Bag-of-Words

Hyun Soo Park

Challenges of Visual Recognition

- **Appearance**
  - DOF: texture, illumniation, material, shading, …
- **Shape**
  - DOF: object category, geometric pose, viewpoint, …



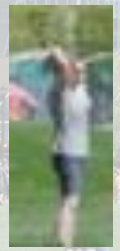
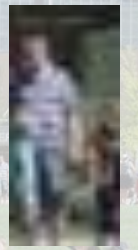




IMAGE CLASSIFICATION

corr [ , ] ? corr [ , ]

IMAGE CLASSIFICATION

Bedroom





Coast

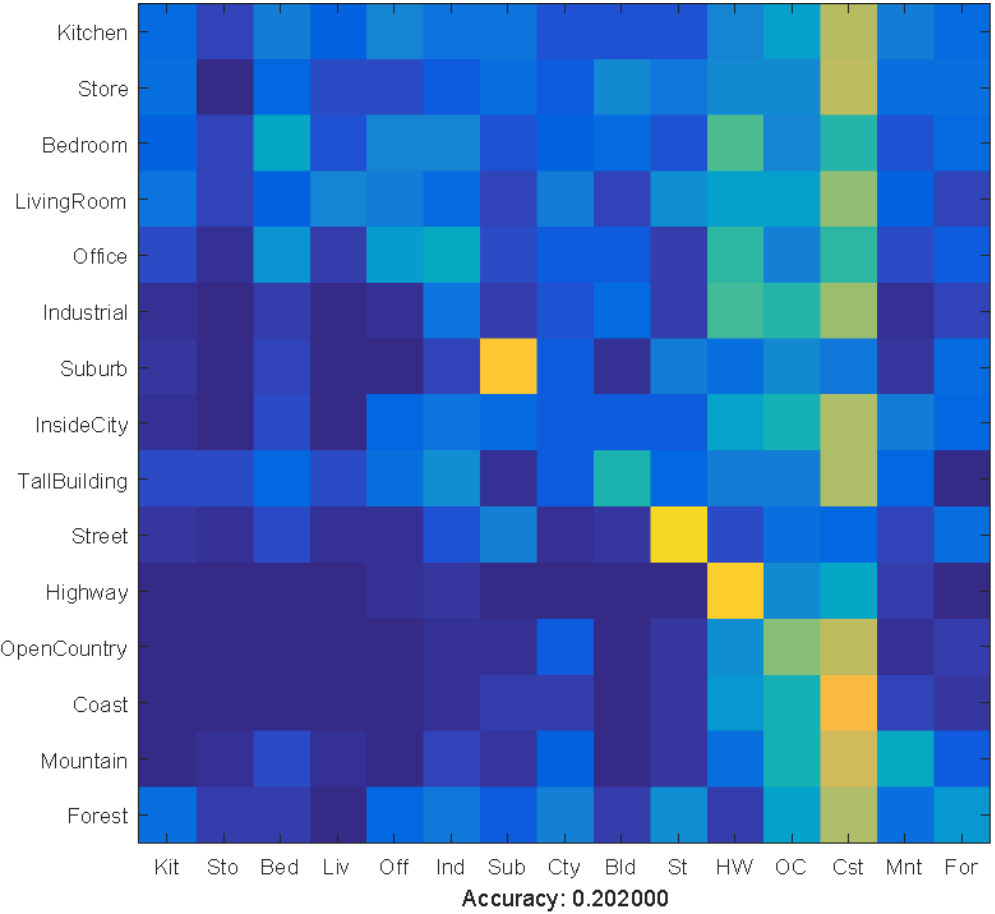






office
kitchen
living room
bedroom
store
industrial
tall building*
inside city*
street*
highway*
coast*
open country*
mountain*
forest*
suburb

TINY IMAGE



Tiny image representation + NN

(a) Image (b) Tiny Image

ACCURACY MEASURE

Prediction

Ground truth label	Bedroom	Beach
Bedroom	0.7	0.3
Beach		

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

Ground truth label	Bedroom	Beach
Bedroom	0.7	0.3
Beach	0.2	0.8

Confusion matrix

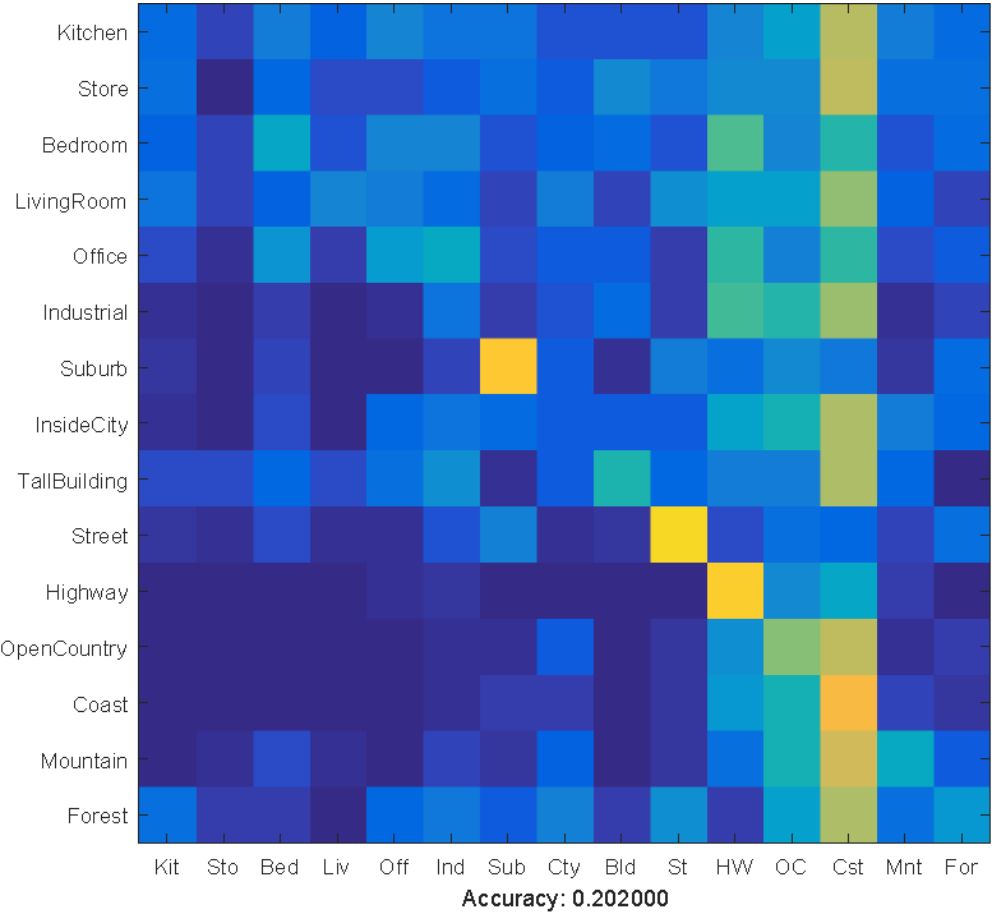
$$\frac{\text{\# of correct prediction on Bedroom data}}{\text{\# of Bedroom data}}$$

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

Accuracy: mean of correct predictions

$(0.7+0.8)/2 = 0.75$

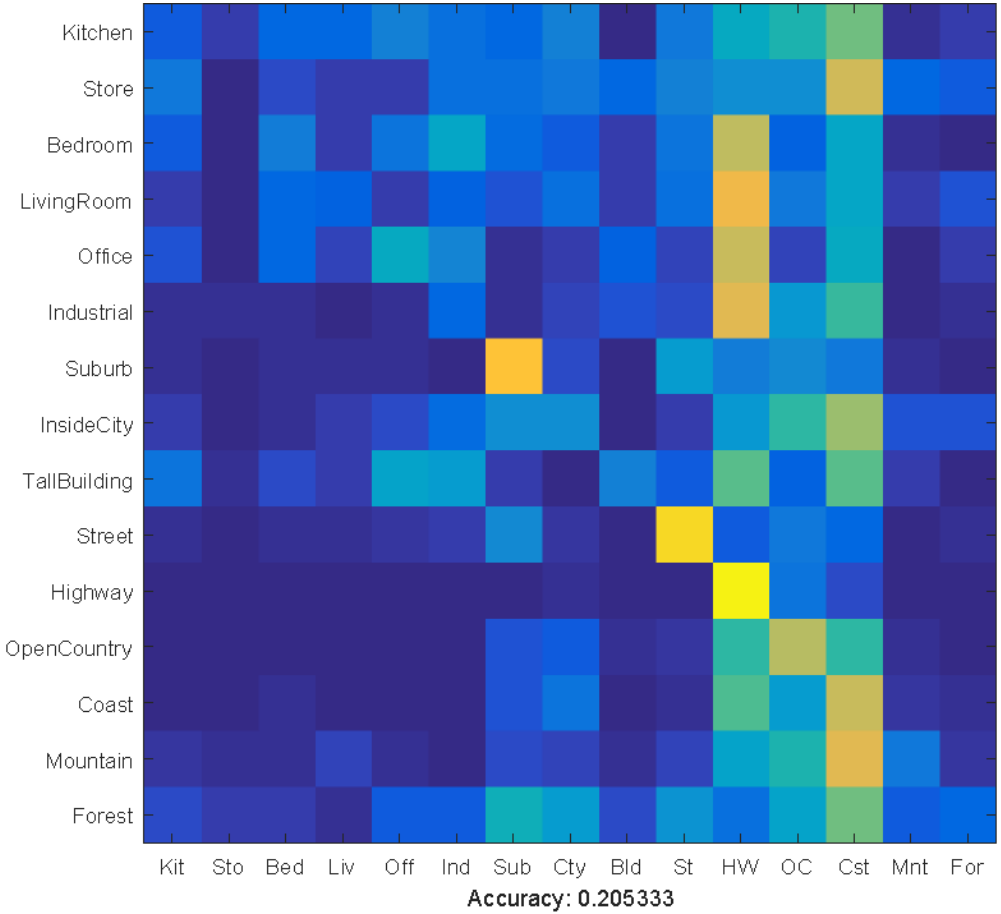
TINY IMAGE



Tiny image representation + NN

(a) Image (b) Tiny Image

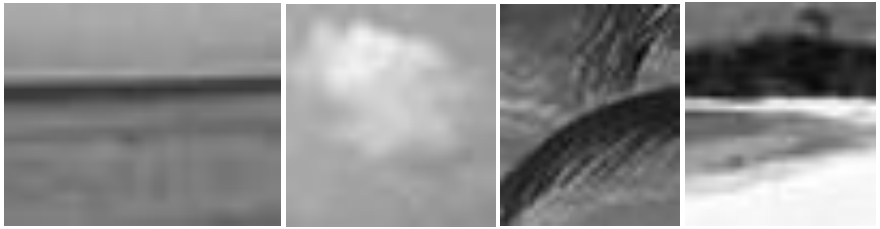
TINY IMAGE



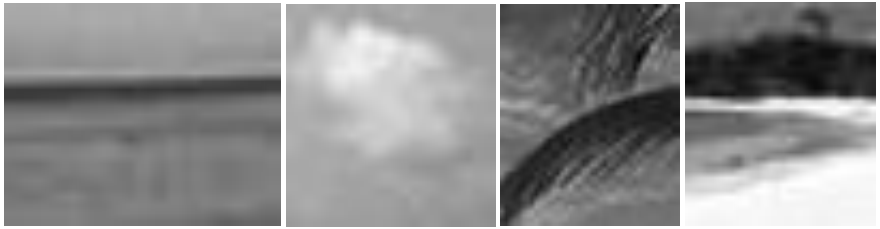
Tiny image representation + KNN (10)

