The Analysis of Patterns 2009 - Pula Science Park, Cagliari, Italy

Pattern Recognition in Computer Vision

Giovanni Maria Farinella

www.dmi.unict.it/farinella

gfarinella@dmi.unict.it

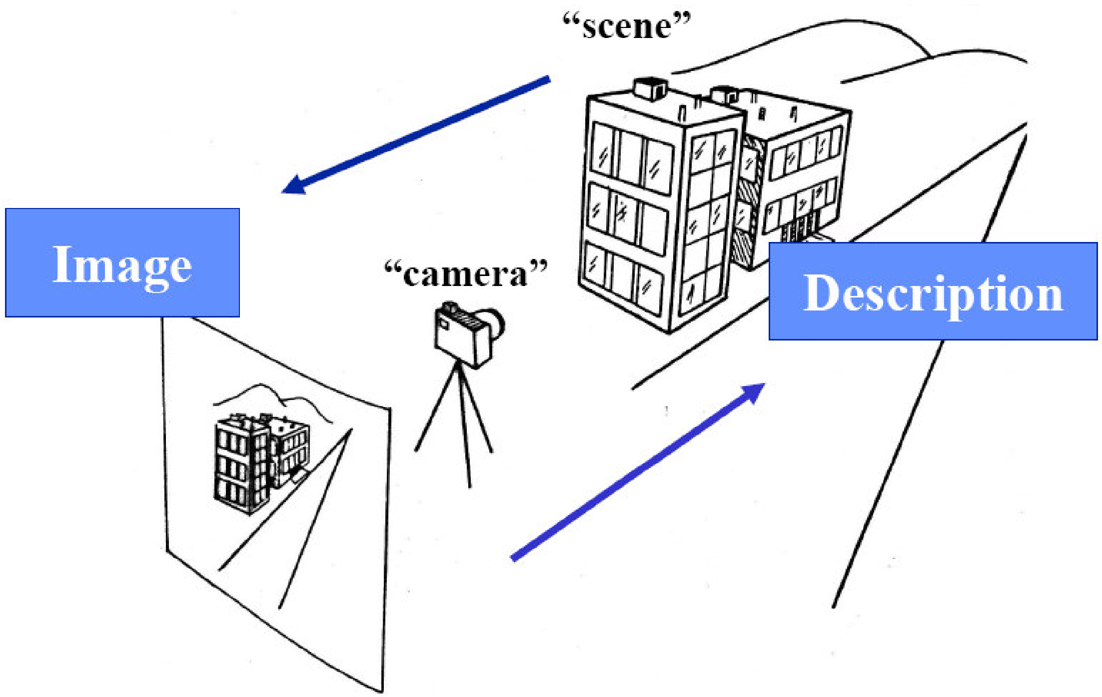
IMAGE PROCESSING LABORATORY

www.dmi.unict.it/iplab

Outline and Goals

- Outline of this seminars:
  - Computer Vision
    - What does it mean?
    - Why it is hard?
  - Recognition in Computer Vision
    - Categorization
    - Identification
    - Parameter Estimation
  - Categorization
    - Bag of Visual Words Model
    - Examples of Application
- Goals of this seminars:
  - Give brief introduction of the field.
  - Show how some PR methods have been used in vision.
  - Provide references and pointers.

What is Vision?

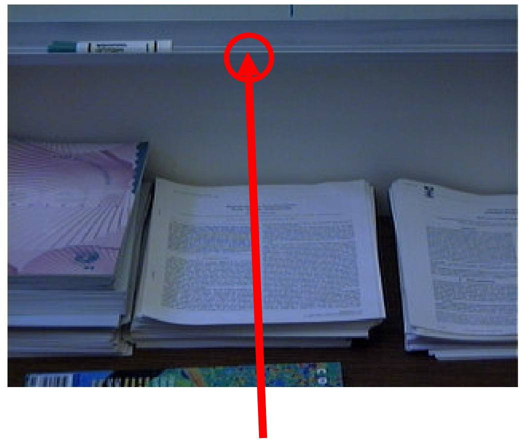


"What does it mean, to see? The plain man's answer (and Aristotle's, too) would be, *to know* ***<u>what</u>*** *is* ***<u>where</u>*** *by looking*."

David Marr, *Vision (1982)*

What do we want?

Vision is the process of discovering from images <u>what</u> is present in the world, and <u>where</u> it is.



Answer #1: *pixel of brightness 243 at position (124,54)*

Answer #2: *looks like bottom edge of whiteboard showing at the top of the image*

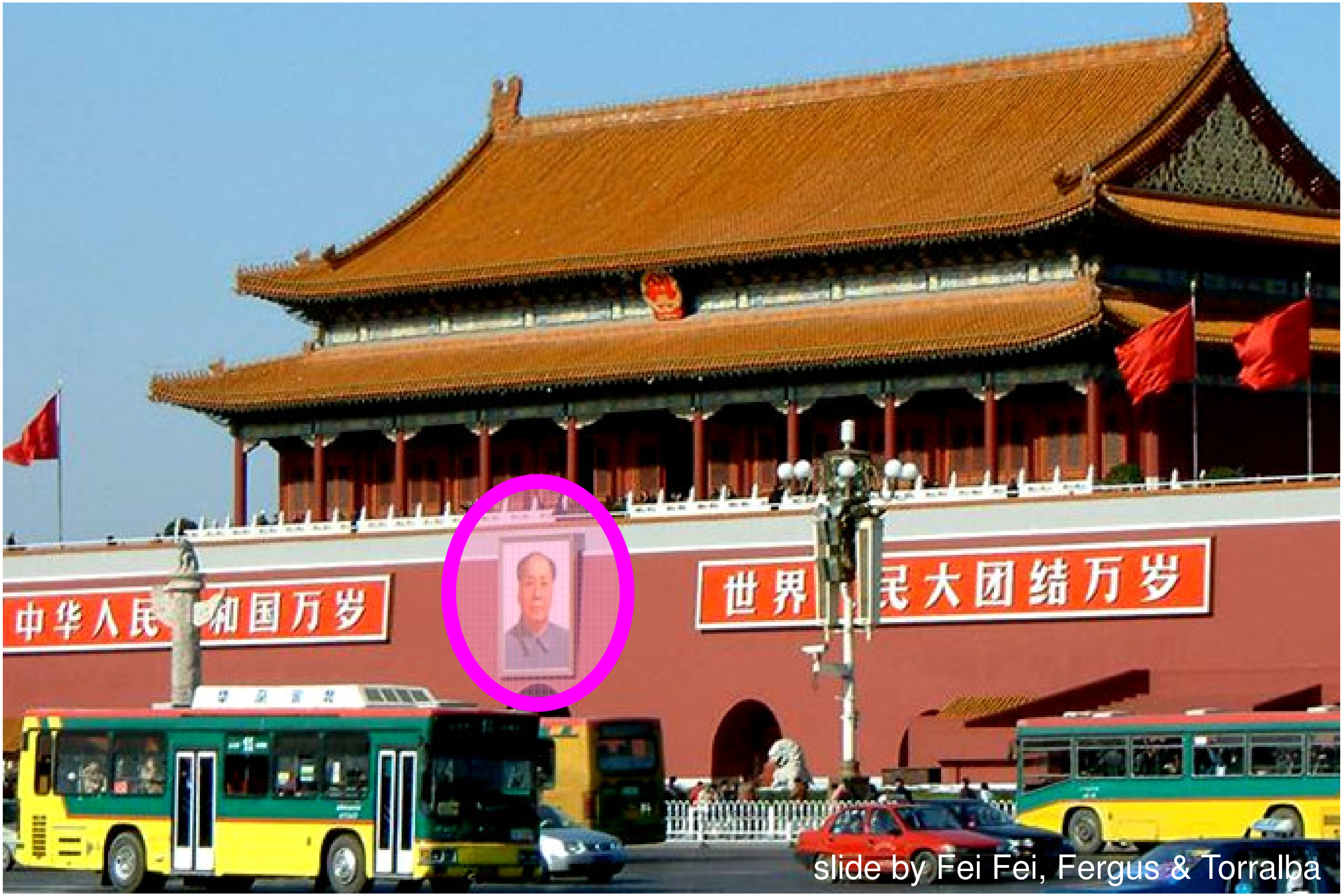
The goals of computer vision (what + where) are in terms of what <u>humans</u> care about.

So what do humans care about?

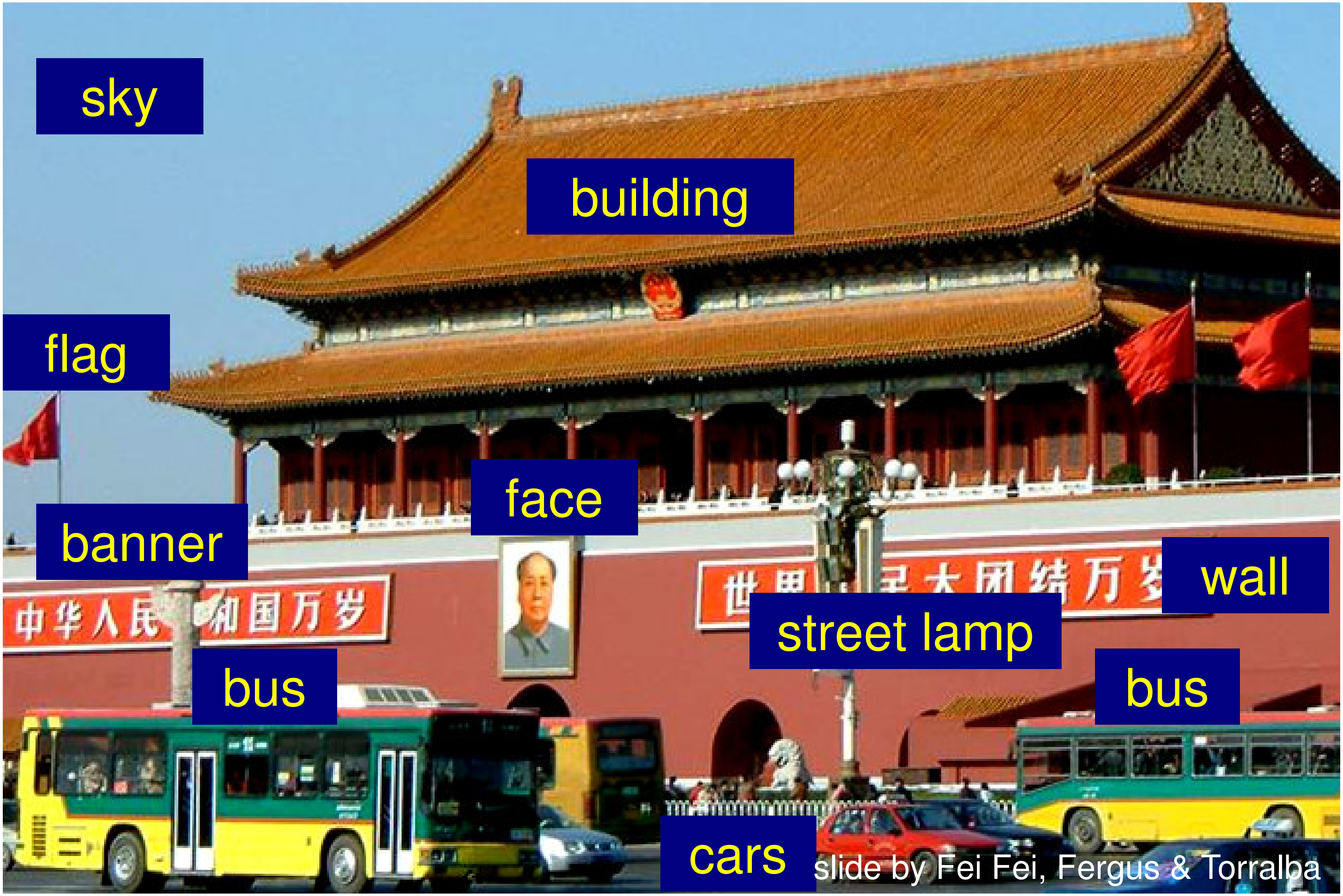
Verification: is that a bus?

Detection: are there cars?

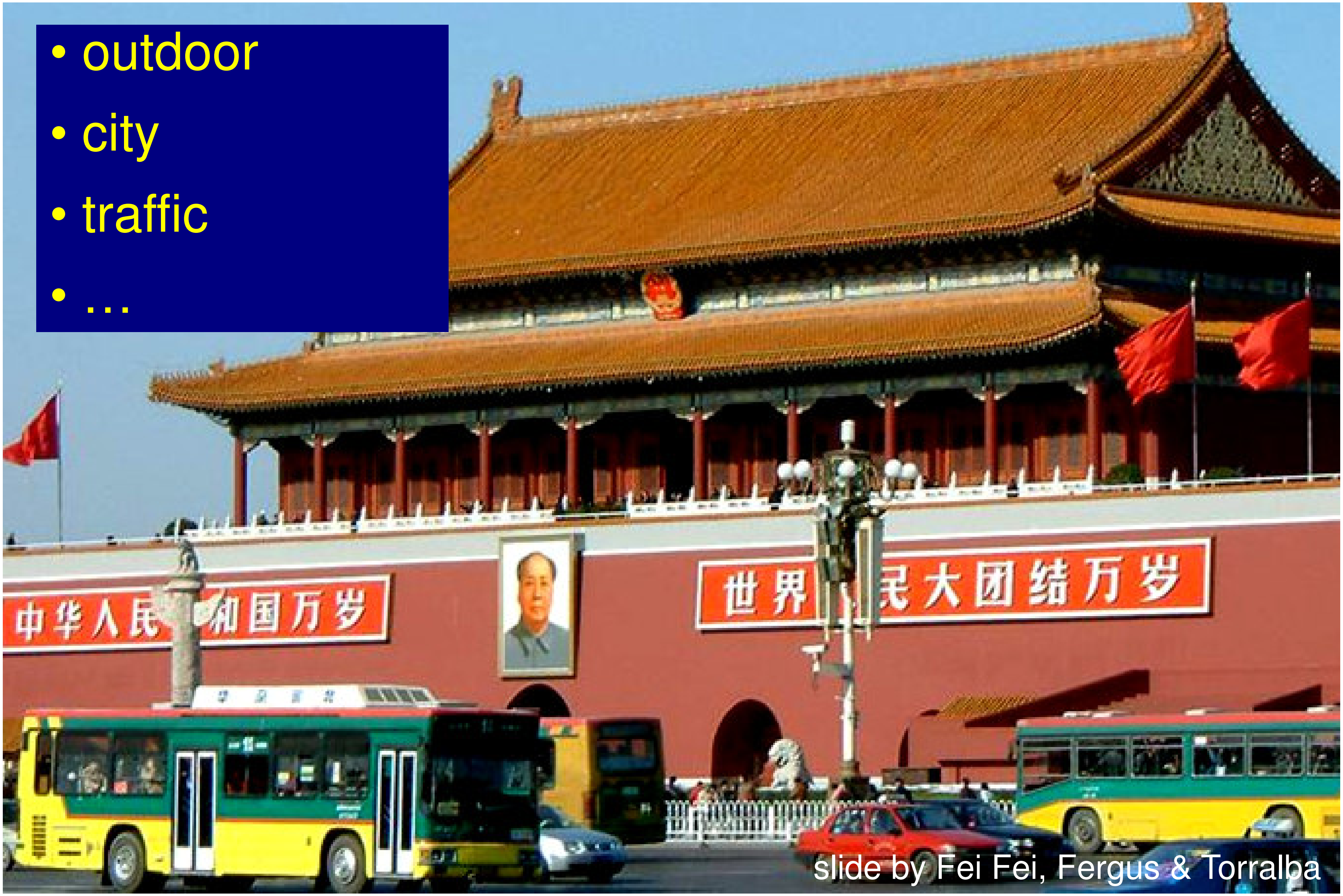
Identification: is that a picture of Mao?



Object categorization



Scene and context categorization



The Computer Vision Industry

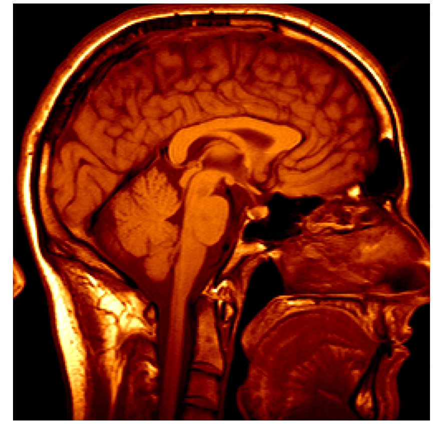
See: http://people.cs.ubc.ca/~lowe/vision.html

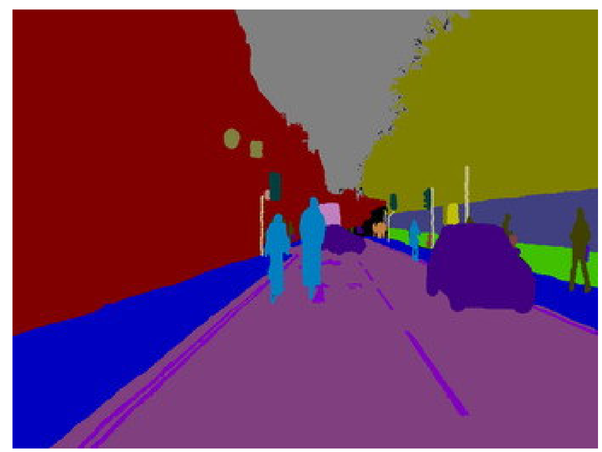
Google images

bing VISUAL SEARCH GALLERIES BETA

flickr

YouTube





Pattern Recognition in Computer Vision

- Humans can understand an observed scene effortlessly, but this is still a daunting challenge for computers-based scene understanding systems.
- Computer Vision aims at devising robust and reusable vision systems.
- Vision systems that learn and adapt represent one of the most important trend in Computer Vision.
- Pattern Recognition is an essential part in the study of Computer Vision.

