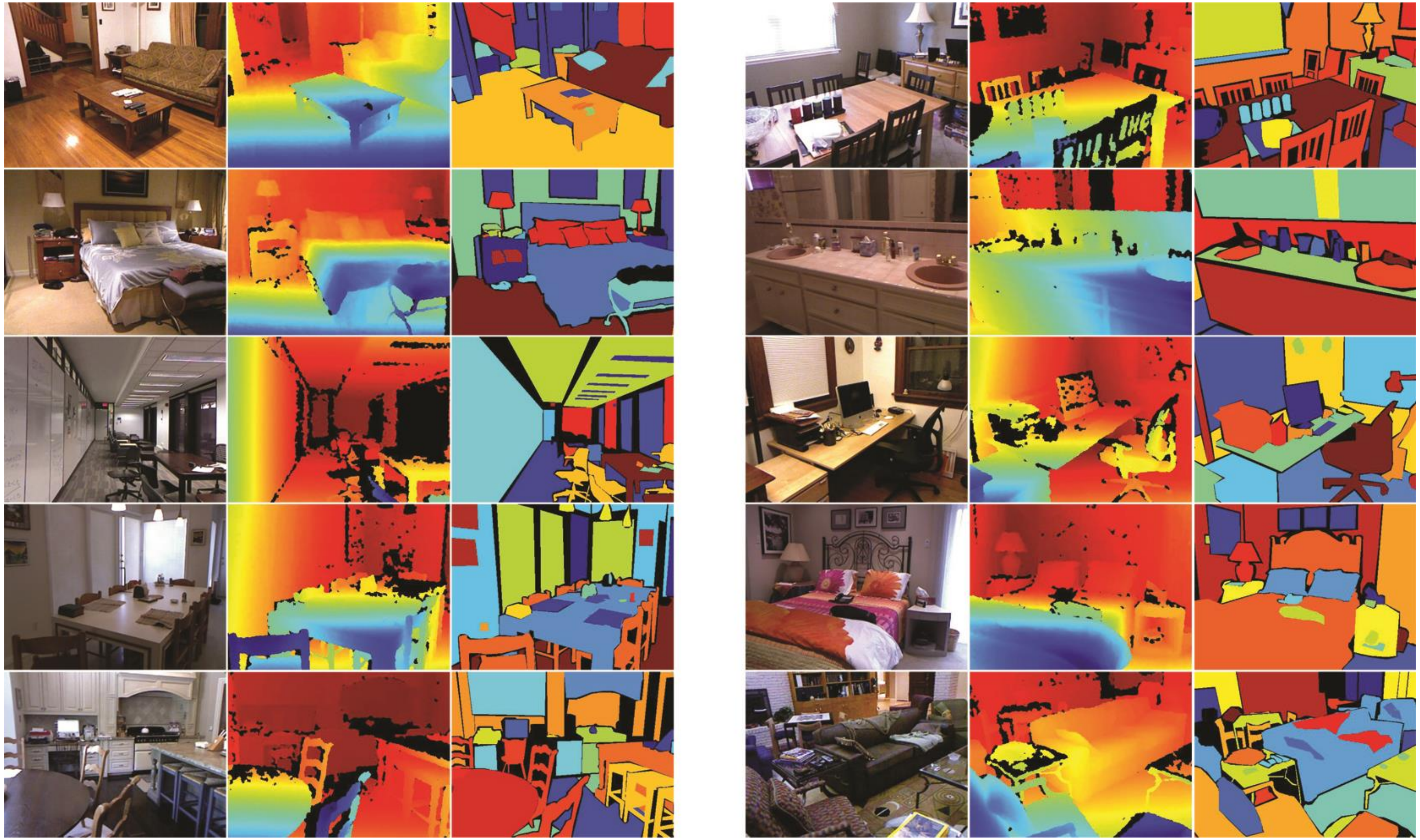
Flickr Style: 80K images covering 20 styles.

Wikipaintings: 85K images for 25 art genres.

[Karayev et al. BMVC 2014]

REGION CATEGORIZATION

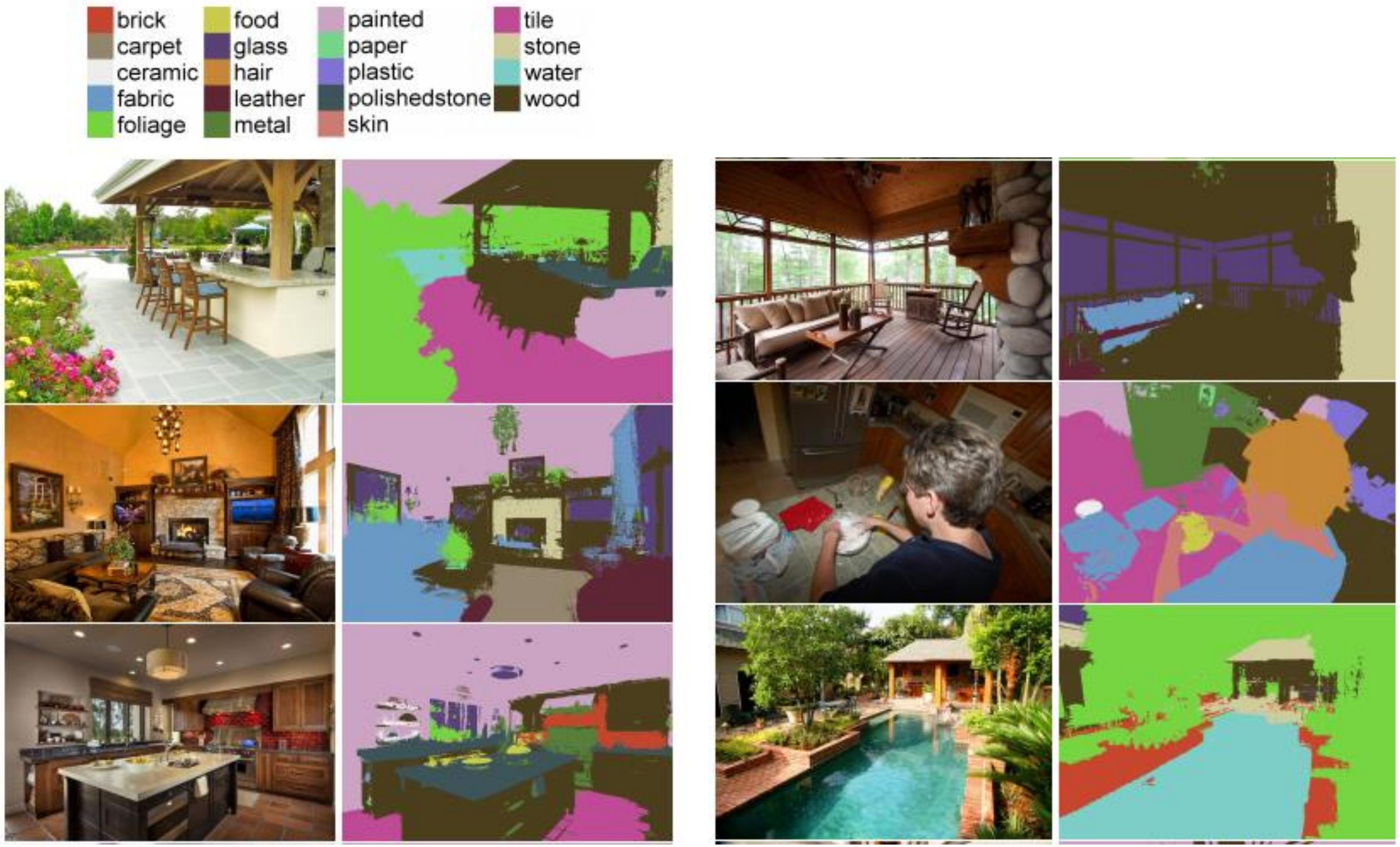
- Semantic segmentation from RGBD images



[Silberman et al. ECCV 2012]

REGION CATEGORIZATION

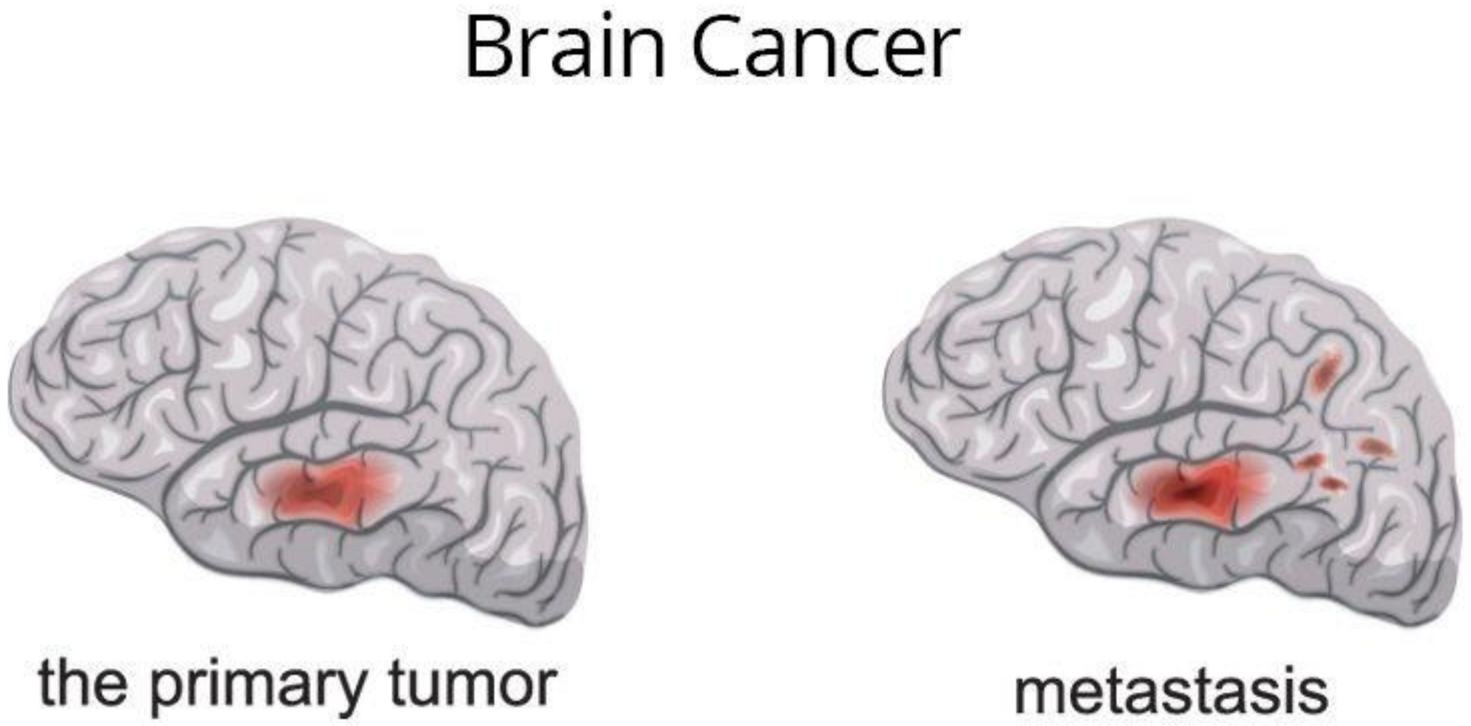
- Material recognition



[Bell et al. CVPR 2015]