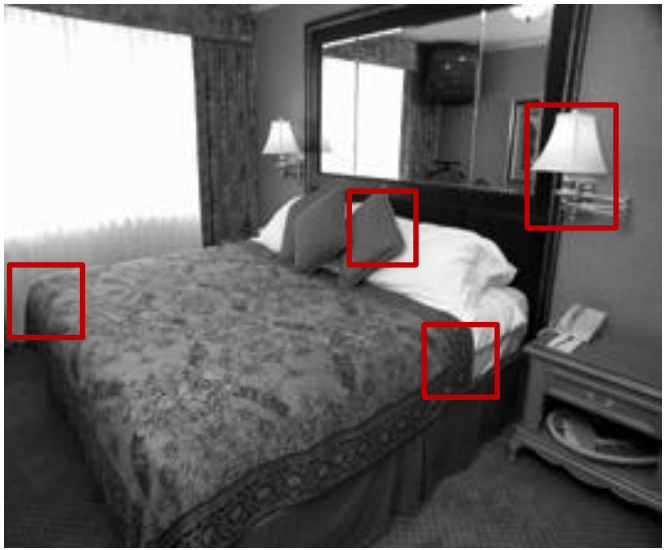
(a) Image (b) Tiny Image

Local Patches



Local Patches





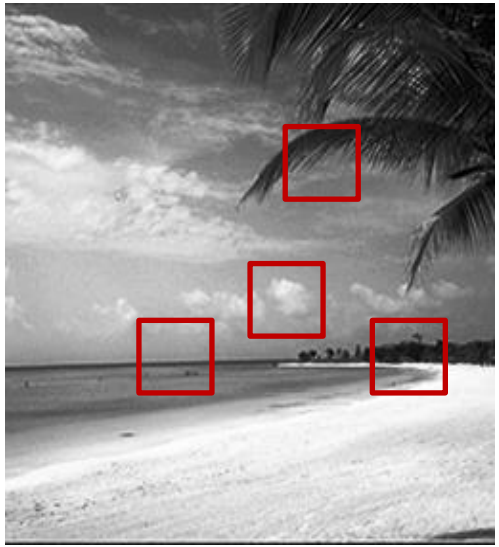


IMAGE CLASSIFICATION

corr [ , ] ? corr [ , ]

corr [ , ] > corr [ , ]

Bag of Words

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]

Bag of Words

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]

Minnesota

From Wikipedia, the free encyclopedia

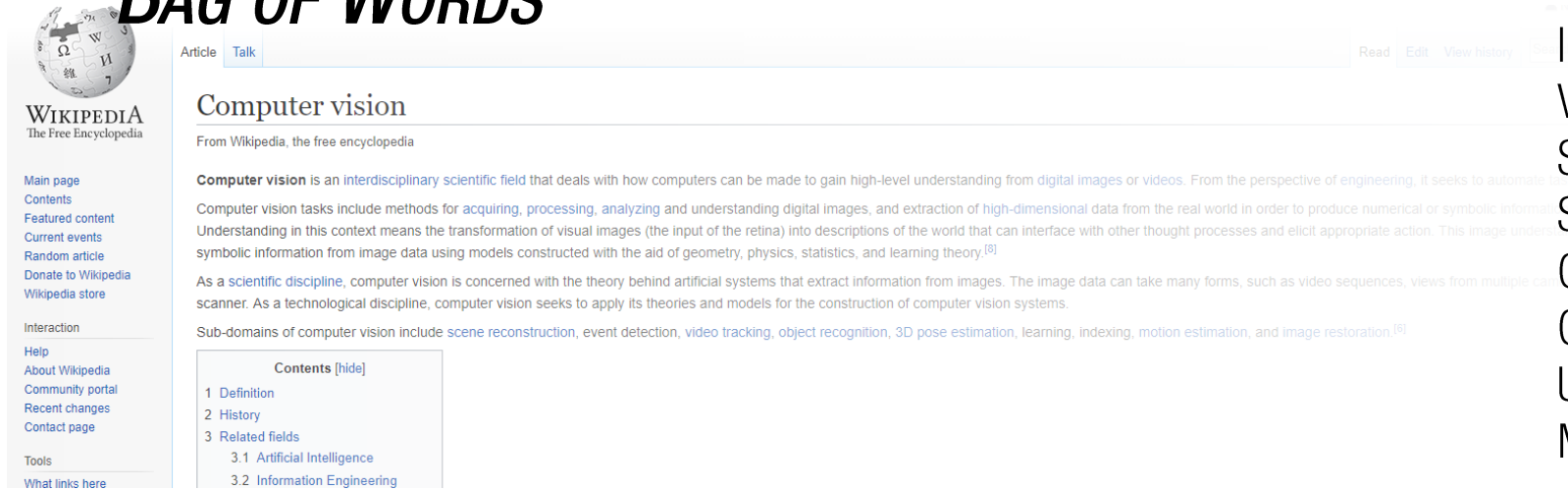
This article is about the U.S. state of Minnesota. For other uses, see Minnesota (disambiguation).

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.

Bag of Words



Article Talk

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

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
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
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
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]

1 Definition
2 History
3 Related fields
3.1 Artificial Intelligence
3.2 Information Engineering

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

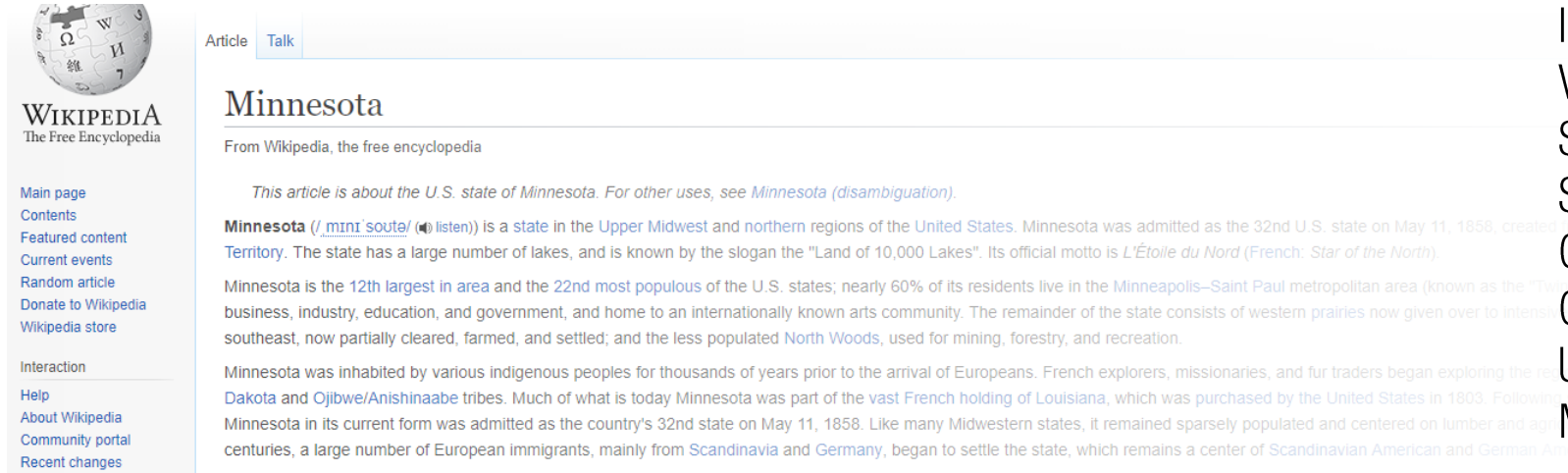
Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

What links here

Image: 145
Video: 13
Science: 7
Space: 6
Camera: 20
Cold: 0
University: 2
Mountain: 0



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/ (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, create
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 "Tw
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 intensi
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 re
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
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 ag
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 Ameri

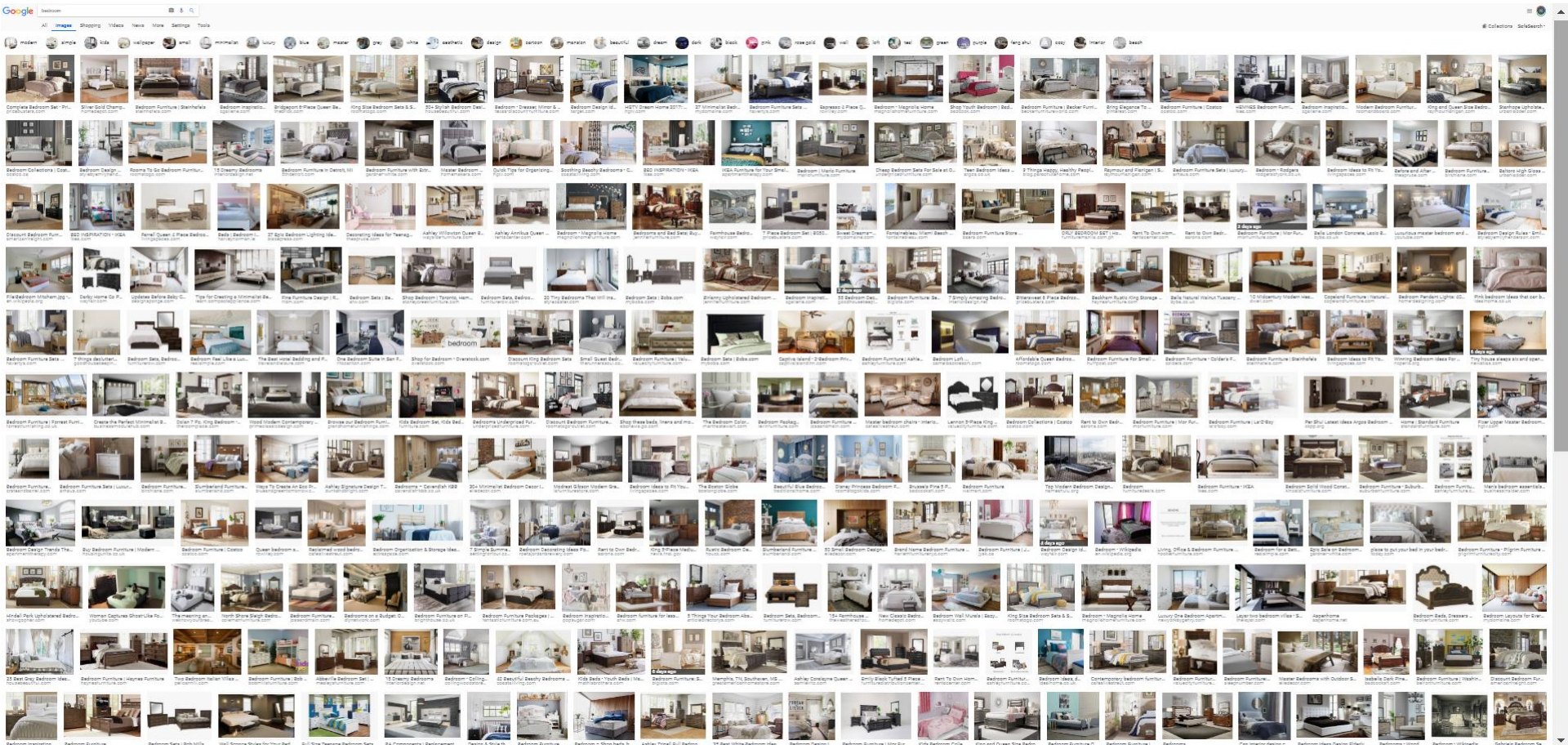
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

