

A photograph of the ruins of the Temple of Concordia in Agrigento, Sicily. The temple is a well-preserved example of ancient Greek architecture, featuring a row of six tall, fluted Doric columns supporting a heavy entablature. The sky is a clear, bright blue, and the foreground shows some dry grass and a few small trees. The text is overlaid on the left side of the image.

BAG-OF-WORDS

HYUN SOO PARK

CHALLENGES OF VISUAL RECOGNITION

- Appearance
 - DOF: texture, illumination, material, shading, ...
- Shape
 - DOF: object category, geometric pose, viewpoint, ...



IMAGE CLASSIFICATION

Bedroom



Coast

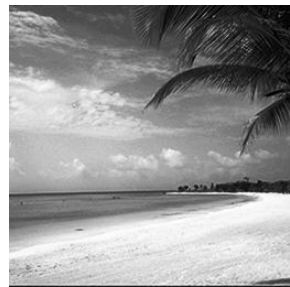


IMAGE CLASSIFICATION

corr



,



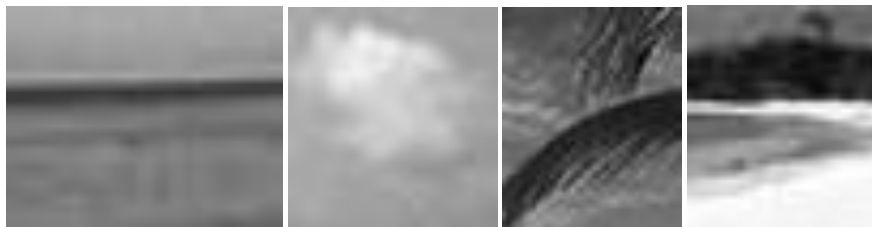
? corr



,



LOCAL PATCHES



LOCAL PATCHES

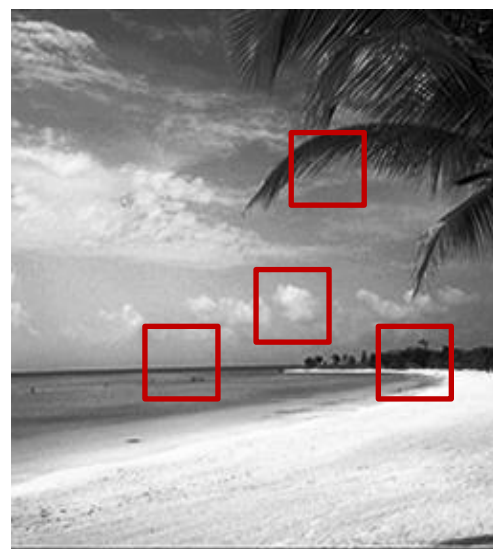
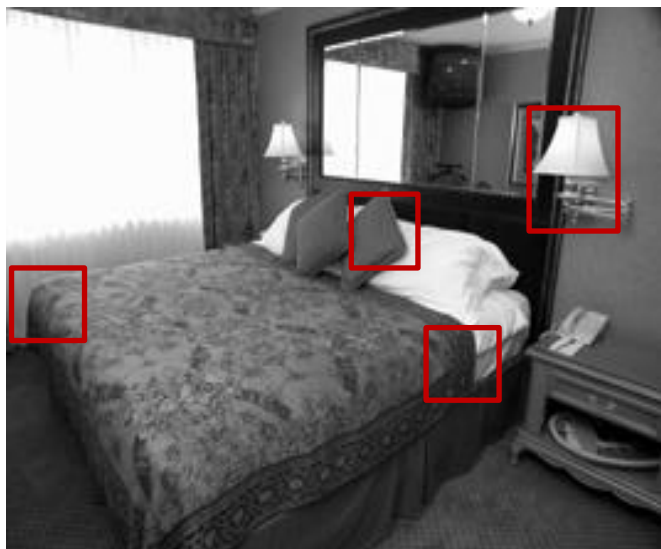
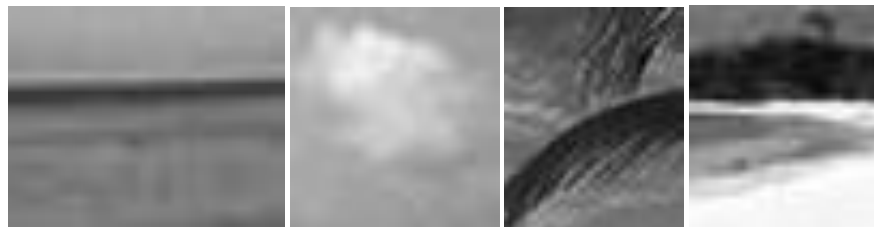
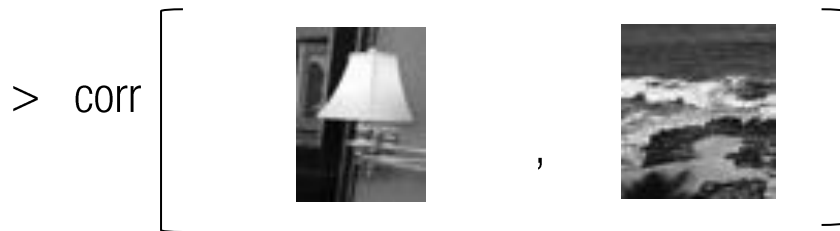
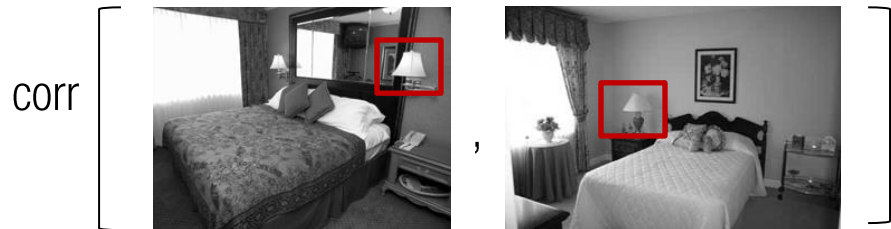
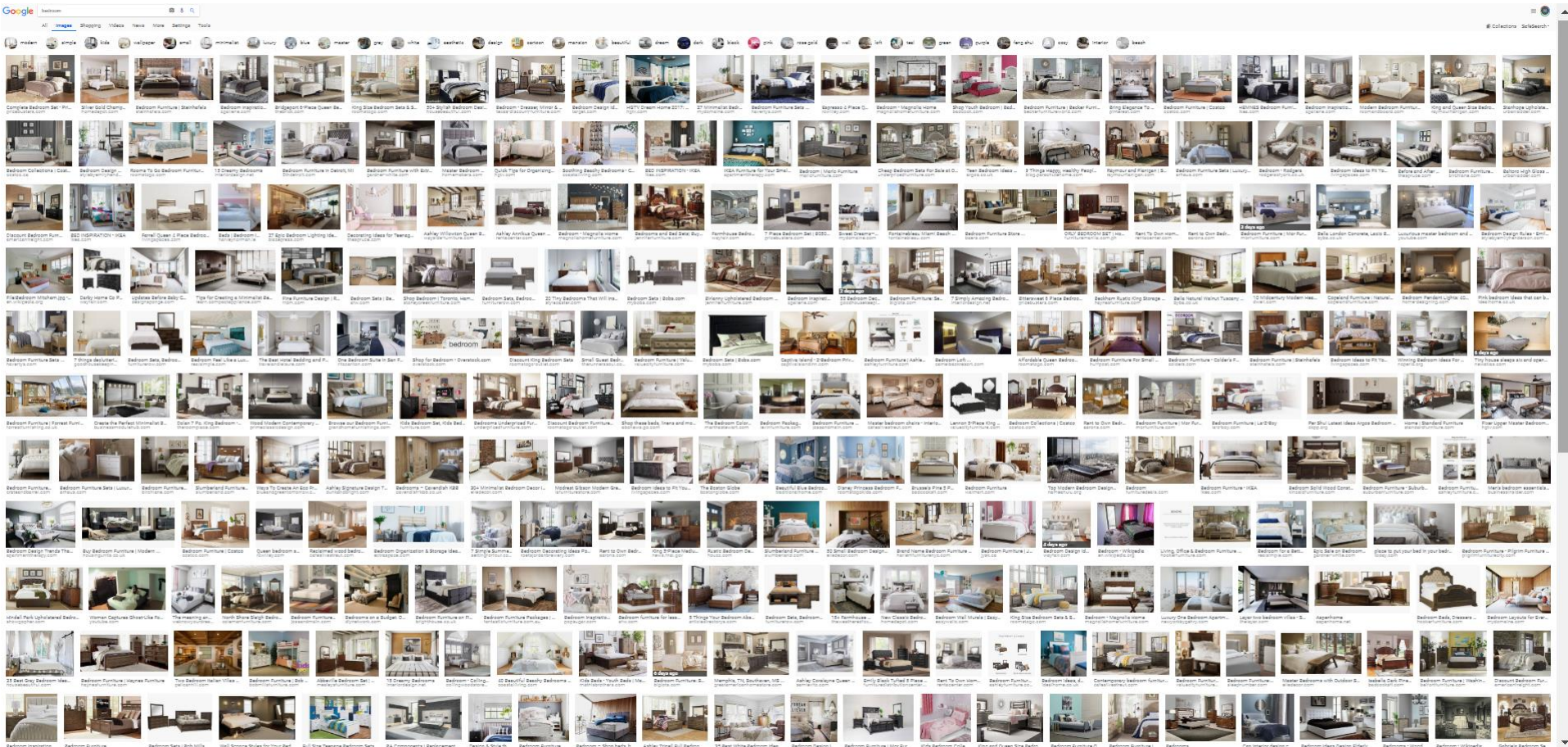


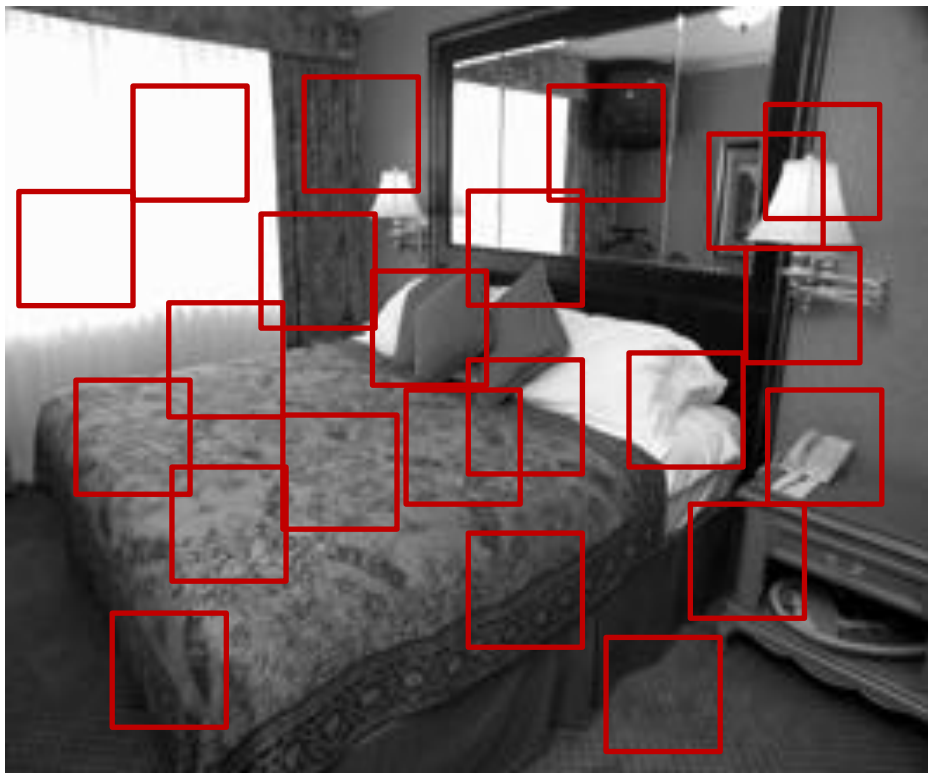
IMAGE CLASSIFICATION



BEDROOM IMAGES

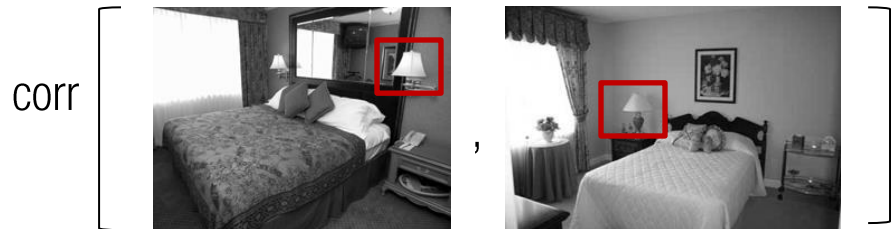


POSSIBLE PATCHES



Millions of patch location and sizes

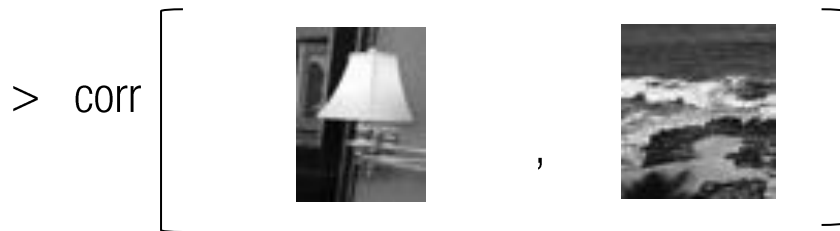
SEARCH SPACE



Search space: $1e+6$



Search space: $1e+12$





WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate to Wikipedia](#)
[Wikipedia store](#)

Interaction

[Help](#)
[About Wikipedia](#)
[Community portal](#)
[Recent changes](#)
[Contact page](#)

[Tools](#)
[What links here](#)



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)
[Current events](#)
[Random article](#)
[Donate to Wikipedia](#)
[Wikipedia store](#)

Interaction

[Help](#)
[About Wikipedia](#)
[Community portal](#)
[Recent changes](#)

BAG OF WORDS

Not logged in Talk Contributions Create account Log in

Article [Talk](#)

[Read](#) [Edit](#) [View history](#)

Computer vision

From Wikipedia, the free encyclopedia

Computer vision is an *interdisciplinary scientific field* that deals with how computers can be made to gain high-level understanding from *digital images* or *videos*. From the perspective of *engineering*, it seeks to automate tasks that the *human visual system* can do.^{[1][2][3]}

Computer vision tasks include methods for *acquiring*, *processing*, *analyzing* and understanding digital images, and extraction of *high-dimensional* data from the real world in order to produce numerical or symbolic information, e.g., in the forms of decisions.^{[4][5][6][7]}

Understanding in this context means the transformation of visual images (the input of the retina) into descriptions of the world that can interface with other thought processes and elicit appropriate action. This image understanding can be seen as the disentangling of symbolic information from image data using models constructed with the aid of geometry, physics, statistics, and learning theory.^[8]

As a *scientific discipline*, computer vision is concerned with the theory behind artificial systems that extract information from images. The image data can take many forms, such as video sequences, views from multiple cameras, or multi-dimensional data from a medical scanner. As a technological discipline, computer vision seeks to apply its theories and models for the construction of computer vision systems.