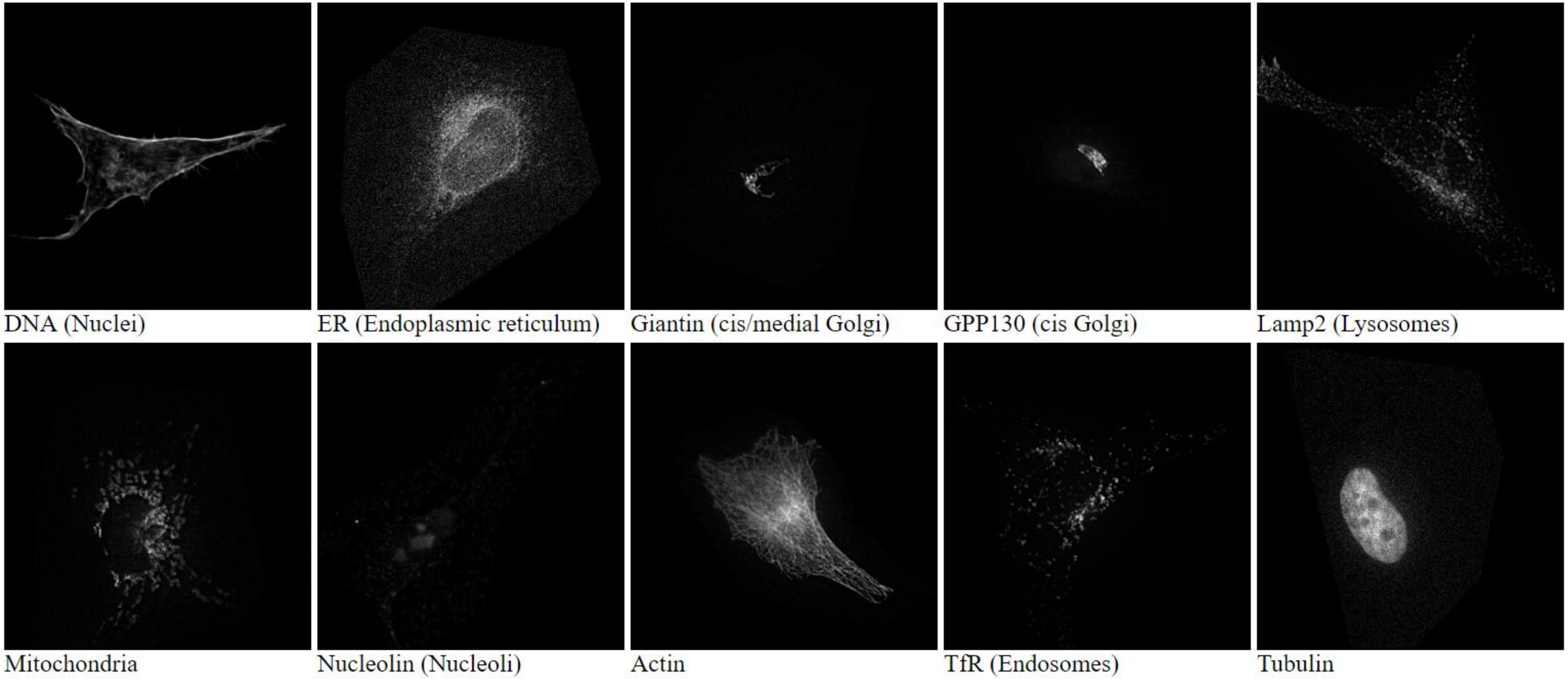
REGION CATEGORIZATION

- Primary Tumor vs Metastasis



REGION CATEGORIZATION

- Identification of sub-cellular organelles



Fluorescence microscopy images of HeLa cells

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

REGION CATEGORIZATION

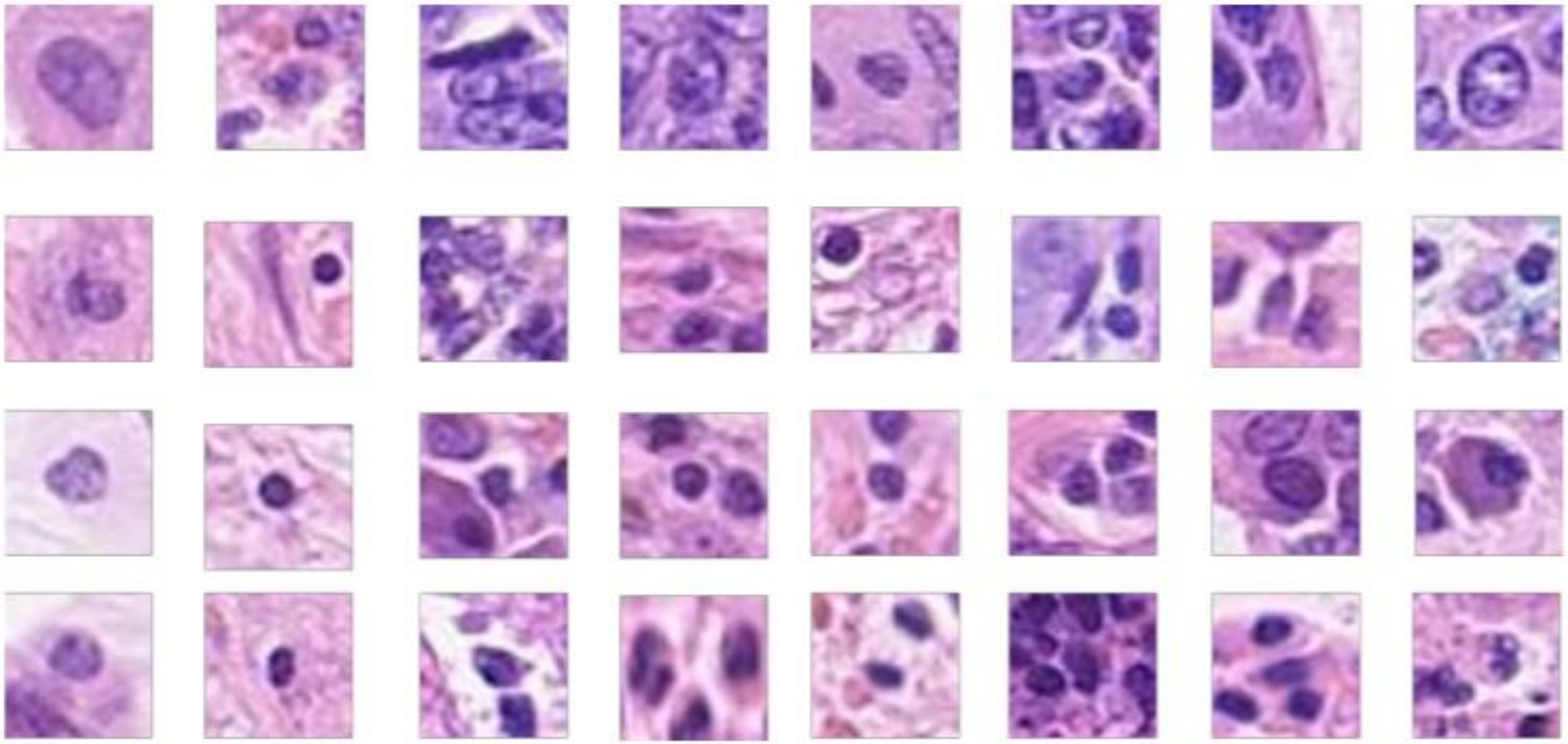
- Colon Cancer Nuclei Classification

'Epithelial'

'Fibroblast'

'Inflammatory'

'Miscellaneous'



"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

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

output — prediction function — Image feature

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

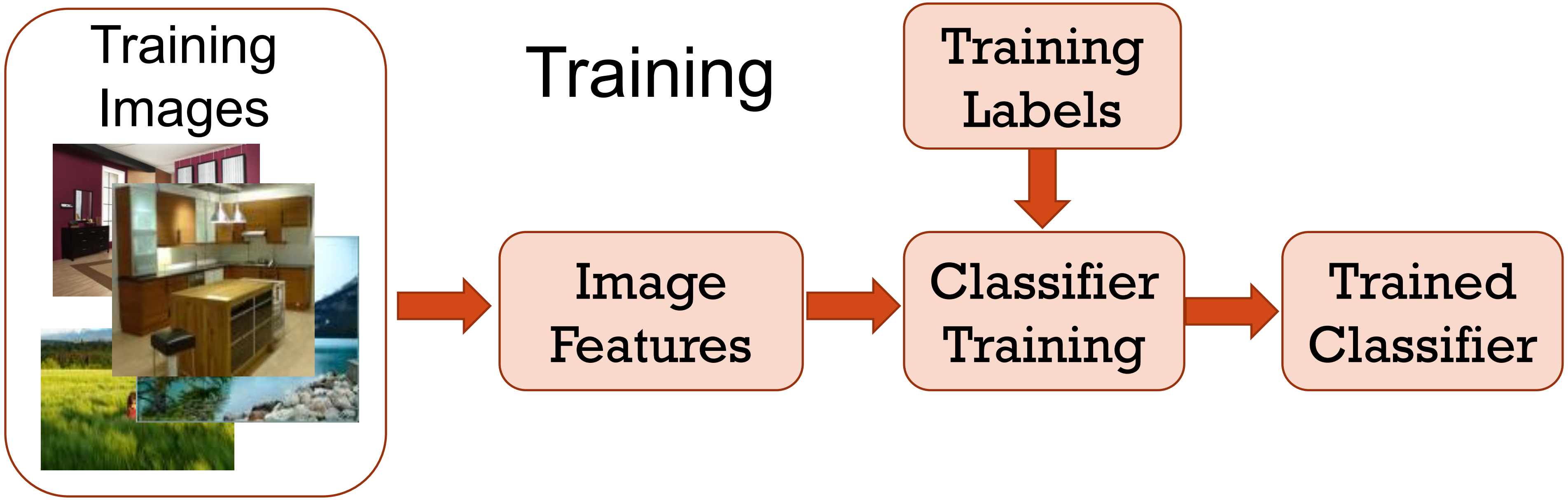
THE STATISTICAL LEARNING FRAMEWORK

$$\underbrace{\mathbf{y}}_{\text{output}} = \underbrace{\mathrm{f}}_{\text{prediction function}}(\underbrace{\mathbf{x}}_{\text{Image feature}})$$

- **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

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

TRAINING PHASE



TRAINING PHASE

Training Images

Training

Training Labels

Image Features → Classifier Training → Trained Classifier

TESTING PHASE

Testing

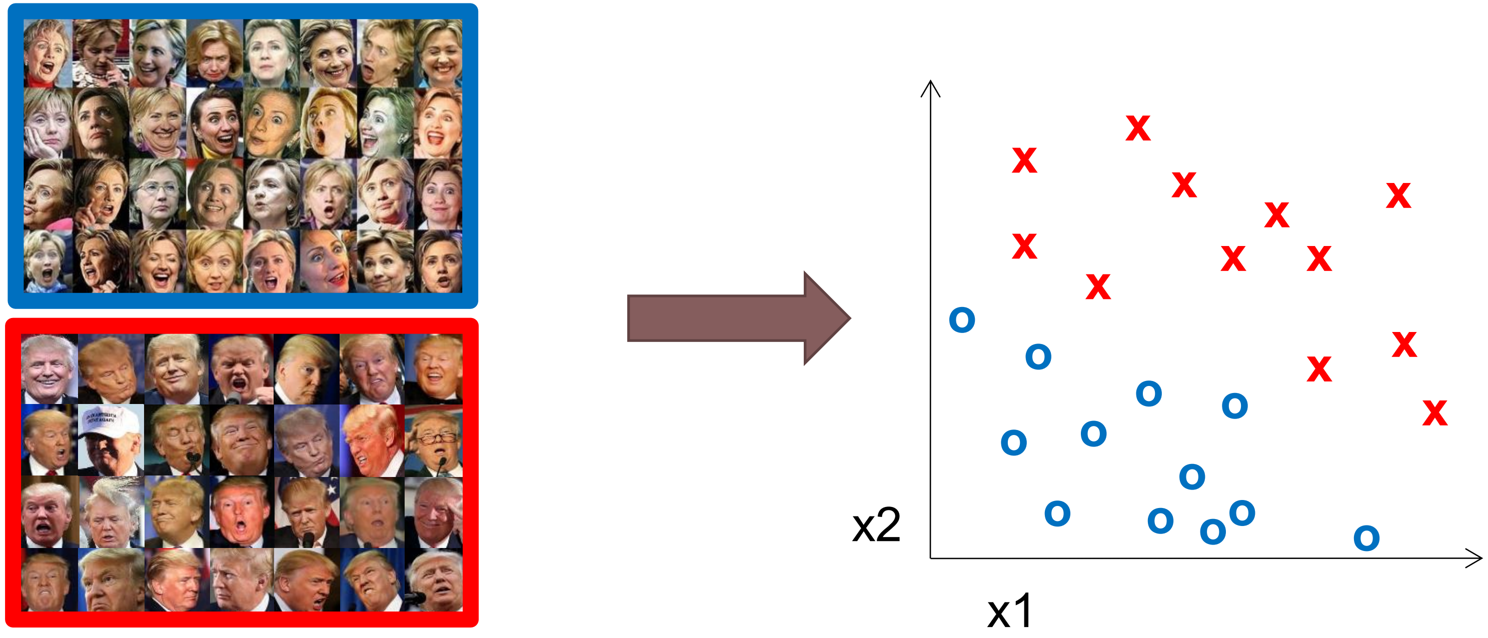
Test Image

Image Features → Trained Classifier → Prediction

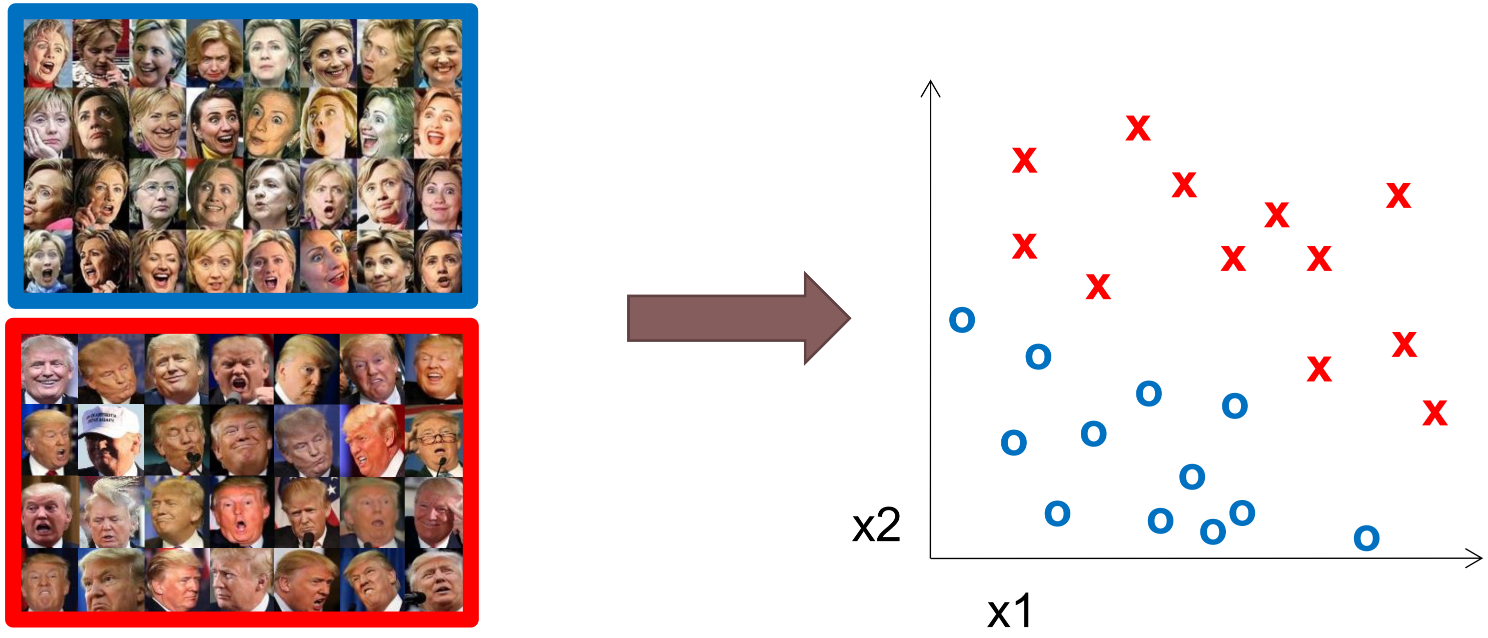
Outdoor

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

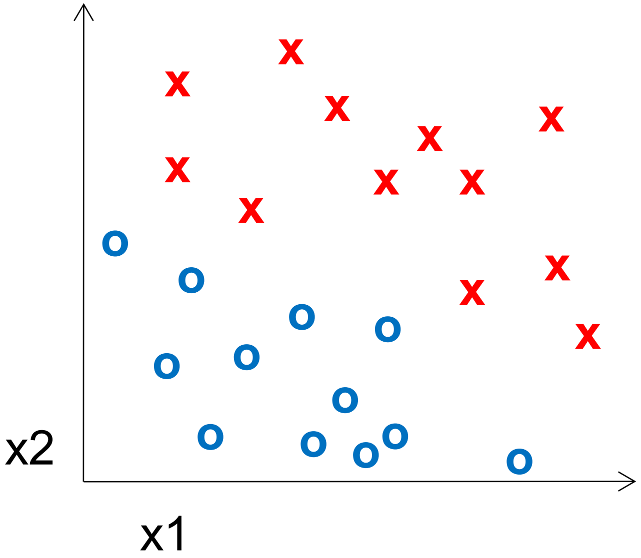
- **Image features**: map images to feature space