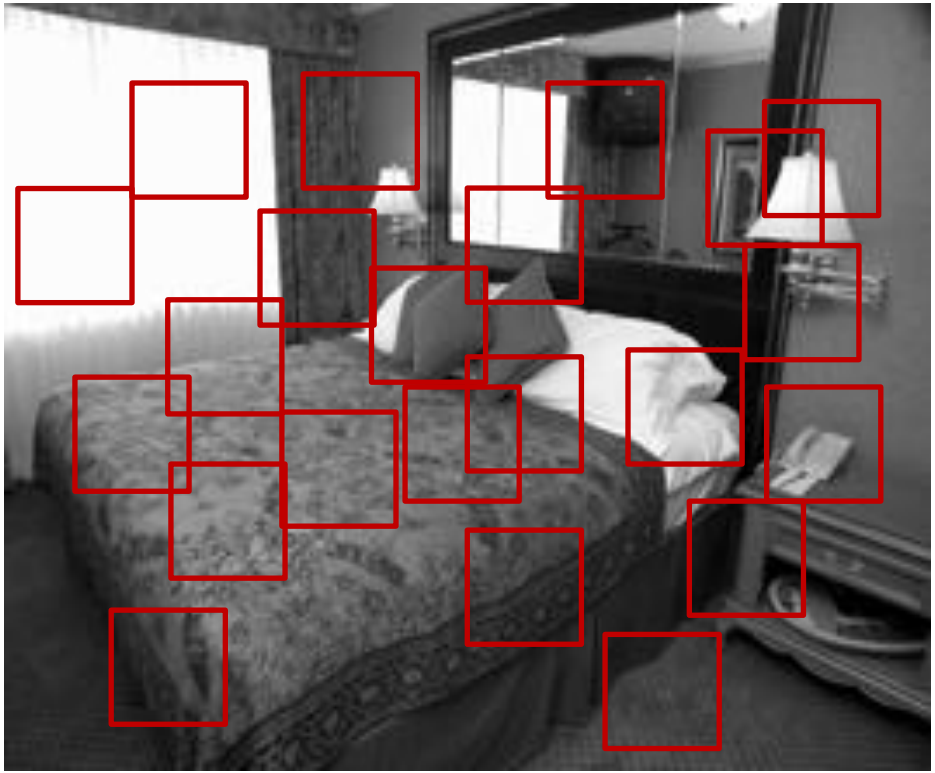
Help
About Wikipedia
Community portal
Recent changes

Image: 0
Video: 0
Science: 3
Space: 0
Camera: 1
Cold: 1
University: 28
Mountain: 5

BEDROOM IMAGES



POSSIBLE PATCHES



Millions of patch location and sizes

SEARCH SPACE

corr [ , ] ? corr [ , ]

corr [ , ] > corr [ , ]