Sub-domains of computer vision include *scene reconstruction*, event detection, *video tracking*, *object recognition*, *3D pose estimation*, learning, indexing, *motion estimation*, and *image restoration*.^[6]

Contents [hide]

- [Definition](#)
- [History](#)
- [Related fields](#)
 - [3.1 Artificial Intelligence](#)
 - [3.2 Information Engineering](#)

Article [Talk](#)

[Read](#) [Edit](#) [View history](#)

Minnesota

From Wikipedia, the free encyclopedia

This article is about the U.S. state of Minnesota. For other uses, see [Minnesota \(disambiguation\)](#).

Minnesota (/ˌmɪnɪˈsoʊtə/ (ⓘ) listen) is a *state* in the *Upper Midwest* and *northern* regions of the *United States*. Minnesota was admitted as the 32nd U.S. state on May 11, 1858, created from the eastern half of the *Minnesota Territory*. The state has a large number of lakes, and is known by the slogan the "Land of 10,000 Lakes". Its official motto is *L'Étoile du Nord* (*French: Star of the North*).

Minnesota is the **12th largest in area** and the **22nd most populous** of the U.S. states; nearly 60% of its residents live in the *Minneapolis–Saint Paul* metropolitan area (known as the "Twin Cities"), the center of transportation, business, industry, education, and government, and home to an internationally known arts community. The remainder of the state consists of western *prairies* now given over to intensive agriculture; *deciduous* forests in the southeast, now partially cleared, farmed, and settled; and the less populated *North Woods*, used for mining, forestry, and recreation.

Minnesota was inhabited by various indigenous peoples for thousands of years prior to the arrival of Europeans. French explorers, missionaries, and fur traders began exploring the region in the 17th century, encountering the *Dakota* and *Ojibwe/Anishinaabe* tribes. Much of what is today Minnesota was part of the *vast French holding of Louisiana*, which was *purchased by the United States* in 1803. Following several territorial reorganizations, Minnesota in its current form was admitted as the country's 32nd state on May 11, 1858. Like many Midwestern states, it remained sparsely populated and centered on lumber and agriculture. During the 19th and early 20th centuries, a large number of European immigrants, mainly from *Scandinavia* and *Germany*, began to settle the state, which remains a center of *Scandinavian American* and *German American* culture.



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes

BAG OF WORDS

Article Talk

Read Edit View history

Computer vision

From Wikipedia, the free encyclopedia

Computer vision is an interdisciplinary scientific field that deals with how computers can be made to gain high-level understanding from digital images or videos. From the perspective of engineering, it seeks to automate the tasks that would be done by the human visual system.

Computer vision tasks include methods for **acquiring**, **processing**, **analyzing** and understanding digital images, and extraction of **high-dimensional** data from the real world in order to produce numerical or symbolic information. Understanding in this context means the transformation of visual images (the input of the retina) into descriptions of the world that can interface with other thought processes and elicit appropriate action. This image understanding is often done by using symbolic information from image data using models constructed with the aid of geometry, physics, statistics, and learning theory.^[R]

As a **scientific discipline**, computer vision is concerned with the theory behind artificial systems that extract information from images. The image data can take many forms, such as video sequences, views from multiple cameras, or images from a scanner. As a technological discipline, computer vision seeks to apply its theories and models for the construction of computer vision systems.

Sub-domains of computer vision include **scene reconstruction**, event detection, **video tracking**, **object recognition**, **3D pose estimation**, learning, indexing, **motion estimation**, and **image restoration**.^[R]

Contents [hide]

- Definition
- History
- Related fields
 - Artificial Intelligence
 - Information Engineering

Article Talk

Minnesota

From Wikipedia, the free encyclopedia

*This article is about the U.S. state of Minnesota. For other uses, see *Minnesota (disambiguation)*.*

Minnesota (/ˌmɪnɪˈsoʊtə/ (hlen listen)) is a state in the Upper Midwest and northern regions of the United States. Minnesota was admitted as the 32nd U.S. state on May 11, 1858, created from the western part of the **Territory**. The state has a large number of lakes, and is known by the slogan the "Land of 10,000 Lakes". Its official motto is *L'Étoile du Nord* (*French: Star of the North*).

Minnesota is the 12th largest in area and the 22nd most populous of the U.S. states; nearly 60% of its residents live in the Minneapolis–Saint Paul metropolitan area (known as the "Twin Cities"), a major business, industry, education, and government, and home to an internationally known arts community. The remainder of the state consists of western prairies now given over to intensive agriculture; the southeastern southeast, now partially cleared, farmed, and settled; and the less populated North Woods, used for mining, forestry, and recreation.

Minnesota was inhabited by various indigenous peoples for thousands of years prior to the arrival of Europeans. French explorers, missionaries, and fur traders began exploring the region in the 17th century, and the state was named after the **Dakota** and **Ojibwe/Anishinaabe** tribes. Much of what is today Minnesota was part of the vast French holding of Louisiana, which was purchased by the United States in 1803. Following the Louisiana Purchase, Minnesota in its current form was admitted as the country's 32nd state on May 11, 1858. Like many Midwestern states, it remained sparsely populated and centered on lumber and agriculture for several centuries, a large number of European immigrants, mainly from Scandinavia and Germany, began to settle the state, which remains a center of Scandinavian American and German American culture.

Image: 145

Video: 13

Science: 7

Space: 6

Camera: 20

Cold: 0

University: 2

Mountain: 0

Image: 0

Video: 0

Science: 3

Space: 0

Camera: 1

Cold: 1

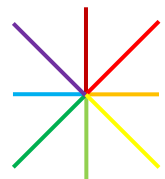
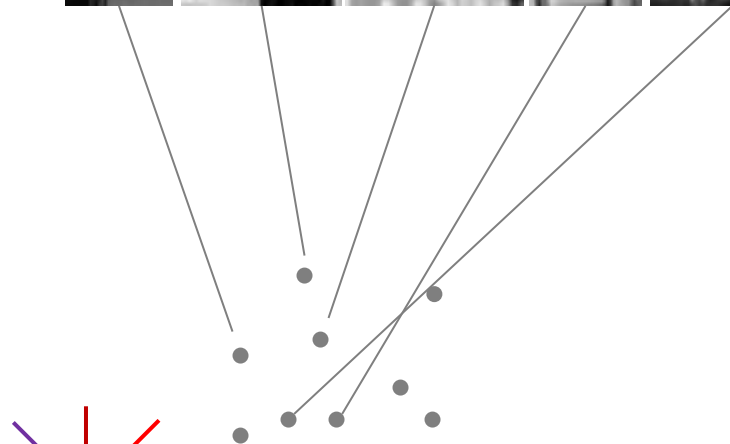
University: 28

Mountain: 5

HOW TO CONSTRUCT VISUAL DICTIONARY



Lamp



\mathbb{R}^d

Visual word descriptor

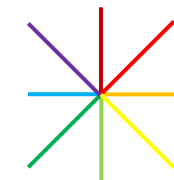
HOW TO CONSTRUCT VISUAL DICTIONARY



Lamp



Mean lamp



\mathbb{R}^d

Visual word descriptor

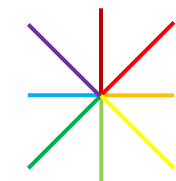
HOW TO CONSTRUCT VISUAL DICTIONARY



Lamp

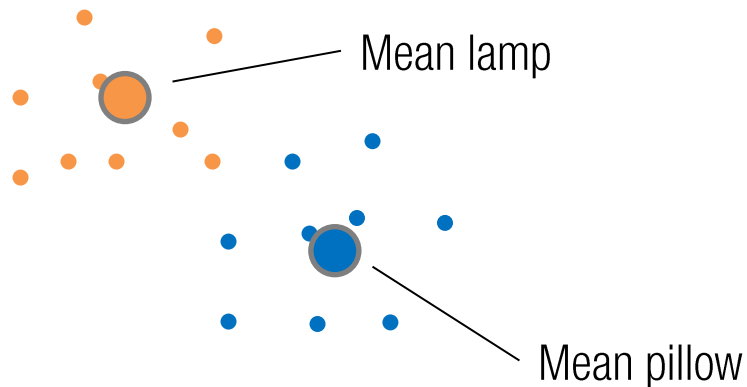


Pillow



\mathbb{R}^d

Visual word descriptor



FEATURE REPRESENTATION OF LOCAL PATCH



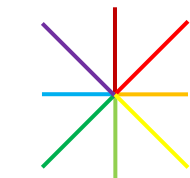
Lamp



Pillow

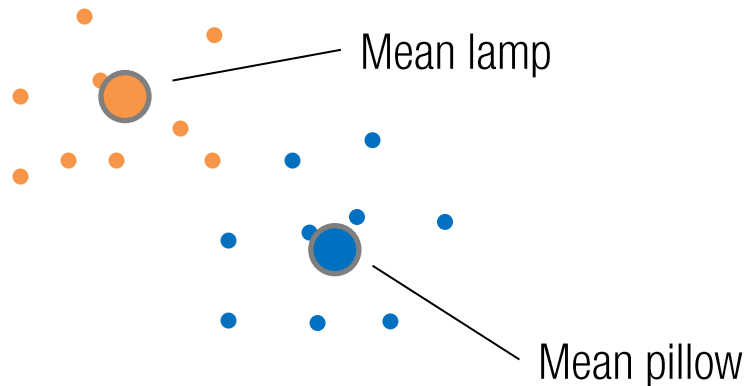


SIFT()