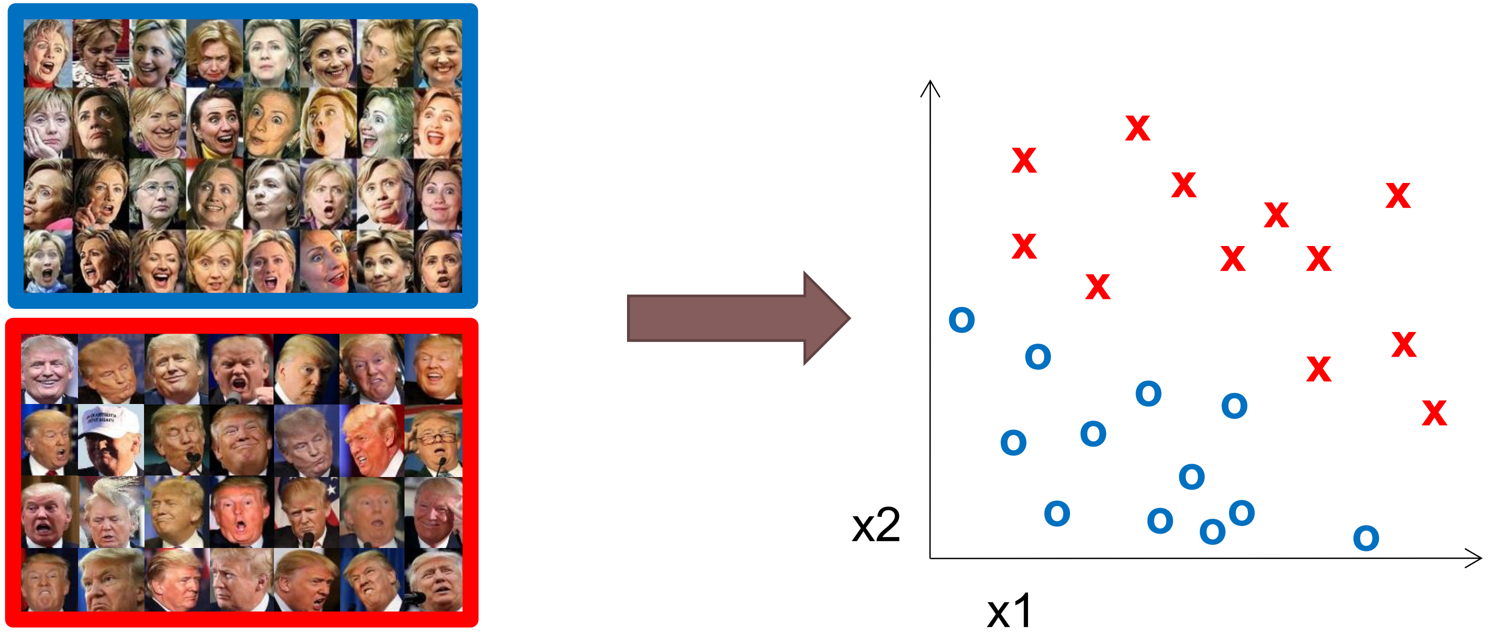
- **Image features**: map images to feature space



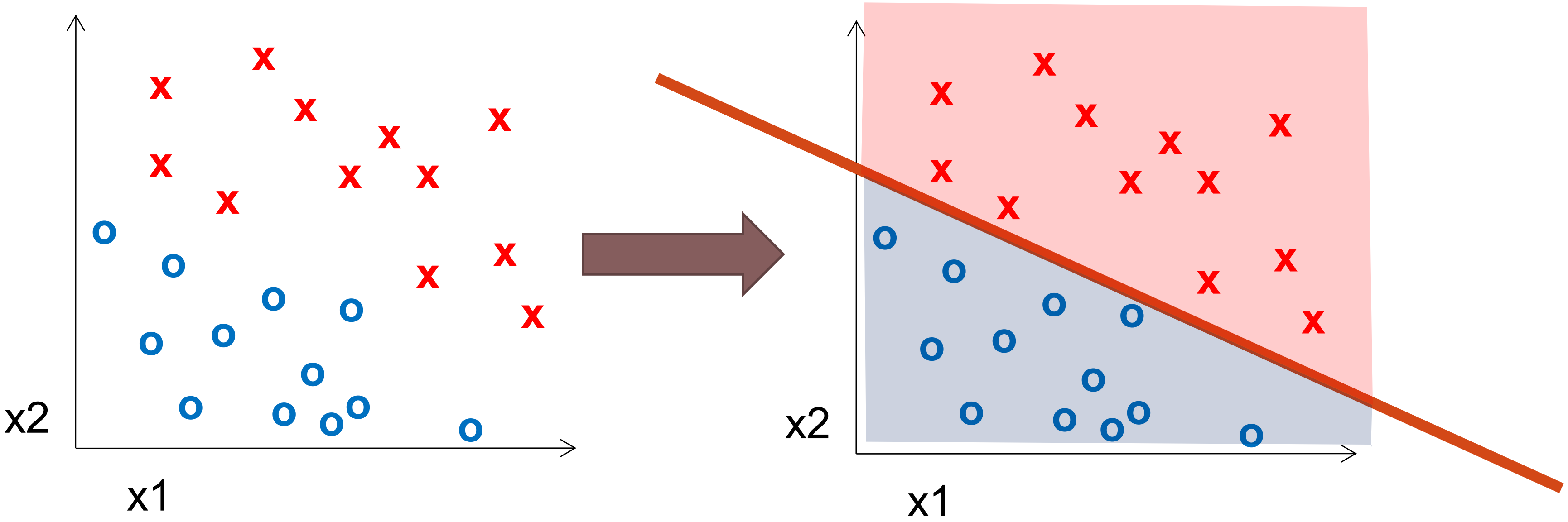
- **Classifiers**: map feature space to label space



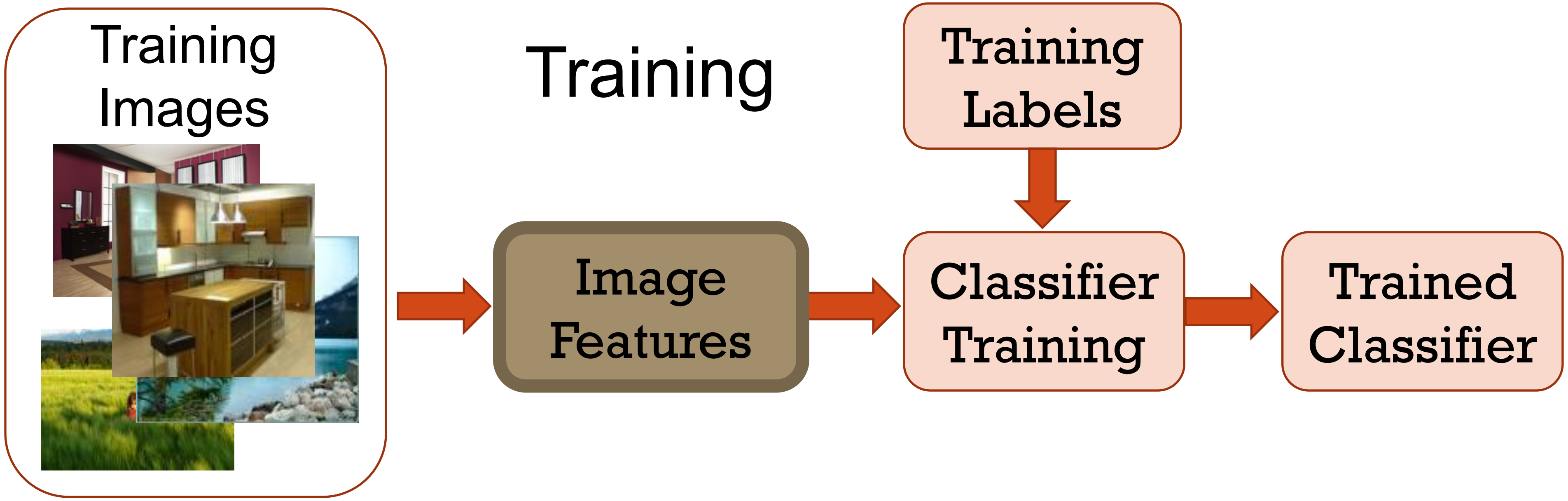
- **Image features**: map images to feature space



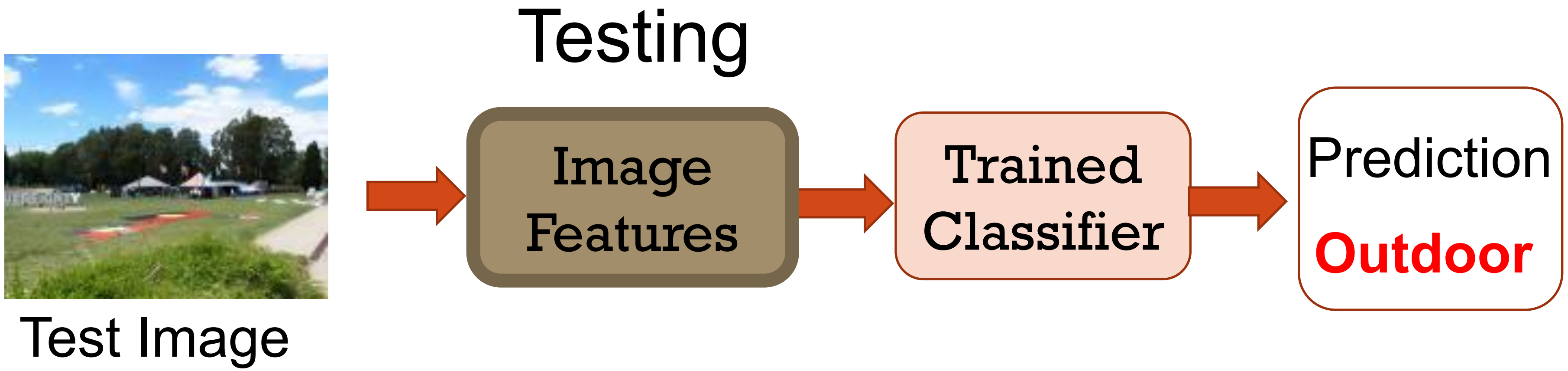
- **Classifiers**: map feature space to label space



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…

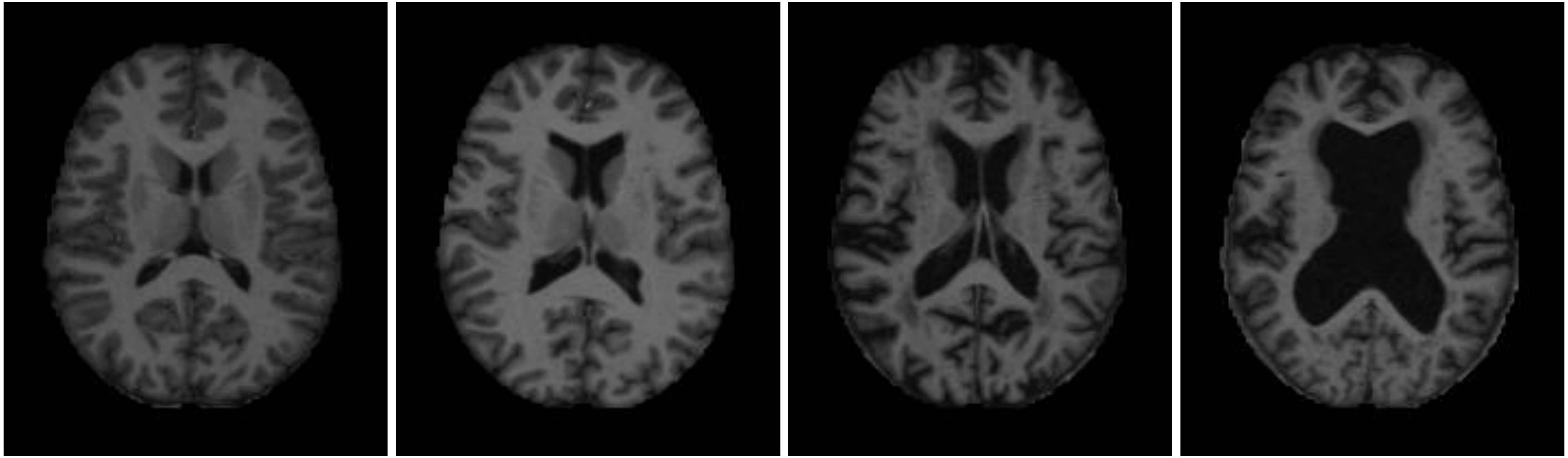
- Recognizing a mug?






Q: WHAT ARE GOOD FEATURES FOR…

- Recognizing the nodule in MRI data?