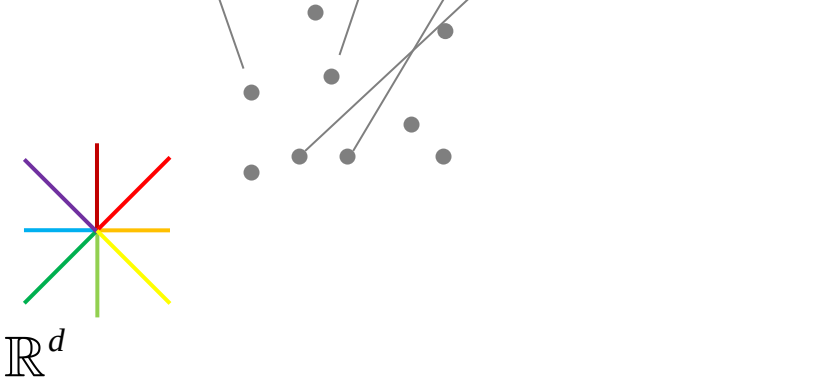
Search space: 1e+12

How to Construct Visual Dictionary





Lamp



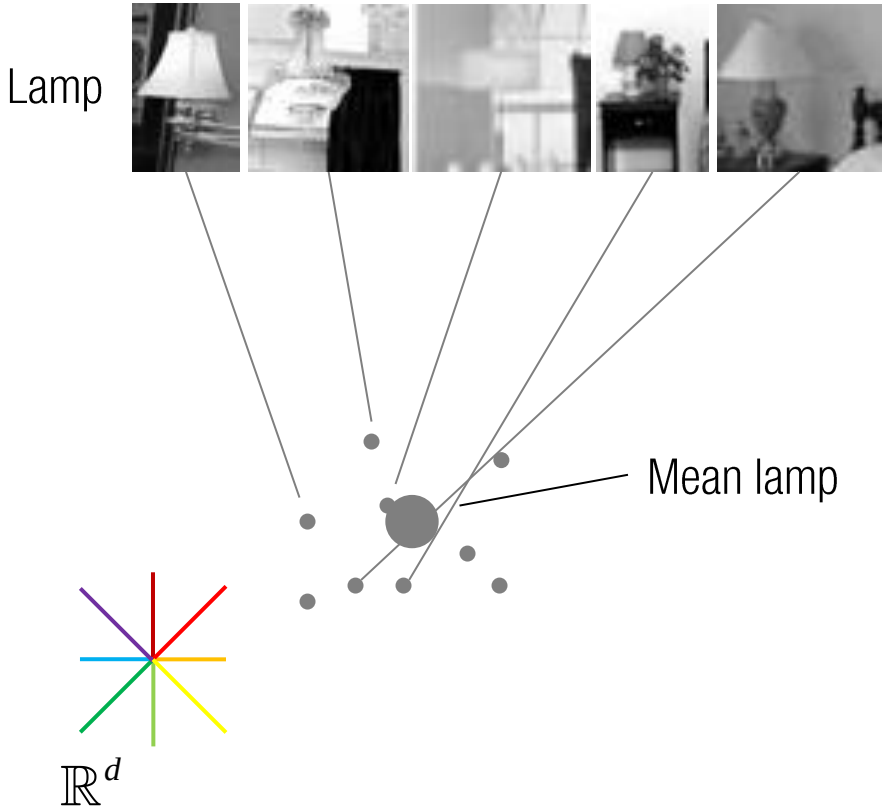
$\mathbb{R}^d$

Visual word descriptor

How to Construct Visual Dictionary



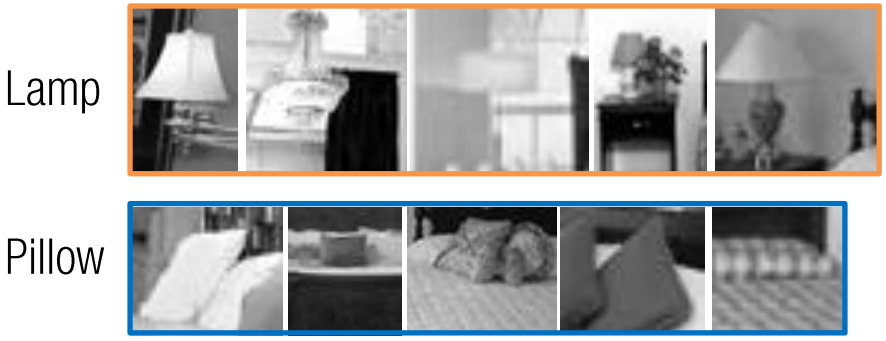


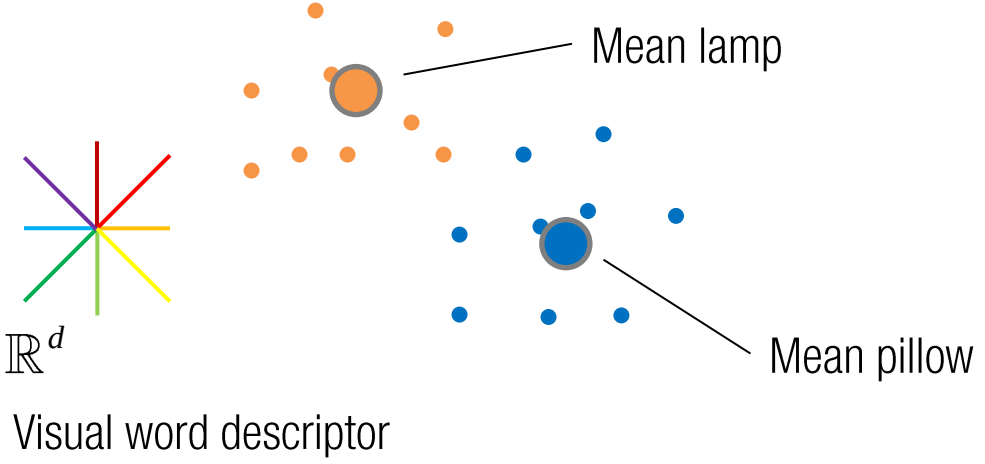


How to Construct Visual Dictionary





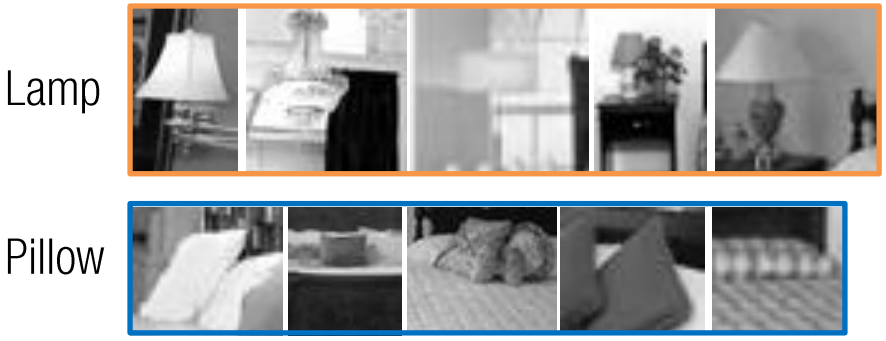


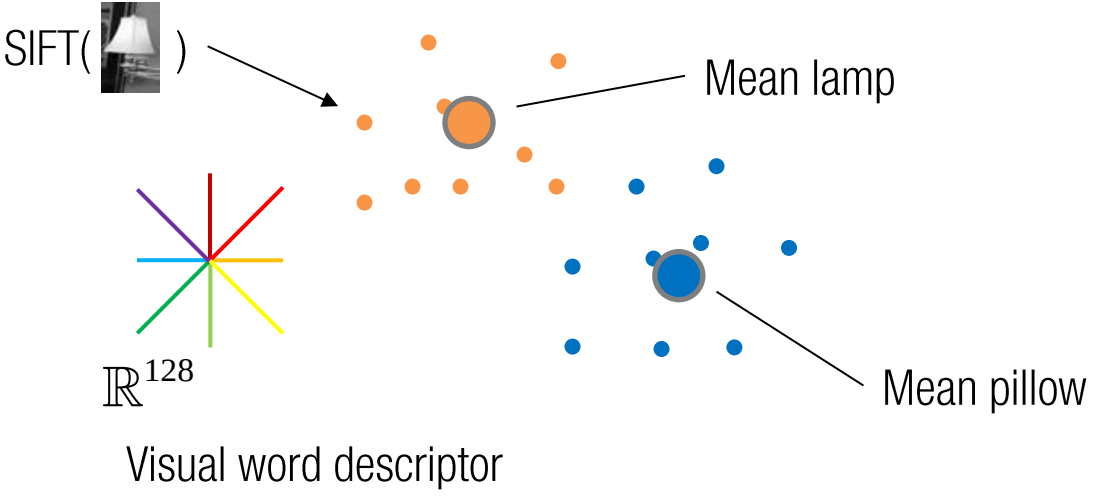


Feature Representation of Local Patch

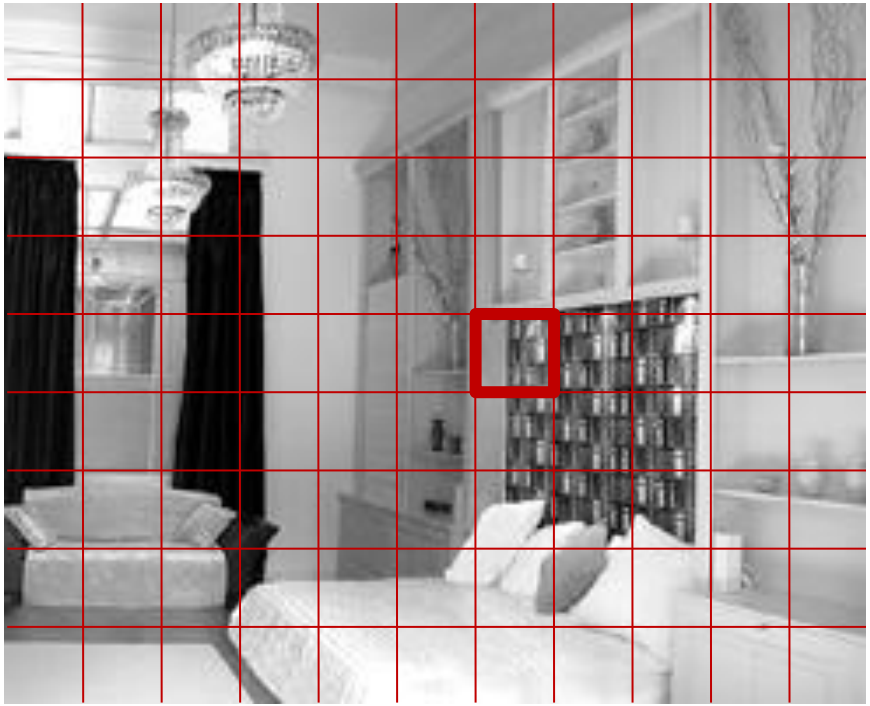








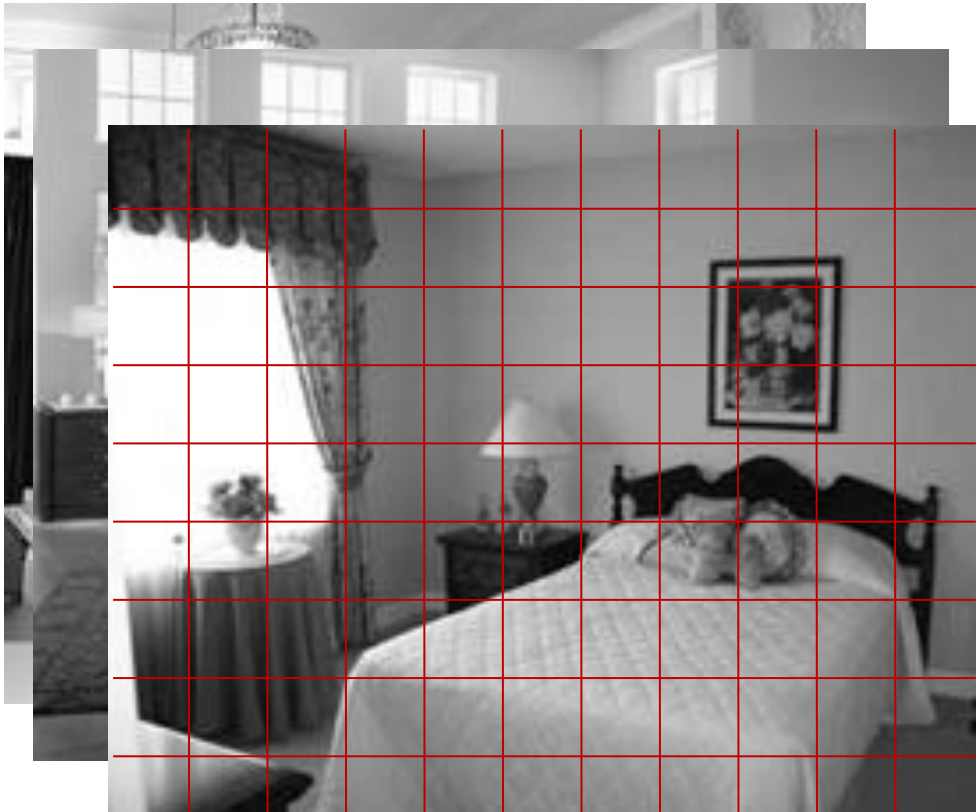
Dictionary Construction from Training Data