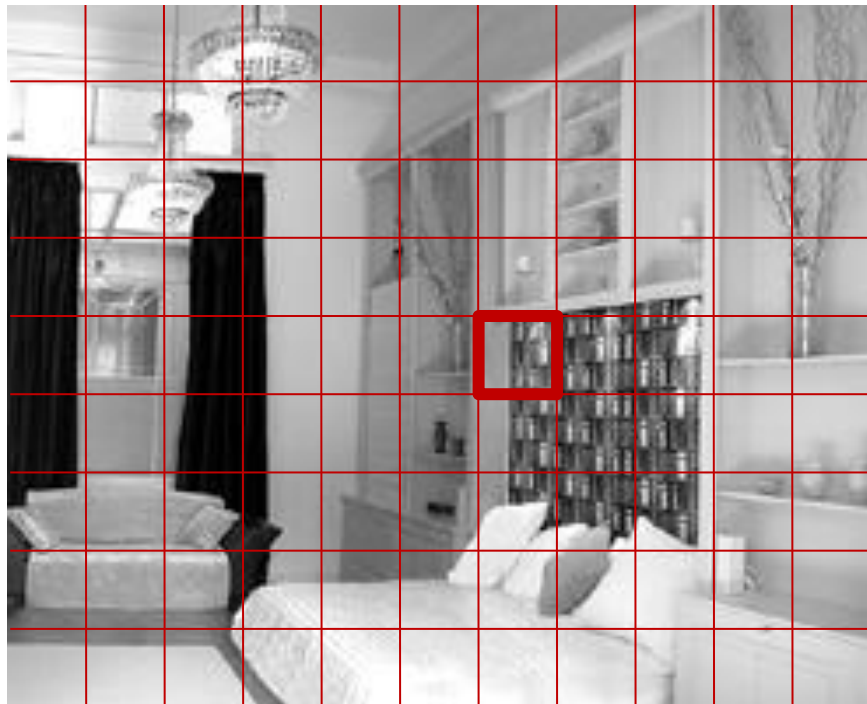


\mathbb{R}^{128}

Visual word descriptor



DICTIONARY CONSTRUCTION FROM TRAINING DATA

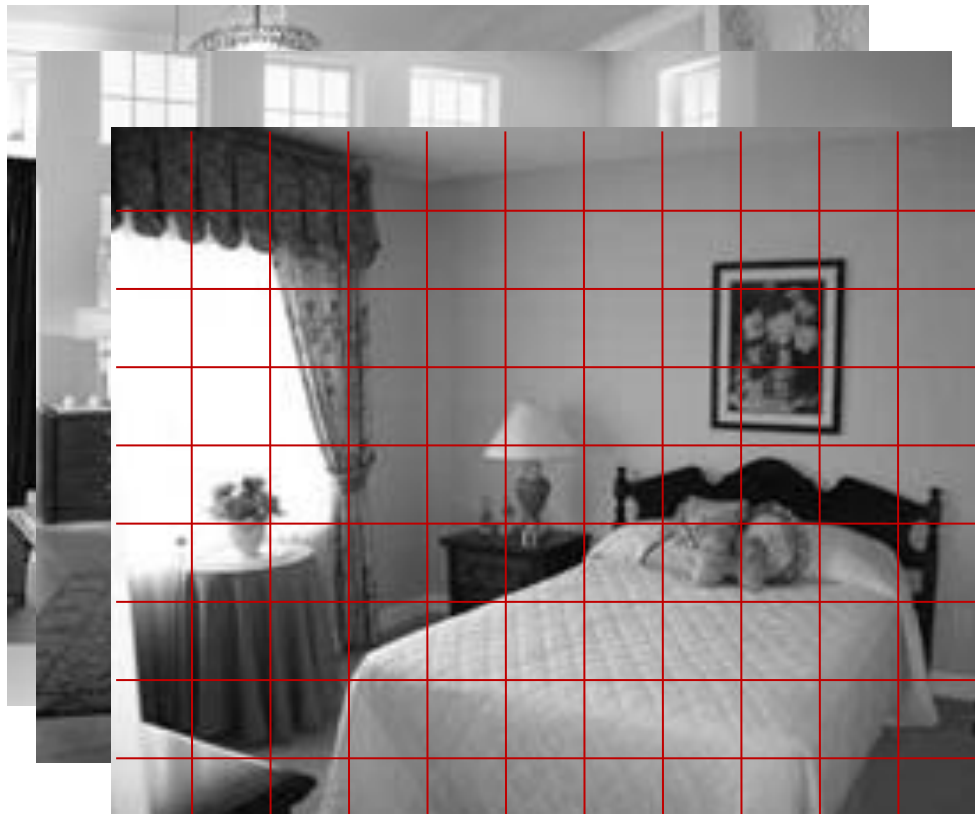


SIFT descriptor



$$x_i \in \mathbb{R}^{128}$$

DICTIONARY CONSTRUCTION FROM TRAINING DATA



SIFT descriptor



$$x_i \in \mathbb{R}^{128}$$

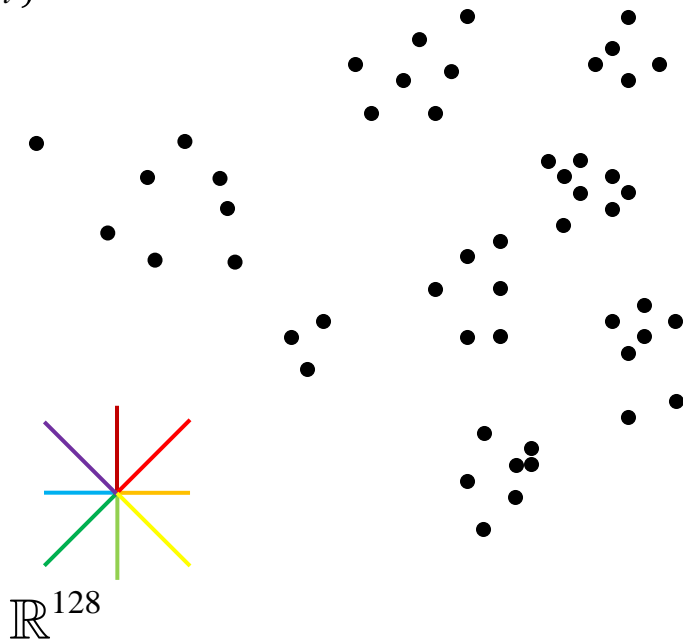
Pool of SIFT descriptors

$$X = \{x_1 \quad \cdots \quad x_n\}$$

DICTIONARY CONSTRUCTION FROM TRAINING DATA

Pool of SIFT descriptors

$$X = \{x_1 \quad \cdots \quad x_n\}$$

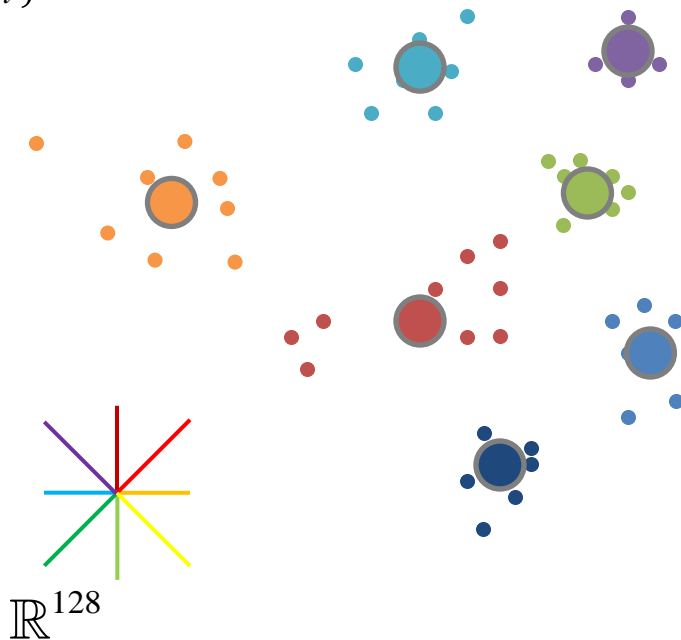


DICTIONARY CONSTRUCTION FROM TRAINING DATA

Pool of SIFT descriptors

$$X = \{x_1 \quad \cdots \quad x_n\}$$

K-means clustering



DICTIONARY CONSTRUCTION FROM TRAINING DATA

Pool of SIFT descriptors

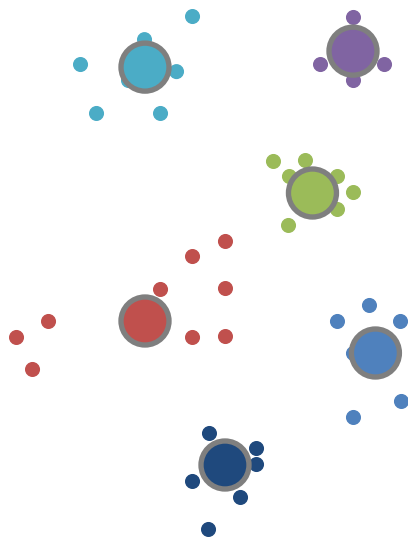
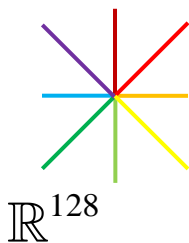
$$X = \{x_1 \quad \cdots \quad x_n\}$$

K-means clustering

Dictionary~centroids

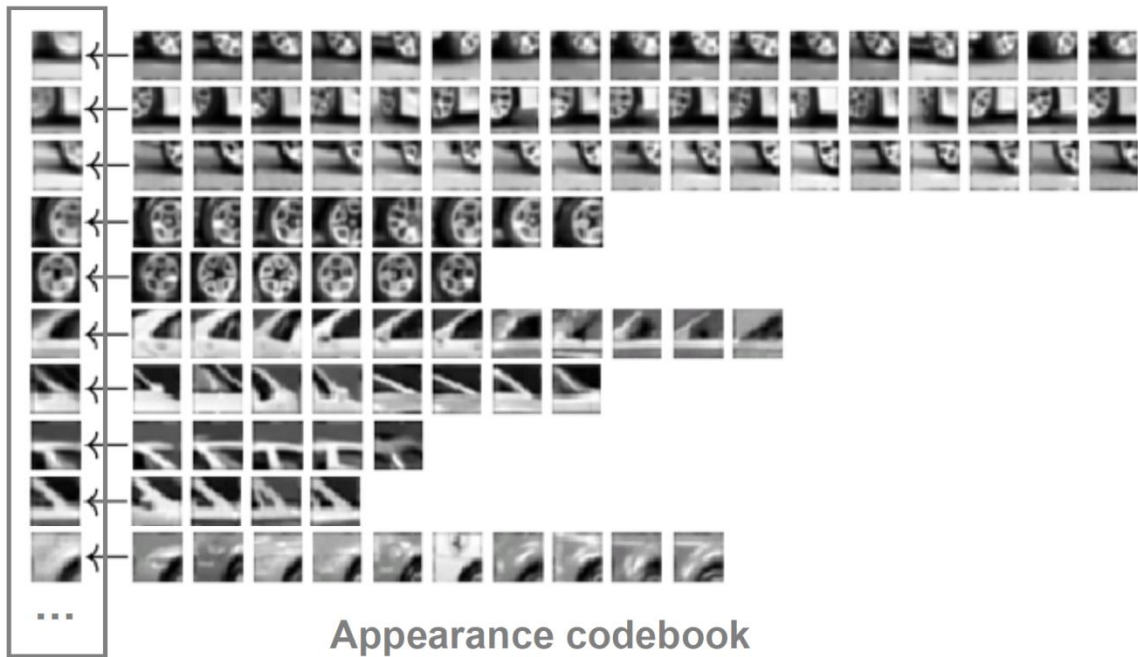
$$Y = \{y_1 \quad \cdots \quad y_k\}$$

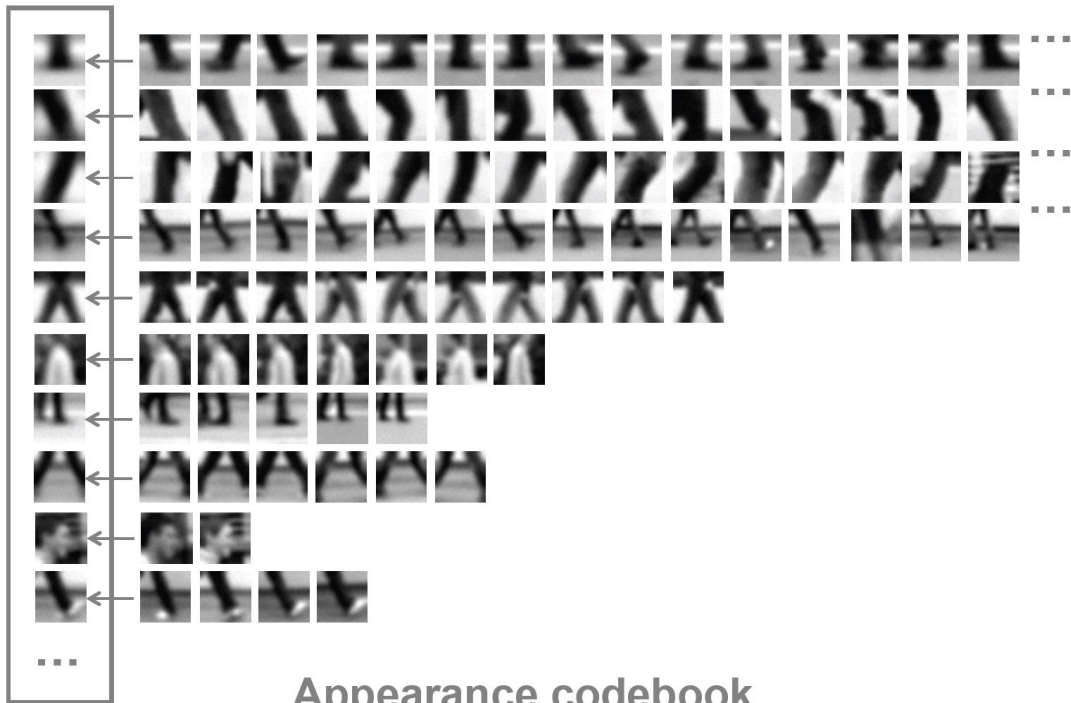
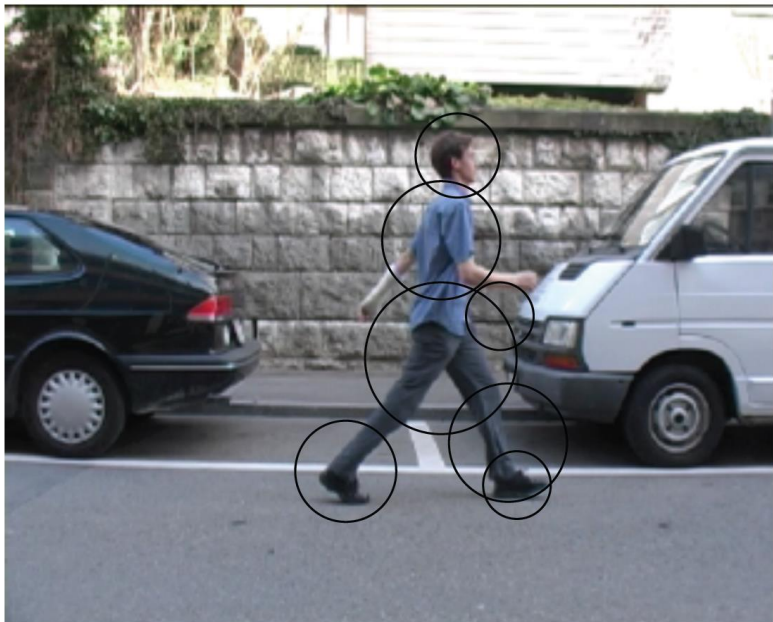
y_i