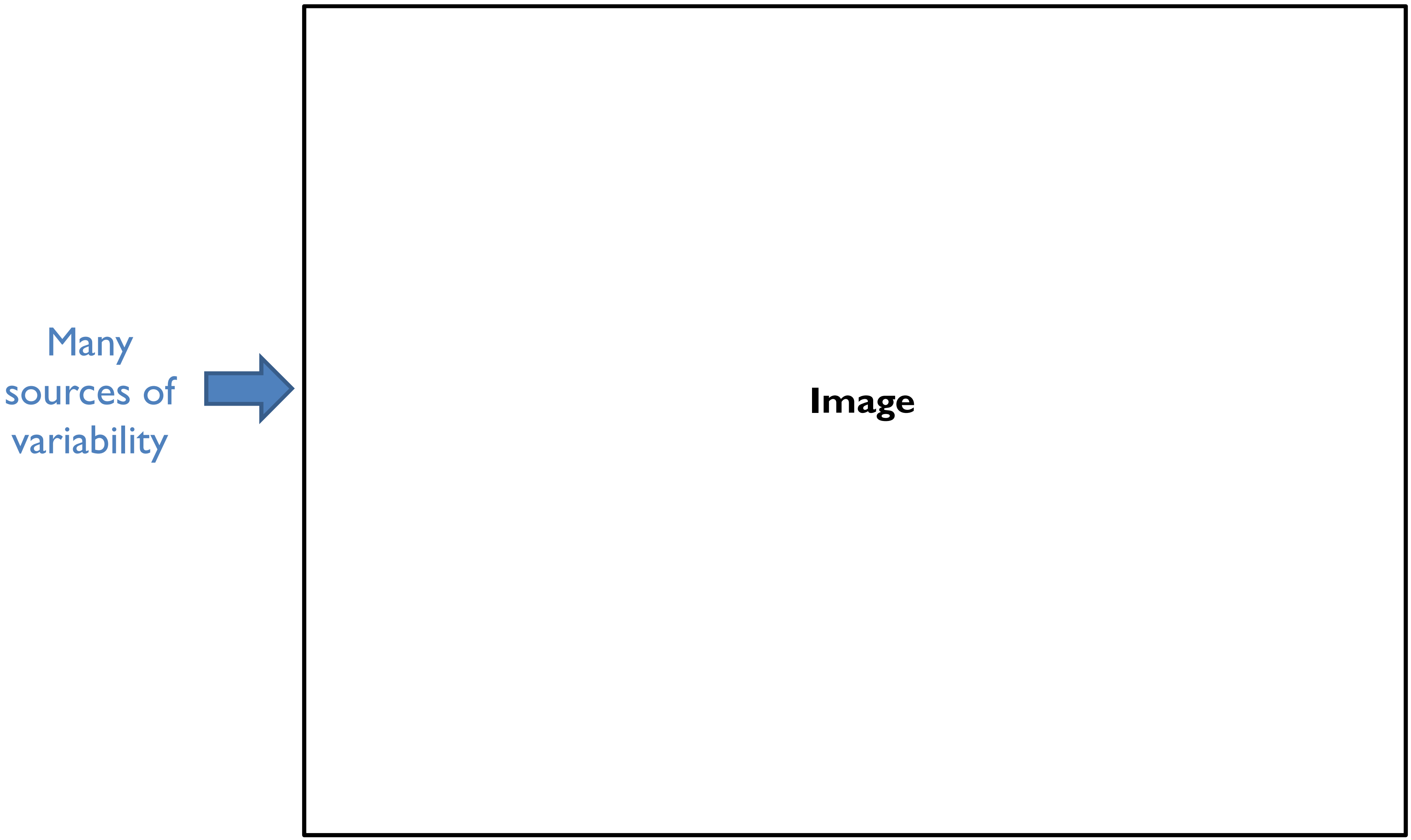
“Pattern Recognition” approach to Computer Vision

- Feature vector representation of an image:
  - invariant or quasi-invariant to some class of transformations, e.g., affine invariant features, histogram (color, gradient)
- Reduction of the space dimensionality
  - e.g., PCA, NMF, Sparse Representation
- Data-driven by using statistical learning and decision-making mechanism
  - Bayesian methods, Discriminative methods, Graphical models, X-RF

Why Computer Vision is so difficult?

- Data: NTSC Video ~ 20 MB/sec
- Degeneracy: Inverting projection is “theoretically“ impossible!
- Knowledge and Context are key component for understanding content of images
- Compoundedness: a pixel value results from many combined factors (atmosphere effect, viewing angle, lighting, materials). Many sources of image variability.