SIFT descriptor

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

Dictionary Construction from Training Data



SIFT descriptor

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

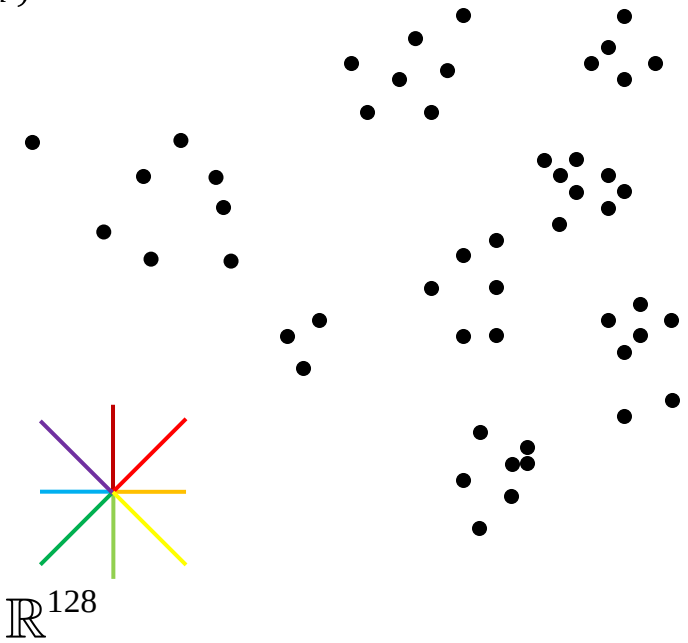
Pool of SIFT descriptors

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

Dictionary Construction from Training Data

Pool of SIFT descriptors

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

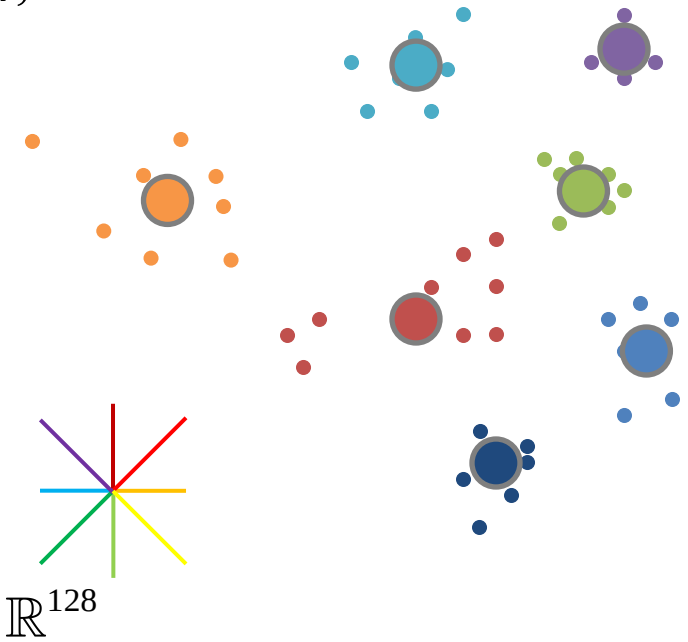


Dictionary Construction from Training Data

Pool of SIFT descriptors

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

K-means clustering



Dictionary Construction from Training Data

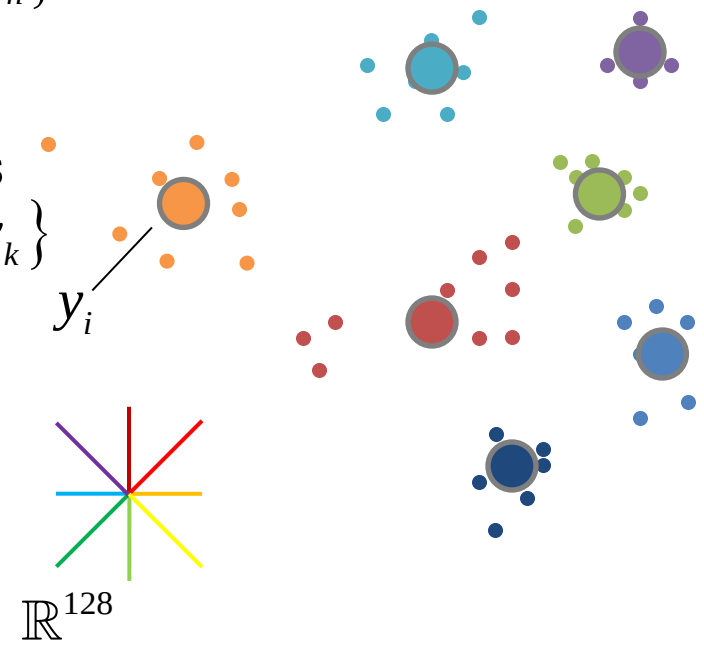
Pool of SIFT descriptors

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

K-means clustering

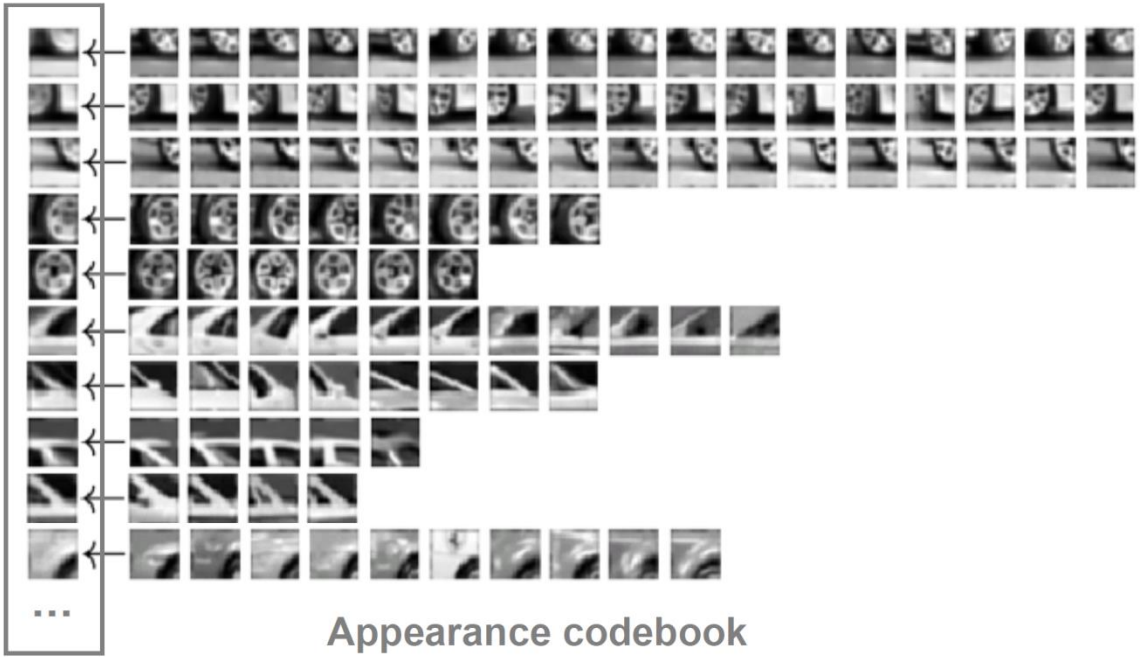
Dictionary~centroids

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

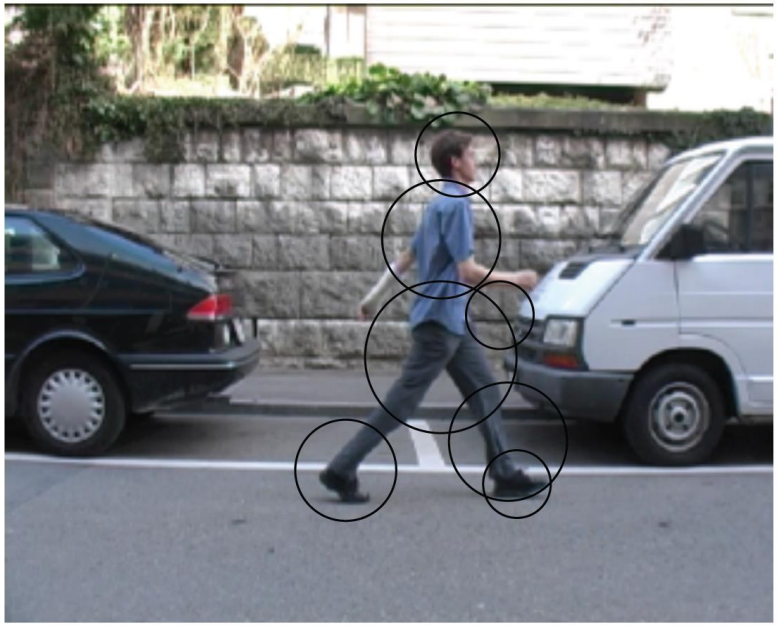


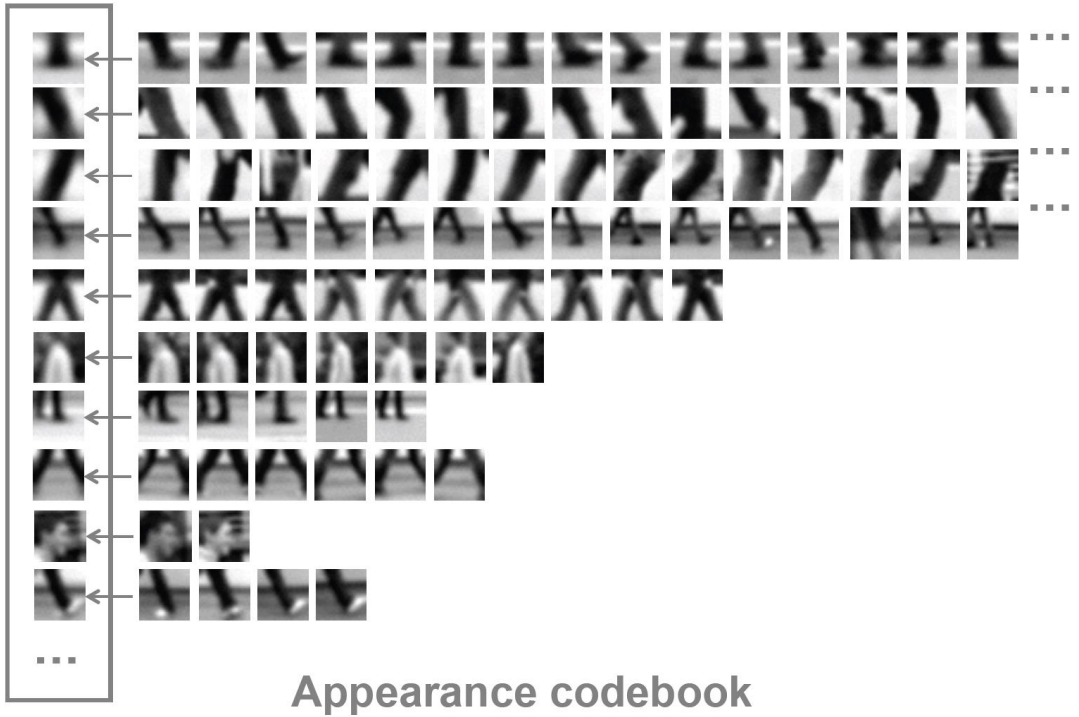

office
kitchen
living room
bedroom
store
industrial
tall building*
inside city*
street*
highway*
coast*
open country*
mountain*
forest*
suburb