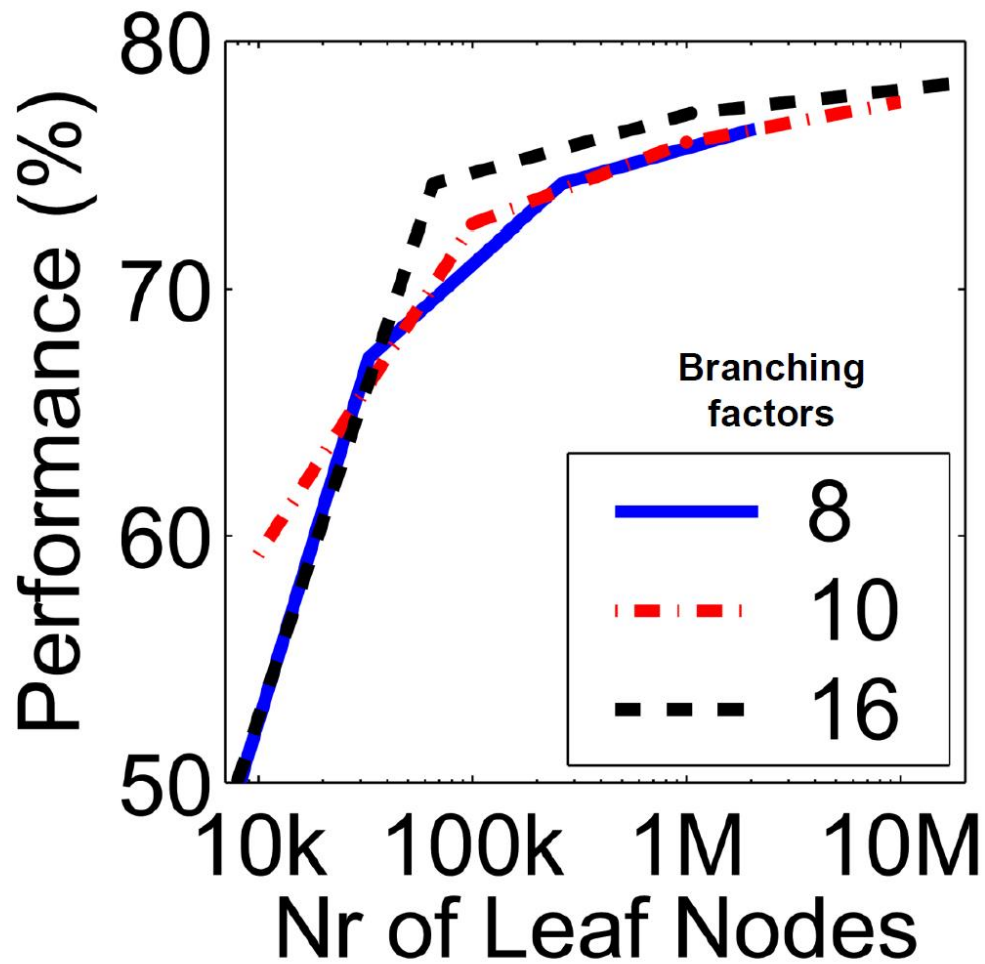
CLUSTERED VISUAL PATCHES USING K-MEANS



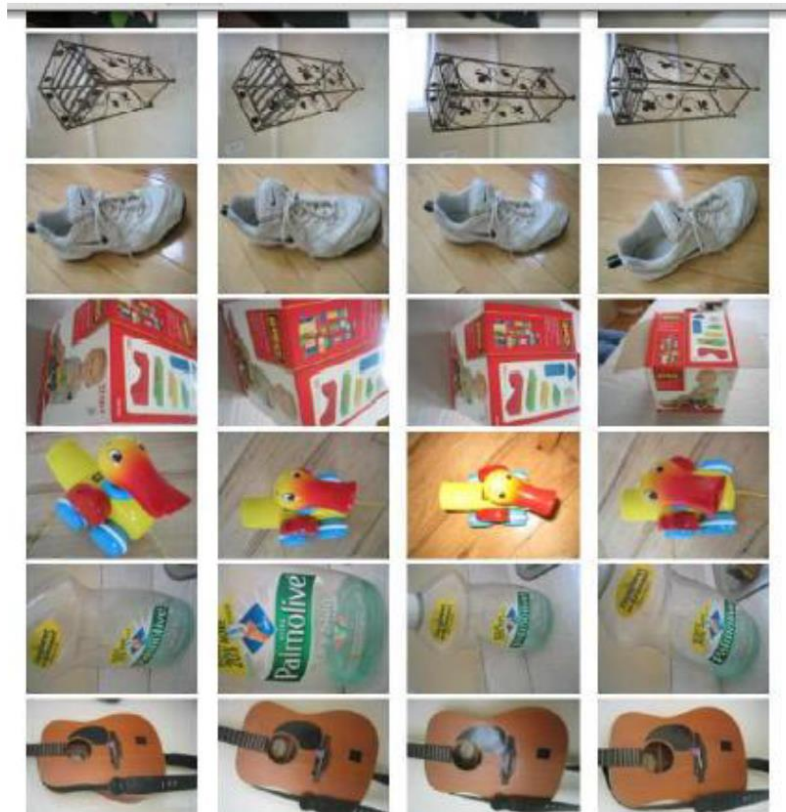




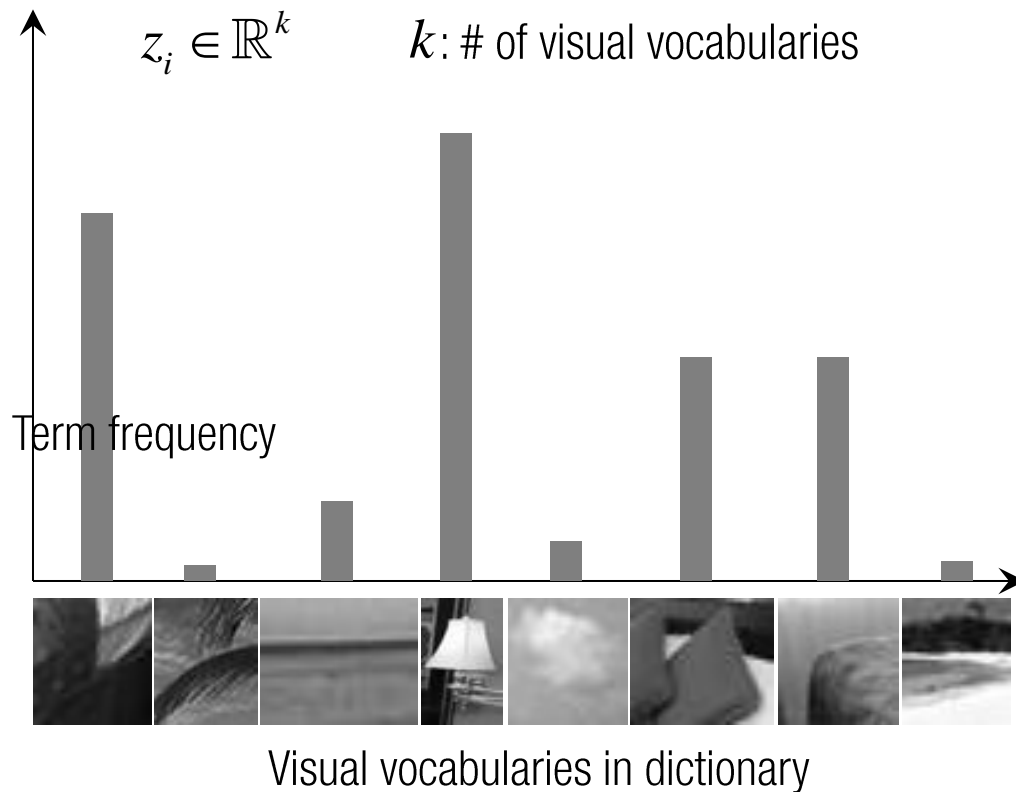
Appearance codebook



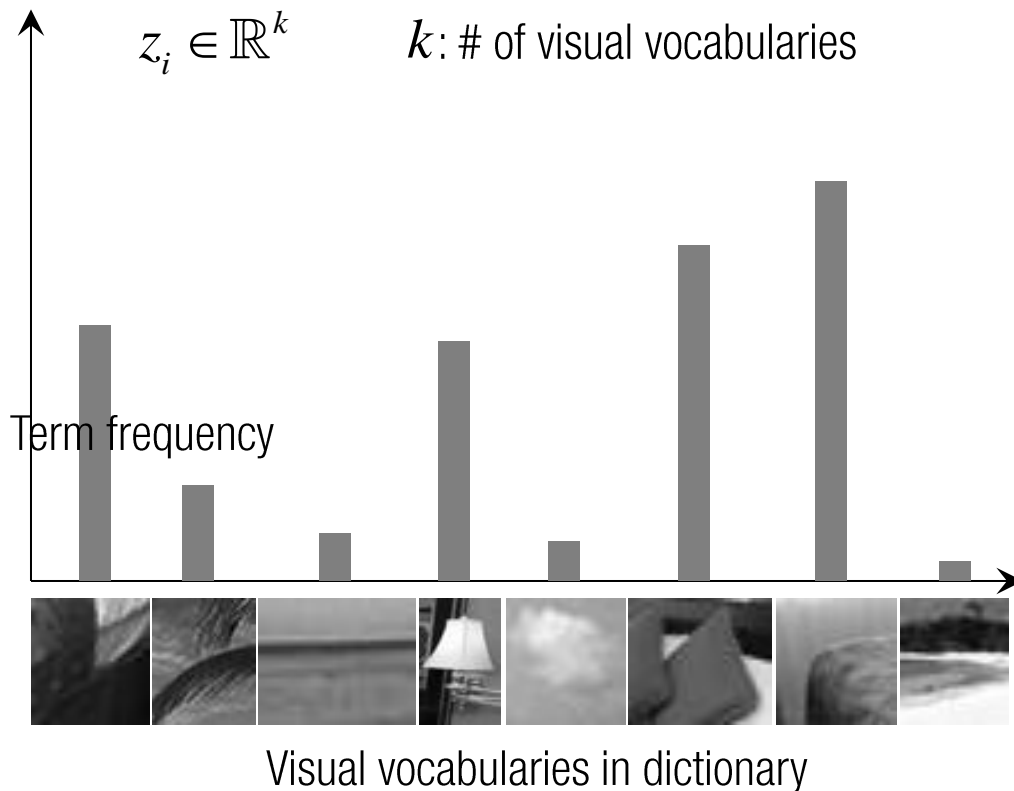
Results for recognition task with 6347 images



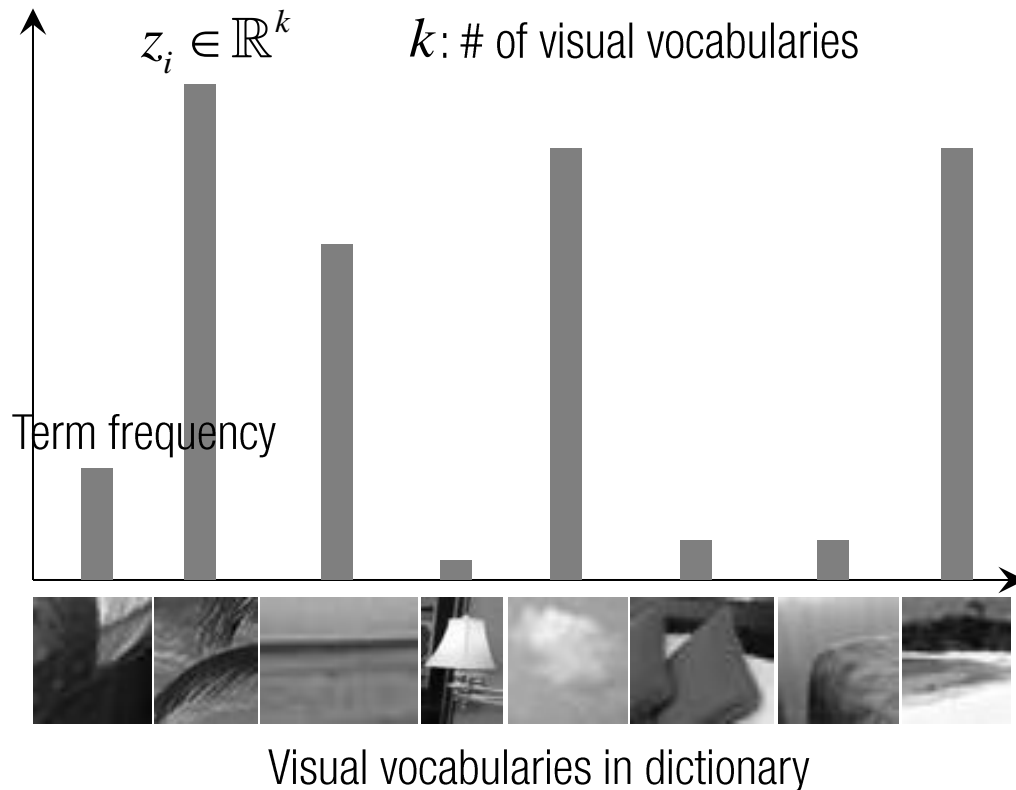
VISUAL BAG-OF-WORD REPRESENTATION



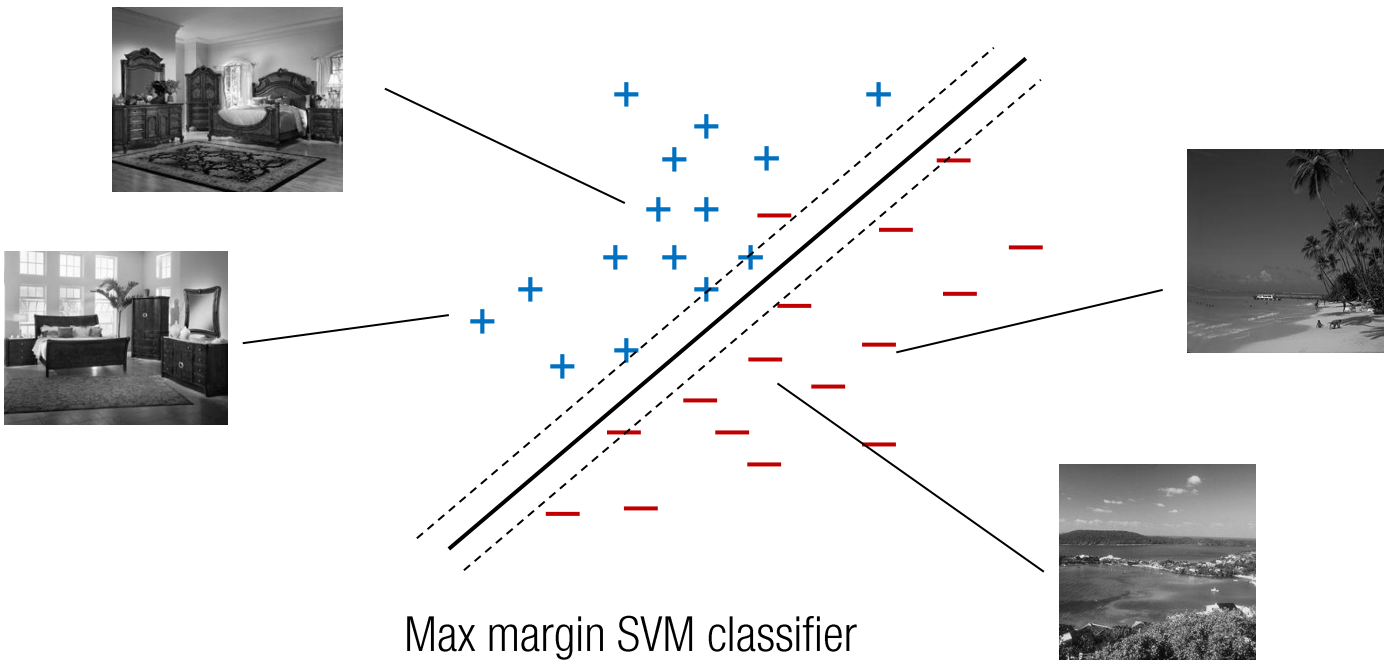
VISUAL BAG-OF-WORD REPRESENTATION



VISUAL BAG-OF-WORD REPRESENTATION



CLASSIFICATION



$$z \cdot w + b > 0 \quad \text{Positive D.}$$

$$z \cdot w + b < 0 \quad \text{Negative D.}$$

ACCURACY MEASURE

	Prediction	
	Bedroom	Beach
Bedroom	0.7	0.3
Beach		

Ground truth label
Confusion matrix

$$\frac{\text{\# of correct prediction on Bedroom data}}{\text{\# of Bedroom data}}$$

$$\frac{\text{\# of incorrect prediction on Bedroom data}}{\text{\# of Bedroom data}}$$

ACCURACY MEASURE

	Prediction	
	Bedroom	Beach
Bedroom	0.7	0.3
Beach	0.2	0.8

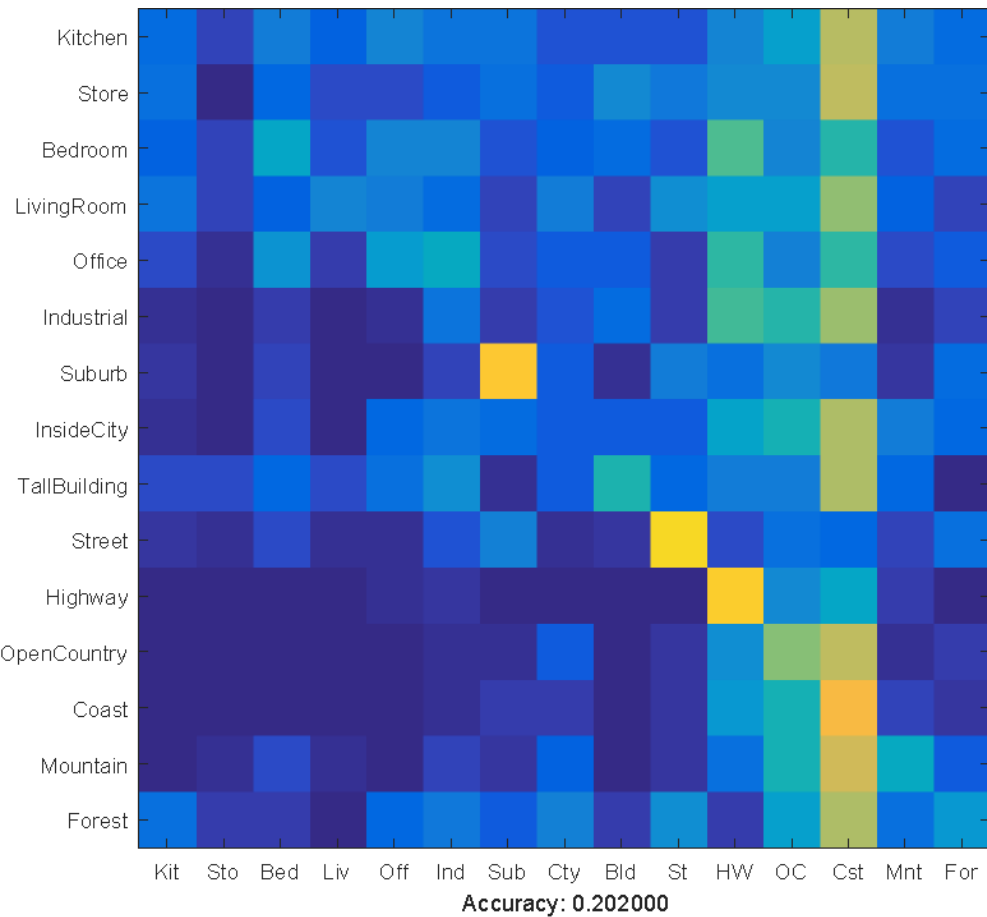
Ground truth label
Confusion matrix

$$\frac{\text{\# of correct prediction on Bedroom data}}{\text{\# of Bedroom data}}$$