Sources of image variability

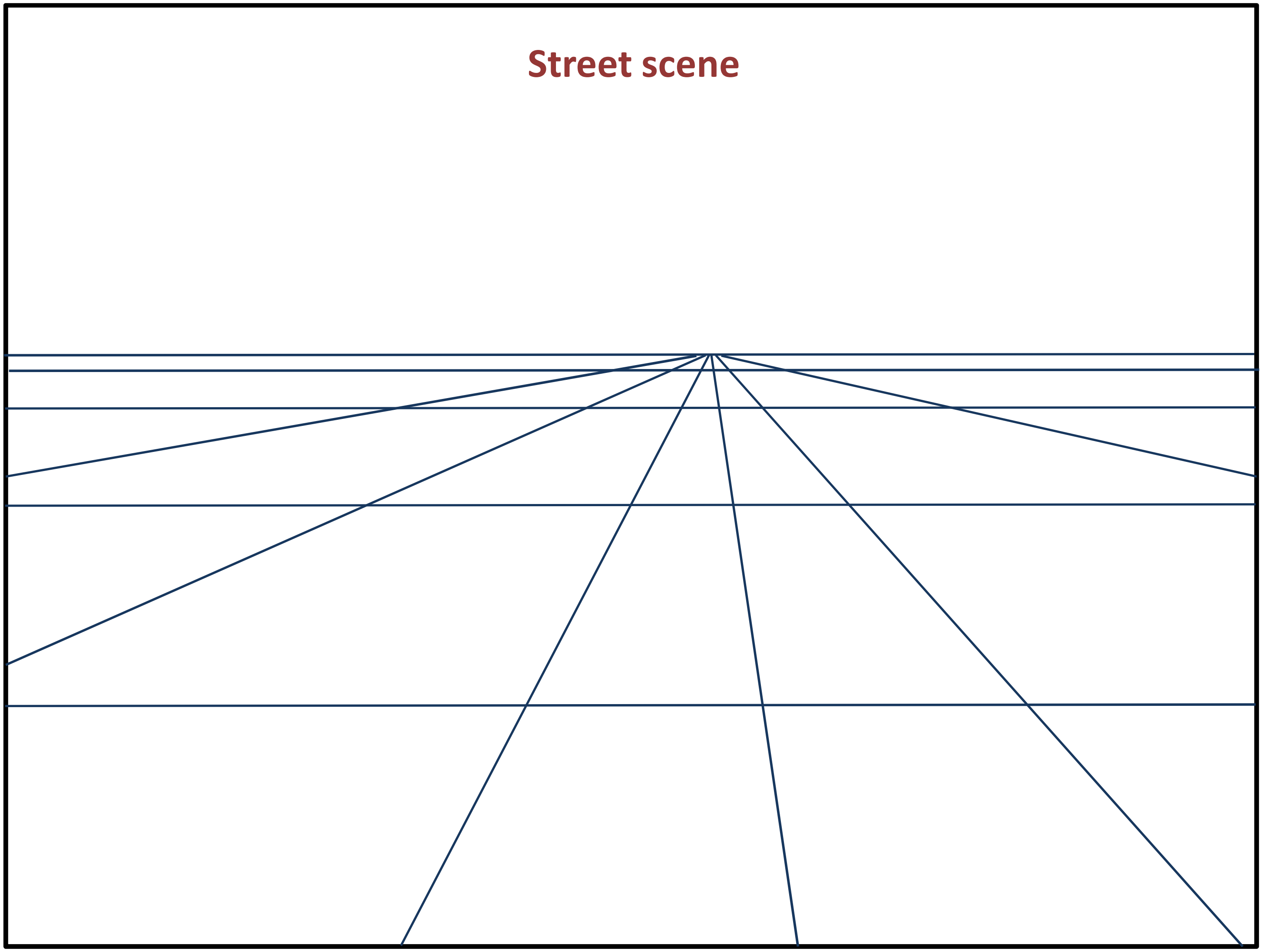


slide by John Winn

Sources of image variability

Scene type

Scene geometry



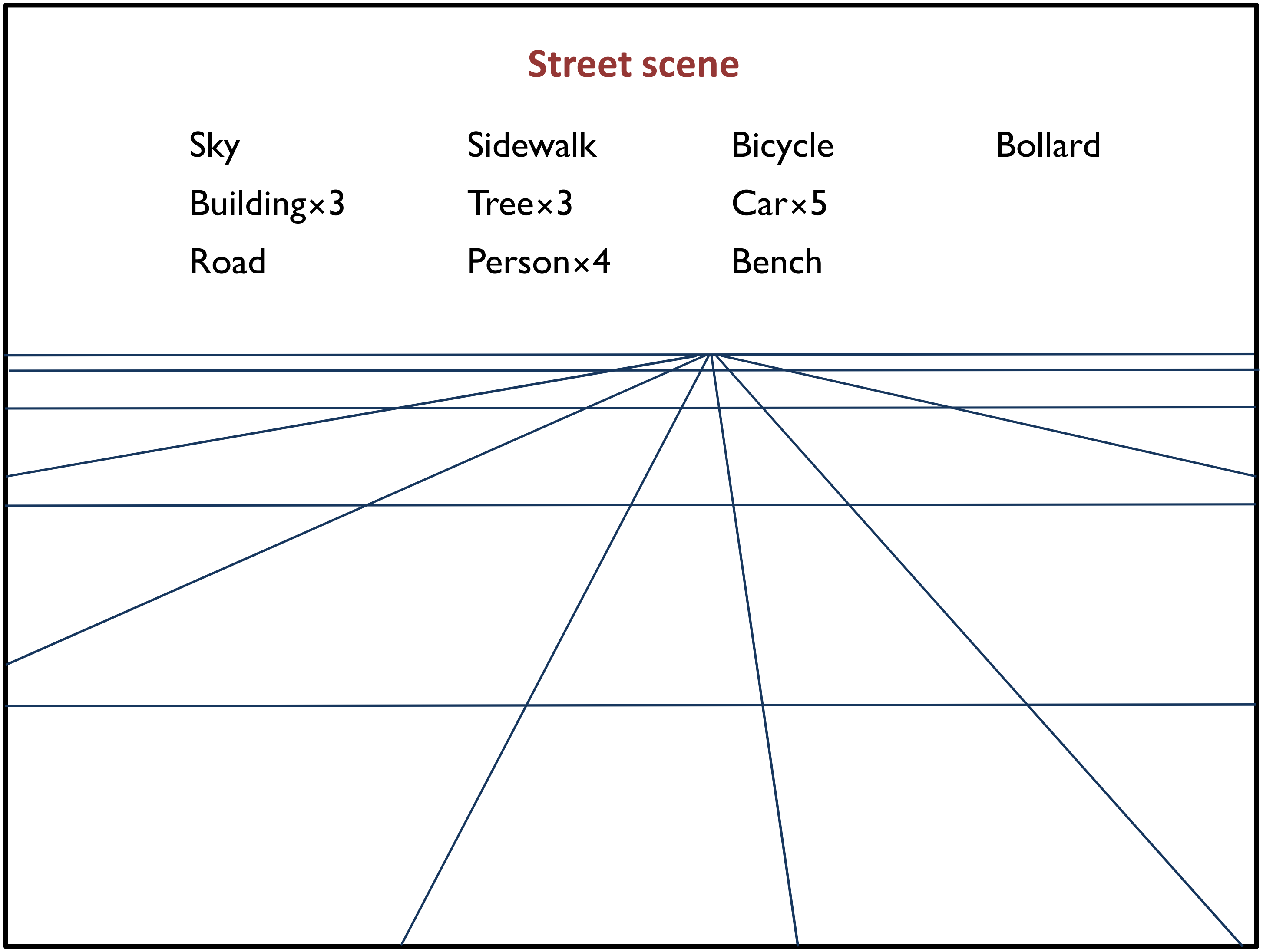
slide by John Winn

Sources of image variability

Scene type

Scene geometry

Object classes



slide by John Winn

Sources of image variability

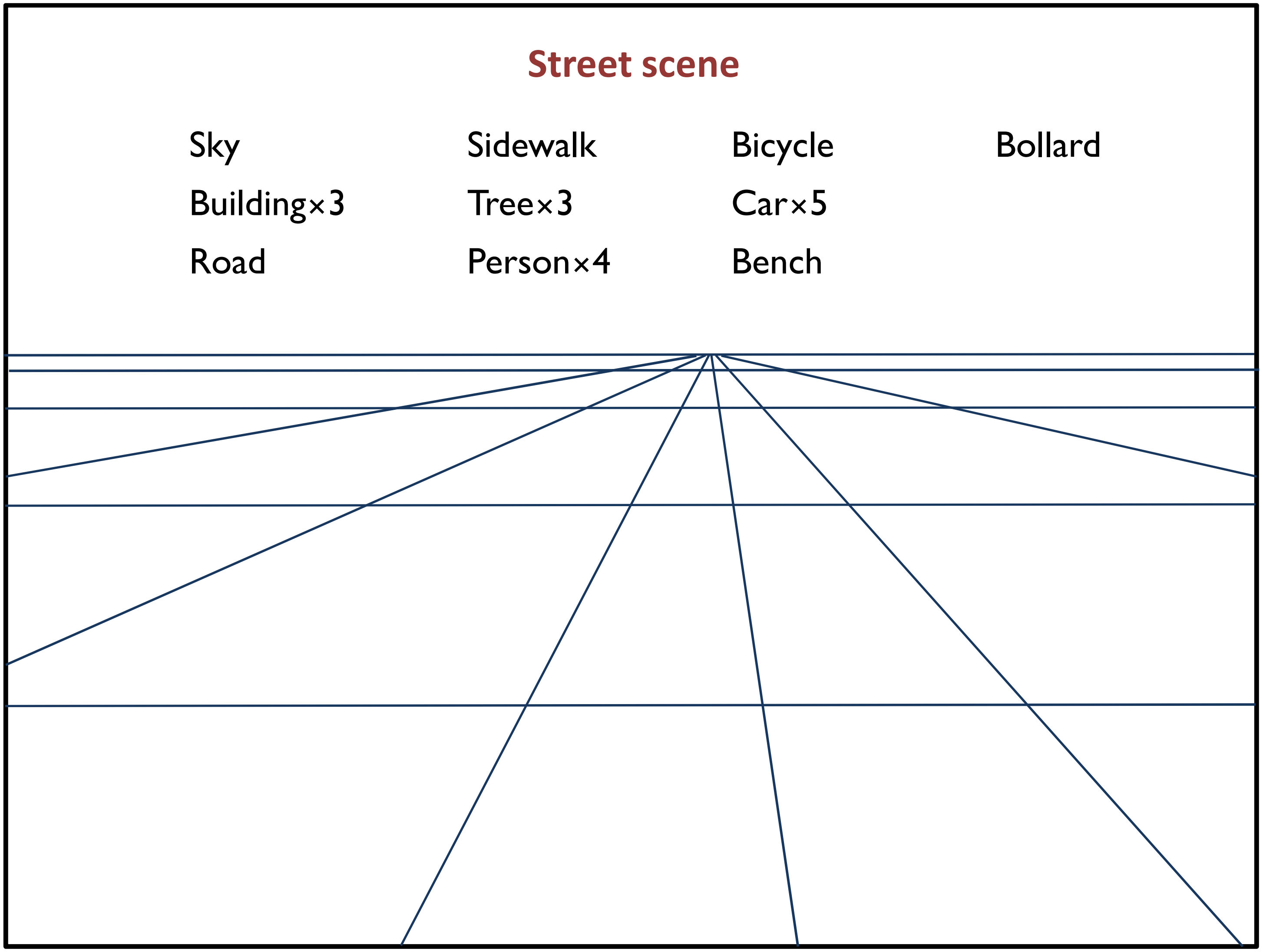
Scene type

Scene geometry

Object classes

Object position

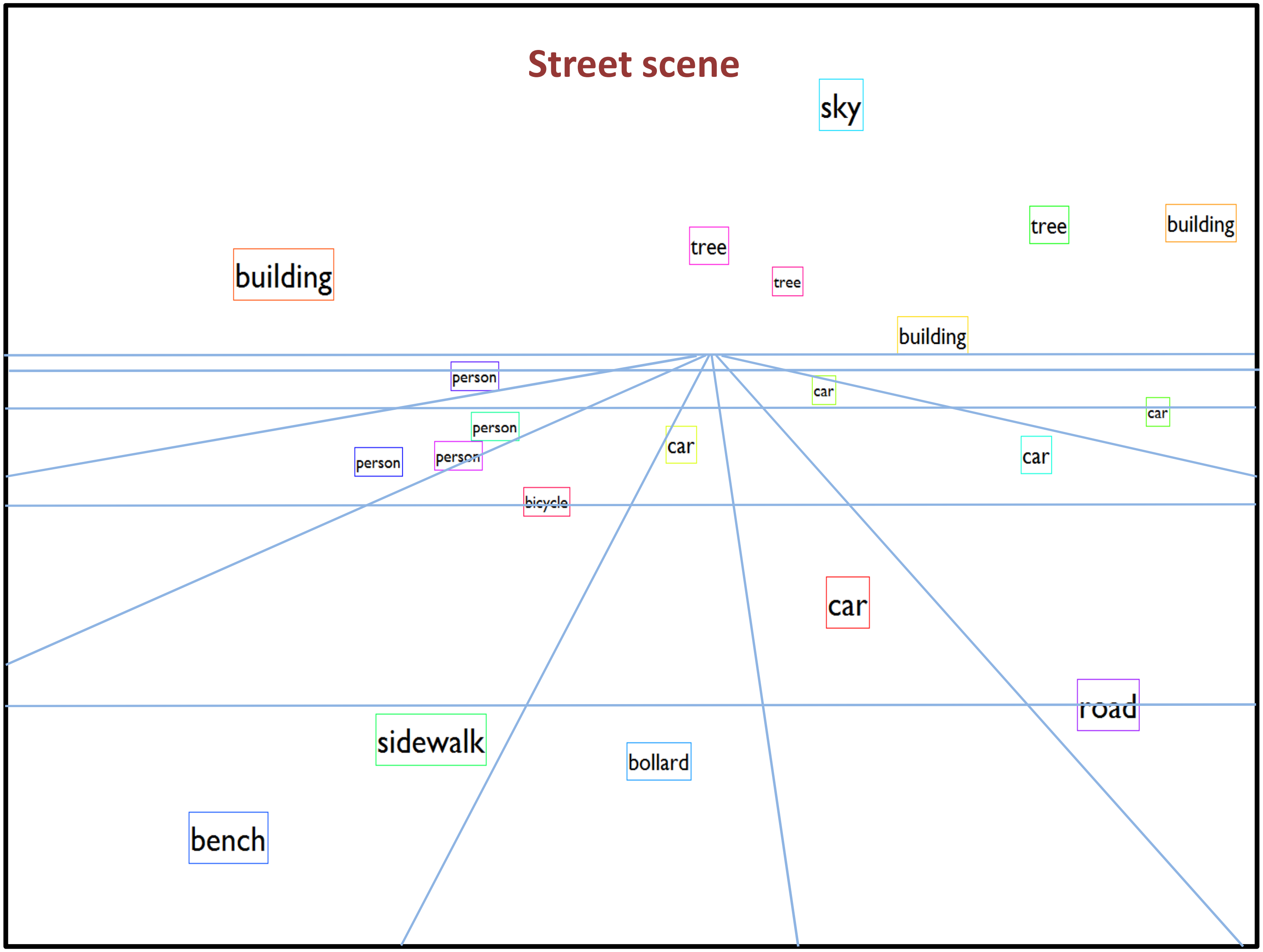
Object orientation



slide by John Winn

Sources of image variability

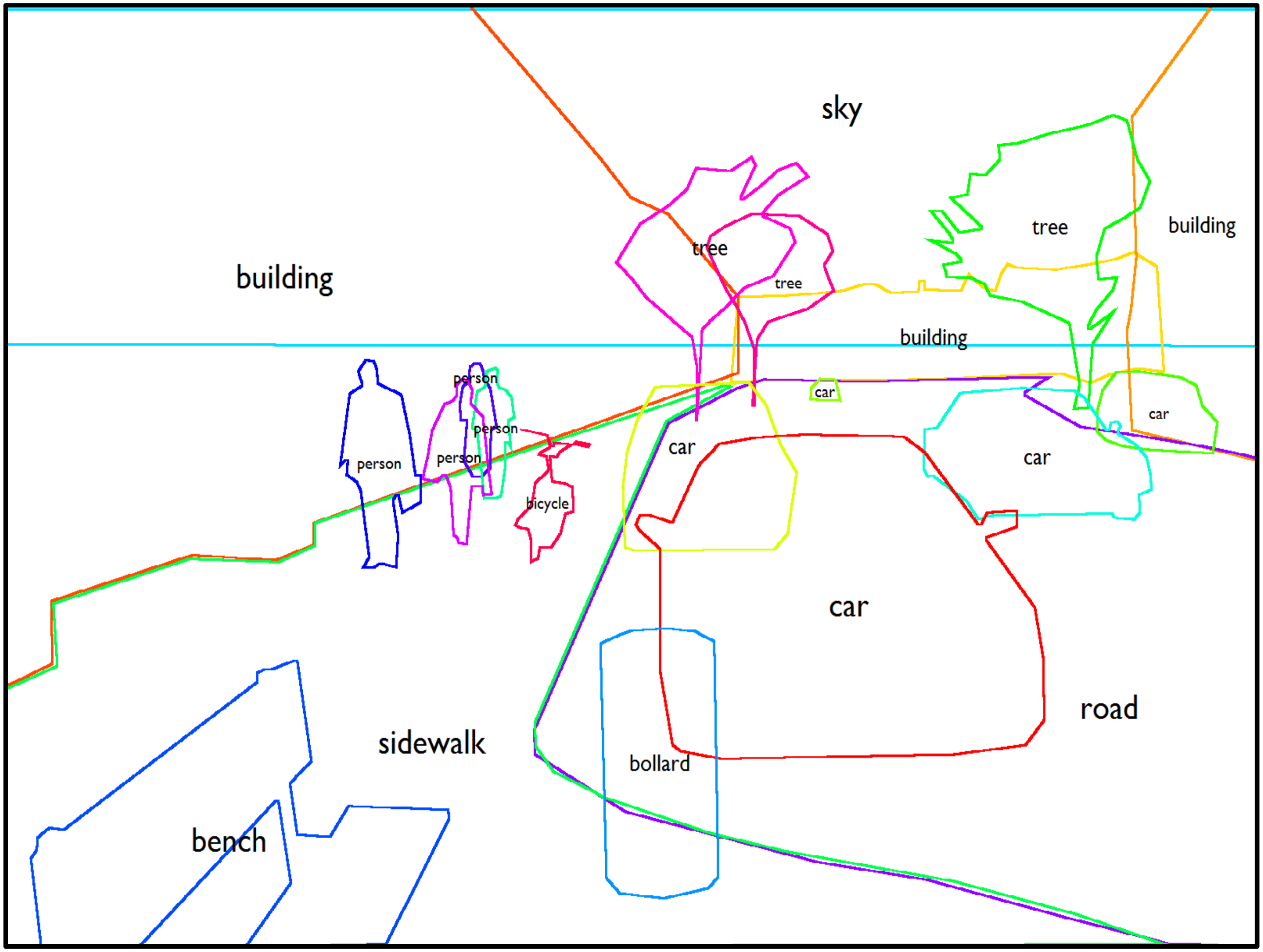
Scene type
Scene geometry
Object classes
Object position
Object orientation
Object shape



slide by John Winn

Sources of image variability

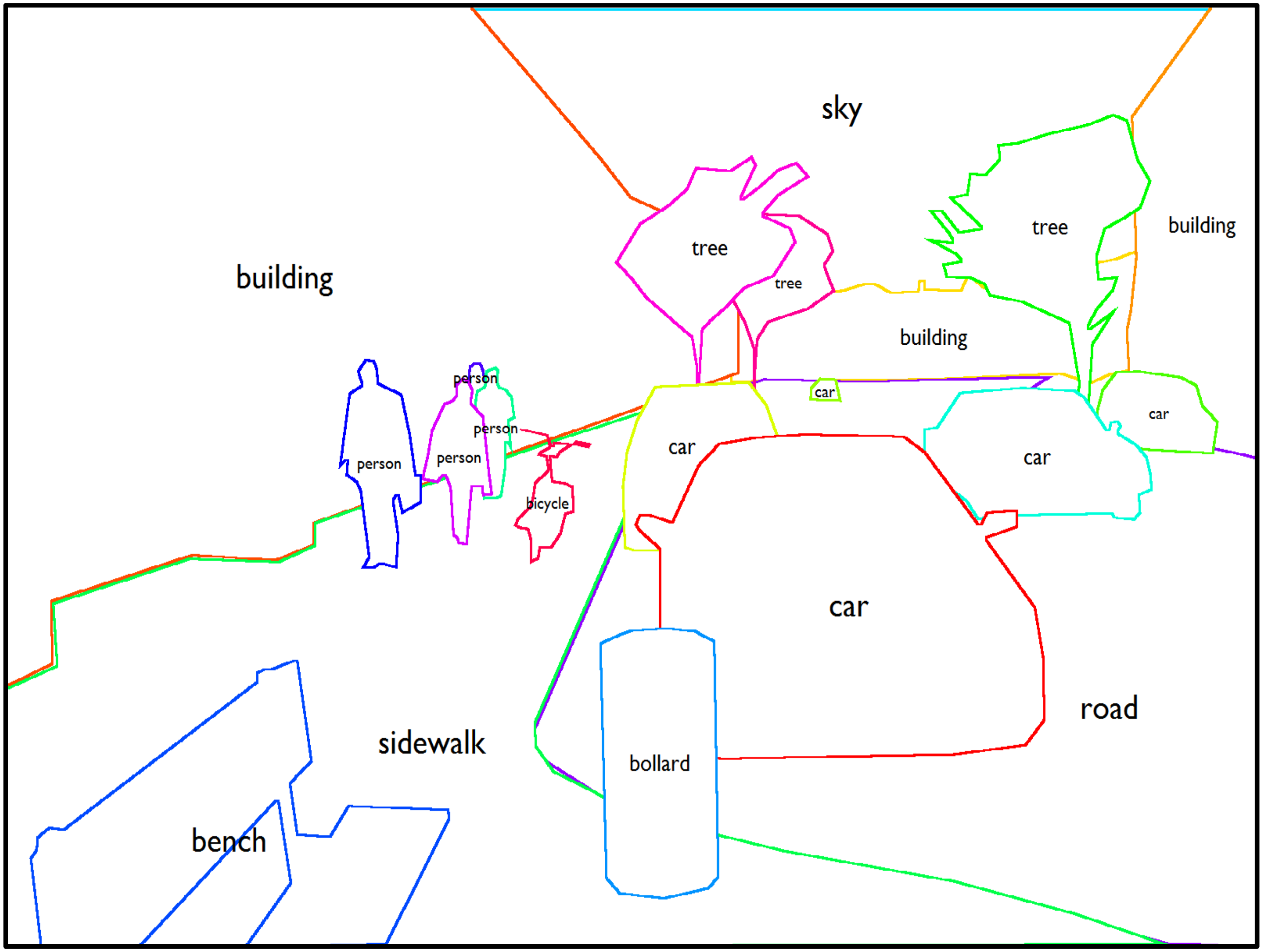
Scene type
Scene geometry
Object classes
Object position
Object orientation
Object shape
Depth/occlusions



slide by John Winn

Sources of image variability

Scene type
Scene geometry
Object classes
Object position
Object orientation
Object shape
Depth/occlusions
Object appearance



slide by John Winn

Sources of image variability

Scene type

Scene geometry

Object classes

Object position

Object orientation

Object shape

Depth/occlusions

Object appearance

Illumination

Shadows

slide by John Winn

Sources of image variability

Scene type
Scene geometry
Object classes
Object position
Object orientation
Object shape
Depth/occlusions
Object appearance
Illumination
Shadows

slide by John Winn

Sources of image variability

Scene type
Scene geometry
Object classes
Object position
Object orientation
Object shape
Depth/occlusions
Object appearance
Illumination
Shadows
Motion blur
Camera effects

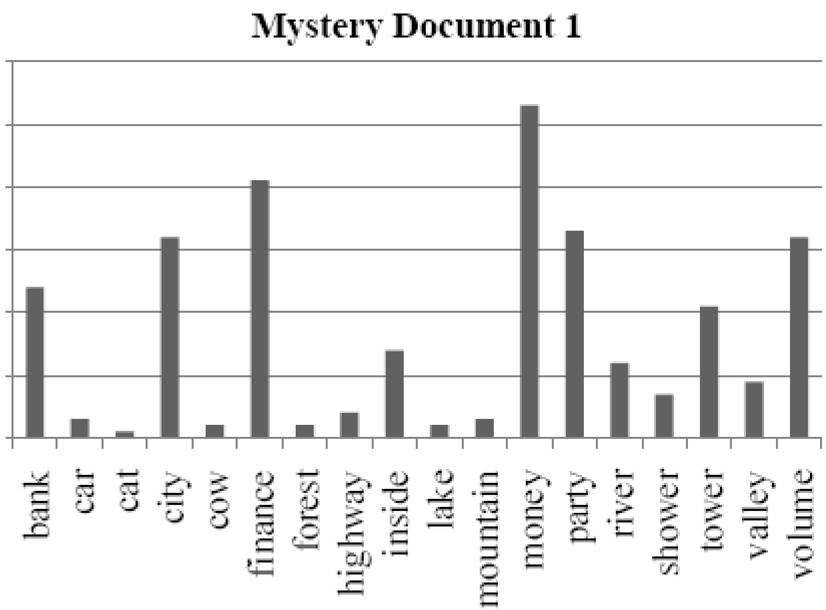
slide by John Winn

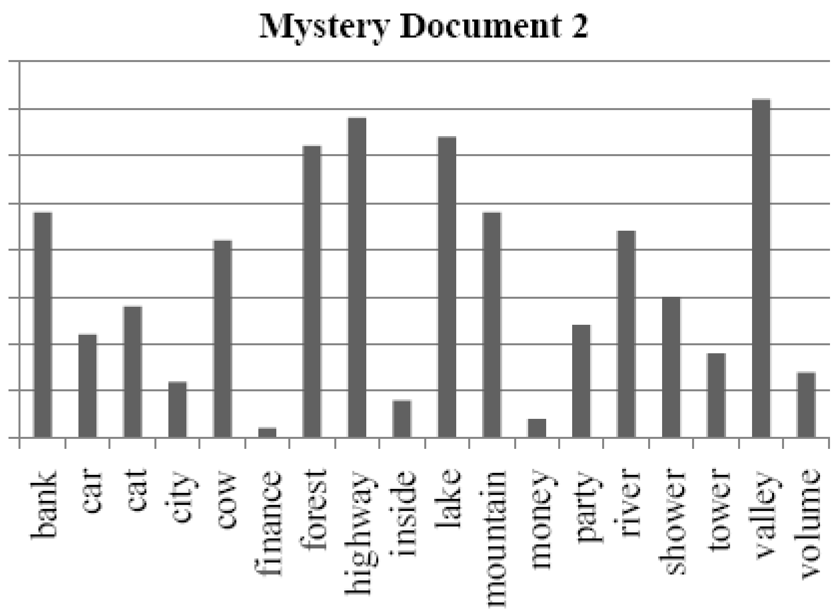
Recognition in Vision

- Recognition is a perceptual and cognitive task fundamental to Vision.
- Three main tasks in Computer Vision:
  - Categorization (or Detection): between-class recognition (e.g. Face Detection: is it a face?)
  - Identification: within-class object recognition (e.g. Face Recognition: is it my friend's face?)
  - Parameter Estimation (e.g. Facial Expression: degree of happiness in a face)

Bag of Words Model

Bag of Words Model



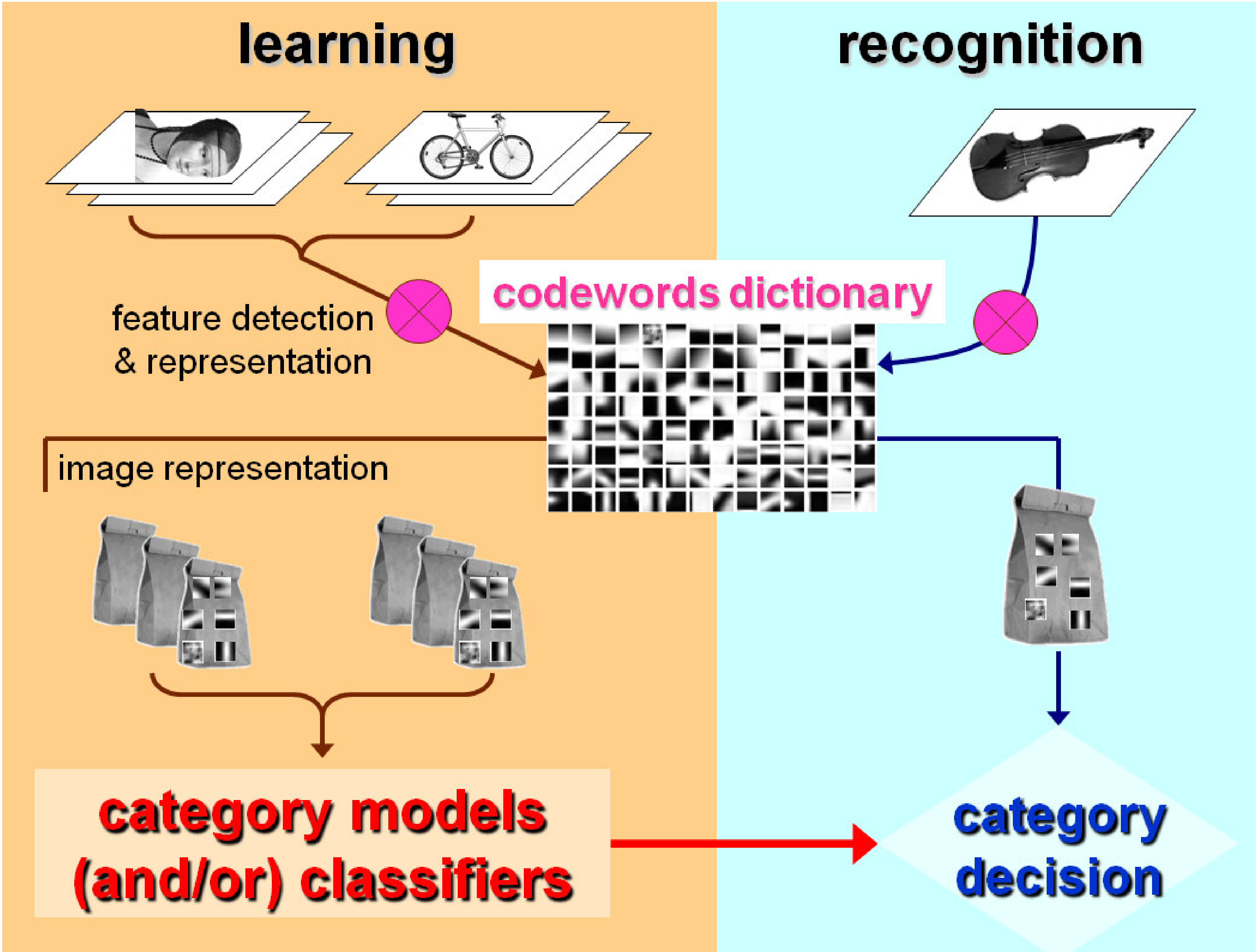


$$tf_i = \frac{n_i}{\sum_k n_i}$$

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

$$tfidf_i = tf_i \times idf_i$$

Object → Bag of ‘words’

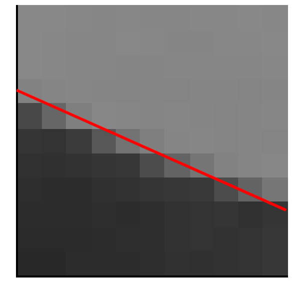

learning
recognition
feature detection
& representation
codewords dictionary
image representation
category models
(and/or) classifiers
category
decision

Bag of Visual Words: Representing Visual Data

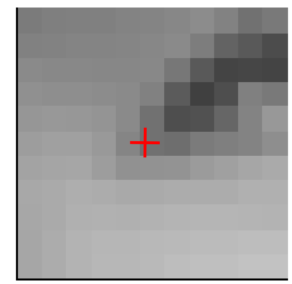
1. Extraction of Local Image feature
   – E.g., Interest points, response of the Filter Banks
2. Descriptors
   – E.g., Orientation Histograms, SIFT, Textons
3. Creation of a Visual Vocabulary
   – Generative Approach (e.g. K-means)
   – Discriminative Approach (e.g. Random Decision Forest)
4. Image Representation
   – E.g., Visual Words distribution (e.g. TF-IDF normalization), Visual Words Co-Occurence distribution, Visual Words Correlograms

Local Image feature: Interest Points

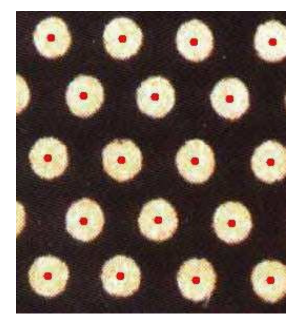
- ***Edges***: an image patch containing the edge reveals an intensity discontinuity in one direction.