CLUSTERED VISUAL PATCHES USING K-MEANS

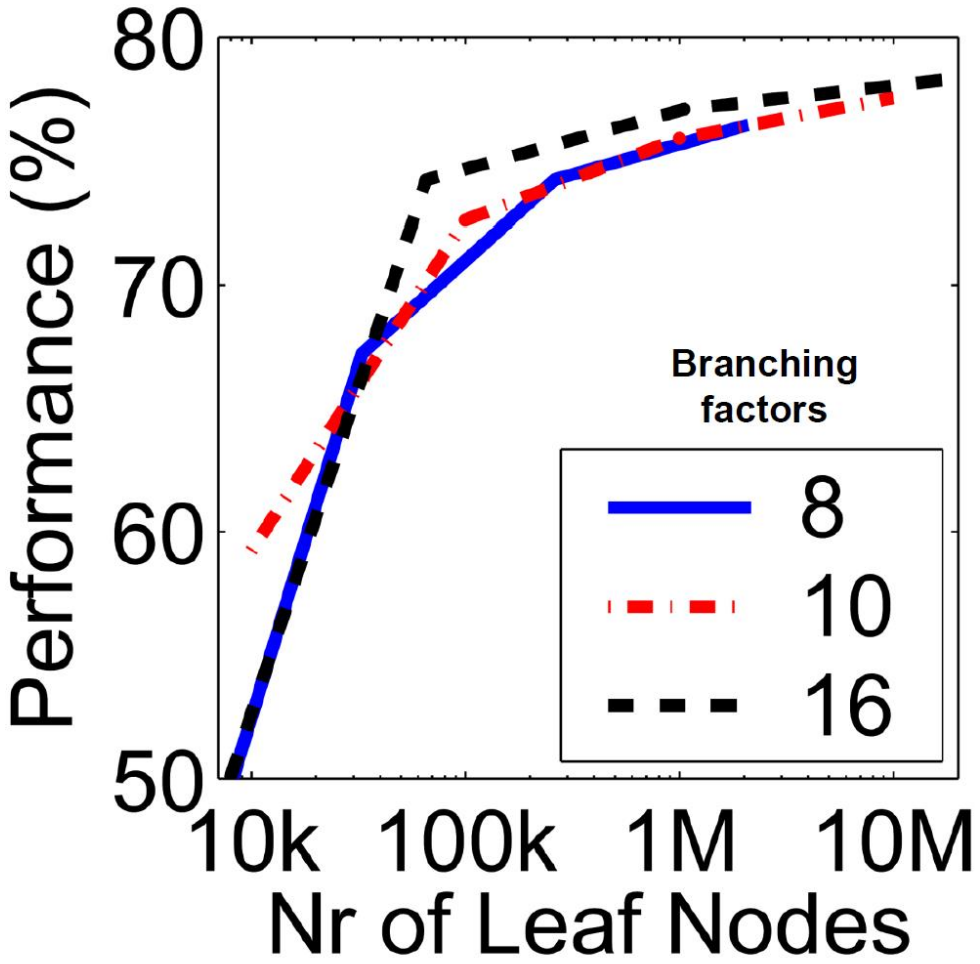


Appearance codebook

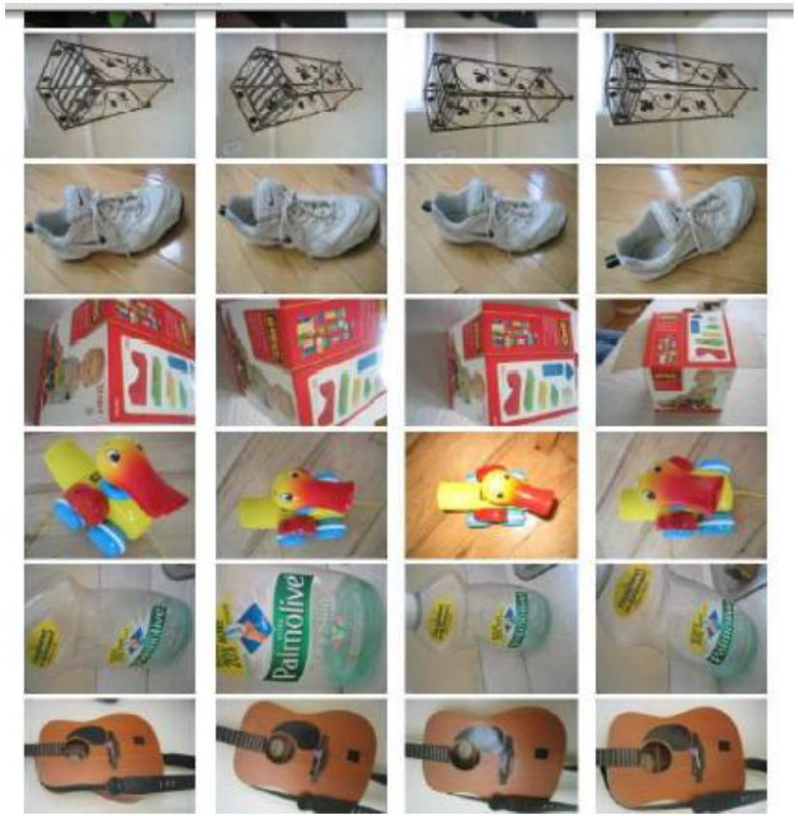




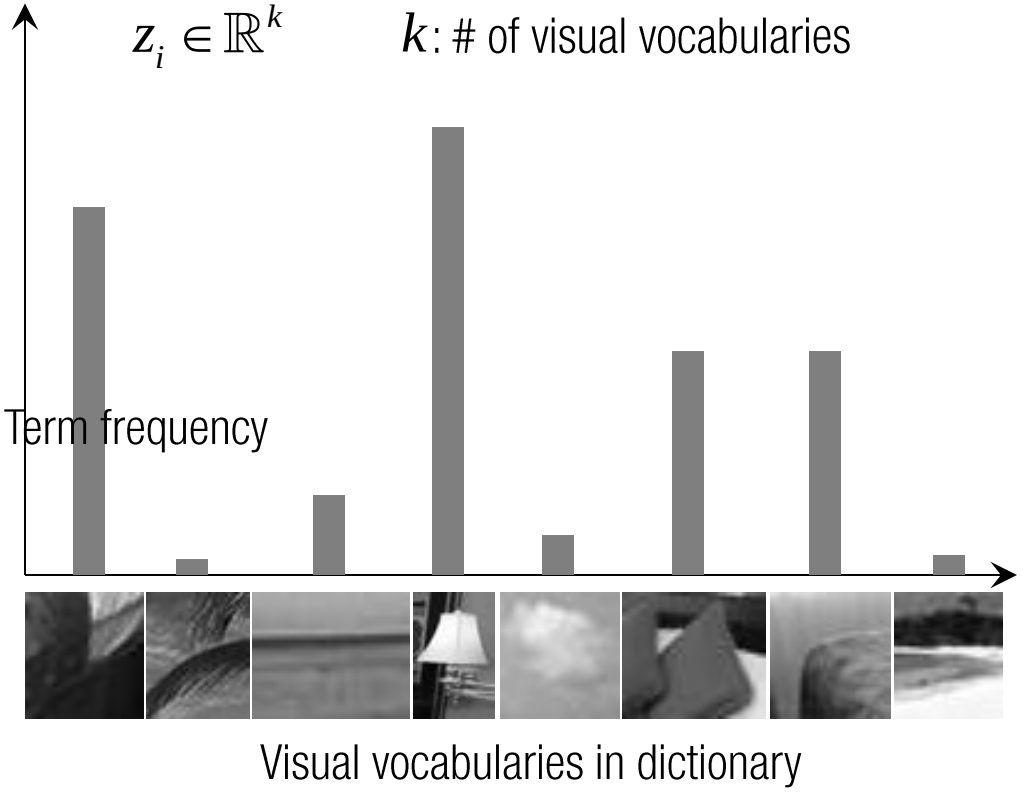
Appearance codebook



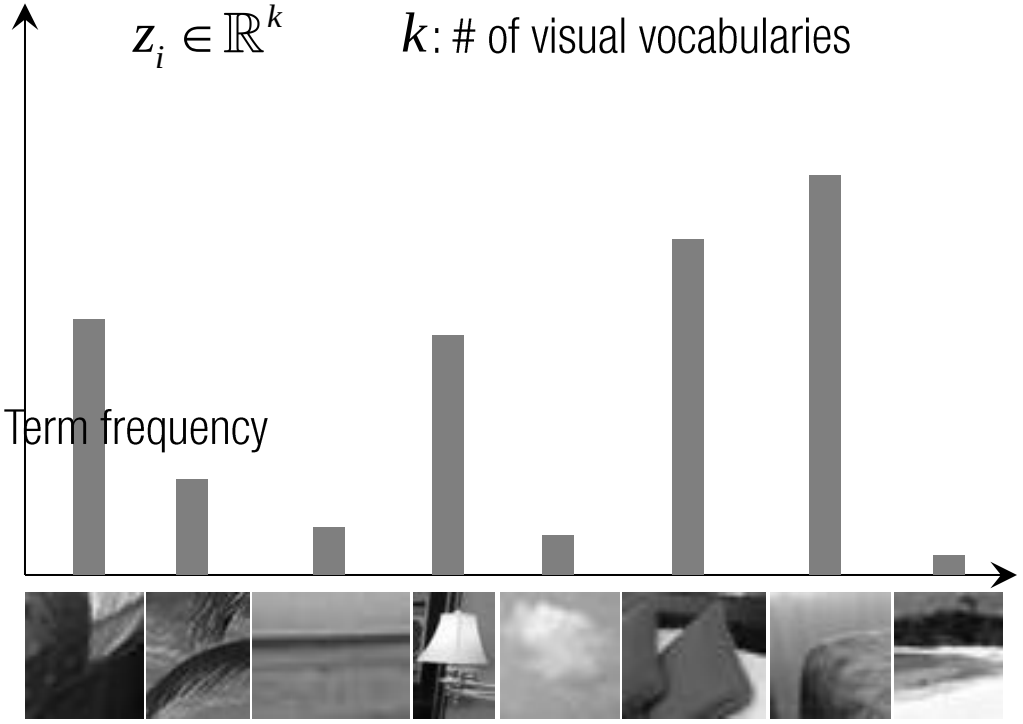
Results for recognition task with 6347 images