Q: WHAT ARE GOOD FEATURES FOR…

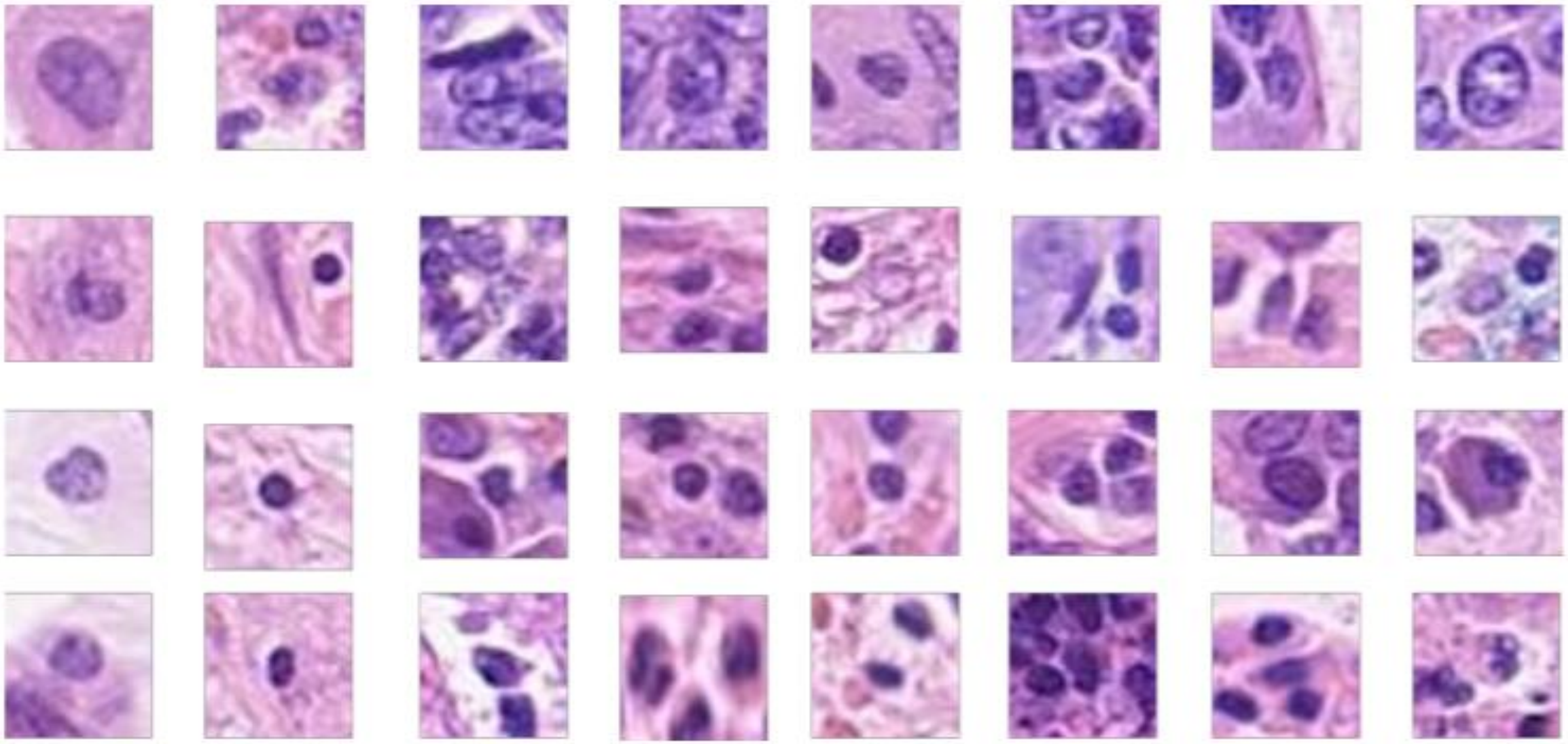
- Recognizing the type of colon cancer?

'Epithelial'

'Fibroblast'

'Inflammatory'

'Miscellaneous'



"CRCHistoPhenotypes" dataset images

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

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, LBP descriptors, etc.

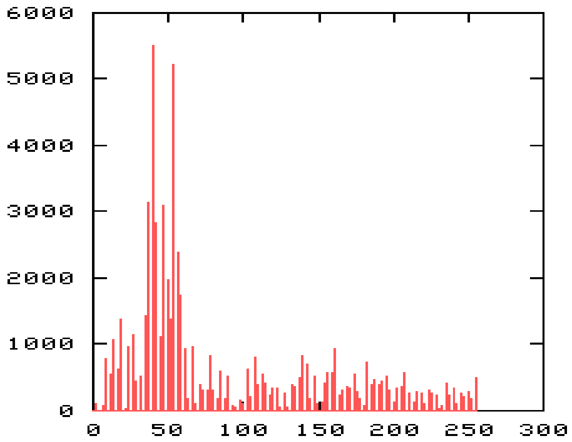
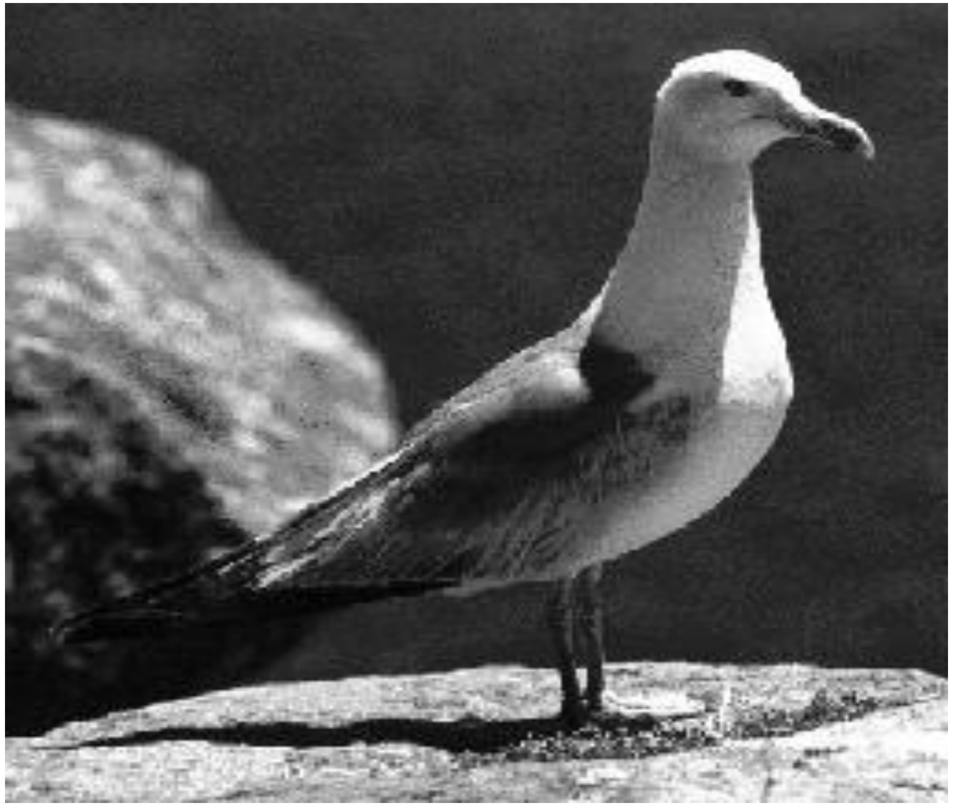
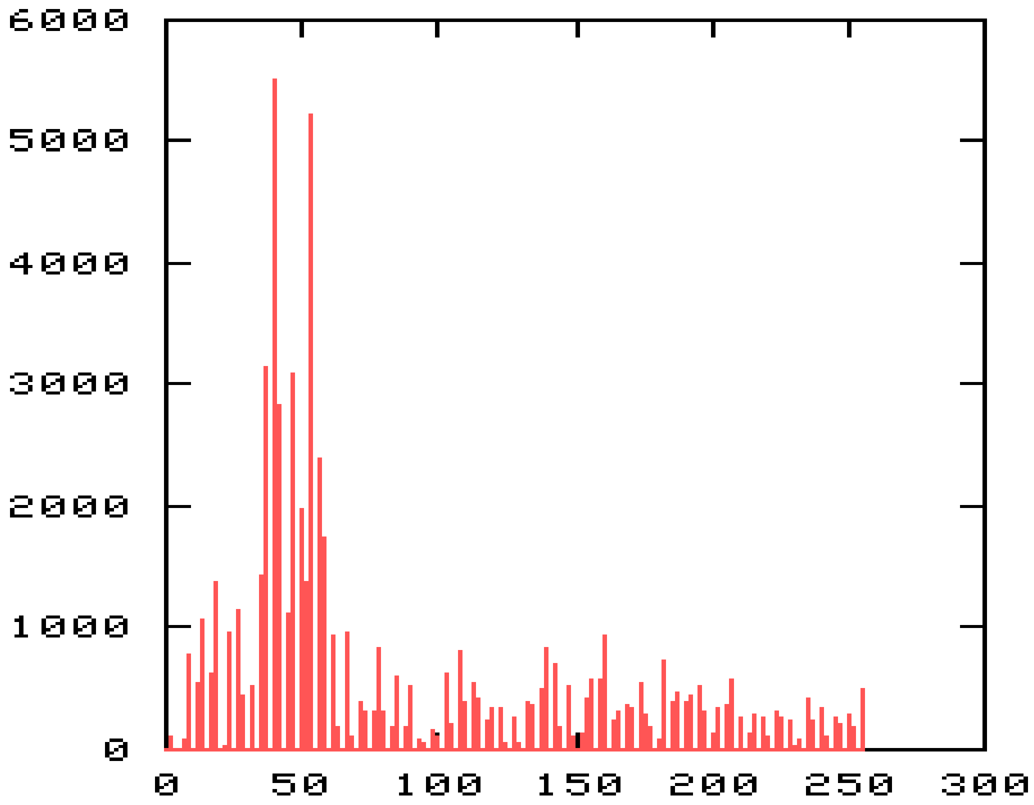


IMAGE REPRESENTATIONS: HISTOGRAMS





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**

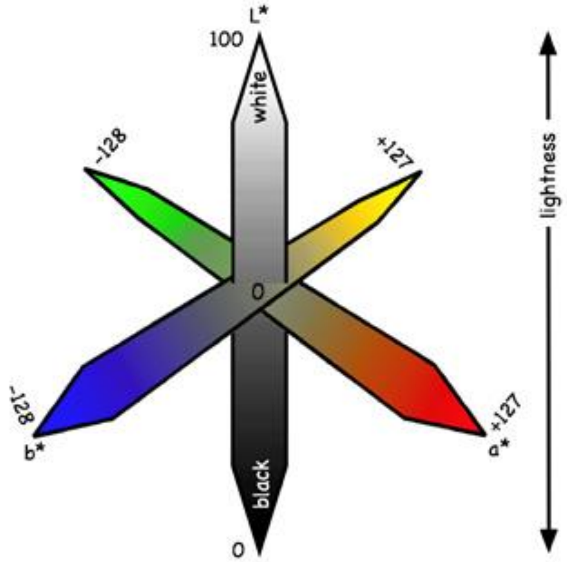
$$\chi^2(h_i, h_j) = \frac{1}{2} \sum_{m=1}^{K} \frac{[h_i(m) - h_j(m)]^2}{h_i(m) + h_j(m)}$$

- **Earth mover's distance**
  - Cross-bin similarity measure
  - Minimal cost paid to transform one distribution into the other

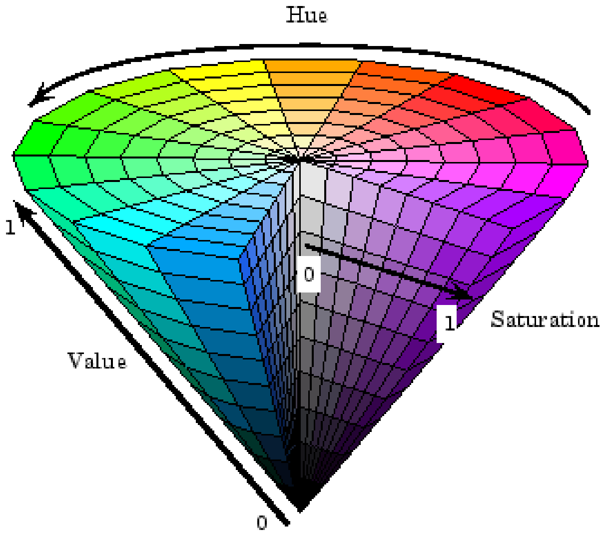
[Rubner et al. The Earth Mover's Distance as a Metric for Image Retrieval, IJCV 2000]

WHAT KIND OF THINGS DO WE COMPUTE HISTOGRAMS OF?

- **Color**

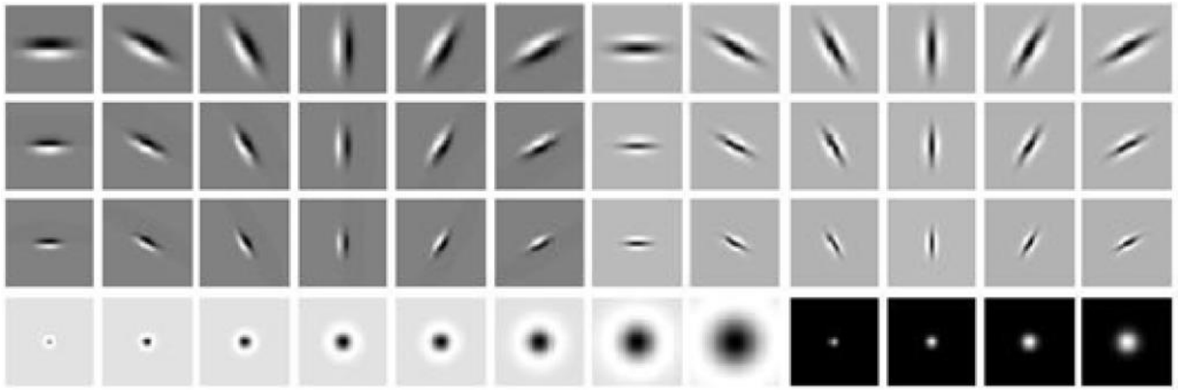


L*a*b* color space



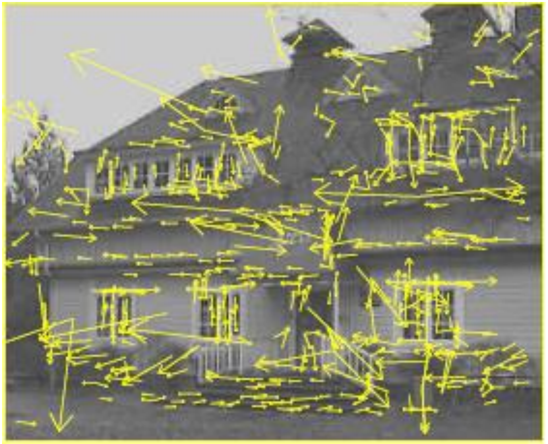
HSV color space

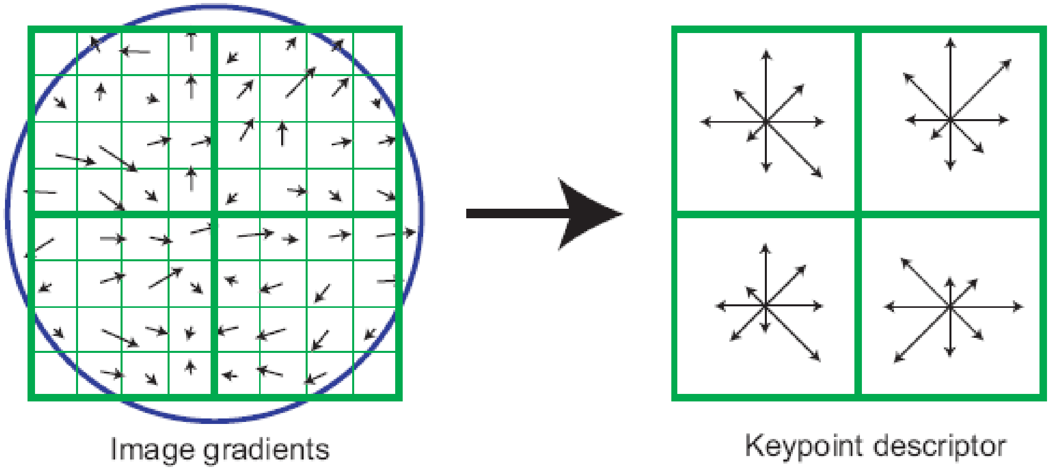
- **Texture** (filter banks or HOG over regions)



WHAT KIND OF THINGS DO WE COMPUTE HISTOGRAMS OF?