- ***Corners***: an image patch containing the corner reveals an intensity discontinuity in two directions.

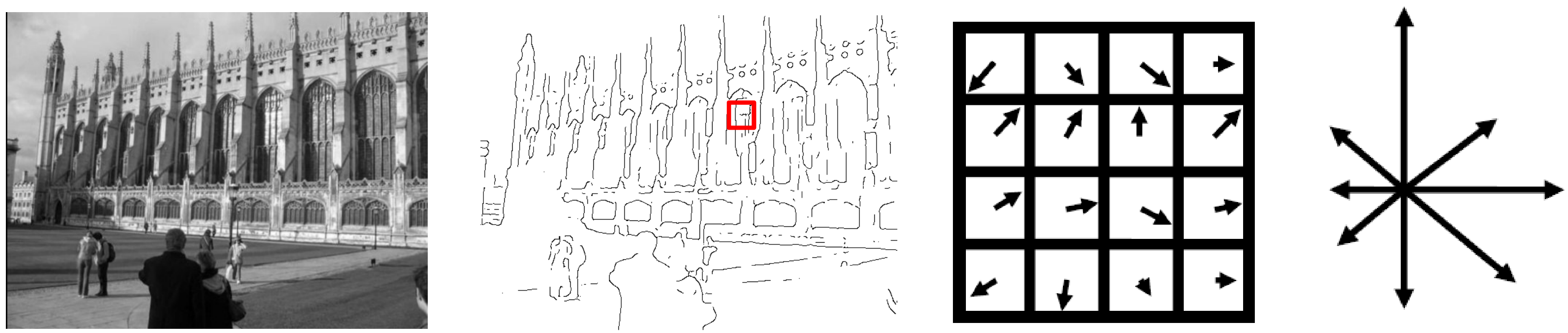


- ***Blobs***: a region of pixels with intensities higher (or lower) than surrounding pixels.

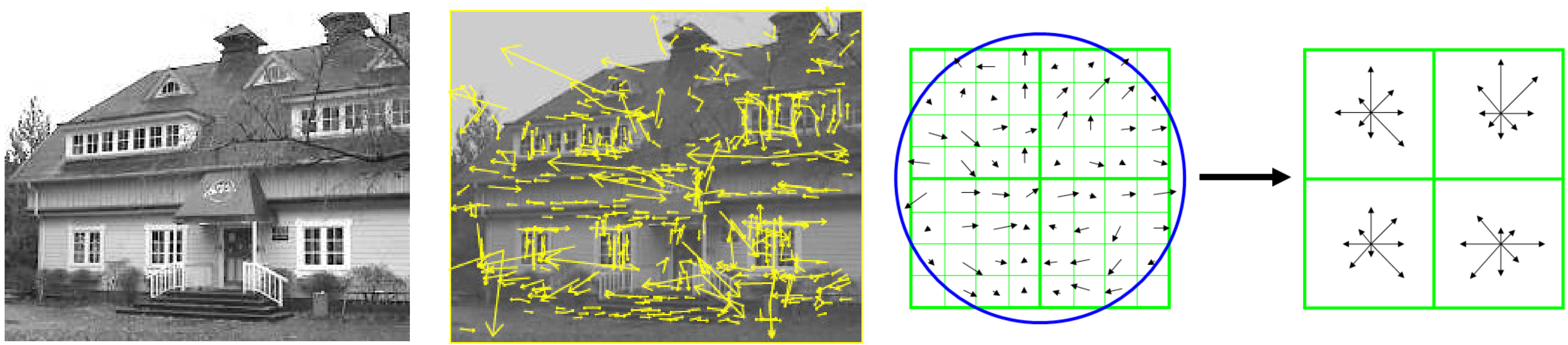


Interest Points Descriptors

Orientation Histograms



SIFT



Interest Point Descriptors

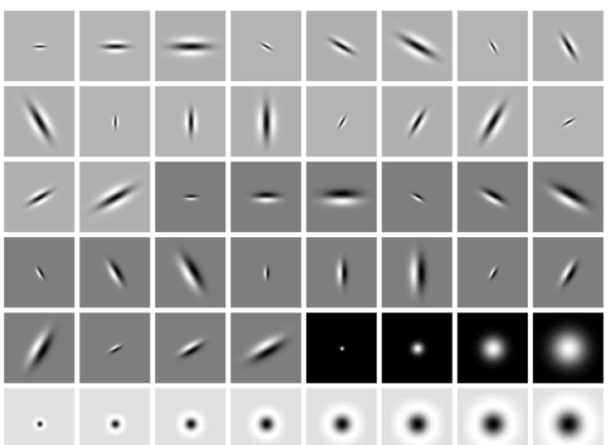
- Many descriptors have been proposed in Computer Vision literature:
  - N-SIFT
  - Colour SIFT
  - Shape Context
  - HoG
  - C-HoG
  - HoF

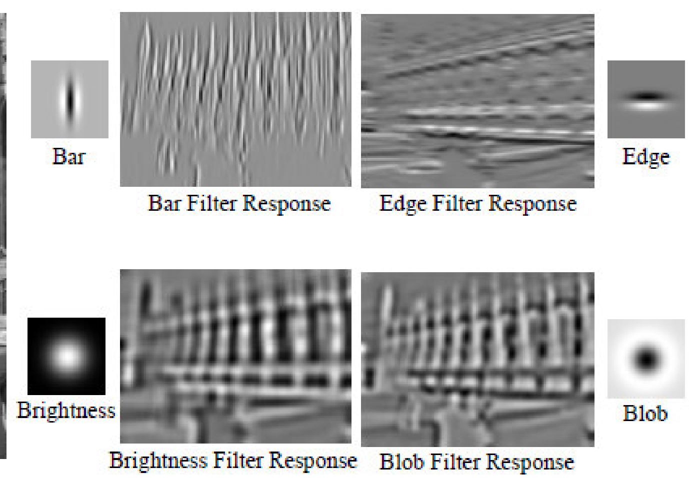
Local Image feature: Texture

Texture



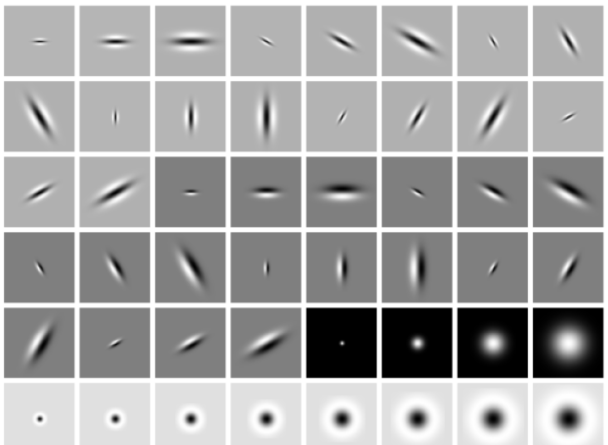

Filter Banks

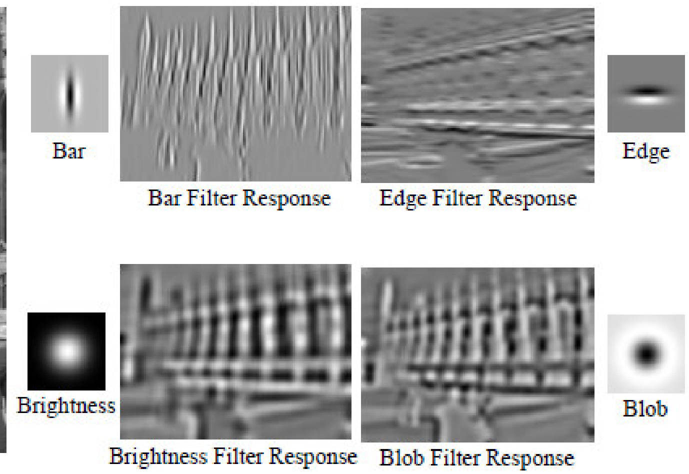


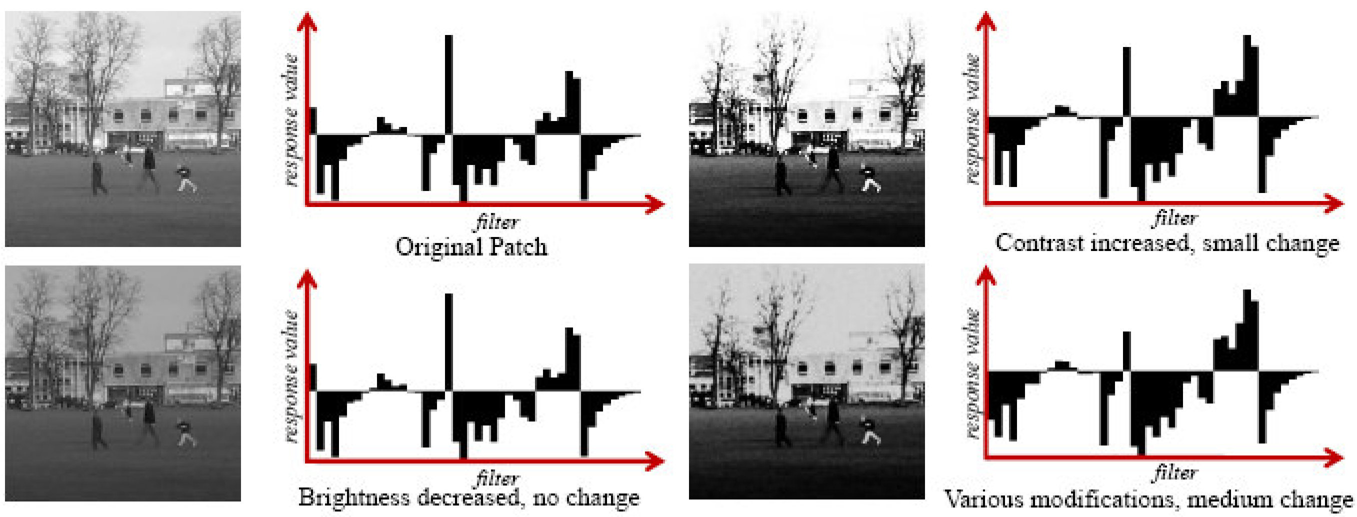


Texture - Descriptors

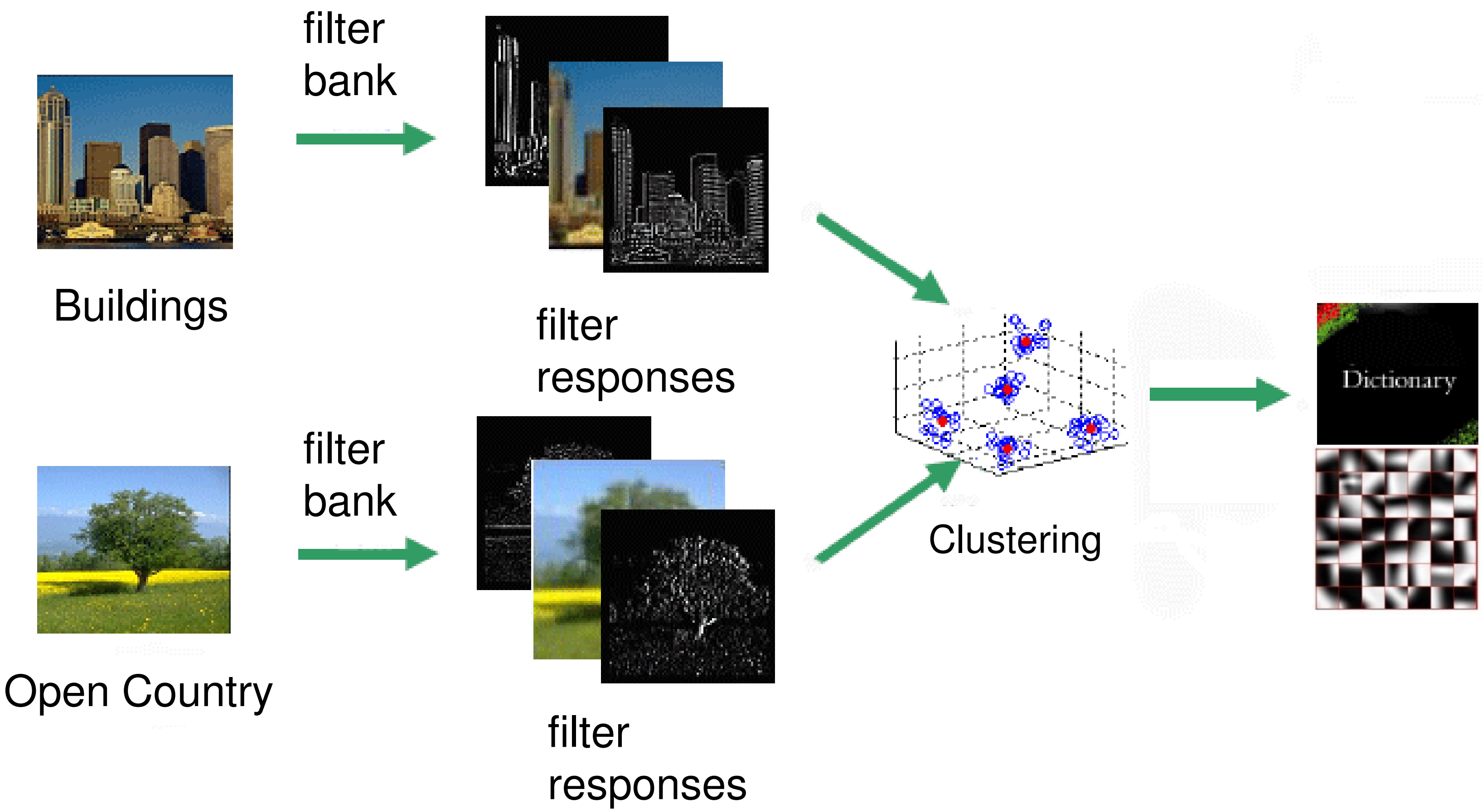
Filter Bank Responses







Creation of a Visual Vocabulary



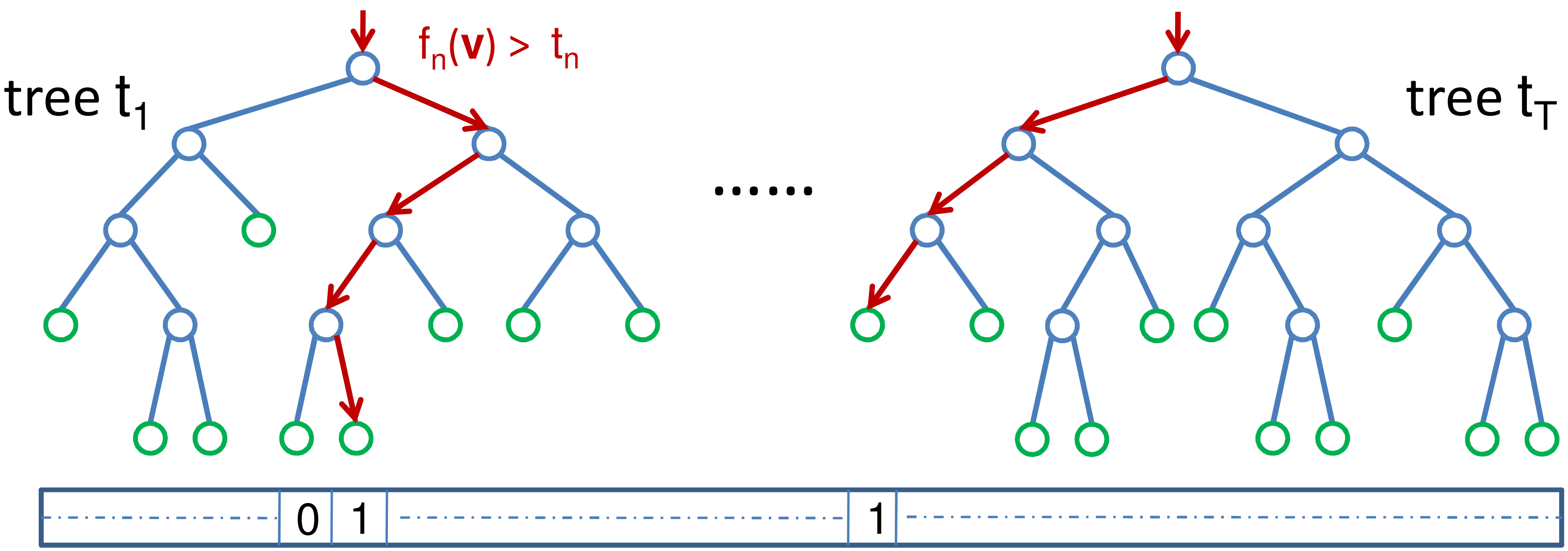
KMeans

- It is the most common used algorithm to build visual vocabularies.
- The algorithm consists of two steps, which are repeated until no vector changes membership.
  1. Compute a cluster center for each cluster as the mean of the cluster members.
  2. Reassign each data point to the cluster whose center is nearest.