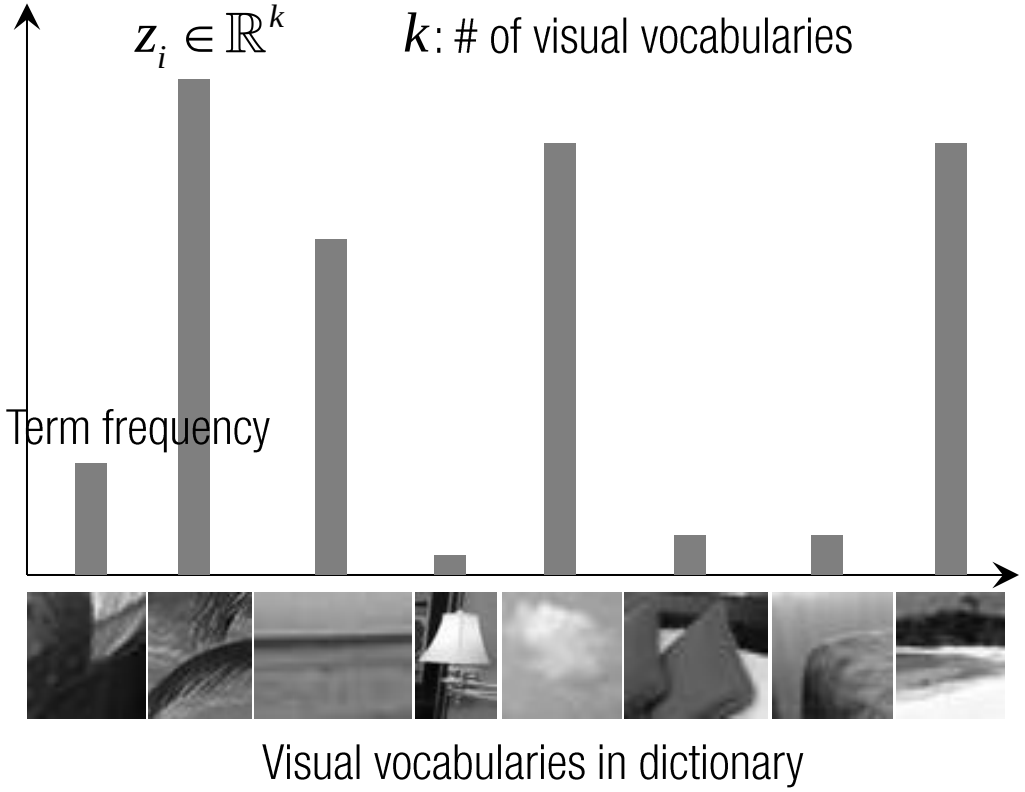
VISUAL BAG-OF-WORD REPRESENTATION



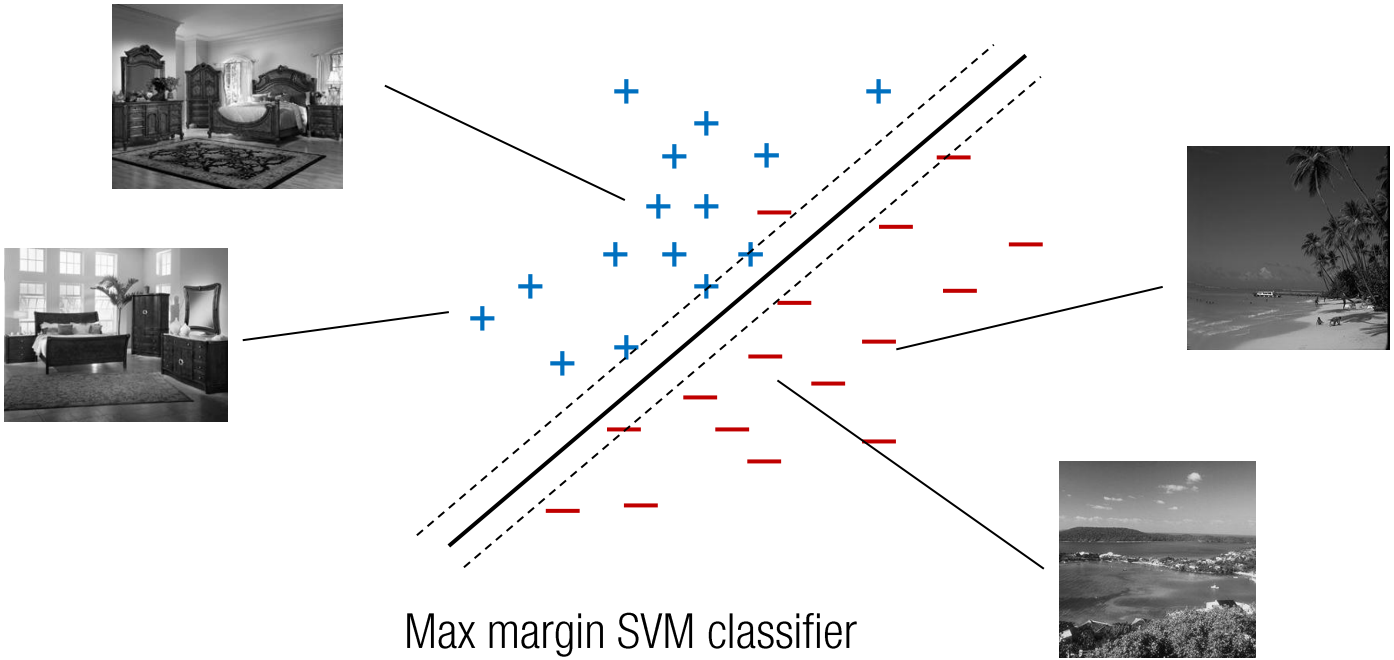
VISUAL BAG-OF-WORD REPRESENTATION



VISUAL BAG-OF-WORD REPRESENTATION



CLASSIFICATION

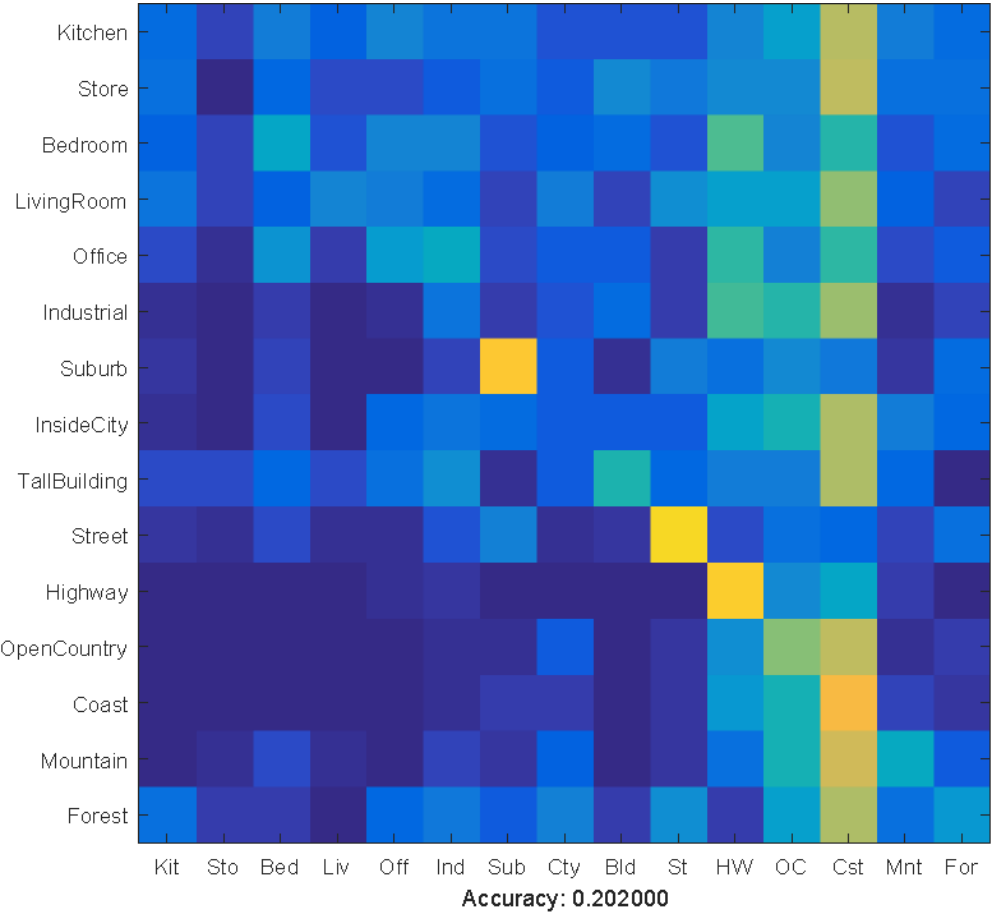


Max margin SVM classifier

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

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

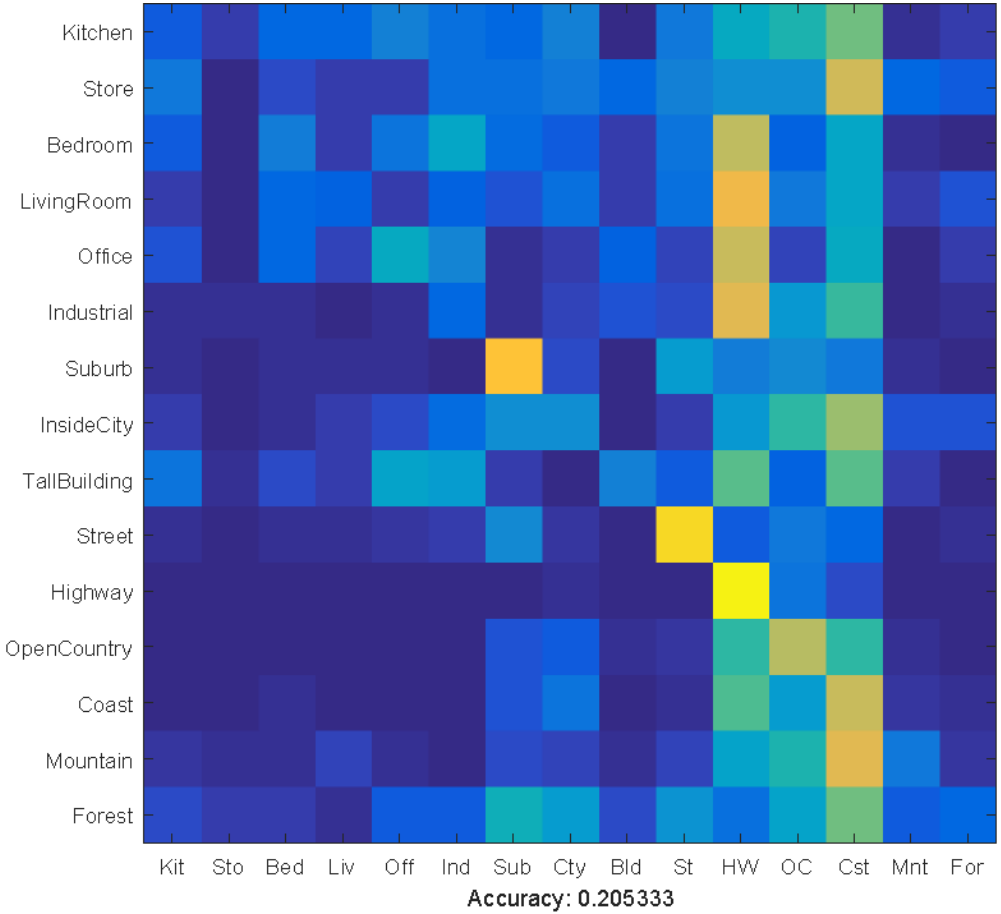
COMPARISON



Tiny image representation + NN

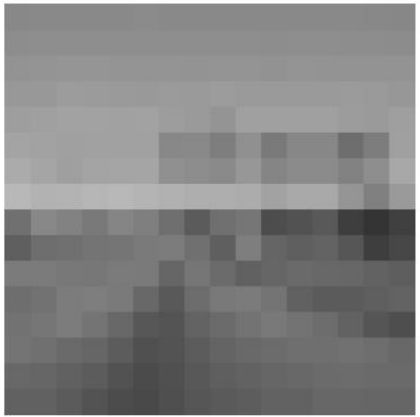
(a) Image (b) Tiny Image

COMPARISON



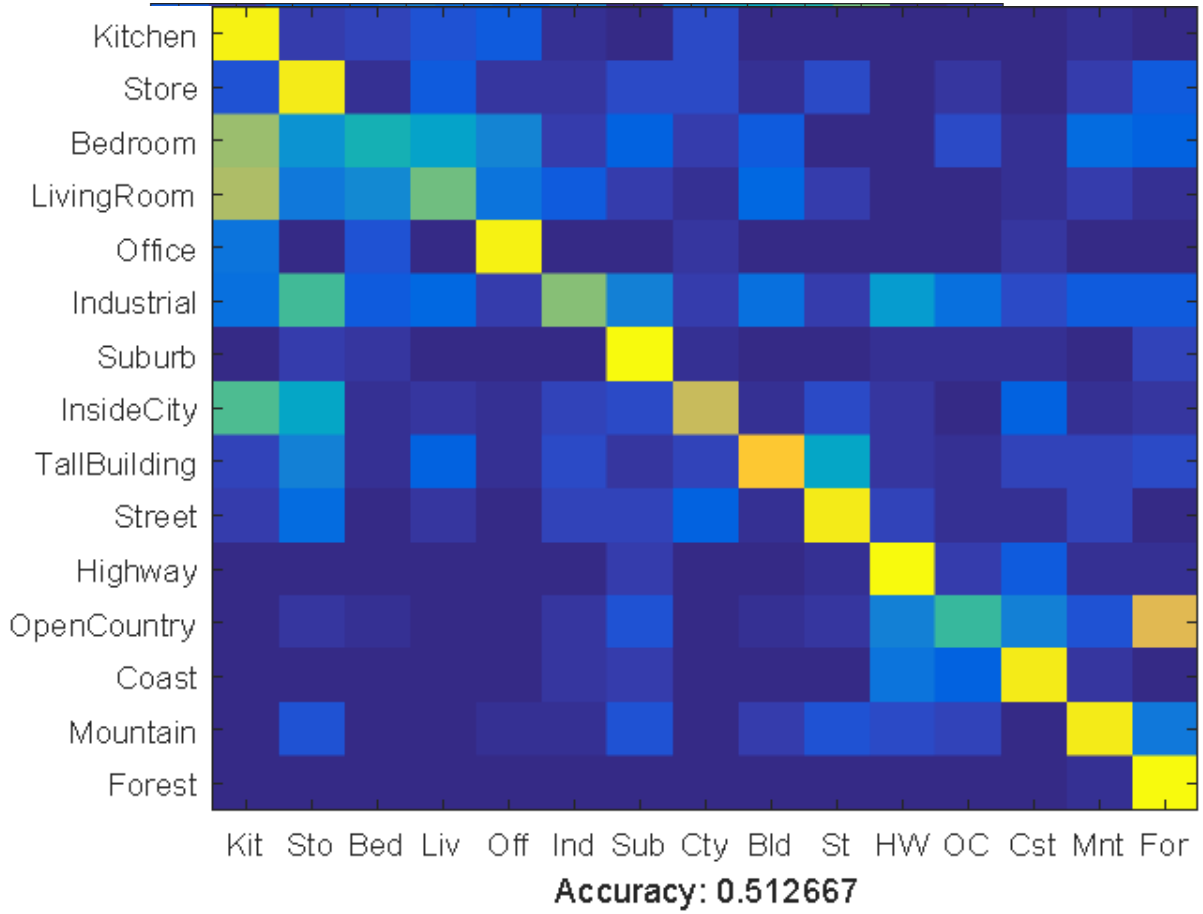
Tiny image representation + KNN (10)