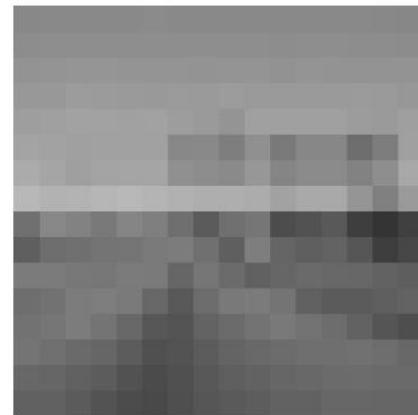
$$\frac{\text{\# of incorrect prediction on Bedroom data}}{\text{\# of Bedroom data}}$$

Accuracy: mean of correct predictions
 $(0.7+0.8)/2 = 0.75$

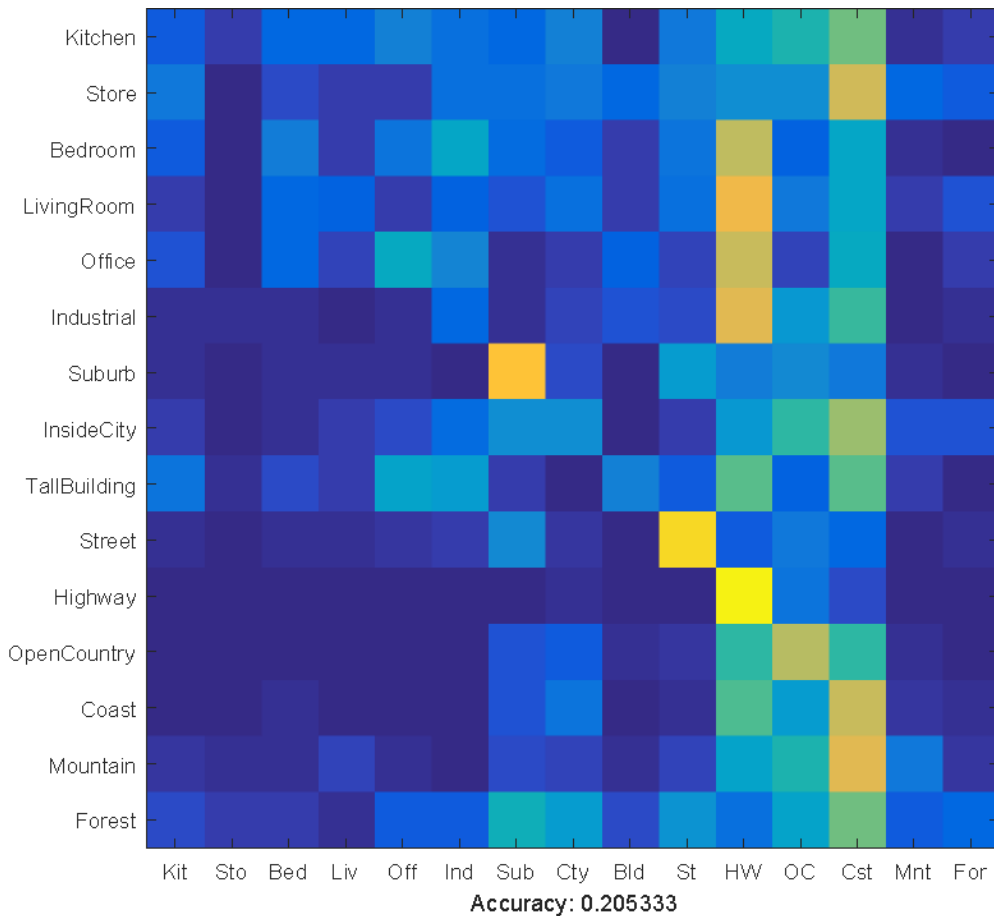
COMPARISON



Tiny image representation + NN



COMPARISON



Tiny image representation + KNN (10)

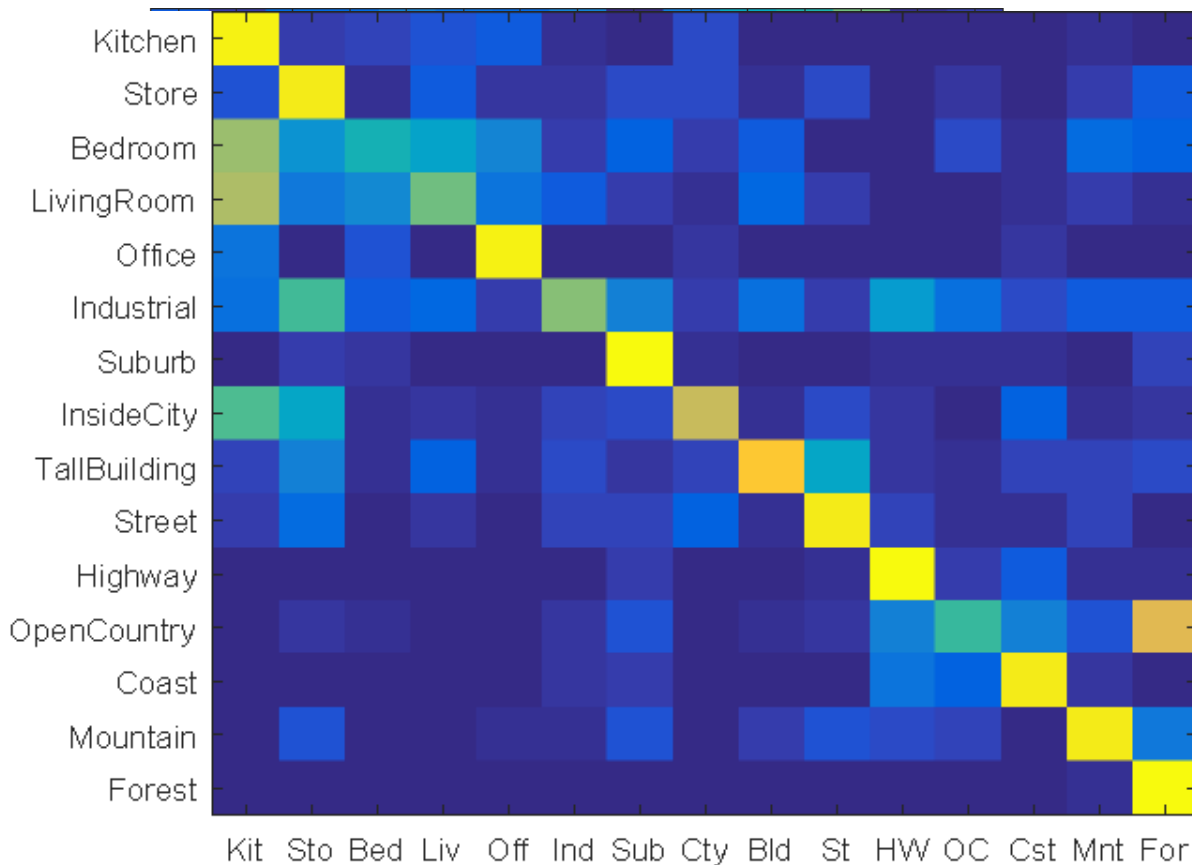


(a) Image



(b) Tiny Image

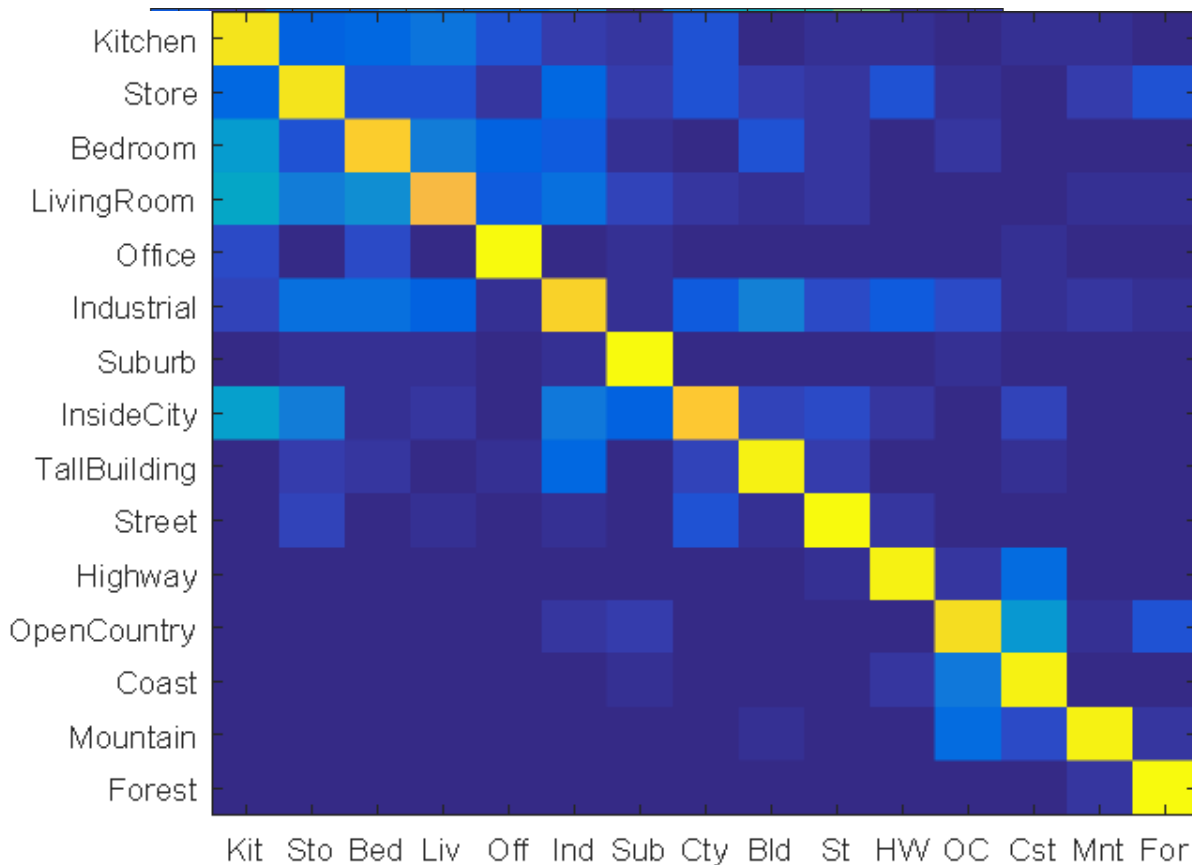
COMPARISON



BoW + KNN (10)

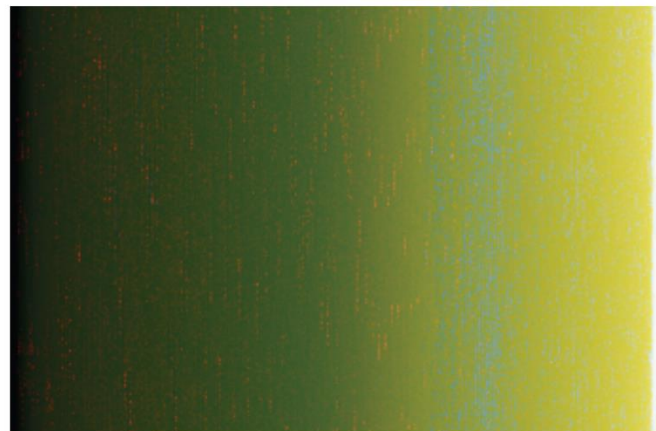
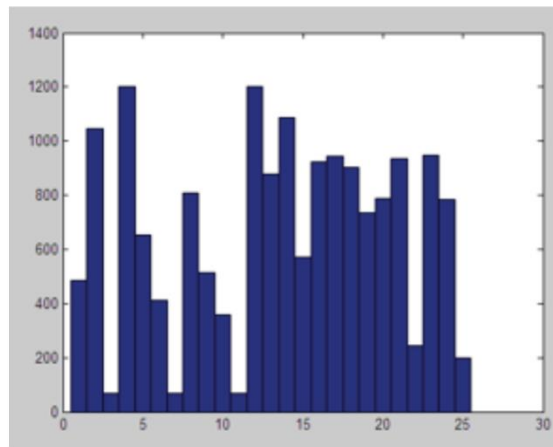
Accuracy: 0.512667