- **Kmeans It is computationally expensive during training and use**
- **Kmeans do not uses the knowledge about the classes**

Randomized Decision Forests

A forest is ensemble of several decision trees



- feature vector $\mathbf{v}$
- split functions $f_n(\mathbf{v})$
- thresholds t_n

left split $I_l = \{i \in I_n \mid f(\mathbf{v}_i) < t\}$

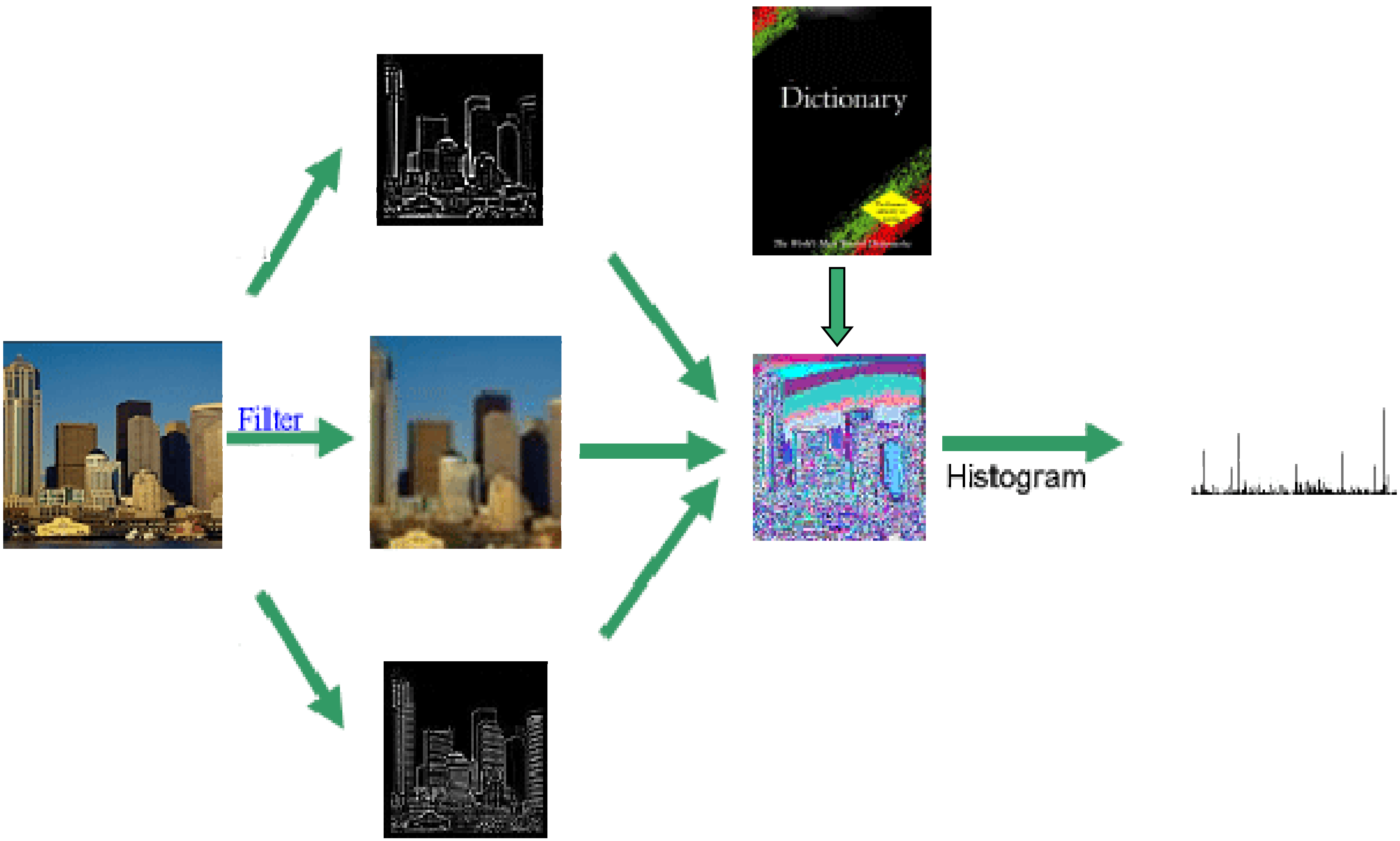
right split $I_r = I_n \setminus I_l$

Features f(v) chosen from feature pool f ϵF

Thresholds t chosen random in range $t \in (\min_i f(\mathbf{v}_i), \max_i f(\mathbf{v}_i))$

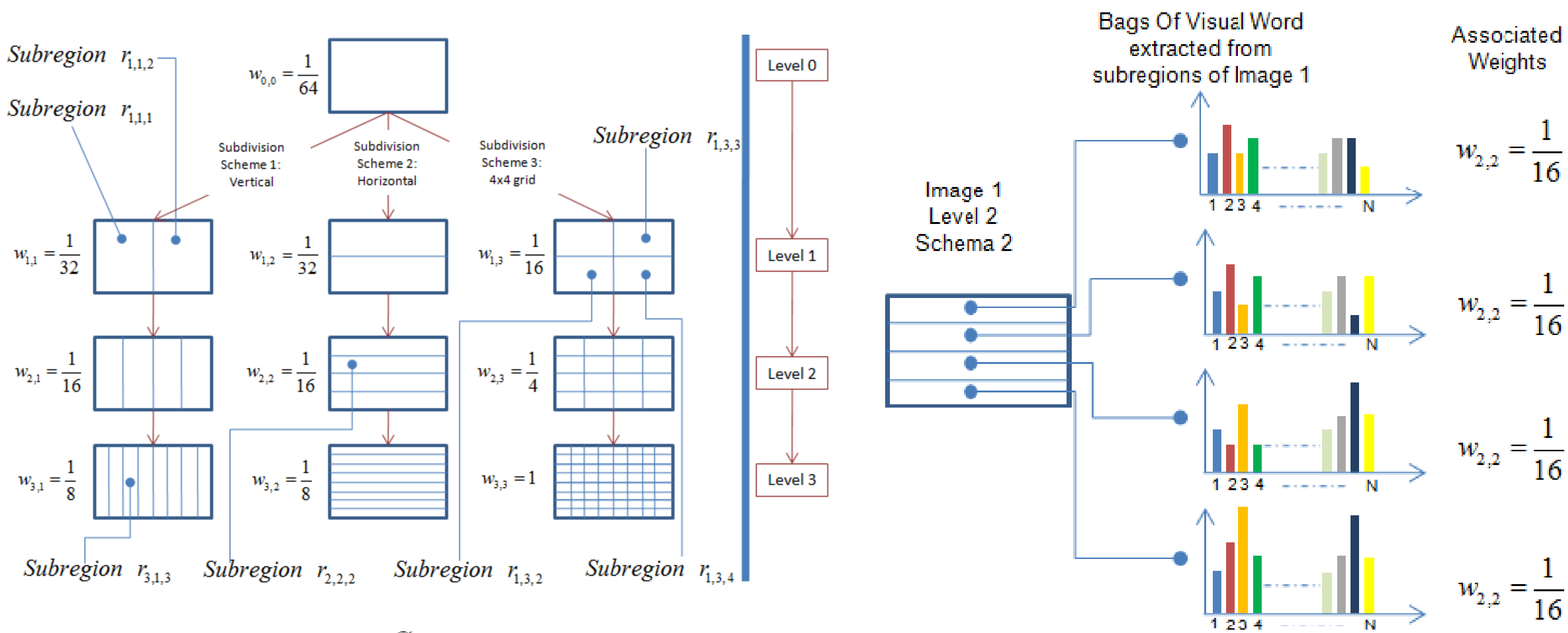
Choose f and t to maximize gain in information $\Delta E = -\frac{|I_l|}{|I_n|}E(I_l) - \frac{|I_r|}{|I_n|}E(I_r)$

Image Representation



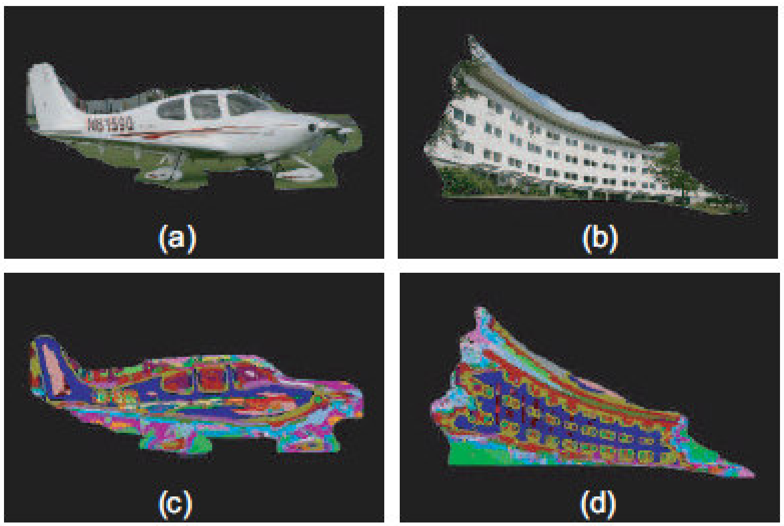
Spatial Hierarchy Representation

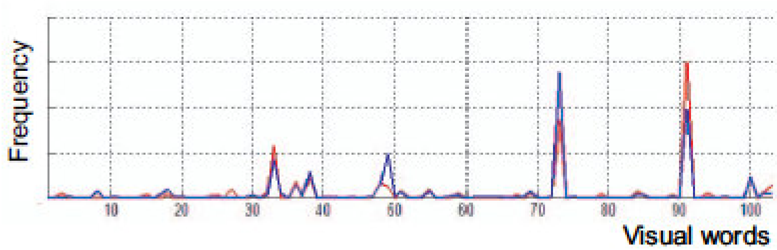
- Extension of a bag of visual words model
- Visual Words representation partitioning the image with different schemes at several levels of resolution

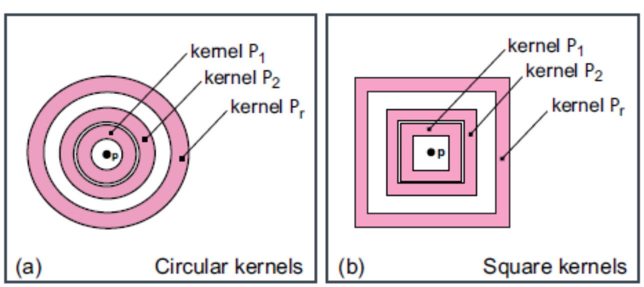


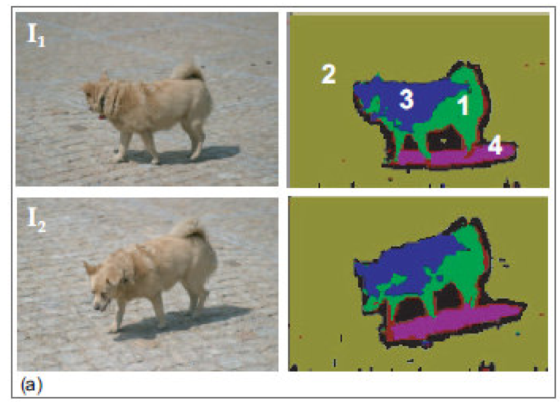
$$w_{l,s} = \frac{S_{l,s}}{\max_{Level,Scheme}(S_{Level,Scheme})}$$

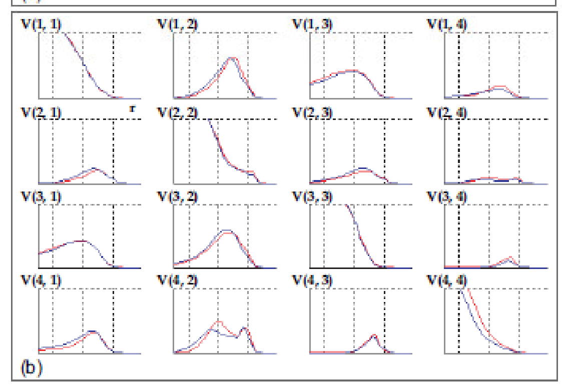
Visual Words Correlograms

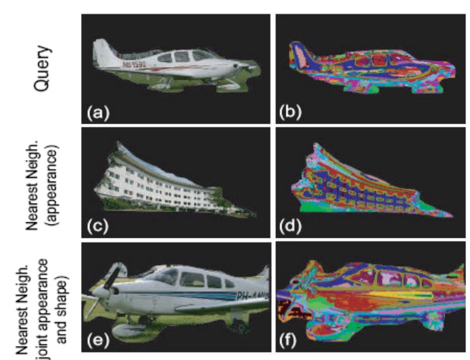


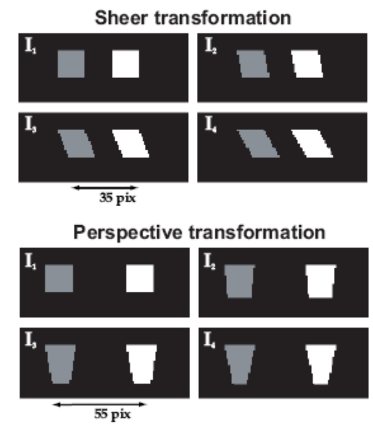


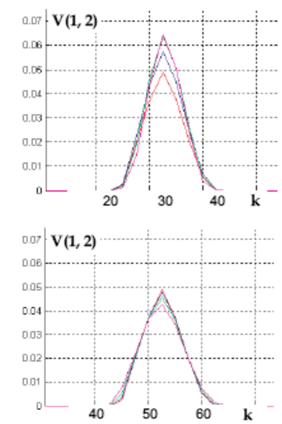


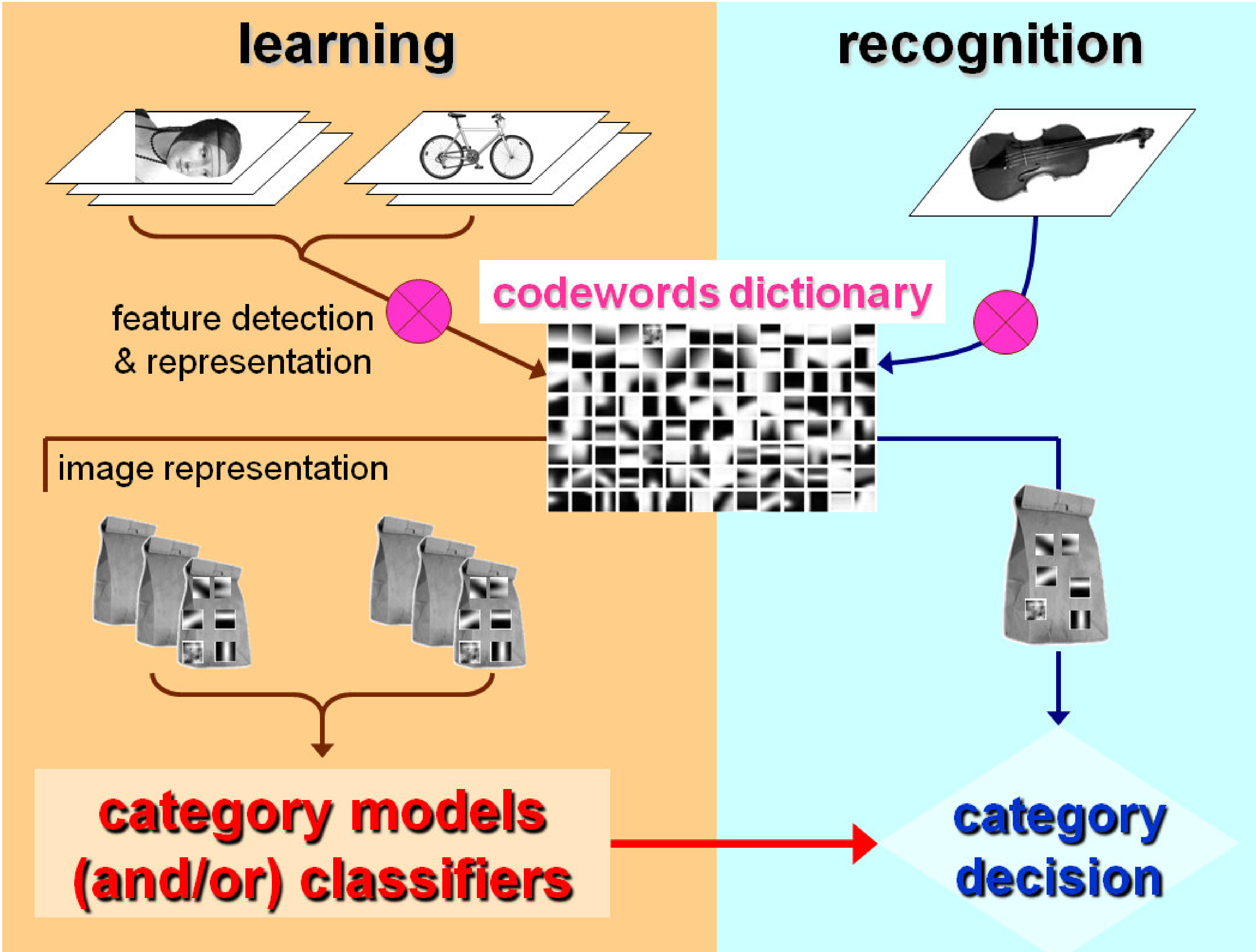











learning
recognition
feature detection
& representation
codewords dictionary
image representation
category models
(and/or) classifiers
category
decision

The Statistical Viewpoint

$$p(zebra \mid image)$$

vs.