Histograms of descriptors

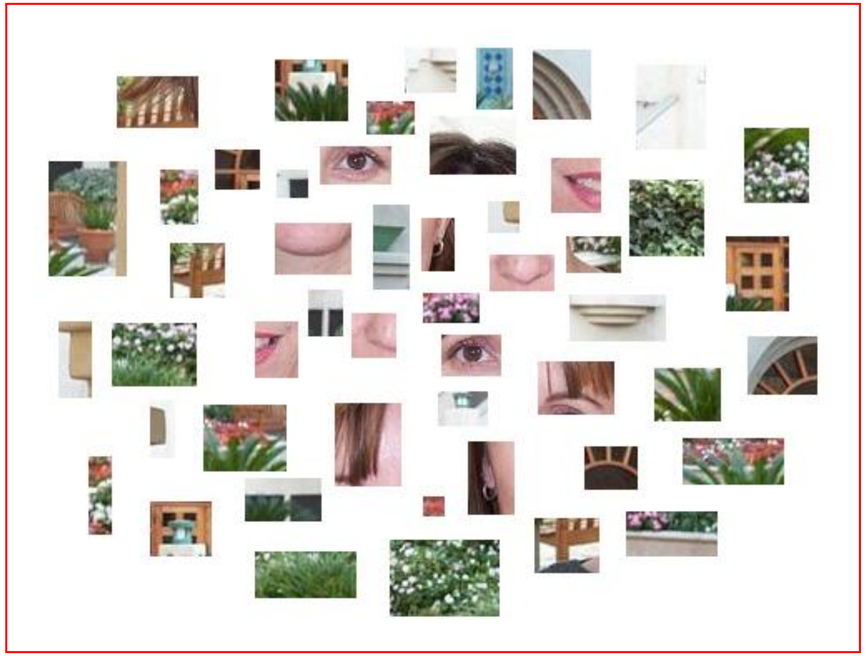




SIFT – [Lowe IJCV 2004]

WHAT KIND OF THINGS DO WE COMPUTE HISTOGRAMS OF?

BAGS OF VISUAL WORDS

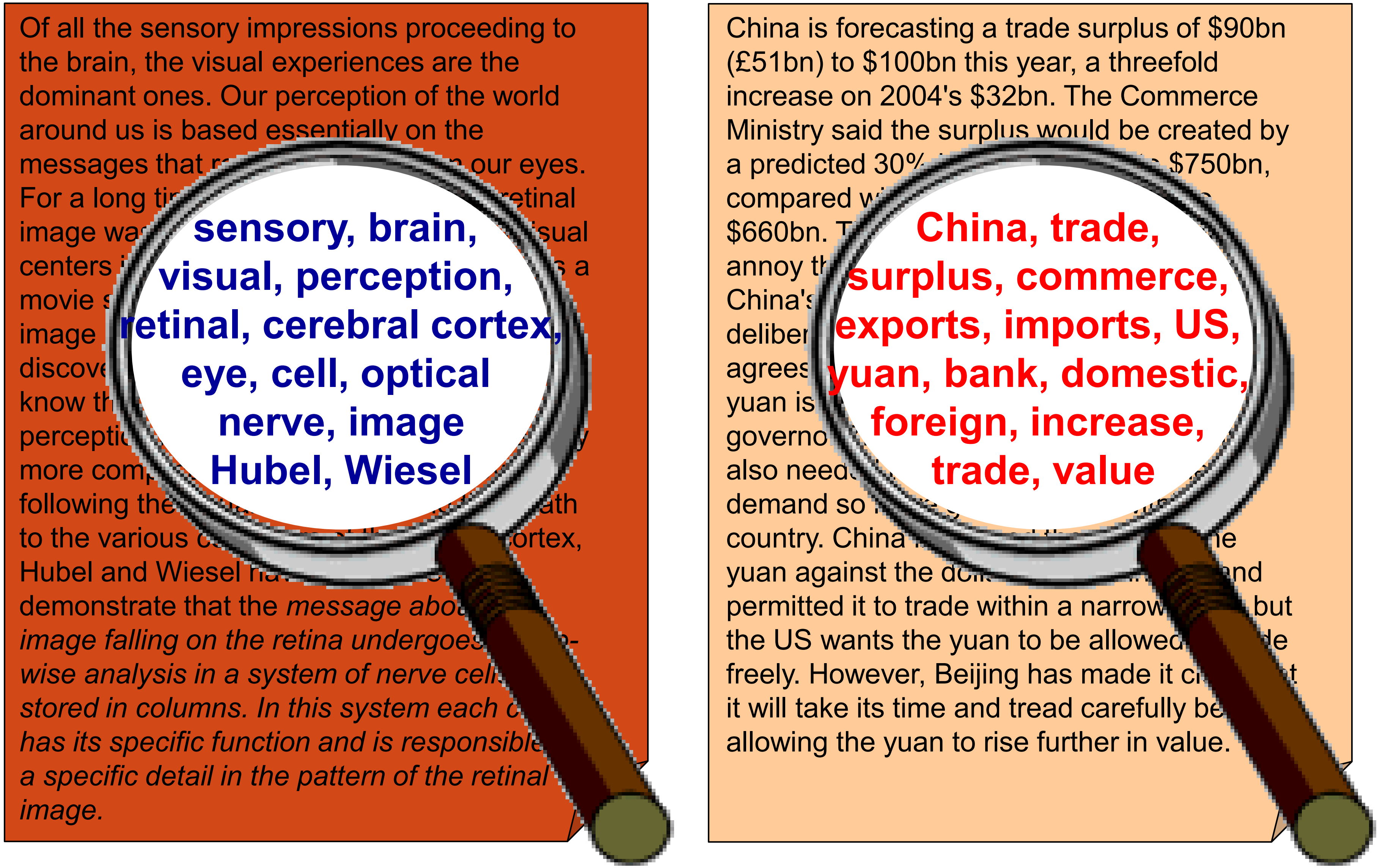


ANALOGY TO DOCUMENTS

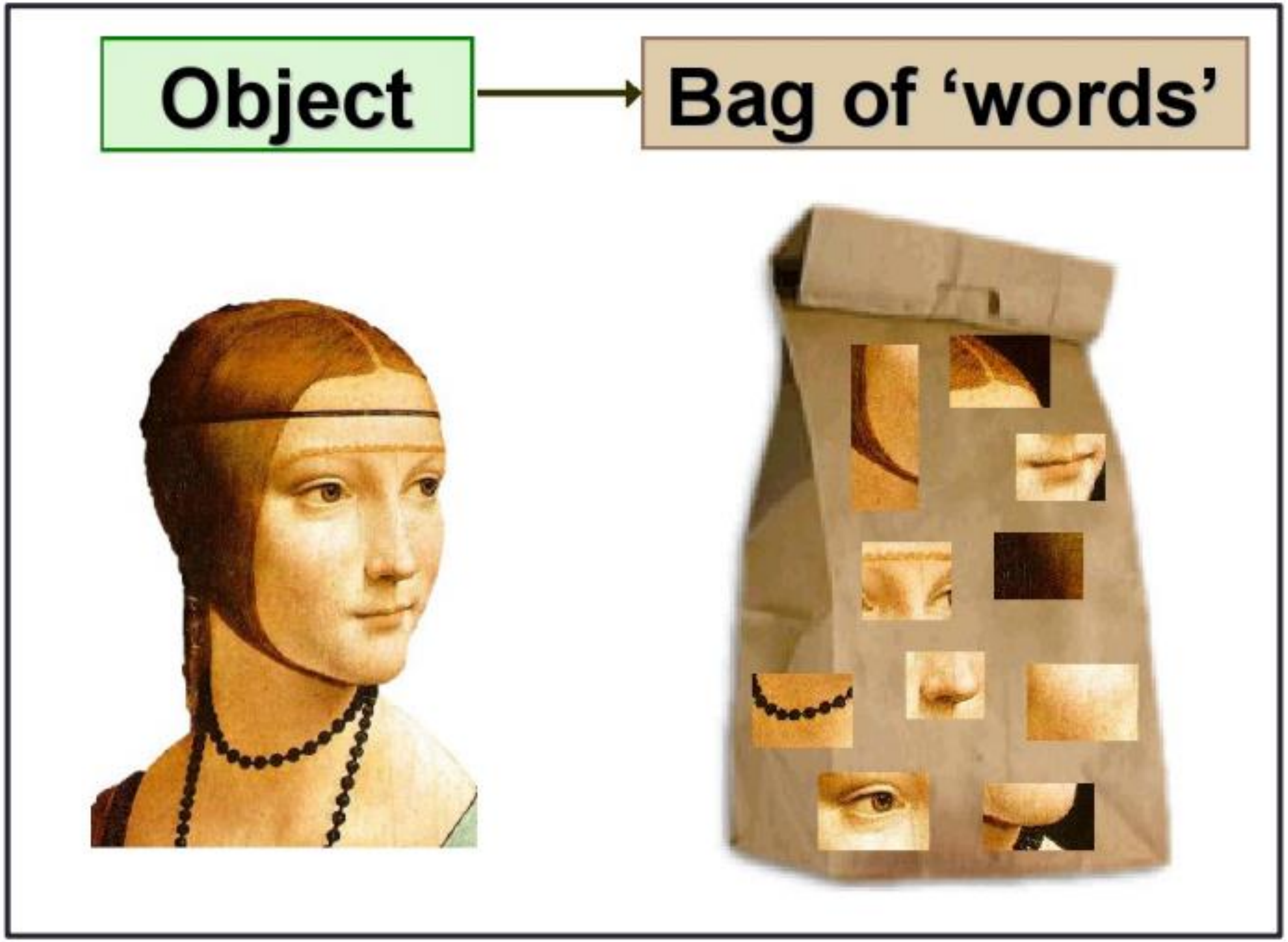
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach the brain from our eyes. For a long time it was thought that the retinal image was transmitted point by point to 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