COMPARISON

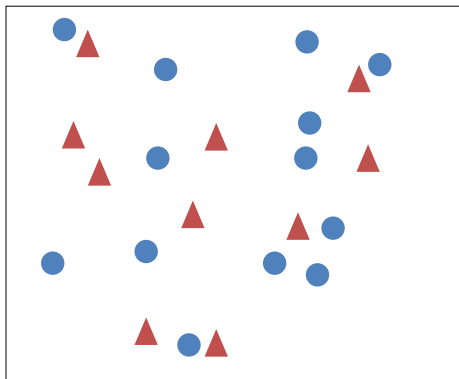


BoW + SVM

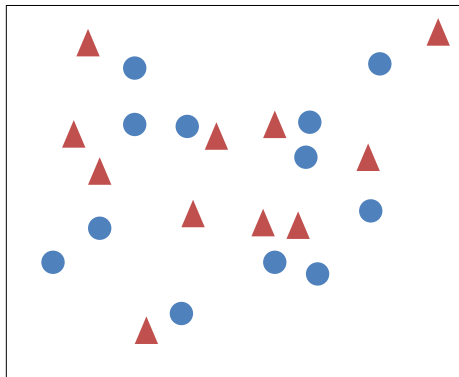
Accuracy: 0.629333



MATCHING

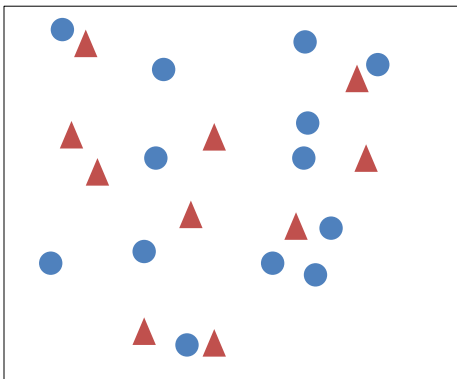


Feature

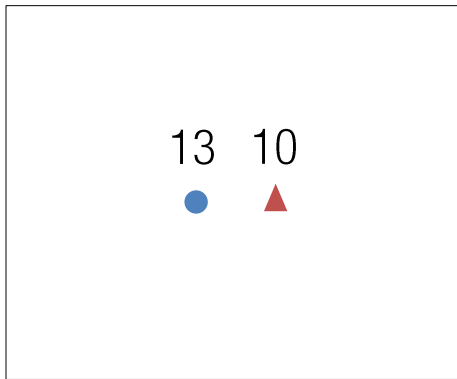


Feature

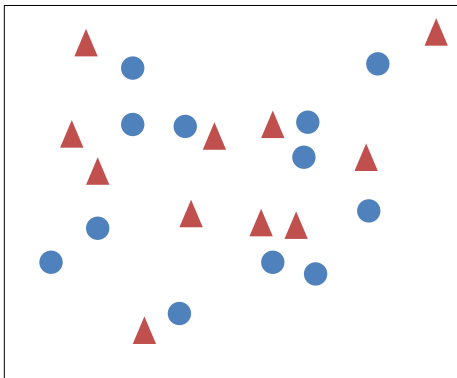
MATCHING



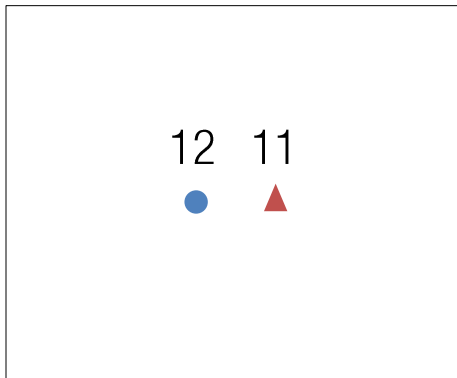
Feature



Histogram

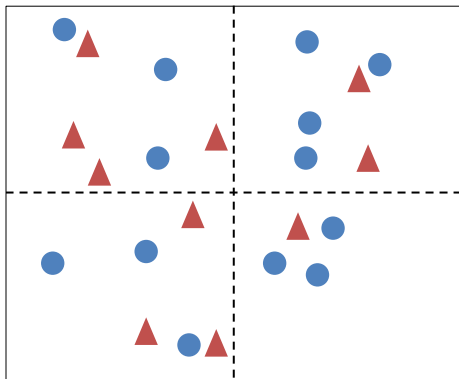


Feature

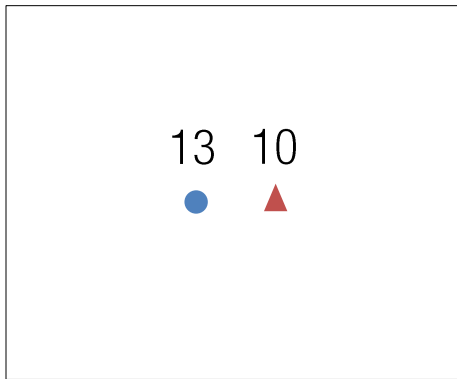


Histogram

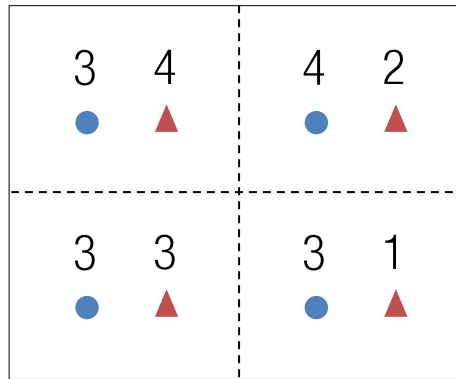
MATCHING



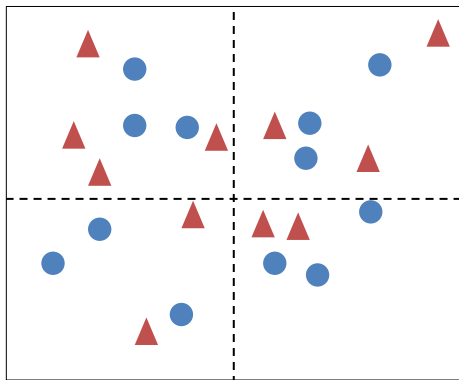
Feature



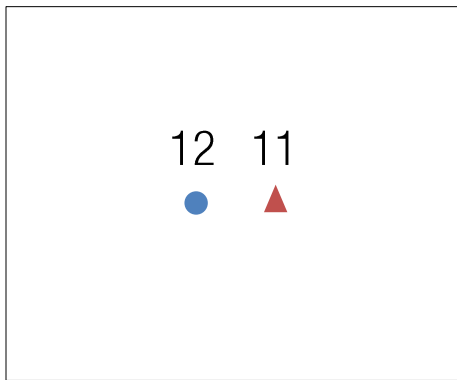
Histogram



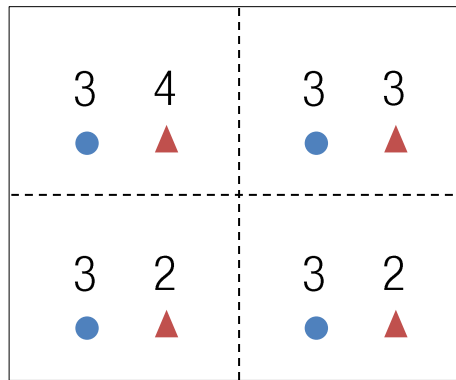
Histogram



Feature

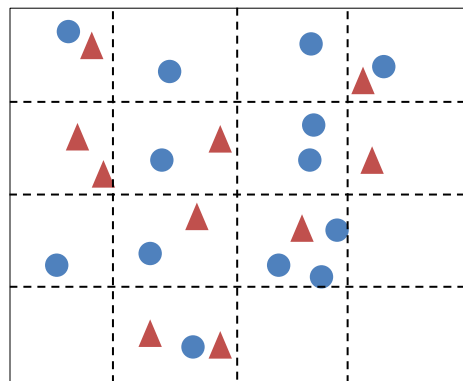


Histogram

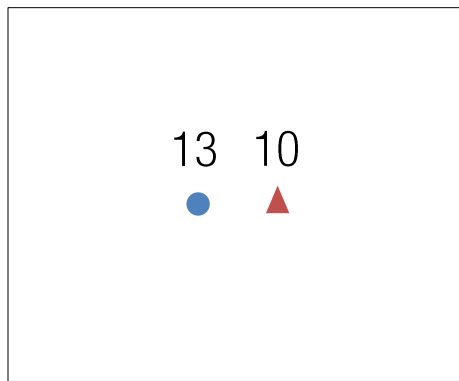


Histogram

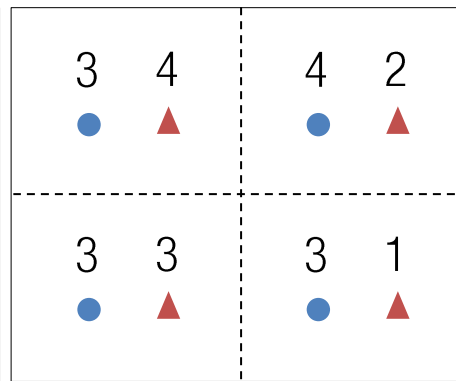
MATCHING



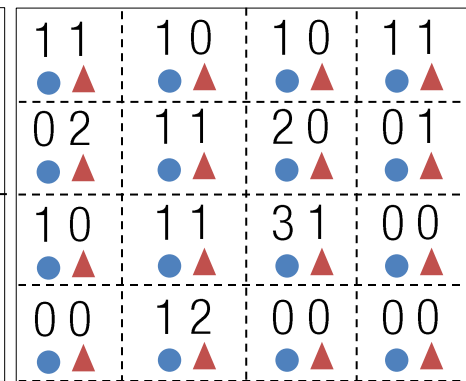
Feature



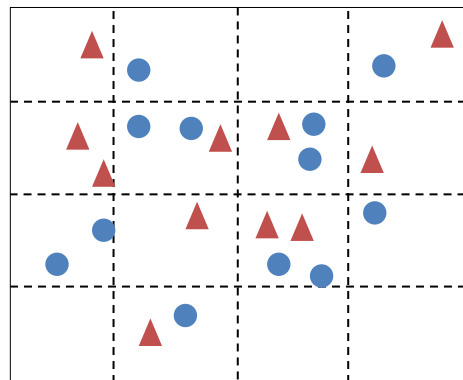
Histogram



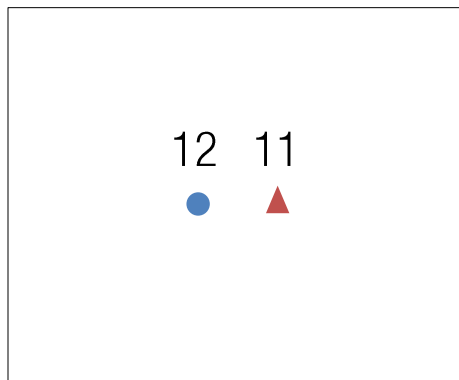
Histogram



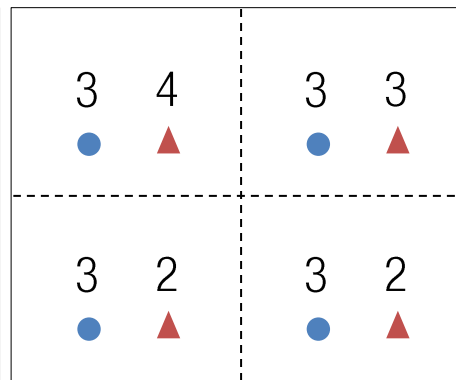
Histogram



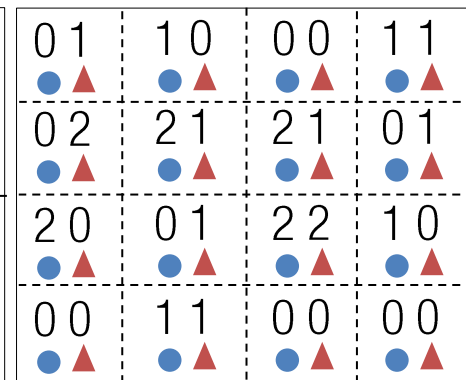
Feature



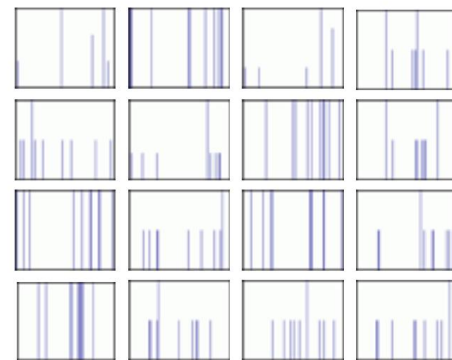
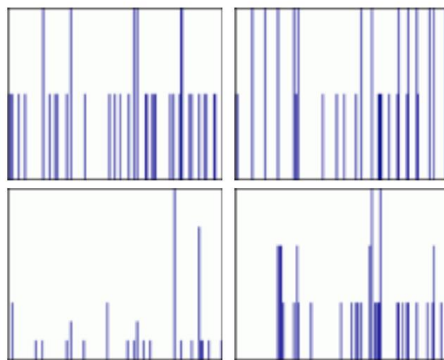
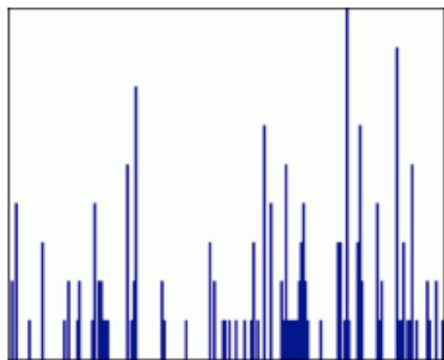
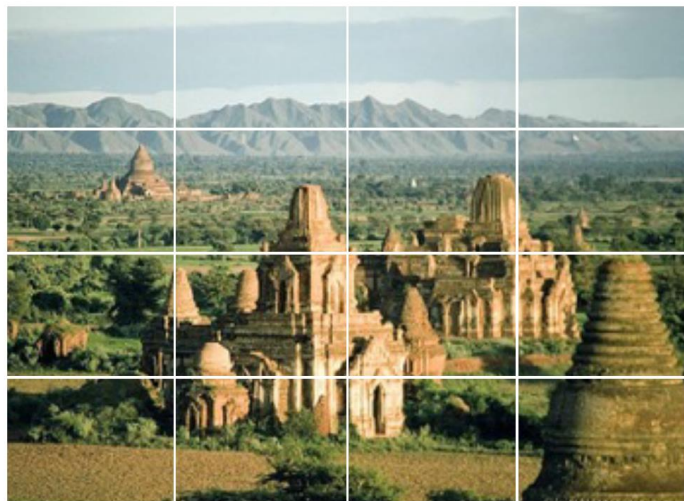
Histogram



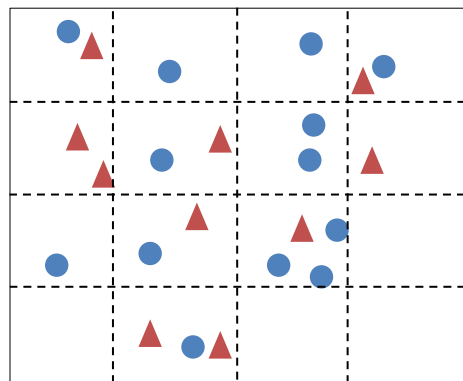
Histogram



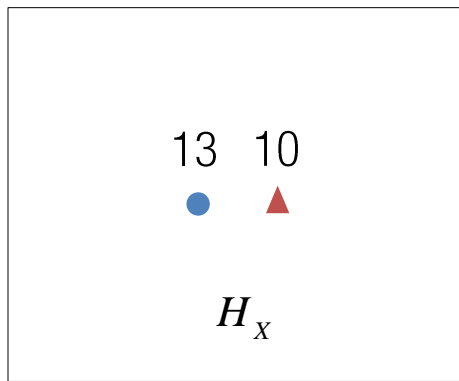
Histogram



SIMILARITY MEASURE ~ PYRAMID MATCHING KERNEL



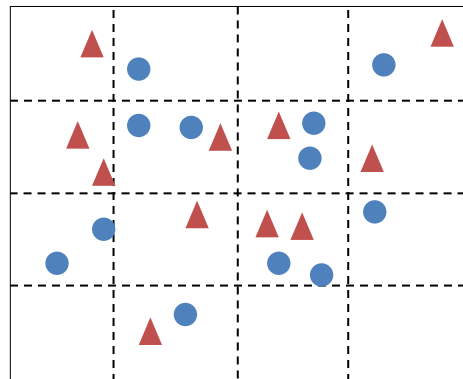
Feature



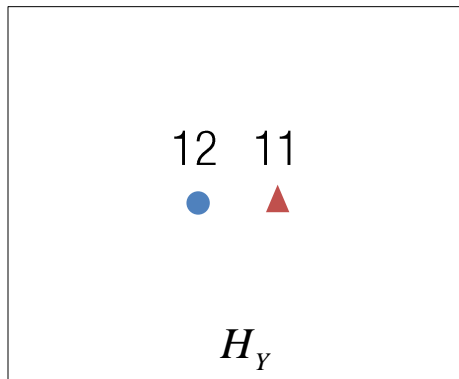
Histogram

Histogram intersection:

$$\mathcal{I}(H_X^0, H_Y^0) = \sum_{d=1}^D \min(\underline{H_X^0(d)}, \underline{H_Y^0(d)})$$