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

- Bayes rule:

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

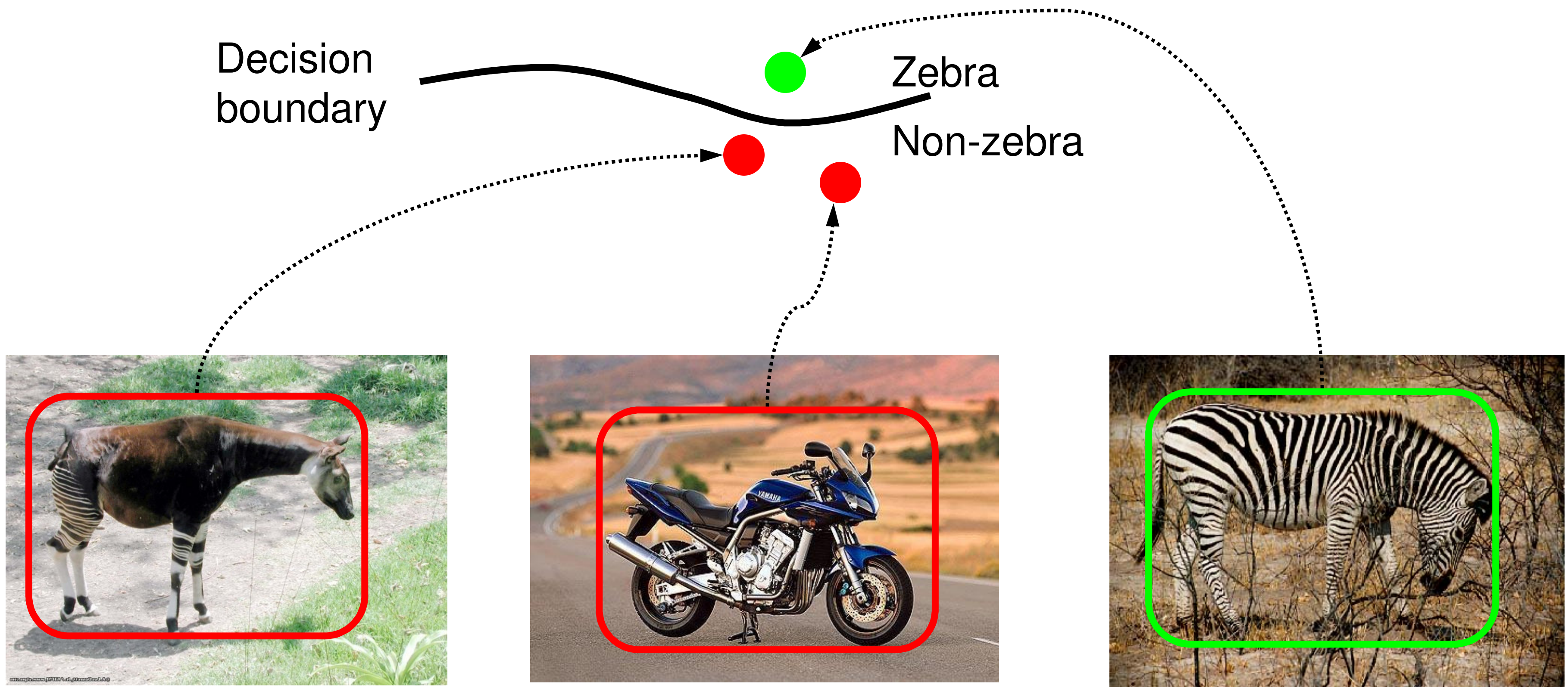
The Statistical Viewpoint

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

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

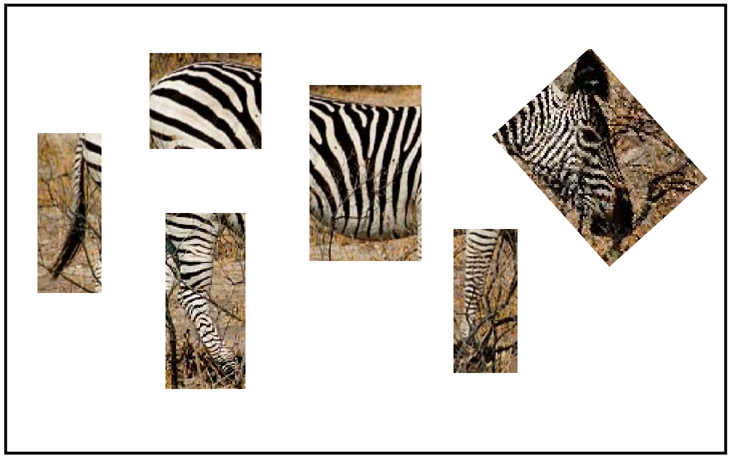
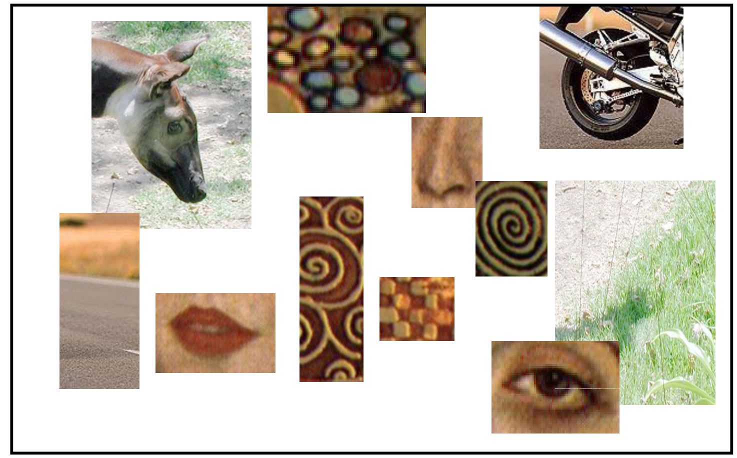
Discriminative

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



Generative

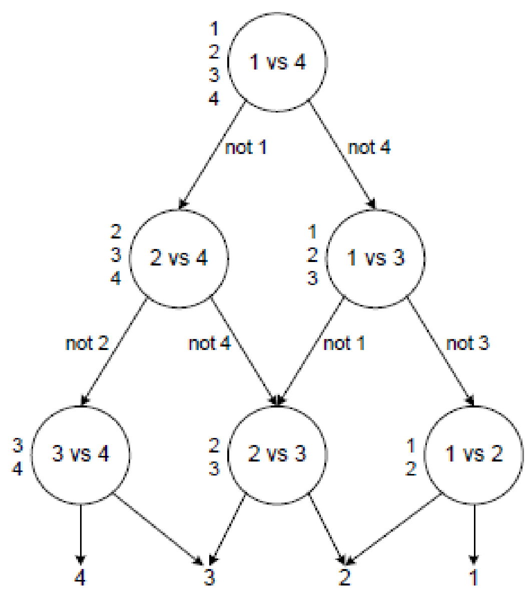
- Model $p(image \mid zebra)$ and $p(image \mid no\ zebra)$

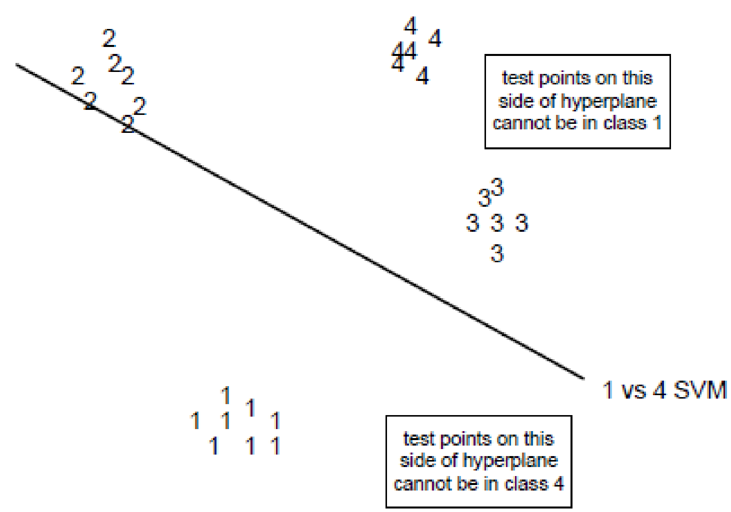



	$p(image \mid zebra)$	$p(image \mid no\ zebra)$
	Low	Middle
	High	Middle→Low

Multi-Class Classification with Binary Classifiers

- One-against-all
- One-against-one
- Decision DAG





Learning and Recognition of Categories

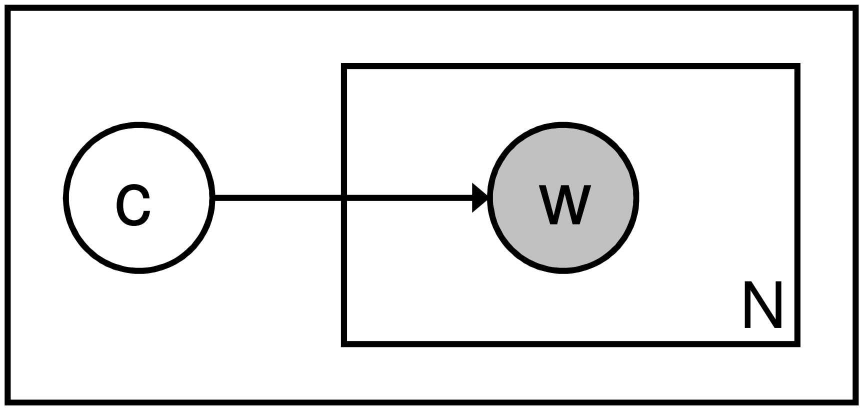
Some of the commonly used techniques are:

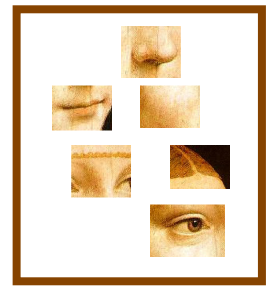
- Generative
  - Naïve Bayes
  - Probabilistic Latent Semantic Analysis (PLSA)
- Discriminative
  - Support Vector Machines
  - Boosting
  - Nearest Neighbour
- Hybrid
  - PLSA + SMV

Notation

- w_n: each visual word in an image
  - $w_n = [0,0,\ldots 1,\ldots,0,0]^T$
- **w:** a collection of all N visual word in an image
  - $\mathbf{w} = [w_1, w_2, \ldots, w_N]$
- d: image in an collection
- c: category of the image
- z: theme or topic of the patch

Naïve Bayes Model





$$c^* = \arg\max_c \, p(c \mid \mathbf{w}_i) \propto p(c)\, p(\mathbf{w}_i \mid c) = p(c) \prod_{n=1}^{N} p(w_n \mid c)^{T(n,i)}$$

class decision

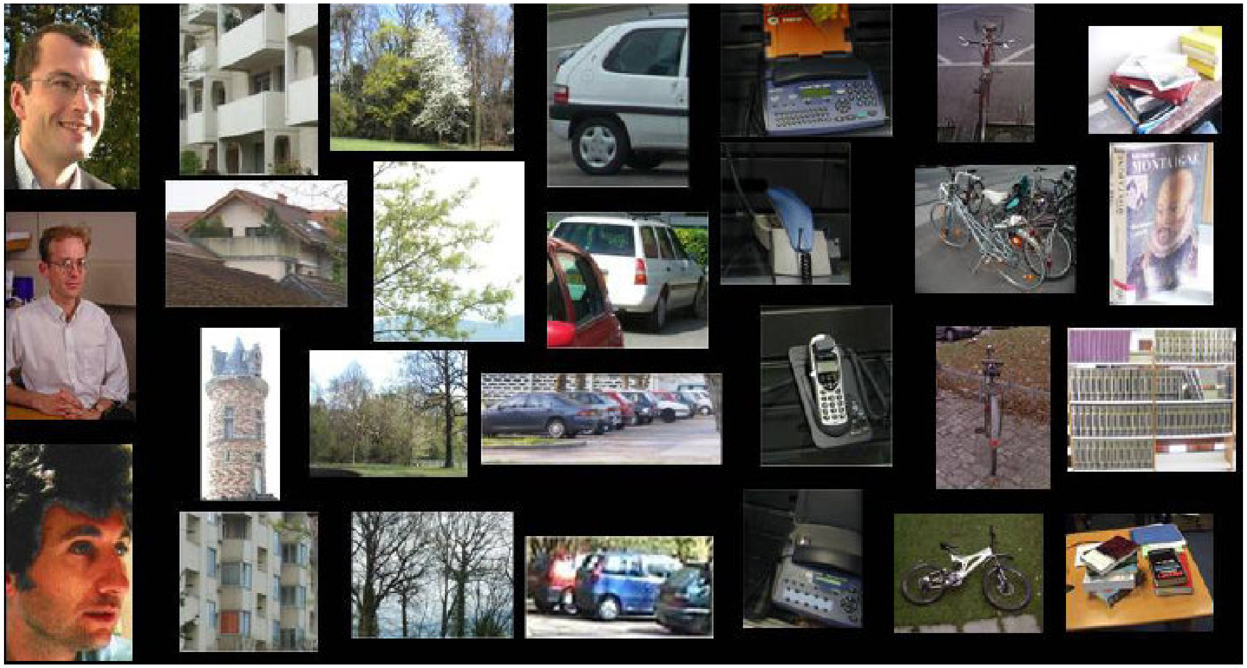
Prior prob. of the classes

Image likelihood given the class

Naïve Bayes Model

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

- 7 object classes
- SIFT
- Kmeans (K=1000)
- Naïve Bayes



Naïve Bayes vs SVM

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

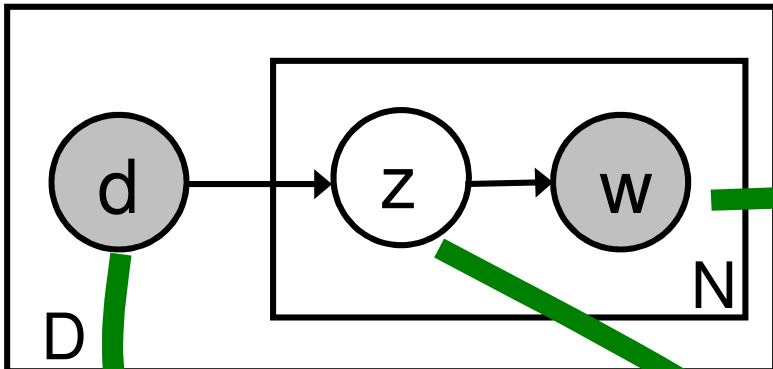
- 7 object classes
- SIFT
- Kmeans (K=1000)
- Naïve Bayes

- 7 object classes
- SIFT
- Kmeans (K=1000)
- SVM (linear kernel)

True classes →	*faces*	*buildings*	*trees*	*cars*	*phones*	*bikes*	*books*
faces	**98**	14	10	10	34	0	13
buildings	1	**63**	3	0	3	1	6
trees	1	10	**81**	1	0	6	0
cars	0	1	1	**85**	5	0	5
phones	0	5	4	3	**55**	2	3
bikes	0	4	1	0	1	**91**	0
books	0	3	0	1	2	0	**73**

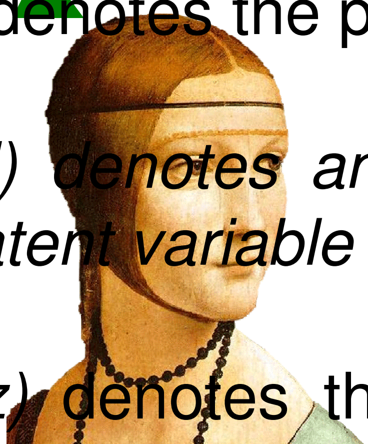
Probabilistic Latent Semantic Analysis

$$P(w,d,z) = P(w \mid z)P(z \mid d)P(d)$$



- Select an image *d* with probability P(d)
- Pick a latent topic *z* with probability P(z|d)
- Generate a word *w* with probability P(w|z)

P(d) denotes the probability of observing a particular image *d*

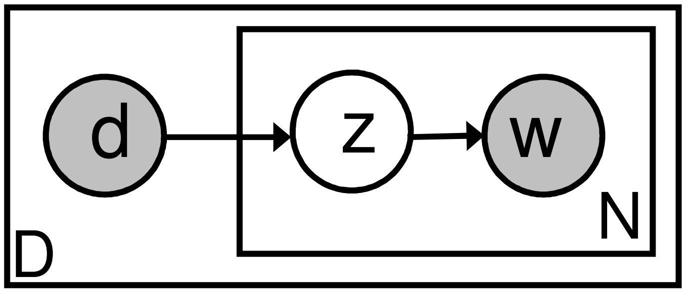


P(z|d) denotes an image specific probability distribution over the latent variable space.

“face”

P(w|z) denotes the conditional probability of a specific word conditioned on the unobserved topic variable *z*

As a result one obtains the observation pair (*w, d), while the latent topic variable z is discarded.*

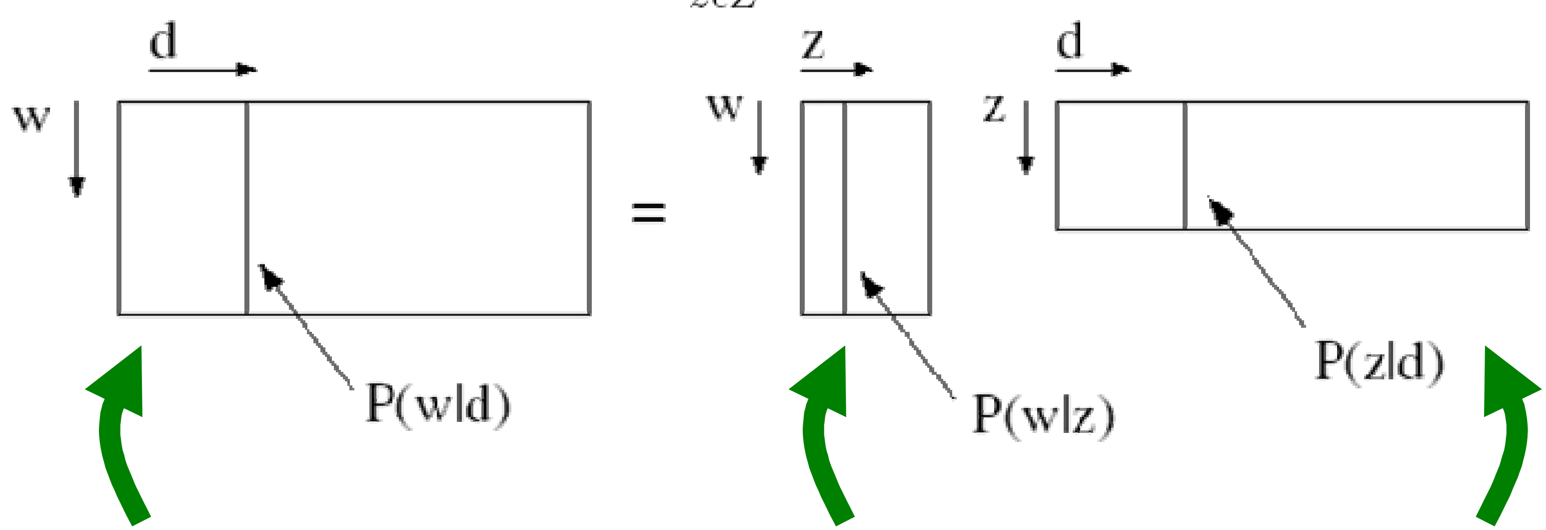


Probabilistic Latent Semantic Analysis

$$P(w, d, z) = P(w|z)P(z|d)P(d)$$

$$P(w, d) = \sum_{z \epsilon Z} P(w, d, z) = P(d) \sum_{z \epsilon Z} P(w|z)P(z|d) \qquad P(w, d) = P(d)P(w|d)$$

$$P(w|d) = \sum_{z \epsilon Z} P(w|z)P(z|d)$$



Observed codeword distributions

Codeword distributions per theme (topic)

Theme distributions per image

PLSA: Learning and Categorization

P(w|z) and P(z|d) are determined by maximizing the likelihood function using EM.