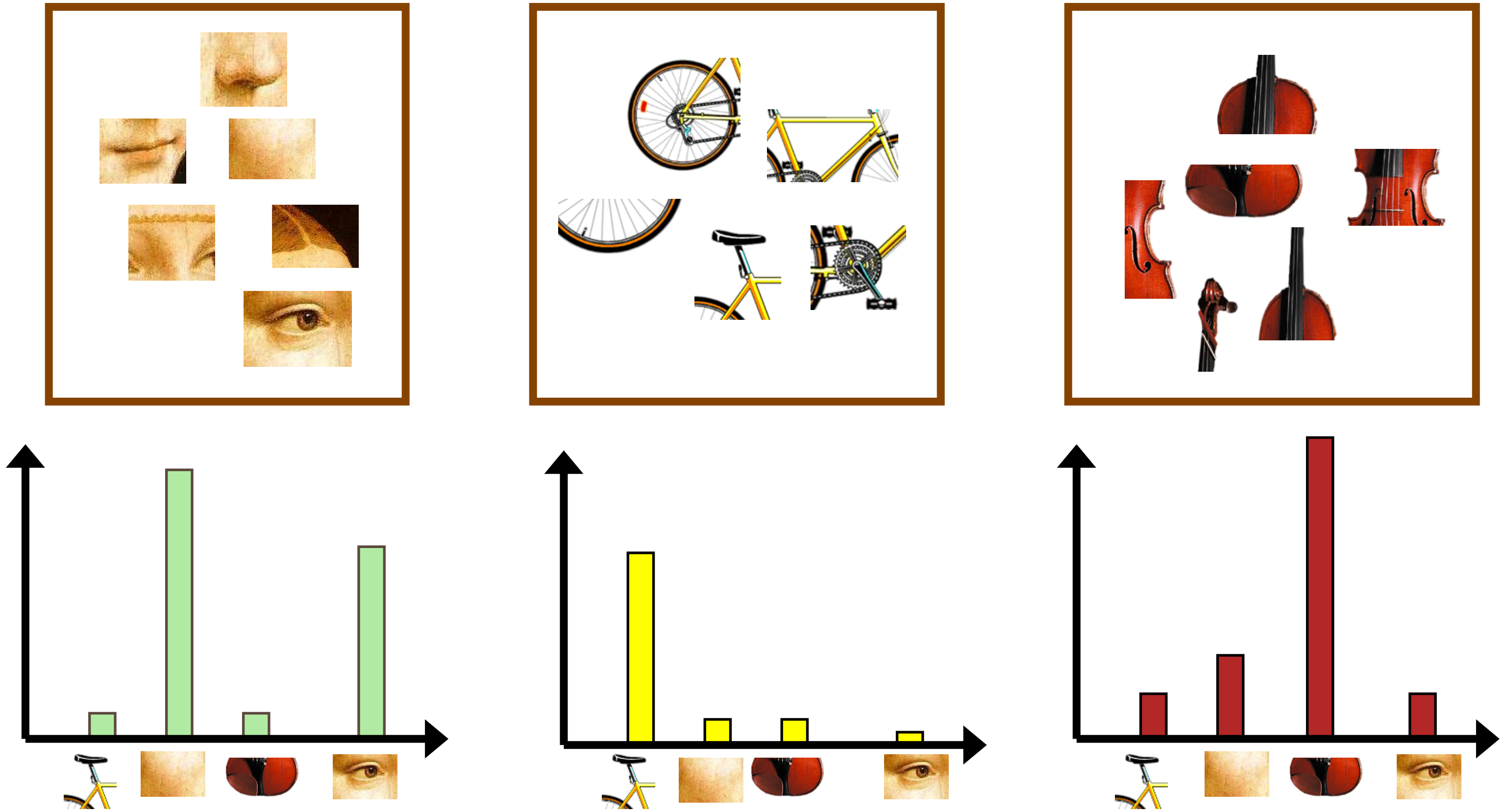
BAGS OF VISUAL WORDS: MOTIVATION



A. Borbick

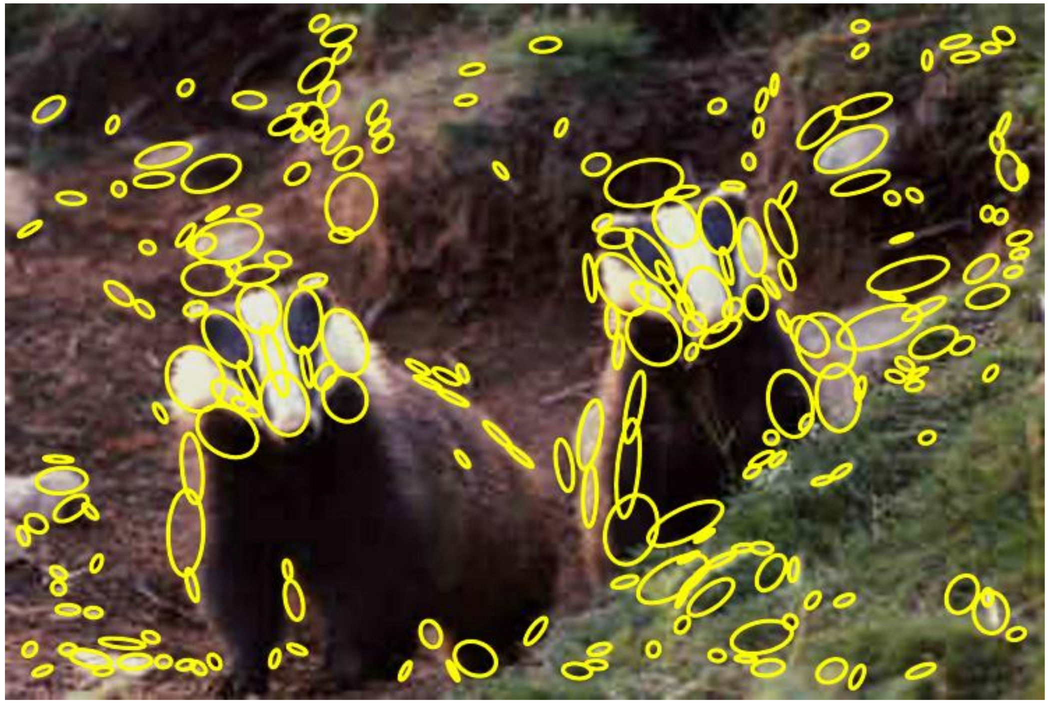
BAGS-OF-VISUAL-WORDS

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

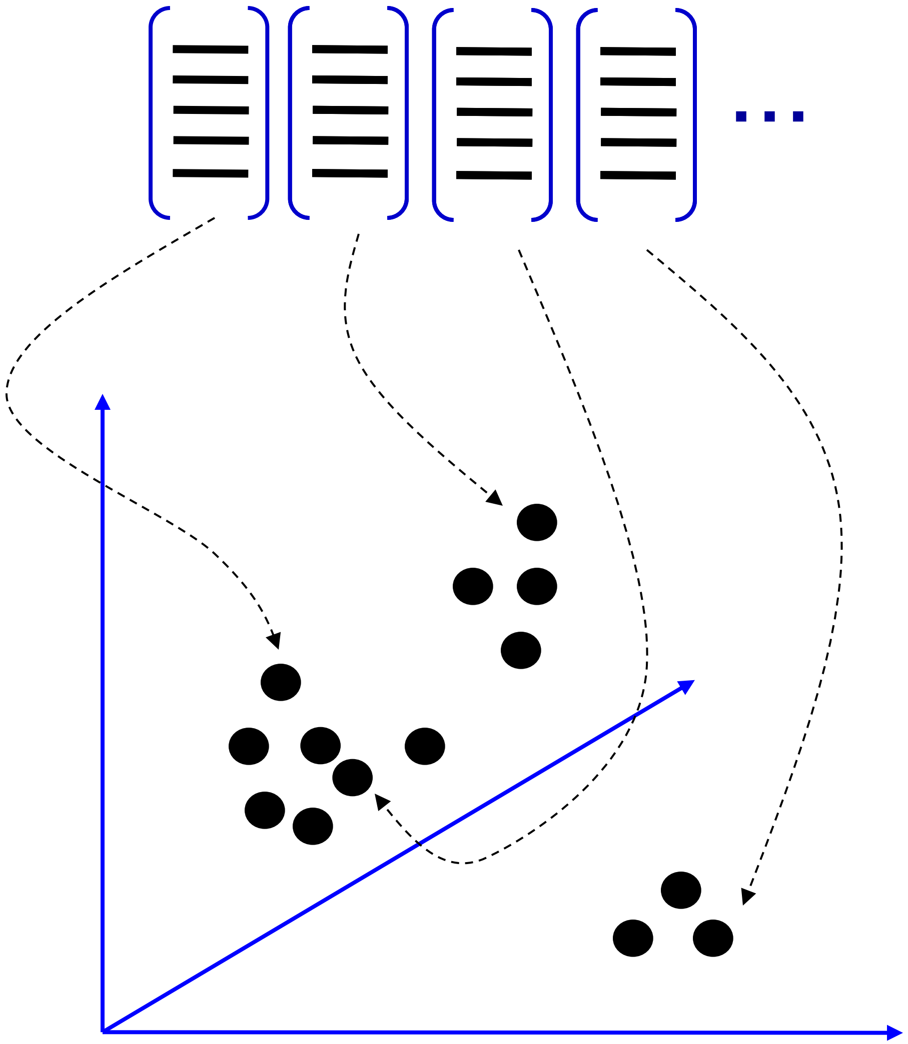


1. LOCAL FEATURE EXTRACTION

Sample patches and extract descriptors



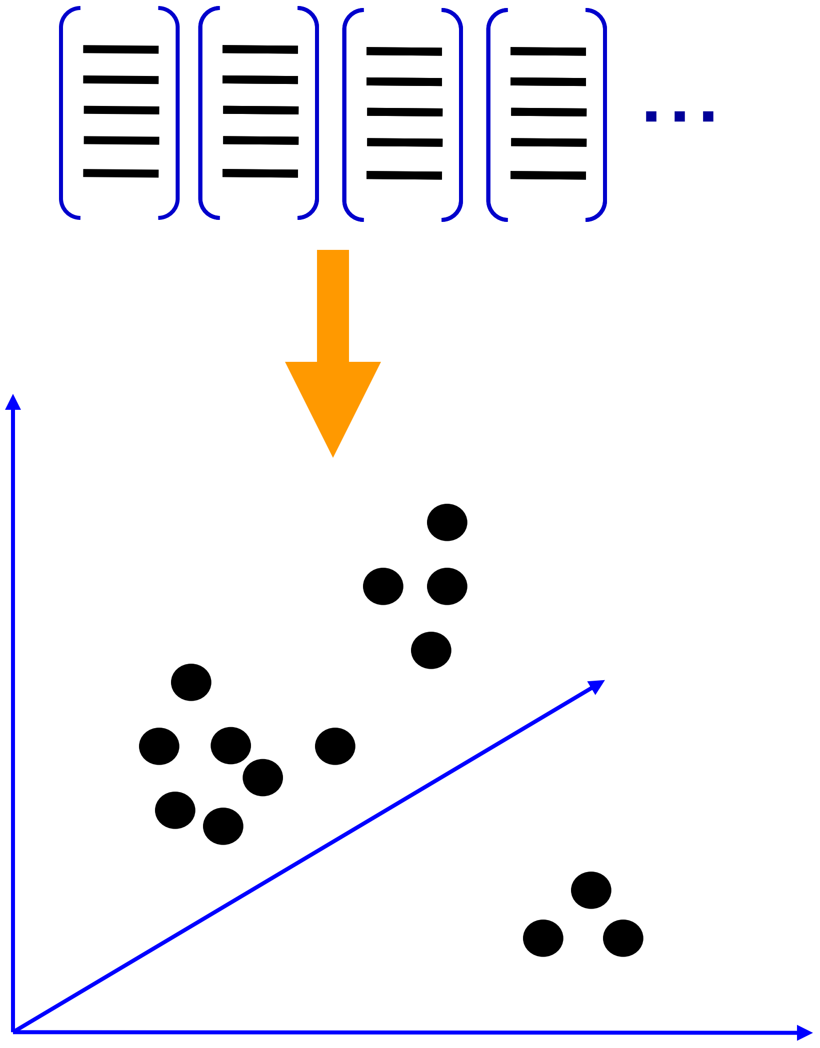
2. LEARNING THE VISUAL VOCABULARY



Extracted descriptors from the training set

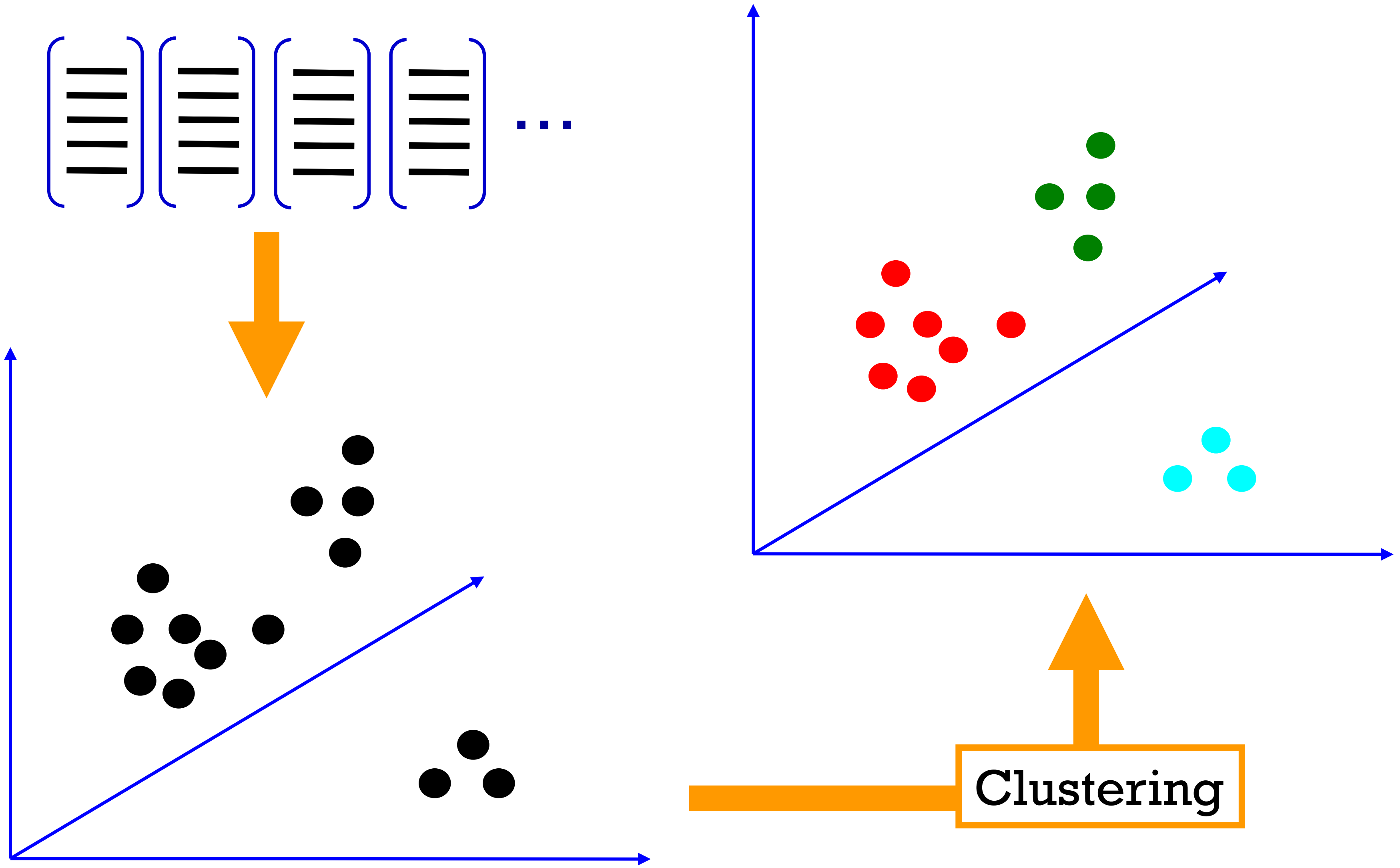
Slide credit: Josef Sivic

2. LEARNING THE VISUAL VOCABULARY



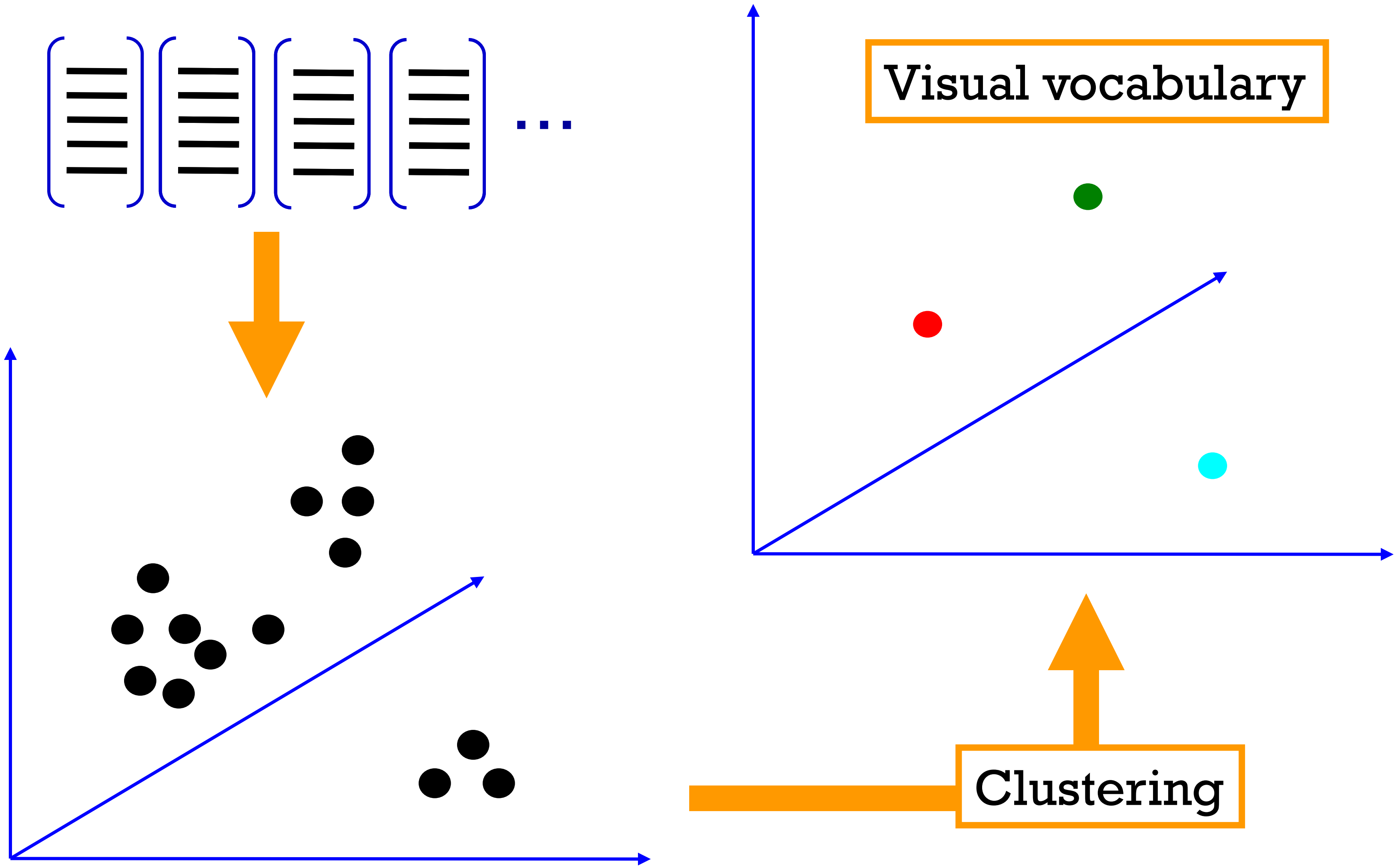
Slide credit: Josef Sivic

2. LEARNING THE VISUAL VOCABULARY



Slide credit: Josef Sivic

2. LEARNING THE VISUAL VOCABULARY



Slide credit: Josef Sivic

REVIEW: K-MEANS CLUSTERING

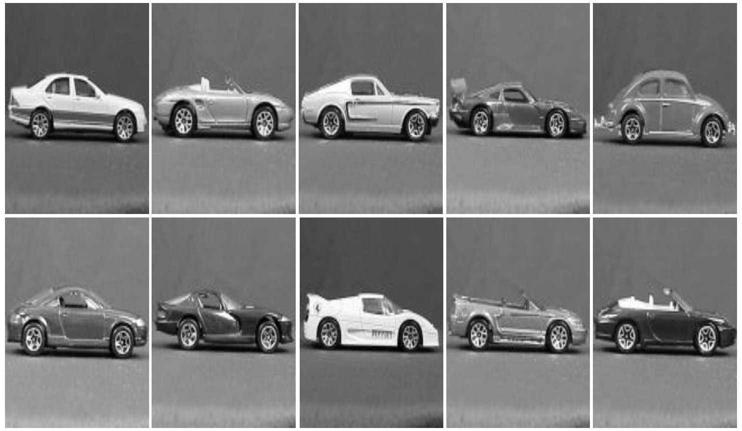