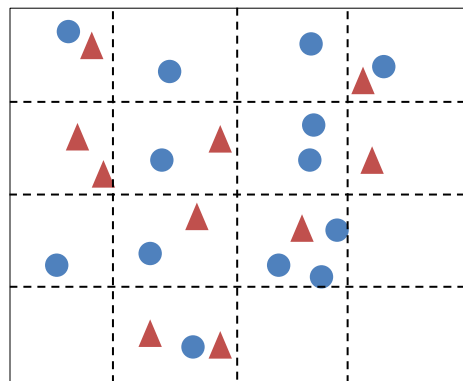
Feature



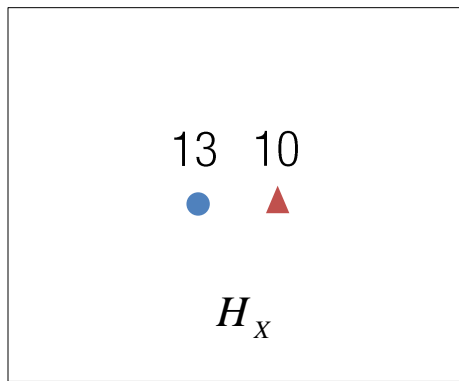
Histogram



SIMILARITY MEASURE ~ PYRAMID MATCHING KERNEL



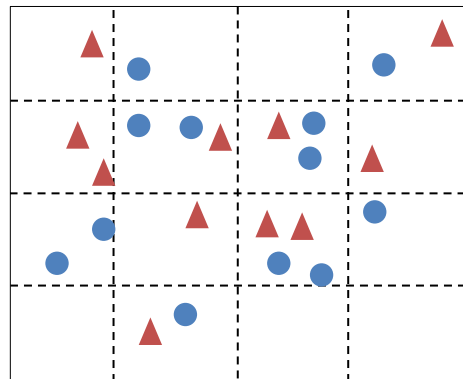
Feature



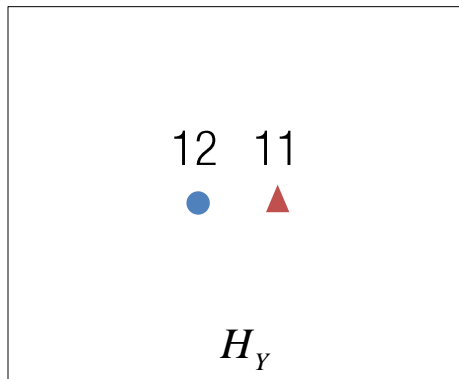
Histogram

Histogram intersection:

$$\mathcal{I}(H_X^0, H_Y^0) = \sum_{d=1}^D \frac{\min(H_X^0(d), H_Y^0(d))}{\max(H_X^0(d), H_Y^0(d))}$$



Feature

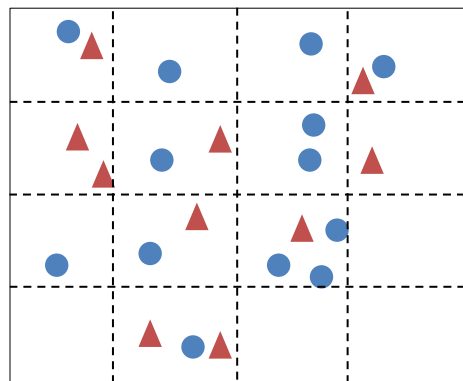


Histogram

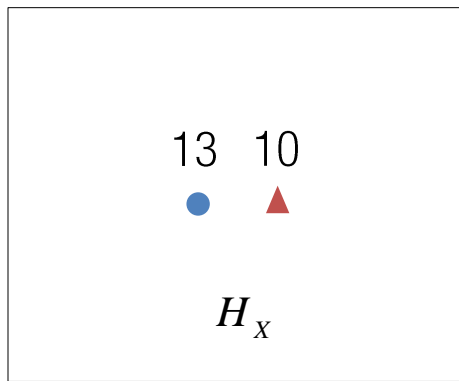
12 10



SIMILARITY MEASURE ~ PYRAMID MATCHING KERNEL



Feature

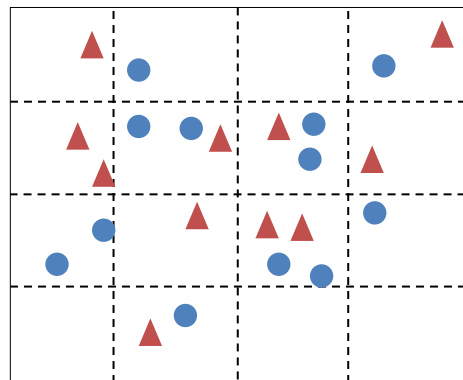


Histogram

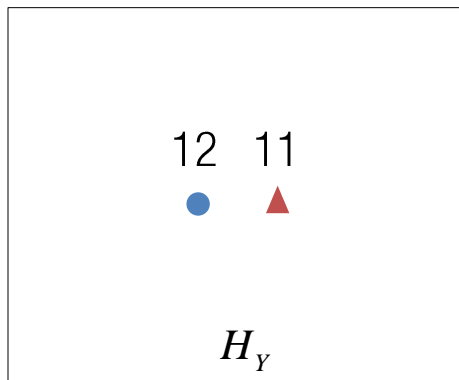
Histogram intersection:

$$\mathcal{I}(H_X^0, H_Y^0) = \frac{\sum_{d=1}^D \min(H_X^0(d), H_Y^0(d))}{22}$$

22

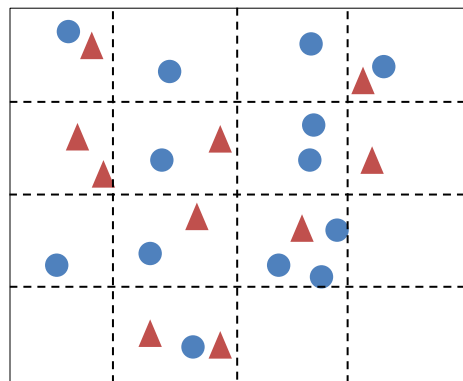


Feature

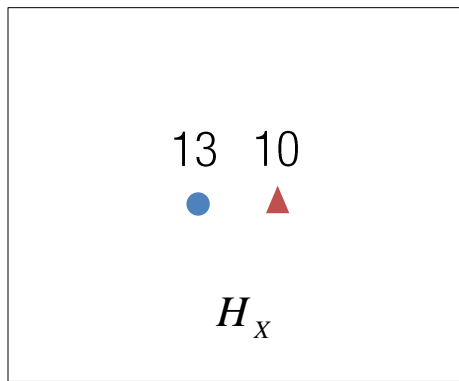


Histogram

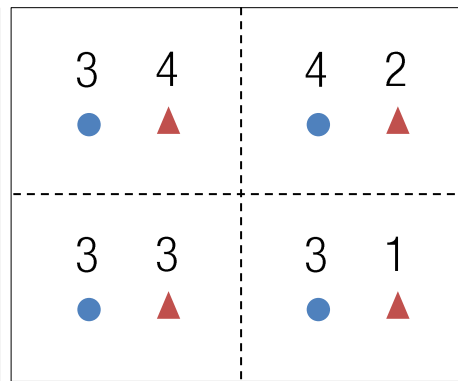
SIMILARITY MEASURE ~ PYRAMID MATCHING KERNEL



Feature



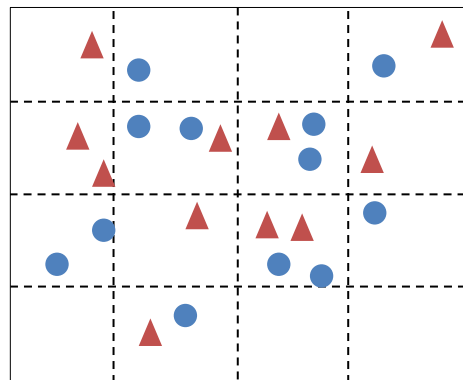
Histogram



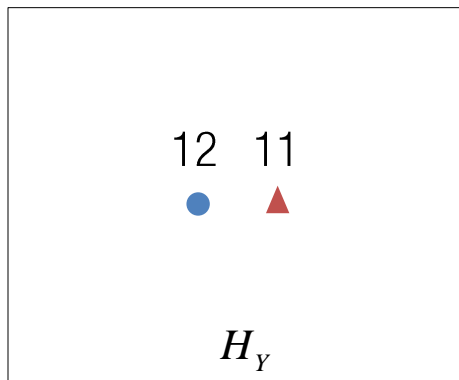
Histogram

Histogram intersection:

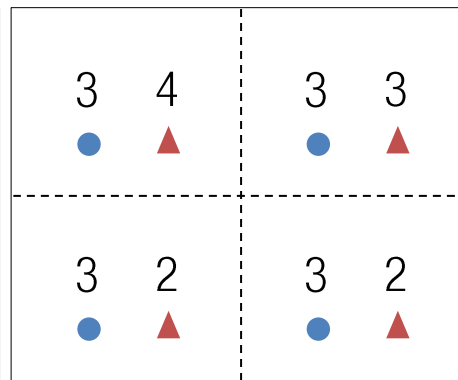
$$\mathcal{I}(H_x^1, H_y^1) = \sum_{d=1}^D \min(H_x^1(d), H_y^1(d))$$



Feature

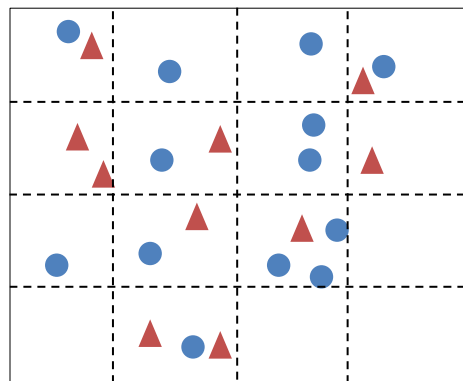


Histogram

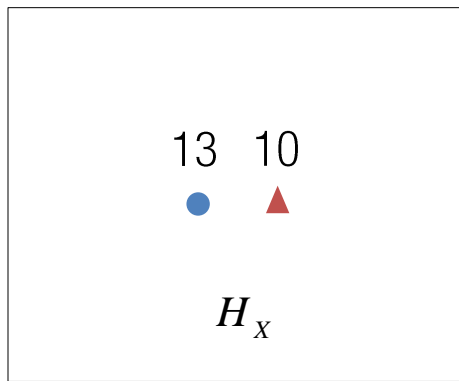


Histogram

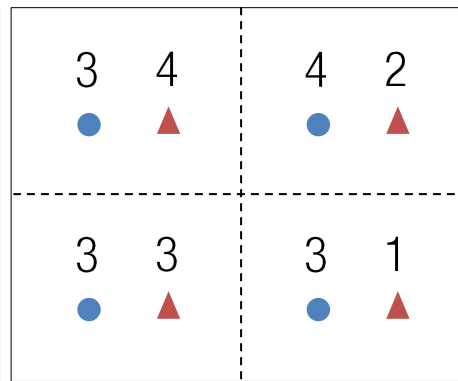
SIMILARITY MEASURE ~ PYRAMID MATCHING KERNEL



Feature



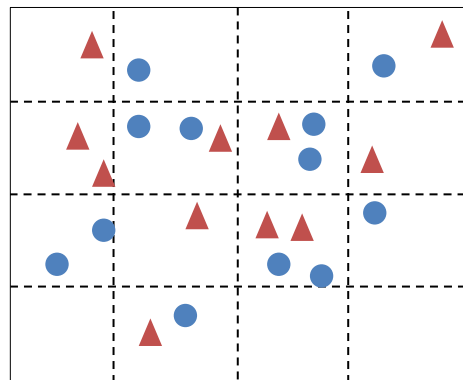
Histogram



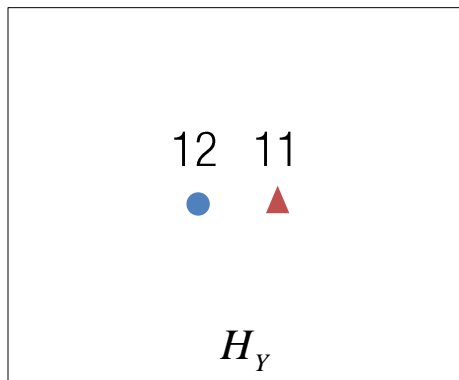
Histogram

Histogram intersection:

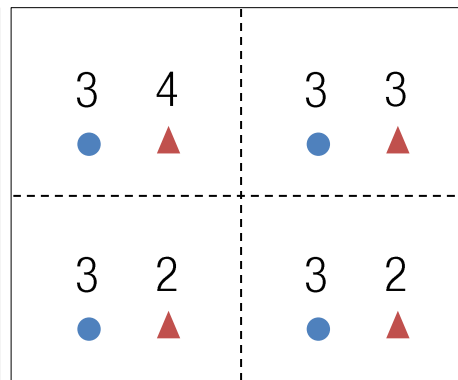
$$\mathcal{I}(H_x^1, H_y^1) = \sum_{d=1}^D \min(H_x^1(d), H_y^1(d))$$