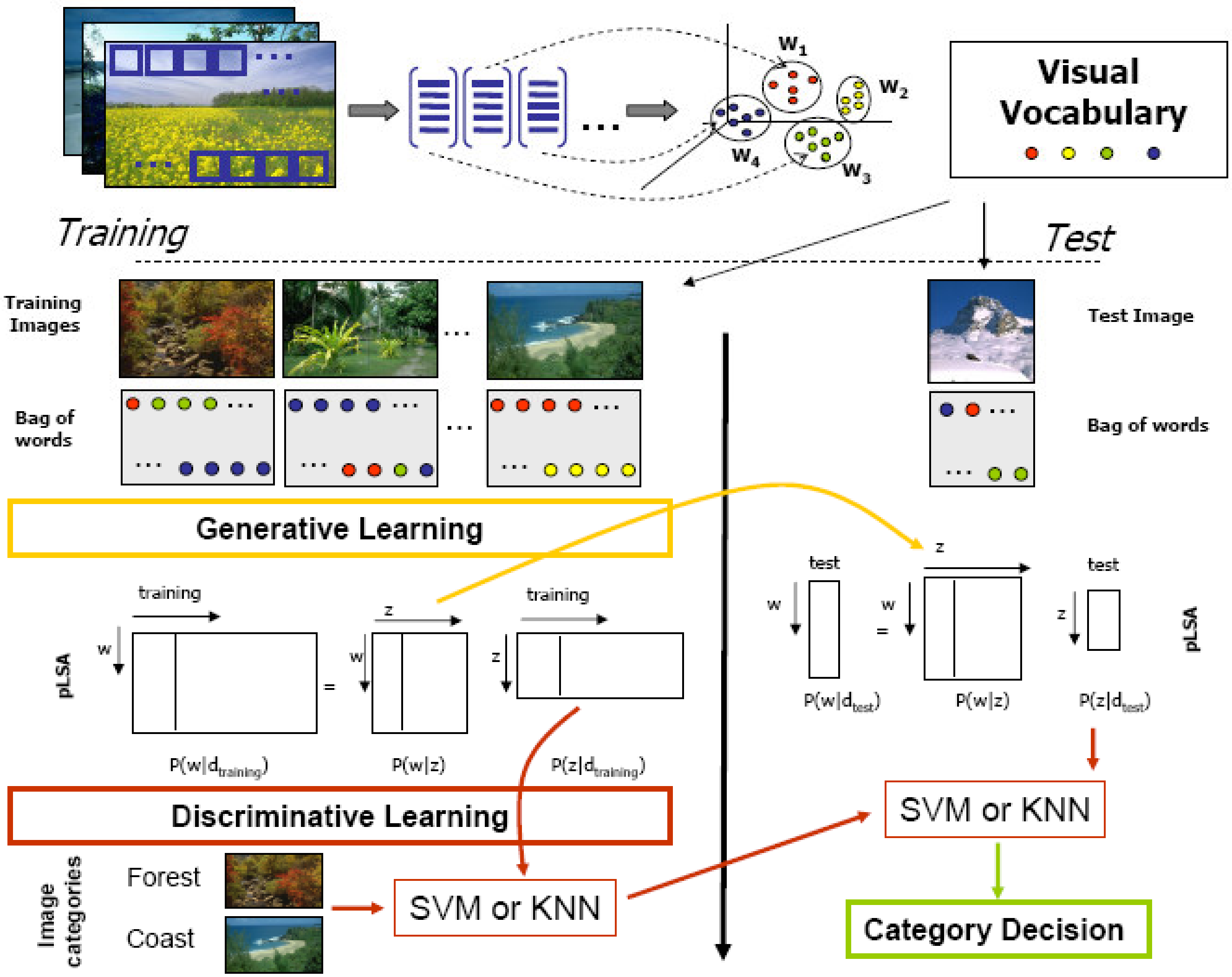
Observed counts of a word in document

$$L = \log P(D,W) = \sum_{d \epsilon D} \sum_{w \epsilon W} n(w,d) \log P(w,d)$$

$$\sum_{z \epsilon Z} P(w|z) P(z|d)$$

$$z^* = \arg\max_z p(z \mid d)$$

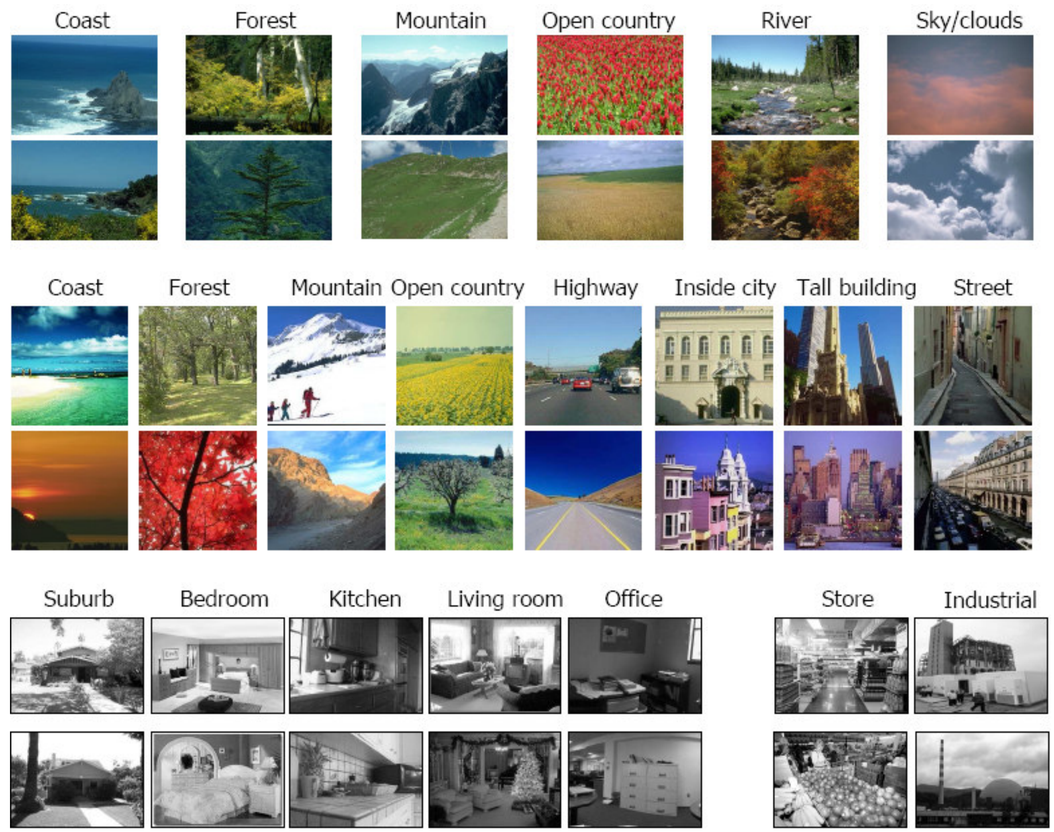
Hybrid generative/discriminative approach



Hybrid generative/discriminative approach

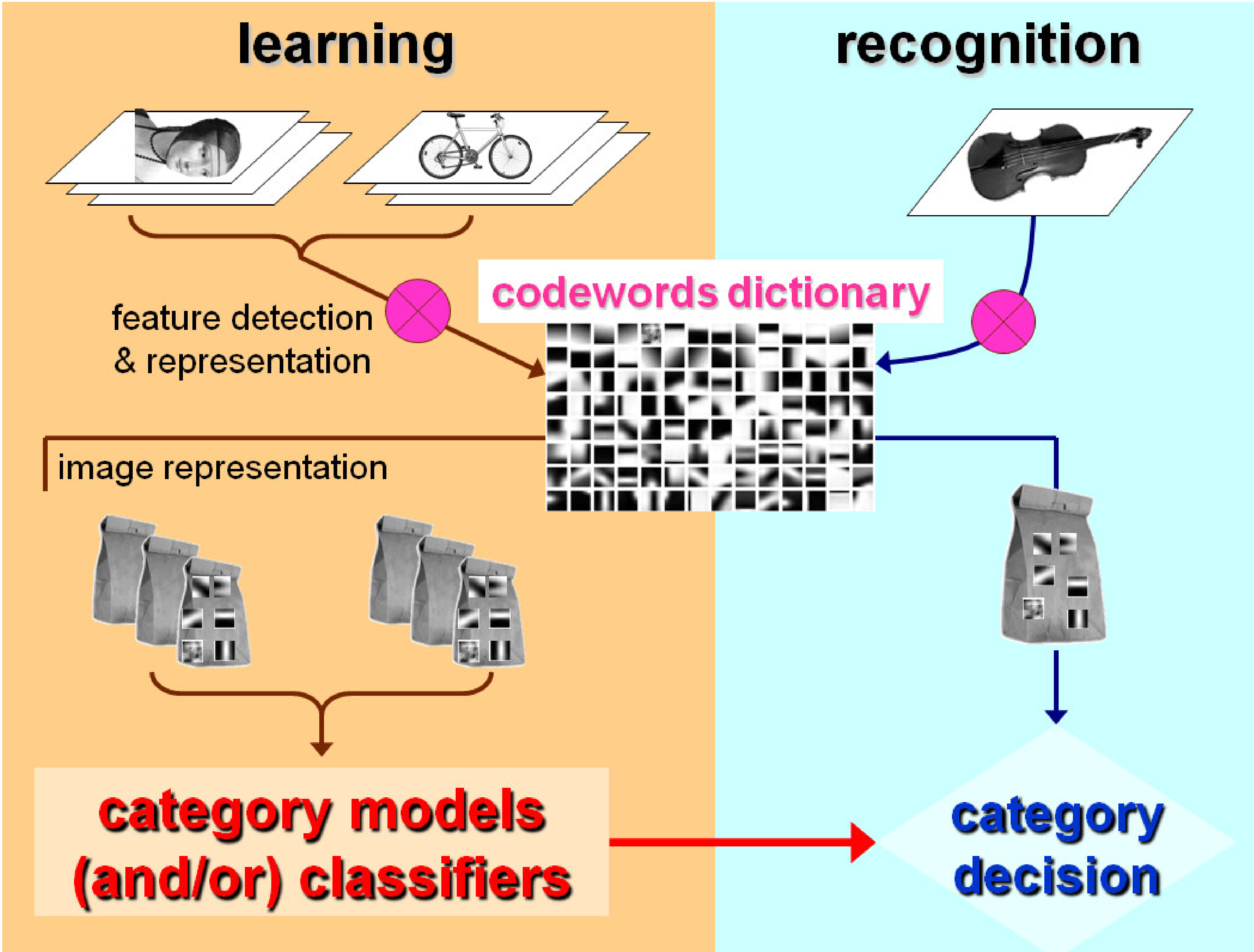
# of categ.	pLSA	SP-pLSA	SPM
8	82.5	**87.8**	87.1
4 Natural	90.7	**93.9**	93.3
4 Man-Made	91.7	**94.8**	94.2
6	87.8	88.3	**88.6**
13	74.3	**85.9**	85.5
15	72.7	**83.7**	83.5

Scene Classification



Hybrid generative/discriminative approach

- Caltech-101 objects data set
- From 31 to 800 images per category
- Large intra class variability
- Mean recognition over 10 tests:
  - 15 training images per class: 59.8
  - 30 training images per class: 67.7%

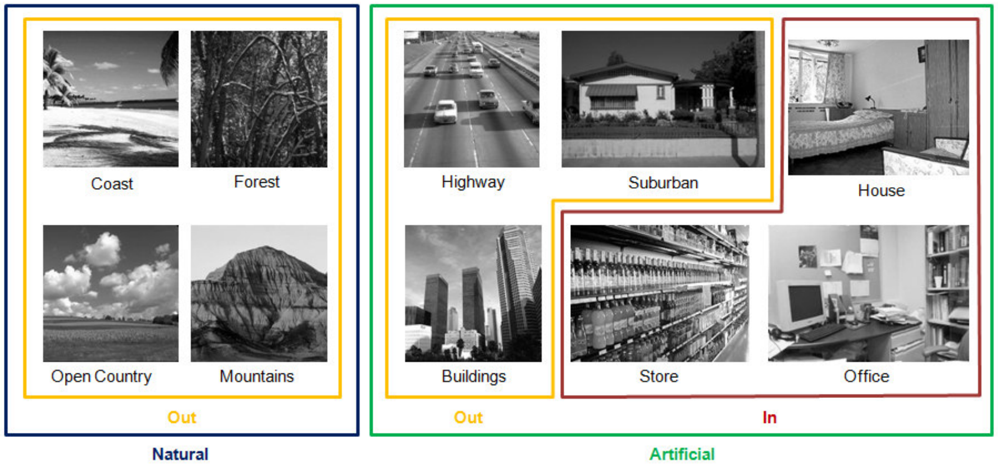

learning
recognition
feature detection
& representation
codewords dictionary
image representation
category models
(and/or) classifiers
category
decision

Examples of Application

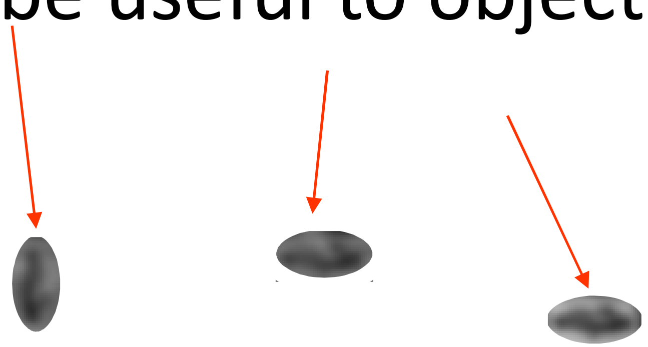
- Scene Classification and Object Classification
- Content Based Image Retrieval
- Semantic Segmentation
- Action Recognition
- Medical Imaging
- Direct Marketing Learning

Scene Classification and Object Recognition

- Given an image we want recognize the context of the image (e.g. Robot Navigation) and/or the objects in that context.

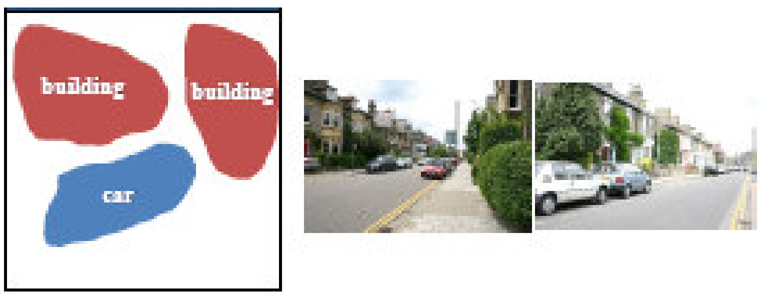


- Context can be useful to object recognition.



Content Based Image Retrieval

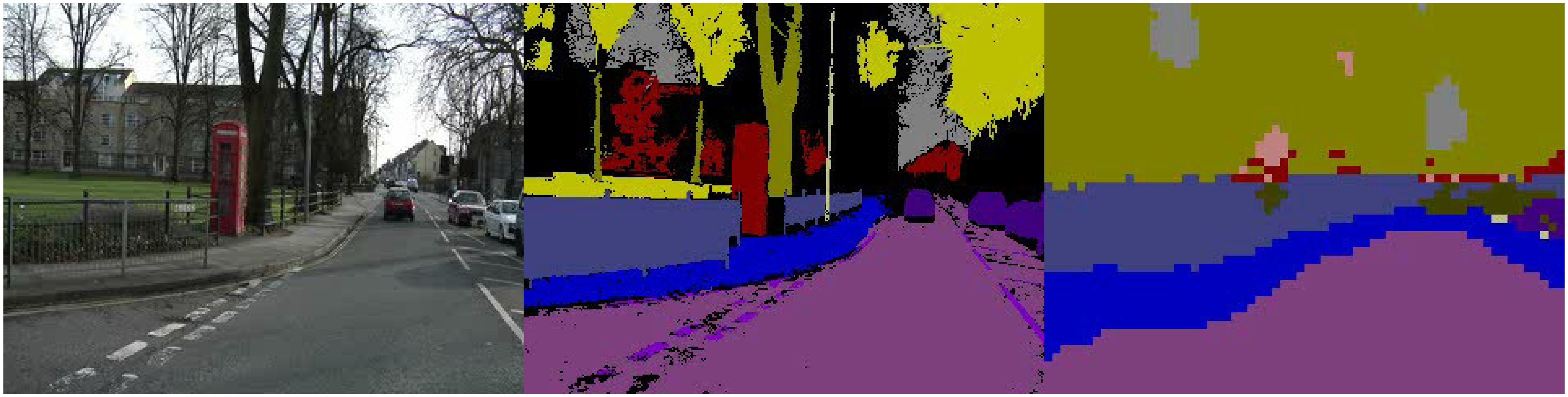
- Given an image we want browse other images of a large image database ranked in terms of visual similarity.



- Given a mental prototype of that image an image retrieval system should rank highly, images which most closely matches that mental prototype.