Want to minimize sum of squared Euclidean distances between features $\mathbf{x}_i$ and their nearest cluster centers $\mathbf{m}_k$

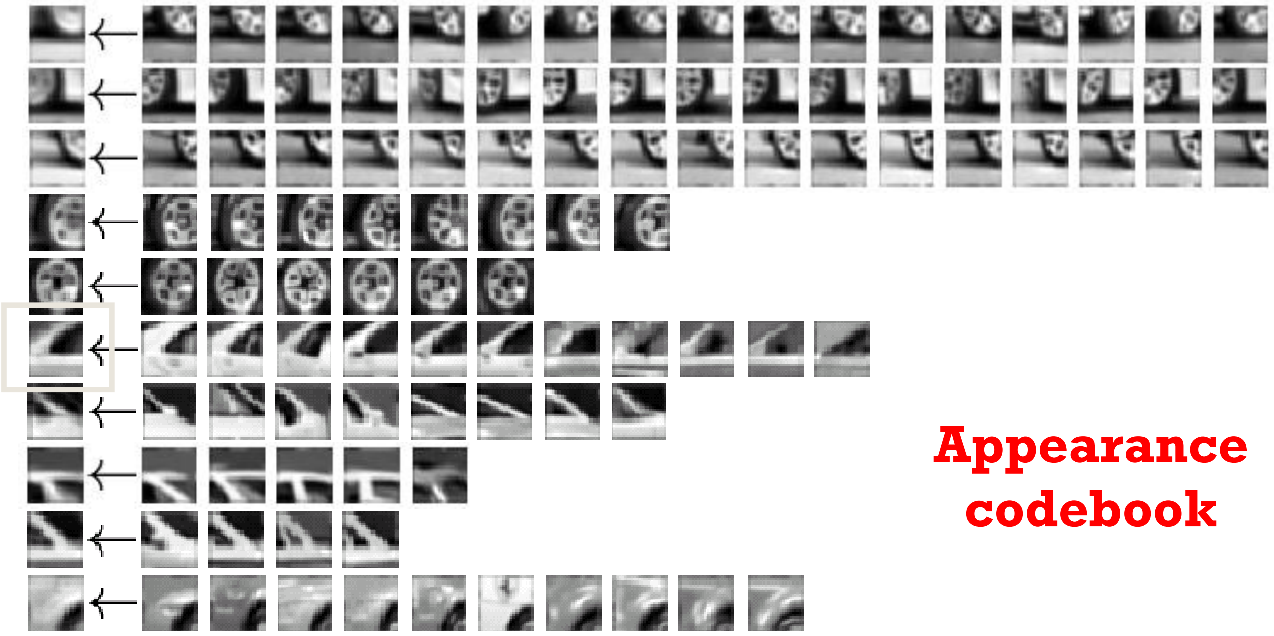
$$D(X,M) = \sum_{\text{cluster } k} \sum_{\substack{\text{point } i \text{ in} \\ \text{cluster } k}} (\mathbf{x}_i - \mathbf{m}_k)^2$$

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





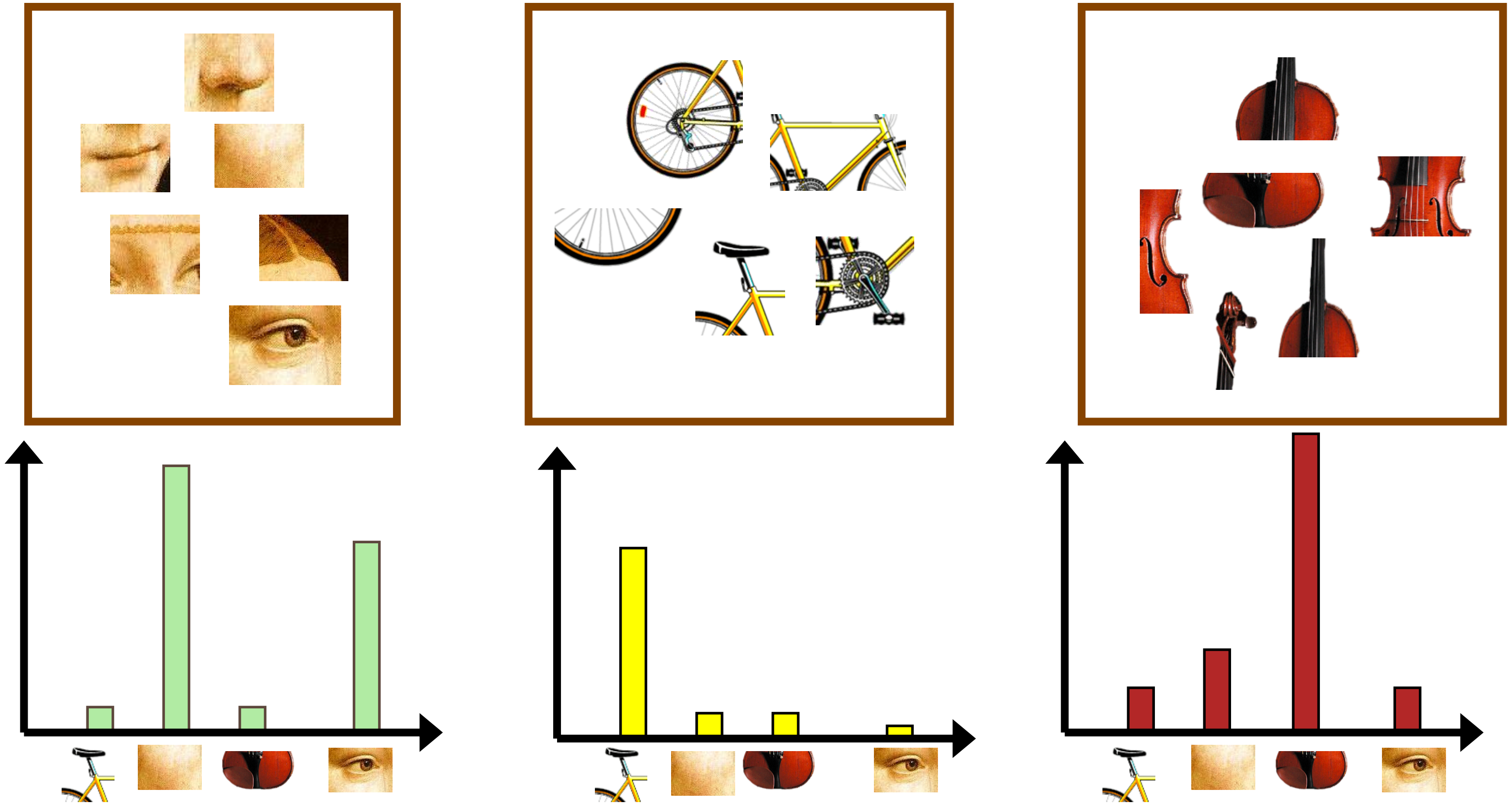
...

Appearance codebook

Source: B. Leibe

BAG-OF-FEATURES STEPS

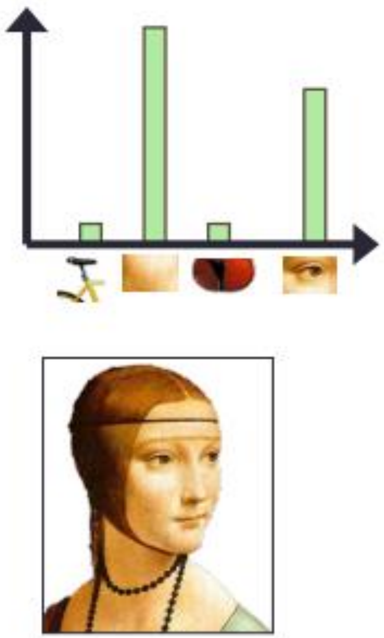
1. Extract local features
2. Learn "visual vocabulary"
3. **Quantize local features using visual vocabulary**
4. **Represent images by frequencies of "visual words"**



COMPARING BAGS OF WORDS

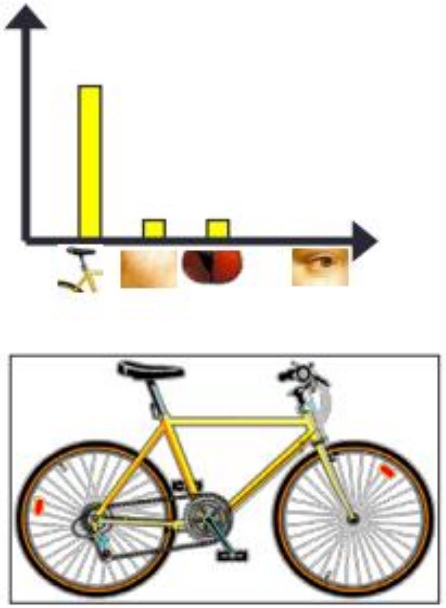
- Rank frames by normalized scalar product between their (possibly weighted) occurrence counts---*nearest neighbor* search for similar images.