Feature



Histogram

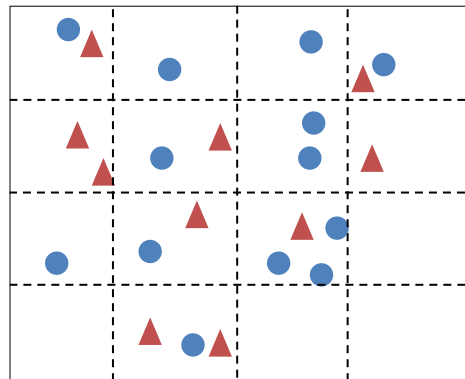


Histogram

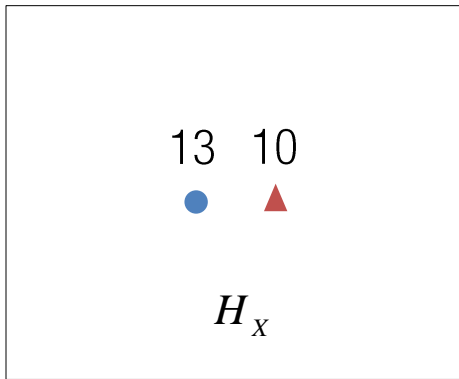
=2

=3

SIMILARITY MEASURE ~ PYRAMID MATCHING KERNEL

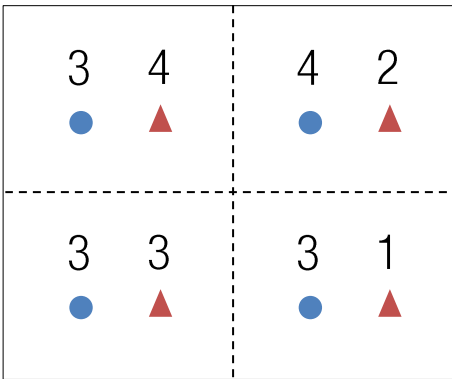


Feature



H_X

Histogram

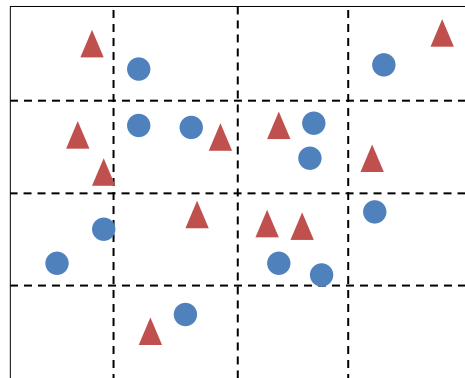


Histogram

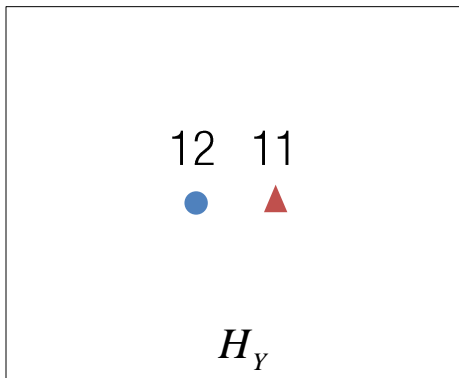
Kernel:

$$\kappa^1(X, Y) = \mathcal{I}^1 - \mathcal{I}^2$$

$$22 \quad 21$$

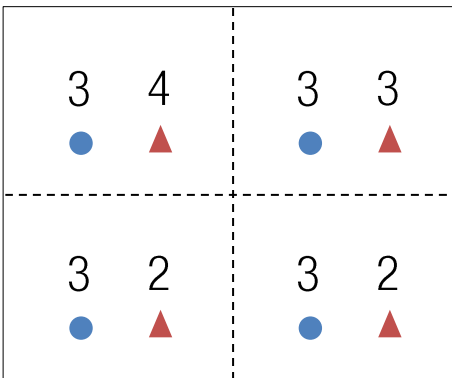


Feature



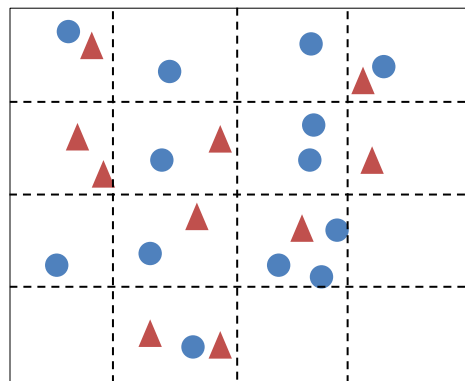
H_Y

Histogram

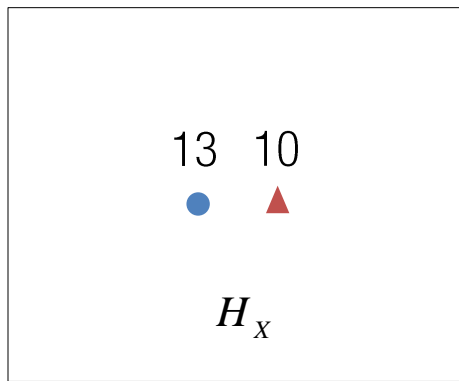


Histogram

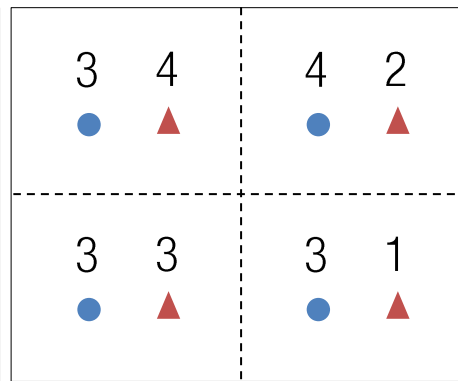
SIMILARITY MEASURE ~ PYRAMID MATCHING KERNEL



Feature



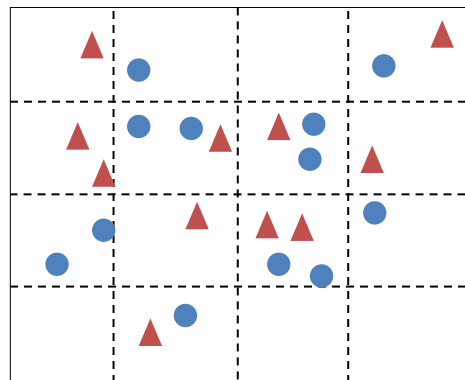
Histogram



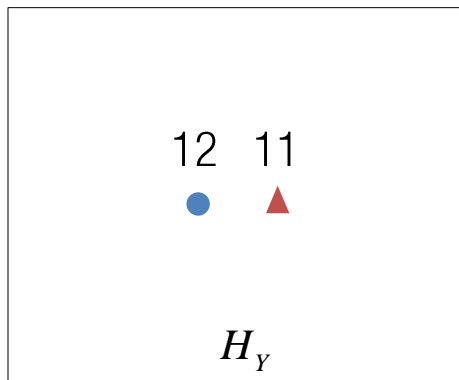
Histogram

Kernel:

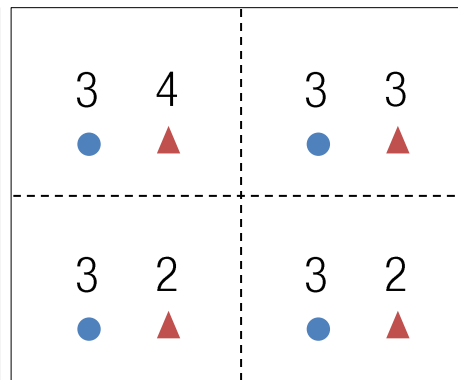
$$\kappa^L(X, Y) = \mathcal{I}^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}} (\mathcal{I}^l - \mathcal{I}^{l+1})$$



Feature

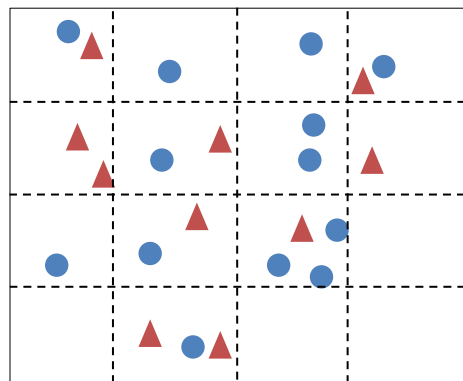


Histogram

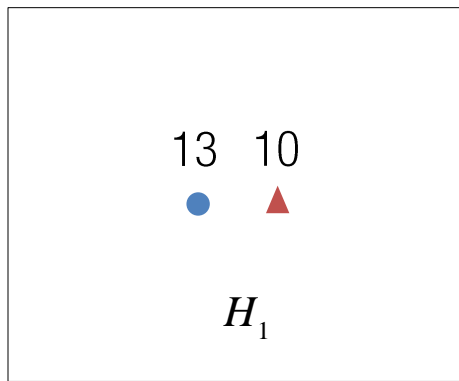


Histogram

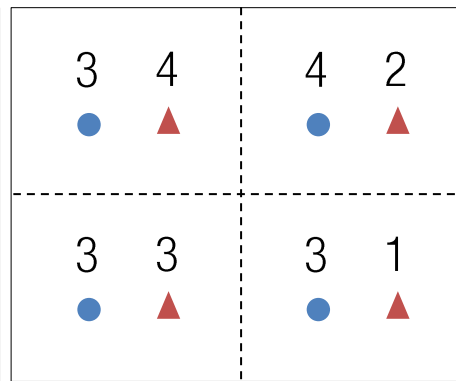
SIMILARITY MEASURE ~ PYRAMID MATCHING KERNEL



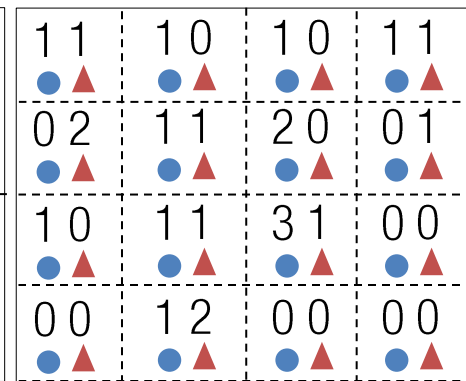
Feature



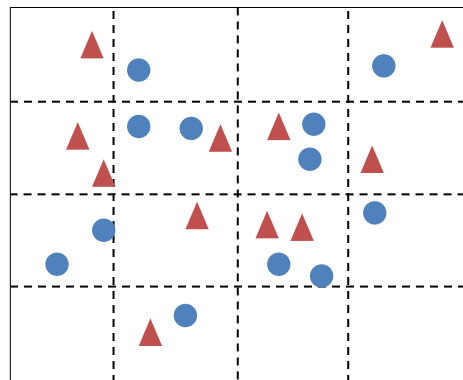
Histogram



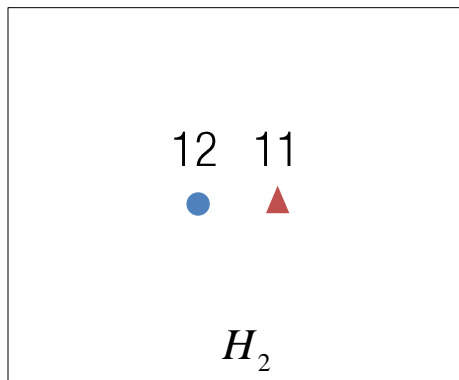
Histogram



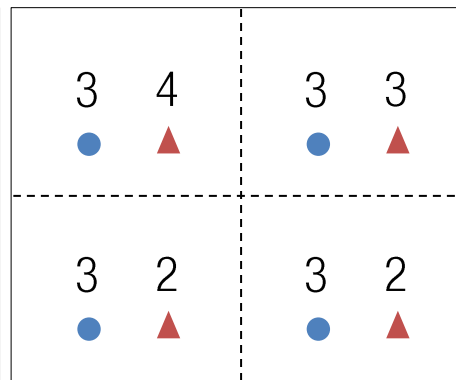
Histogram



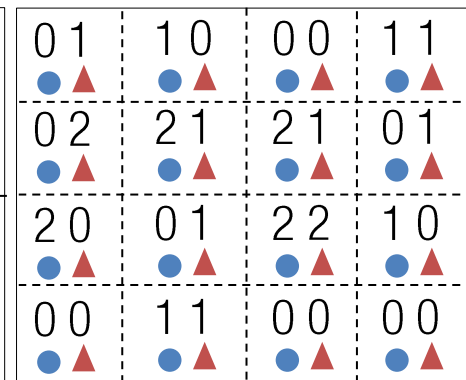
Feature



Histogram



Histogram



Histogram

Original images



Feature histograms:

Level 3



\cap

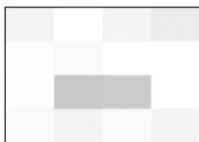


$= \mathcal{I}_3$

Level 2



\cap



$= \mathcal{I}_2$

Level 1



\cap



$= \mathcal{I}_1$

Level 0



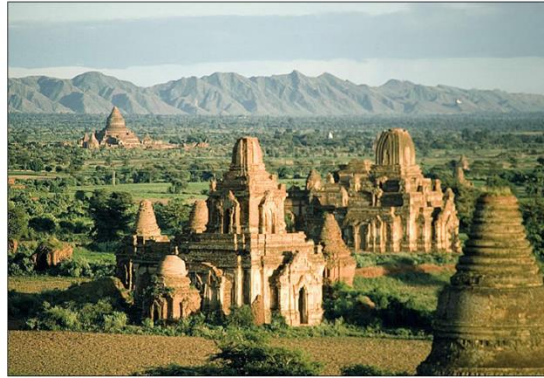
\cap



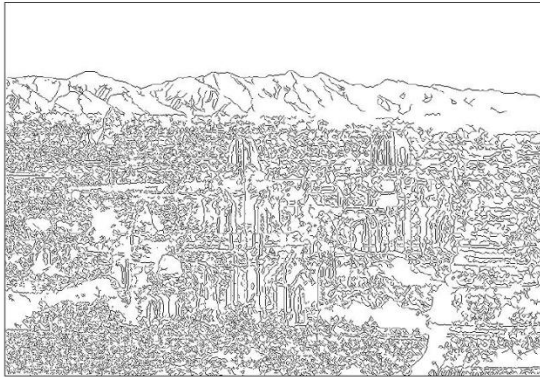
$= \mathcal{I}_0$

Total weight (value of *pyramid match kernel*): $\mathcal{I}_3 + \frac{1}{2}(\mathcal{I}_2 - \mathcal{I}_3) + \frac{1}{4}(\mathcal{I}_1 - \mathcal{I}_2) + \frac{1}{8}(\mathcal{I}_0 - \mathcal{I}_1)$

Feature extraction



Weak features



Edge points at 2 scales and 8 orientations
(vocabulary size 16)

Strong features

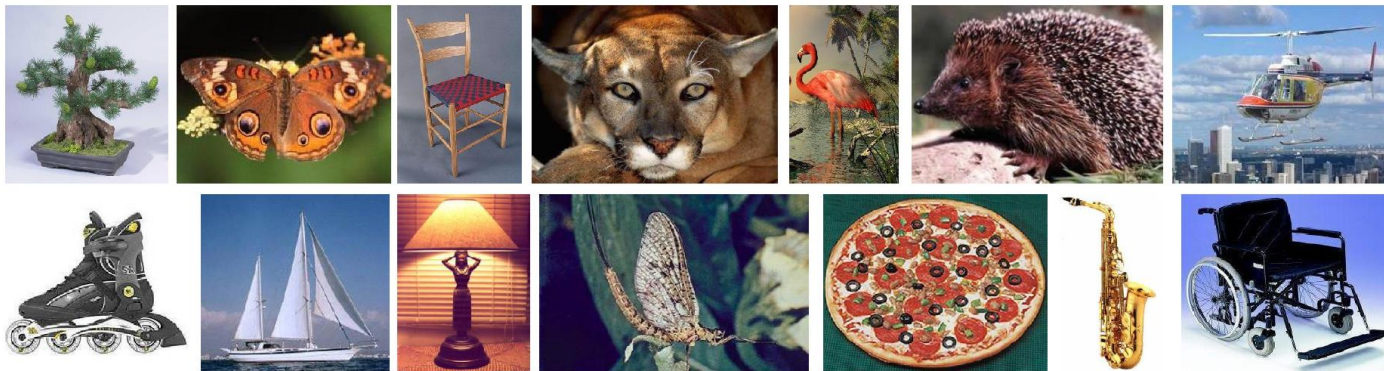


SIFT descriptors of 16x16 patches sampled
on a regular grid, quantized to form visual
vocabulary (size 200, 400)

Caltech101 dataset

Fei-Fei et al. (2004)

http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html



Multi-class classification results (30 training images per class)

	Weak features (16)		Strong features (200)	
Level	Single-level	Pyramid	Single-level	Pyramid
0	15.5 \pm 0.9		41.2 \pm 1.2	
1	31.4 \pm 1.2	32.8 \pm 1.3	55.9 \pm 0.9	57.0 \pm 0.8
2	47.2 \pm 1.1	49.3 \pm 1.4	63.6 \pm 0.9	64.6 \pm 0.8
3	52.2 \pm 0.8	54.0 \pm 1.1	60.3 \pm 0.9	64.6 \pm 0.7