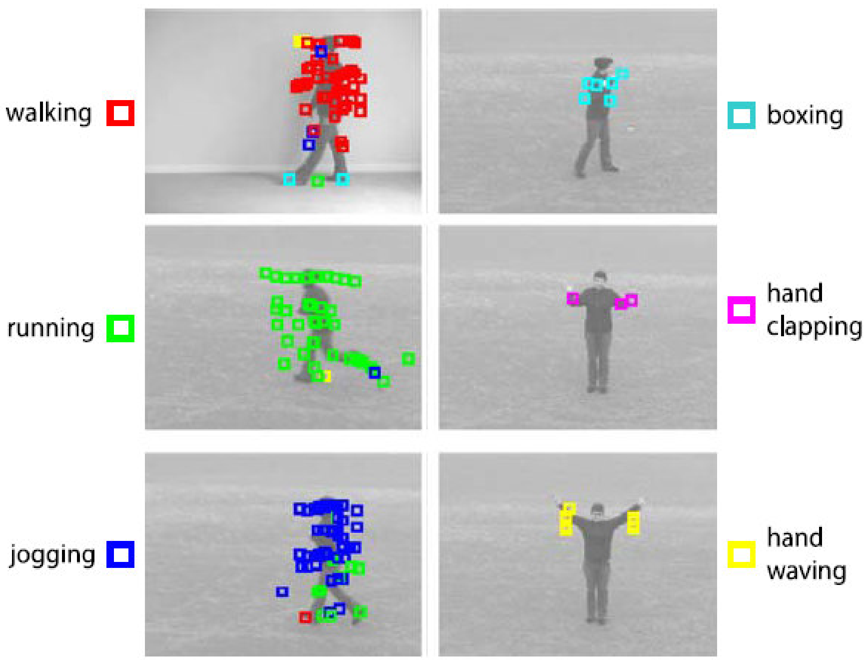
Semantic Segmentation

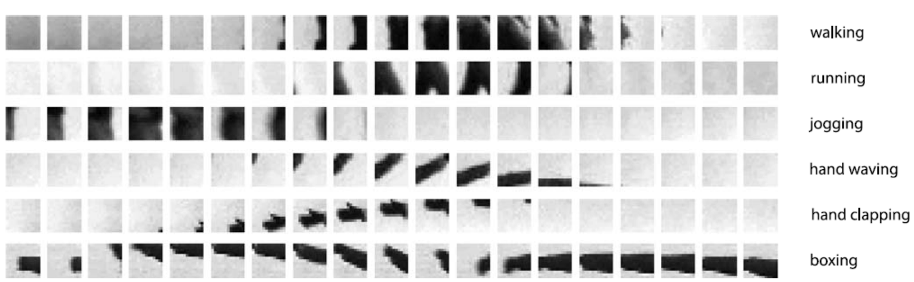
- The semantic segmentation of an image aims in grouping pixels together by common semantic meaning.

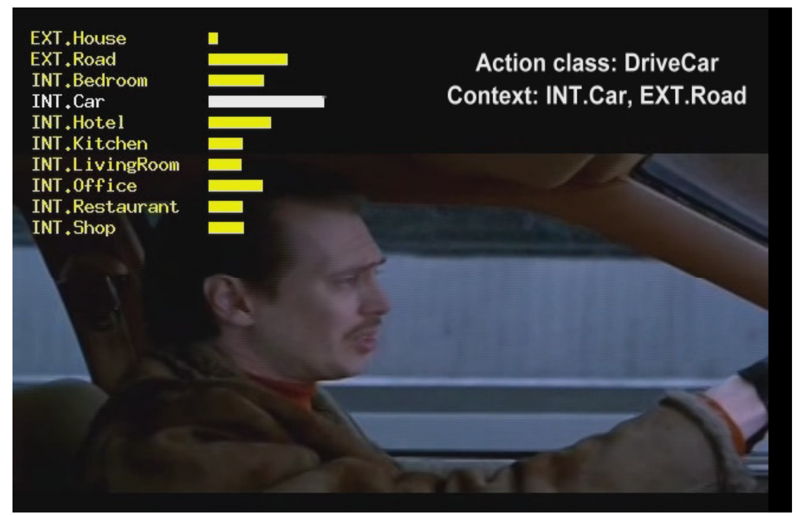


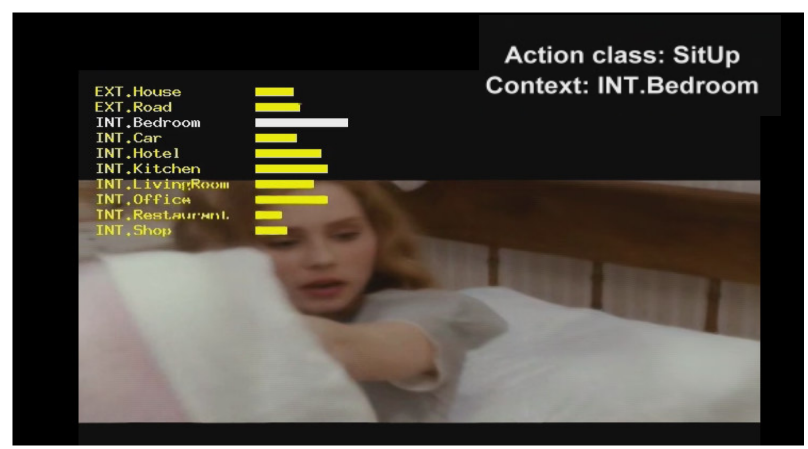
Action Recognition

- Automatic classification or localization of different actions in video sequence.



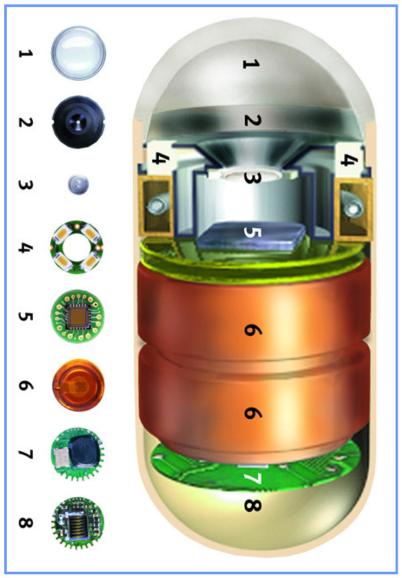




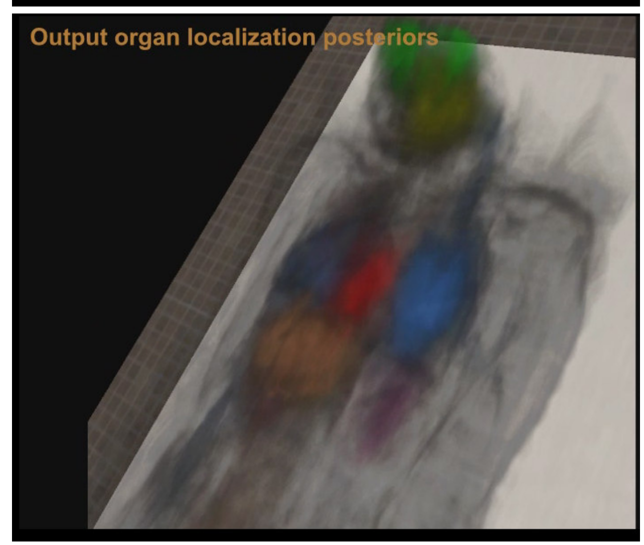


Medical Imaging

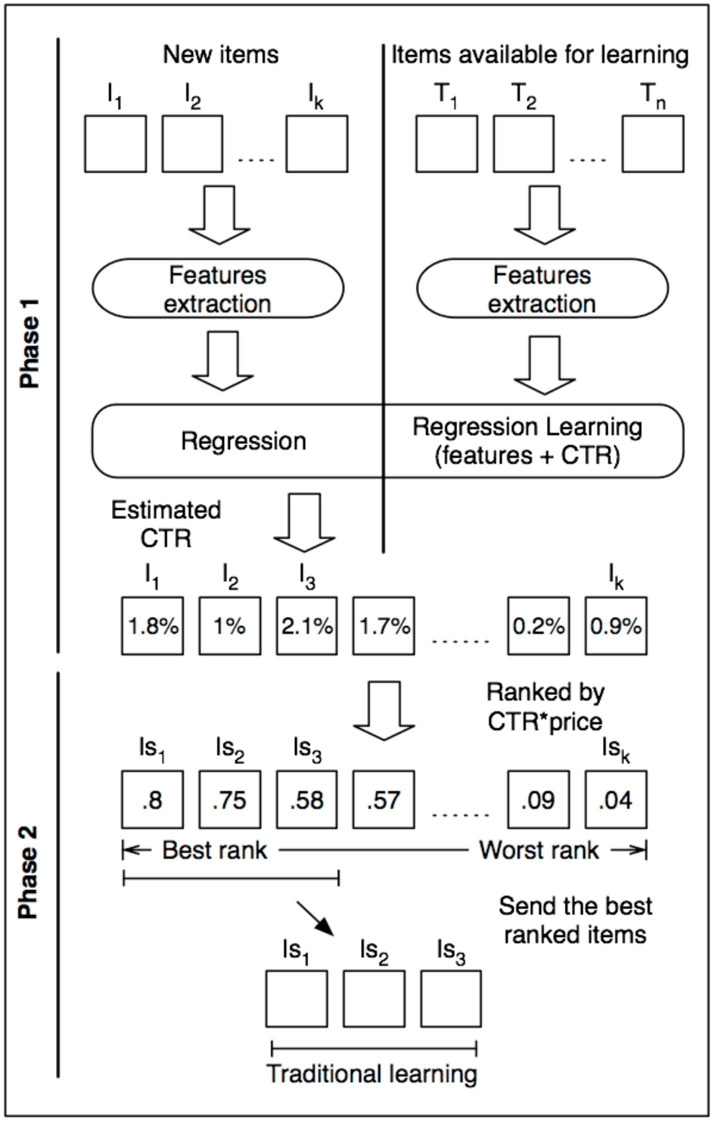
Wireless Capsule Endoscopy Video Analysis

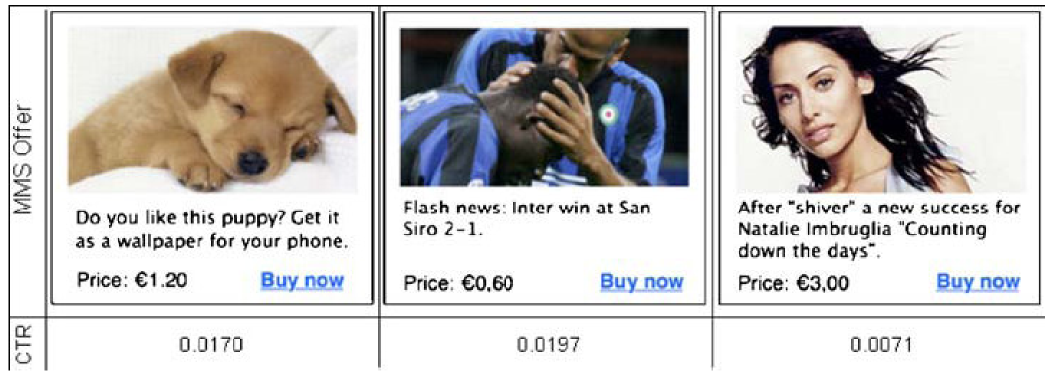


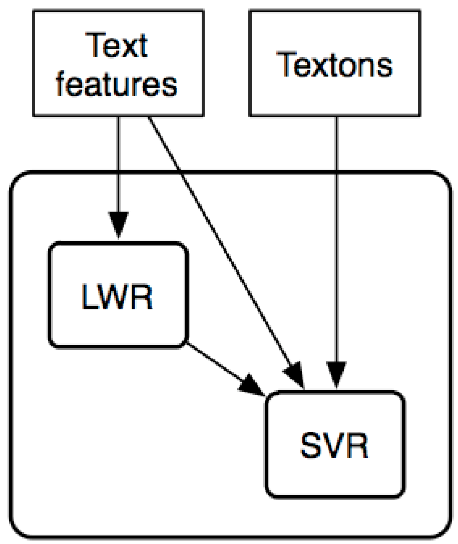
Organs Localization in CT Data

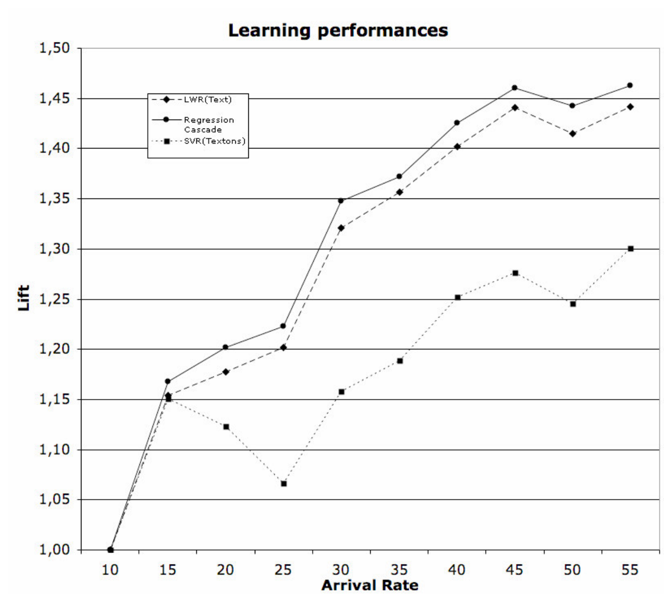


Direct Marketing Learning









Conclusion

From ICVSS 2009 web site: http://www.dmi.unict.it/icvss

- Computer vision researchers are increasingly using algorithms from pattern recognition and machine learning to help build robust and reusable vision systems.

- Just as learning is an essential component of biological visual systems, the design of machine vision systems that learn and adapt represent an important challenge in modern computer vision research.

References and Further Readings

Vision

- D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. New York: Freeman, 1982

Pattern Recognition and Machine Learning in Computer Vision

- Heisele, B., A. Verri and T. Poggio. Learning and Vision Machines, Proceedings of the IEEE, Visual Perception: Technology and Tools, 2002

Local Image Features and Descriptors

- B. Julesz, Textons, the elements of texture perception, and their interactions, Nature, 1981.
- J. Malik and P. Perona, Preattentive texture discrimination with early vision mechanisms J. Optical Society of America A, 12(7):629–639, July 1990
- T. Tuytelaars and K. Mikolajczyk , Local Invariant Feature Detectors - Survey. In CVG, 3(1):1-110, 2008
- David G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 2004
- Song-Chun Zhu, Cheng-en Guo, Yizhou Wang and Zijian Xu, What are Textons?, International Journal of Computer Vision, 2005

Bag of Visual Word Model

- Leung, T. and Malik, J. Recognizing surface using three-dimensional textons, In Proc. of 7th ICCV, Corfu, Greece, 1999.
- Josef Sivic, Andrew Zisserman, Video Google: A Text Retrieval Approach to Object Matching in Videos, International conference on computer vision, 2003
- Gabriela Csurka, Chris Dance, Jutta Willamowski, Lixin Fan, Cedric Bray, Visual categorization with bags of keypoints, ECCV International Workshop on Statistical Learning in Computer Vision, 2004
- Varma, M. and Zisserman, A., A statistical approach to texture classification from single images International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis, to appear in 2005.
- M. Johnson. Semantic Segmentation and Image Search. Phd Thesis, University of Cambridge, 2008

References and Further Readings

Building Visual Vocabularies

- J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In Proc. Int. Conf. on Computer Vision, volume 2, pages 1800–1807, Beijing, China, October 2005.
- F. Moosmann, B. Triggs, F. Jurie, Fast Discriminative Visual Codebooks using Randomized Clustering Forests, Neural Information Processing Systems, 2006
- J. Mairal, F. Bach, Jean Ponce, Guillermo Sapiro, Andrew Zisserman: Discriminative learned dictionaries for local image analysis, International Conference on Computer Visision and Pattern Recognition, 2008
- J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek. Visual word ambiguity. IEEE Trans. Pattern Analysis and Machine Intelligence, in press, 2009.

Scene Classification

- L. Fei-Fei and P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories, IEEE International Conference on Computer Visision and Pattern Recognition, 2005
- A.Bosch, A.Zisserman, X.Muñoz. Scene Classification via PLSA, European Conference on Computer Vision, 2006.
- S. Lazebnik, C. Schmid, and J. Ponce, Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, Computer Visision and Pattern Recognition, 2006
- A.Bosch, A.Zisserman, X.Muñoz. Scene classification using a hybrid generative/discriminative approach, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008.
- S. Battiato, G. M. Farinella, G. Gallo, D. Ravì, Scene Categorization Using Bags of Textons on Spatial Hierarchy, In Proceedings of IEEE International Conference on Image Processing, 2008
- A. Quattoni, and A.Torralba. Recognizing Indoor Scenes. Computer Vision and Pattern Recognition, 2009.

Direct Marketing Learning

- S. Battiato, G. M. Farinella, G. Giuffrida, C. Sismeiro, G. Tribulato, Exploiting visual and text features for direct marketing learning in time and space constrained domains, Pattern Analysis and Applications, 2009
- S. Battiato, G. M. Farinella, G. Giuffrida, C. Sismeiro, G. Tribulato, Using visual and text features for direct marketing on multimedia messaging services domain, Multimedia Tools and Applications, 2009

References and Further Readings

Object Classification

- S. Savarese, A. Criminisi, J. Winn, Discriminative Object Class Models of Appearance and Shape by Correlatons, IEEE Computer Vision and Pattern Recognition, 2006
- K. Grauman and T. Darrell. The Pyramid Match Kernel: Efficient Learning with Sets of Features. Journal of Machine Learning Research, 725-760, 2007.
- J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Appearance, Shape and Context., in Intl. Journal on Computer Vision (IJCV), special issue., Springer Verlag, 2009

Content Based Image Retrieval

- J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007

Semantic Segmentation

- J. Shotton, M. Johnson, R. Cipolla, Semantic Texton Forests for Image Categorization and Segmentation, International Conference on Computer Vision and Pattern Recognition, June 2008.

Action Recognition

- J.C. Niebles, H. Wang and L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, International Journal of Computer Vision, 2008
- M. Marszałek, I. Laptev, C. Schmid, Actions in Context , International Conference on Computer Vision and Pattern Recognition, 2009

Medical Imaging

- G. Gallo, E. Granata, G. Scarpulla, Wireless Capsule Endoscopy Video Segmentation, IEEE International Workshop on Medical Measurement and Applications, 2009
- B. André, T. Vercauteren, A. Perchant, M. B. Wallace, A. M. Buchner, n. Ayache. Endomicroscopic image retrieval and classification using invariant visual features. In Proceedings of the Sixth IEEE International Symposium on Biomedical Imaging, 2009
- A. Criminisi, J. Shotton, S. Bucciarelli, Decision Forests with Long-Range Spatial Context for Organ Localization in CT Volumes, in MICCAI workshop on Probabilistic Models for Medical Image Analysis, 2009

Tutorials (some slides of this seminar are taken form the following tutorials)

- L.Fei-Fei, R. Fergus, A. Torralba, Recognizing and Learning Object Categories, Computer Vision and Pattern Recognition,2007
- J. Winn, Probabilistic models for understanding images, International Computer Vision Summer School, 2009
- L. Fei-Fei, Generative Models for Visual Objects and Object Recognition via Bayesian Inference, Machine Learning over Text & Images - Autumn School, 2006
- T. Kanade, Computer Vision and Machine Learning: the past, the present, the future, International Computer Vision Summer School, 2009

Thank you!