[1 8 1 4]



$\vec{d}_j$

[5 1 1 0]



$\vec{q}$

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

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

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

Testing

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

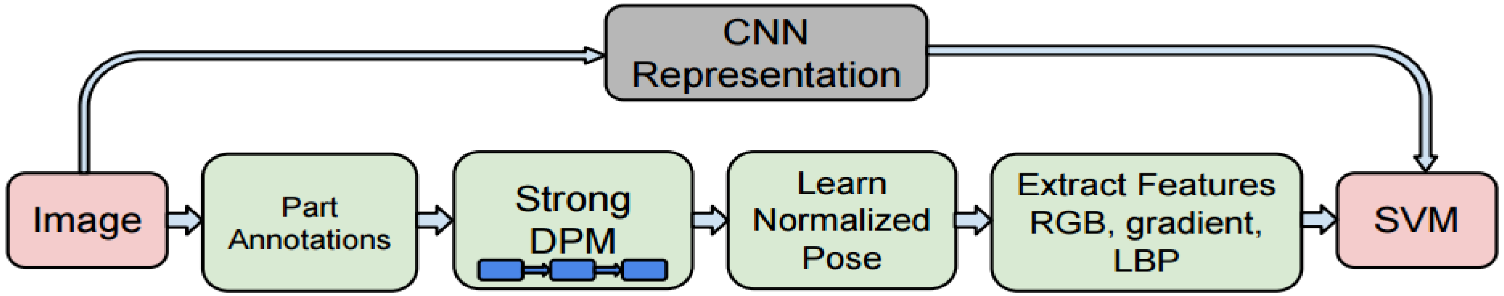
OBJECT CLASSIFICATION WITH BAG OF WORDS

- 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**

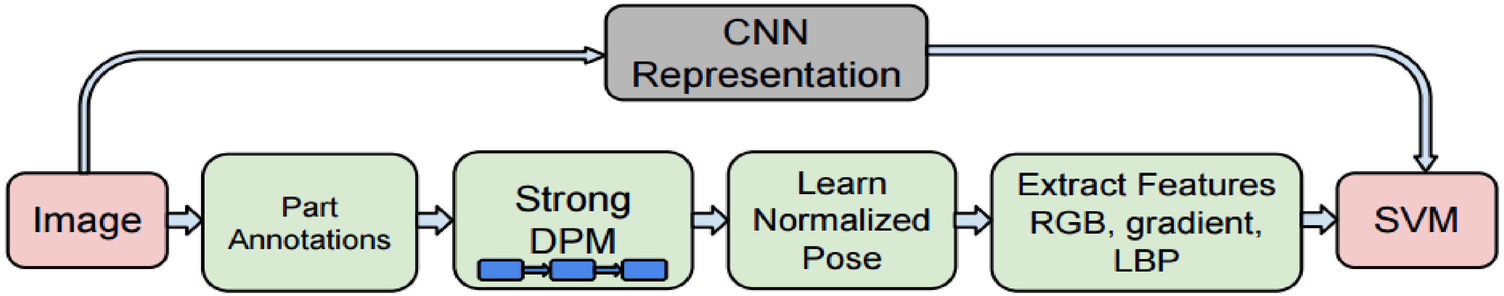
[Csurka et al., ‘04]

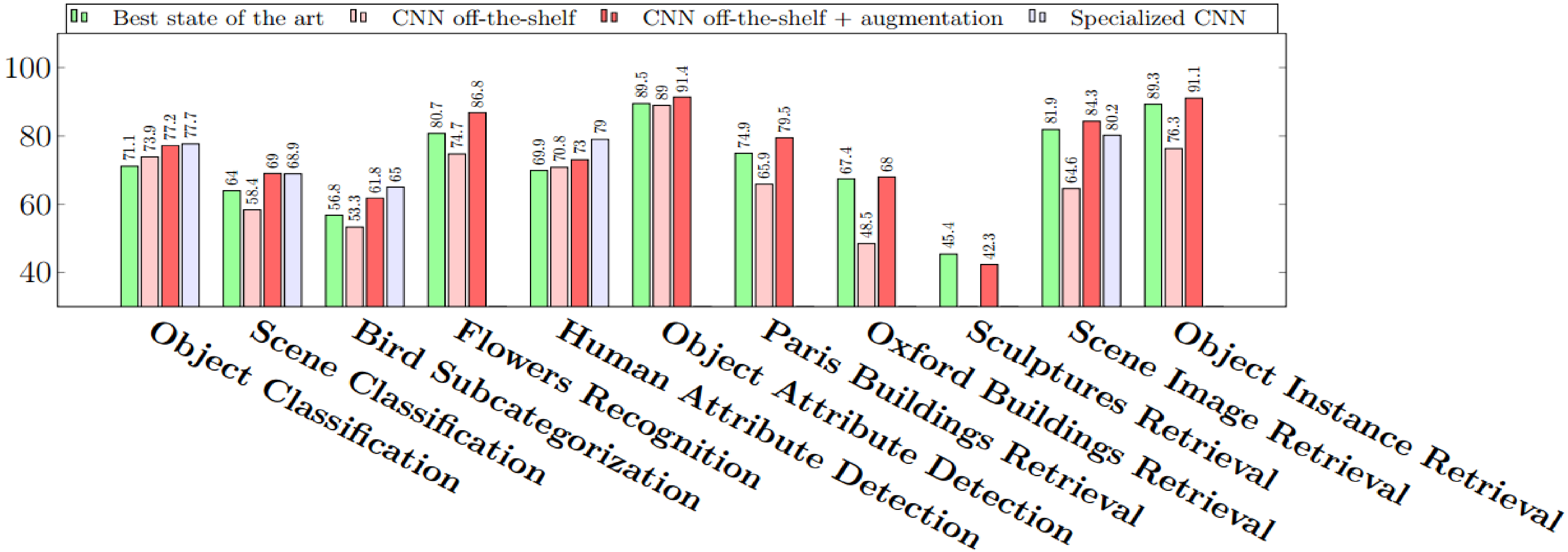
CONVOLUTIONAL ACTIVATION FEATURES



CNN Features off-the-shelf: an Astounding Baseline for Recognition [Razavian et al. 2014]

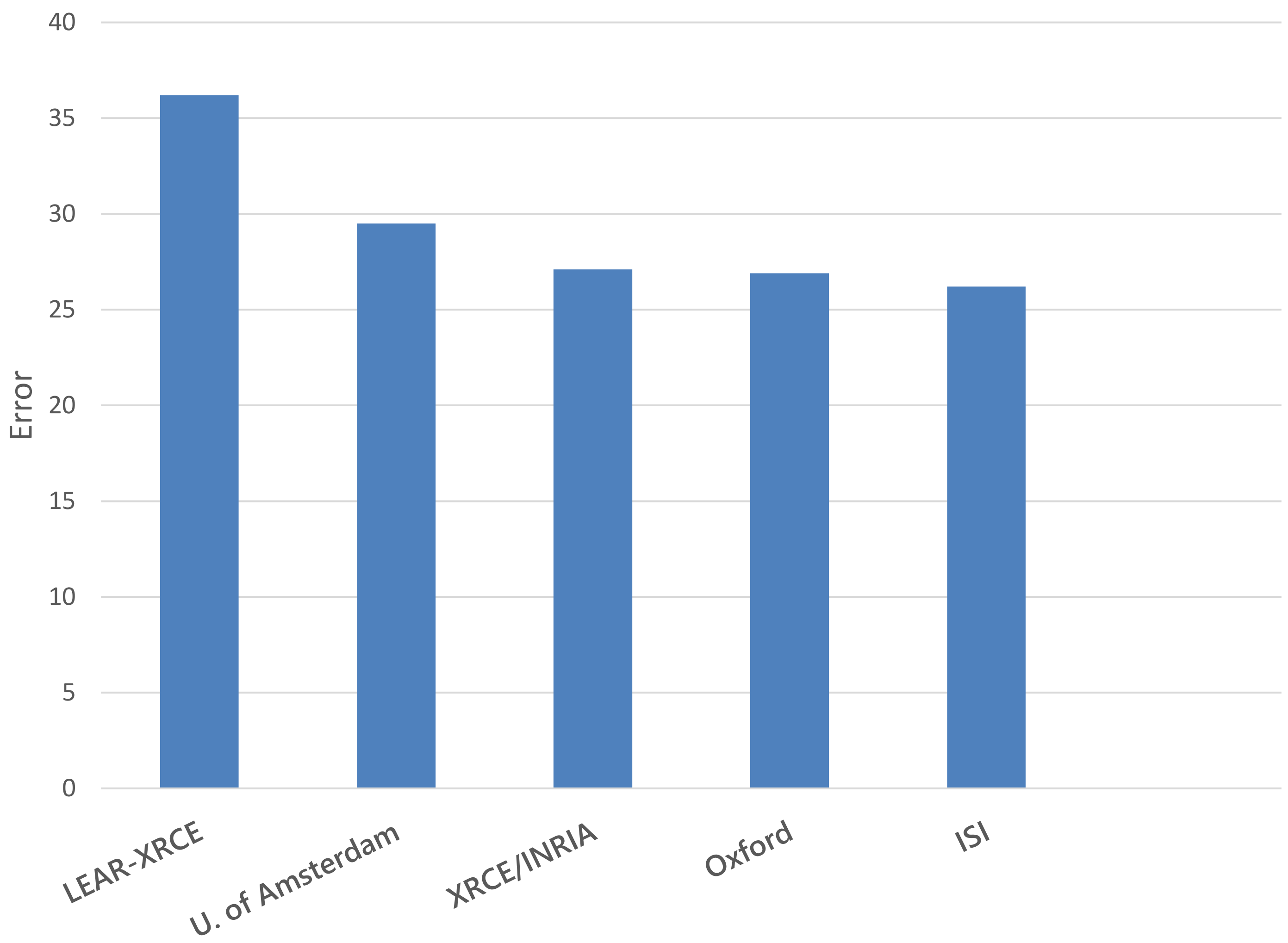
CONVOLUTIONAL ACTIVATION FEATURES





CNN Features off-the-shelf: an Astounding Baseline for Recognition [Razavian et al. 2014]

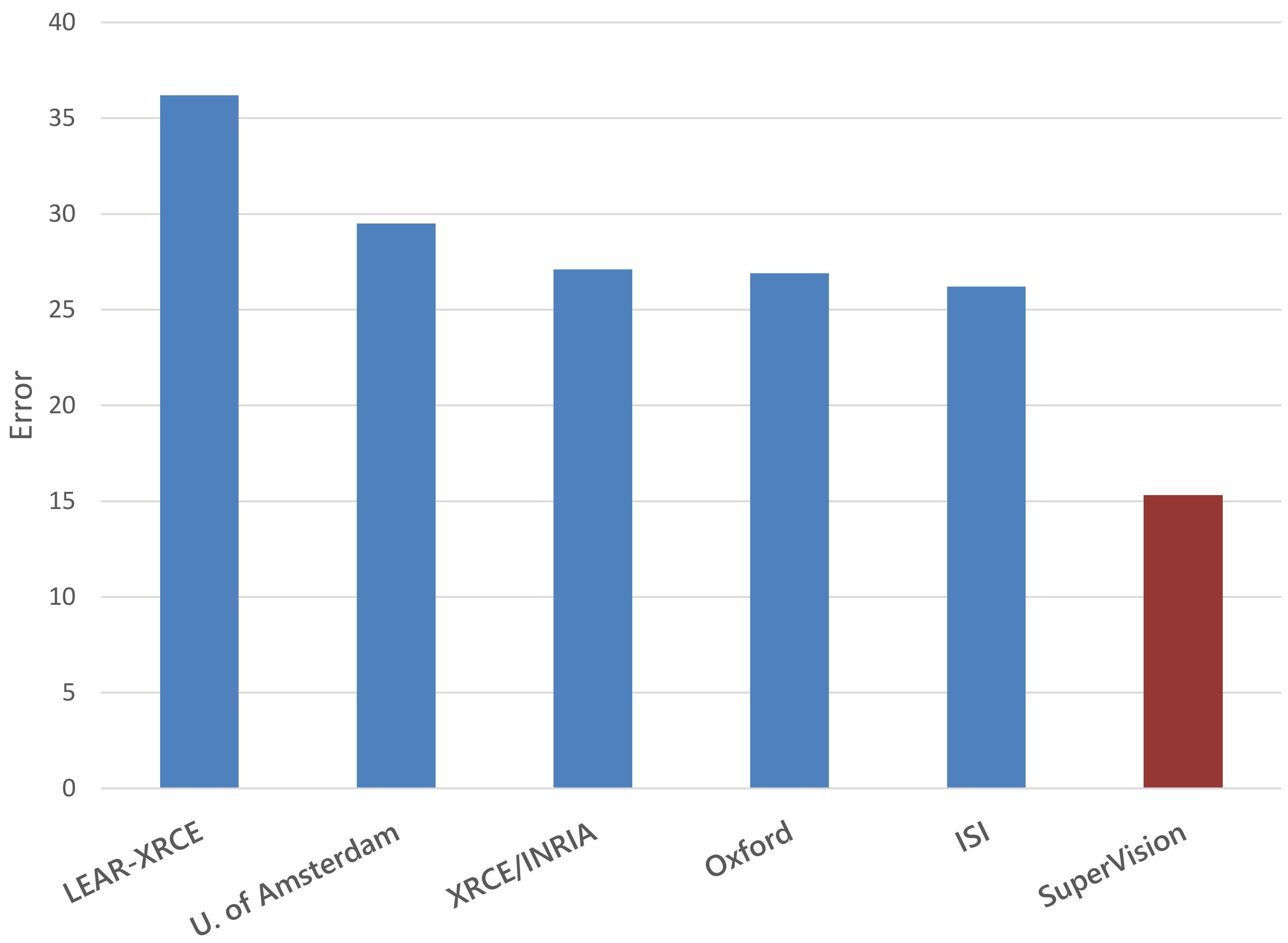
ImageNet 1K



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

ImageNet 1K

(Fall 2012)



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

ACKNOWLEDGEMENT

Thanks to the following courses and corresponding researchers for making their teaching/research material online

- Convolutional Neural Networks for Visual Recognition, Stanford University
- Deep Learning, Stanford University
- Introduction to Deep Learning, University of Illinois at Urbana-Champaign
- Introduction to Deep Learning, Carnegie Mellon University
- Natural Language Processing with Deep Learning, Stanford University
- And Many More Publicly Available Resources ……

Questions?