(a) Image



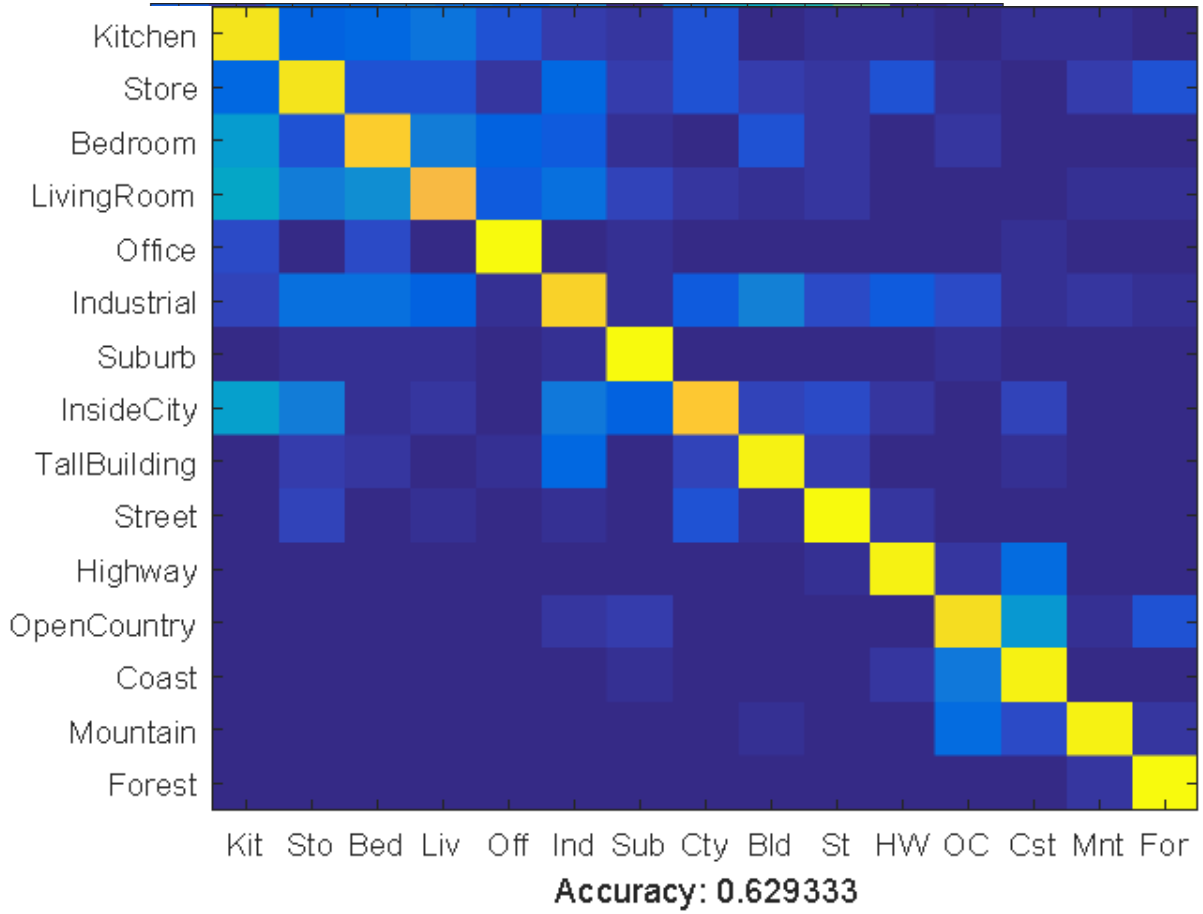
(b) Tiny Image

COMPARISON



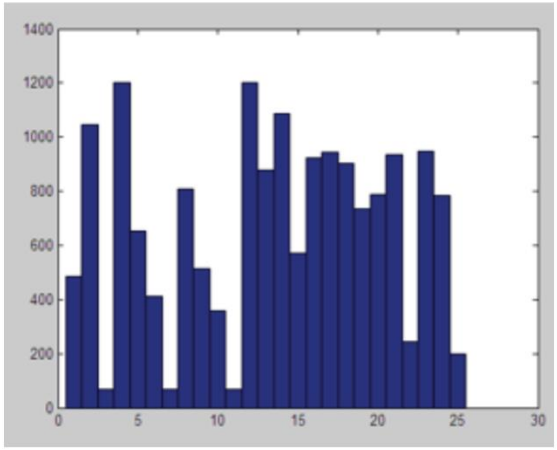
BoW + KNN (10)

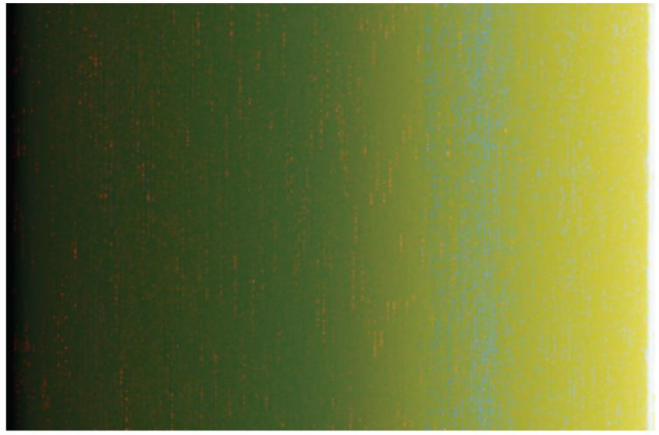
COMPARISON



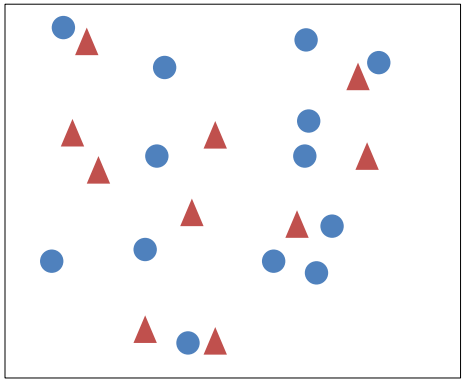
BoW + SVM

https://www.youtube.com/watch?v=Y6J6C_-mgRw

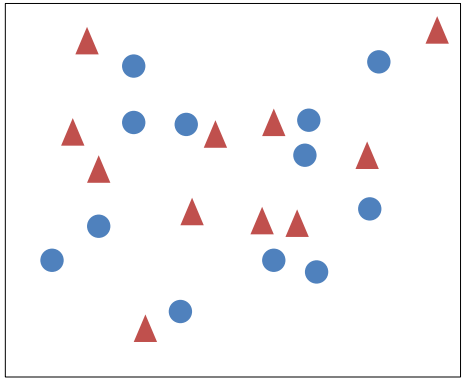

1400
1200
1000
800
600
400
200
0
0
5
10
15
20
25
30



MATCHING

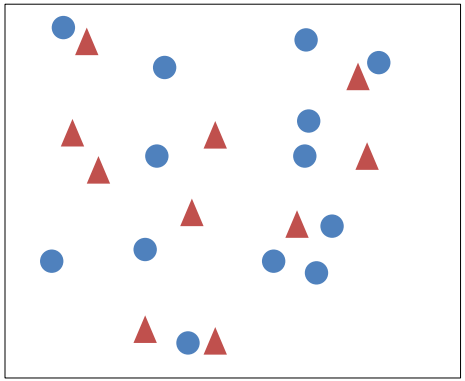


Feature

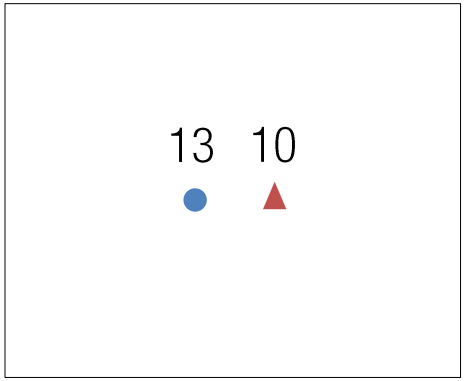


Feature

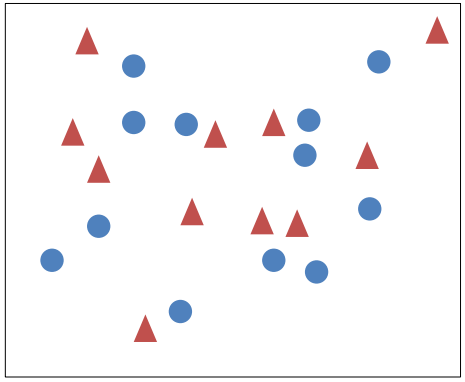
MATCHING



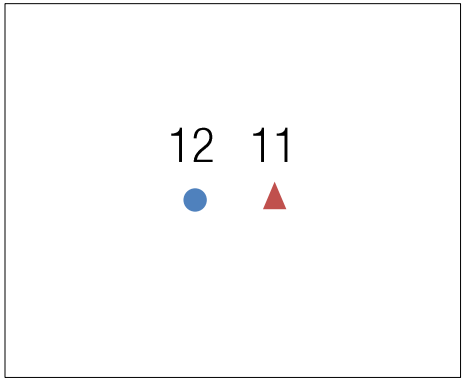
Feature



Histogram

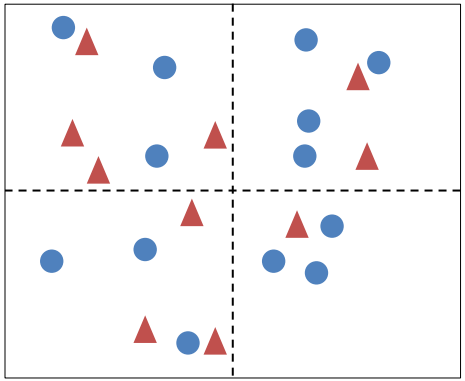


Feature

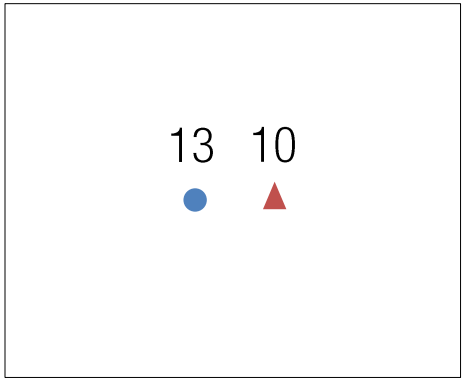


Histogram

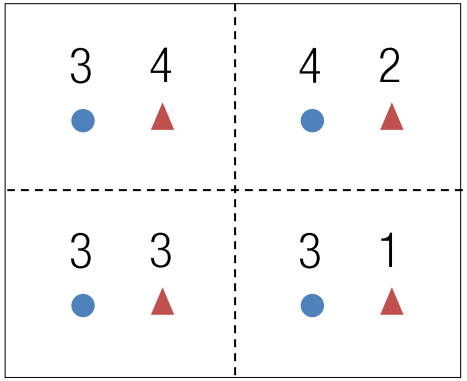
MATCHING



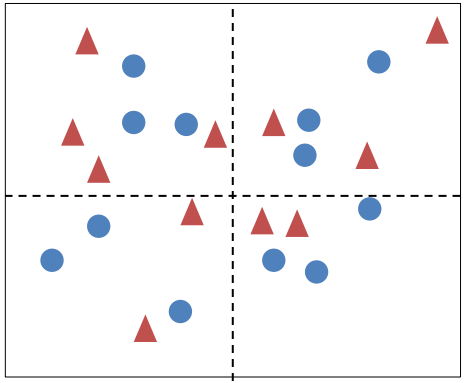
Feature



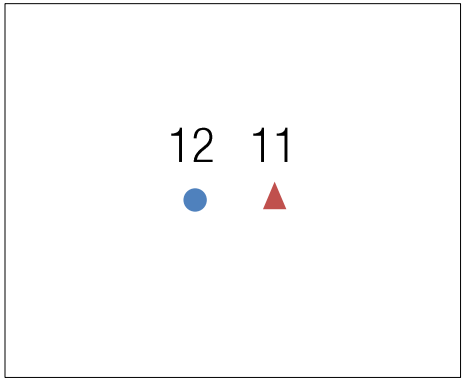
Histogram



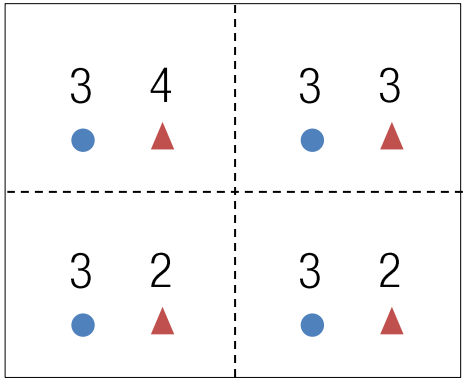
Histogram



Feature

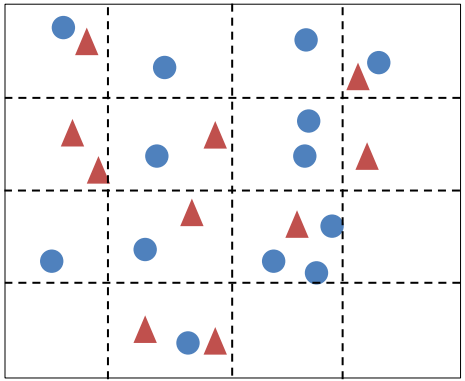


Histogram

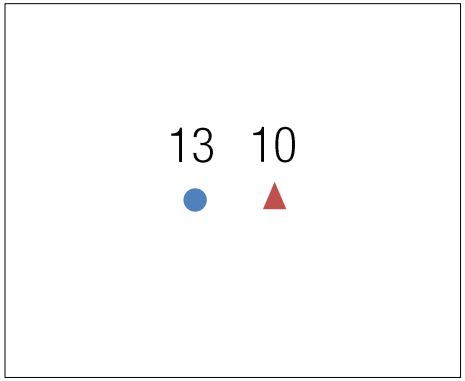


Histogram

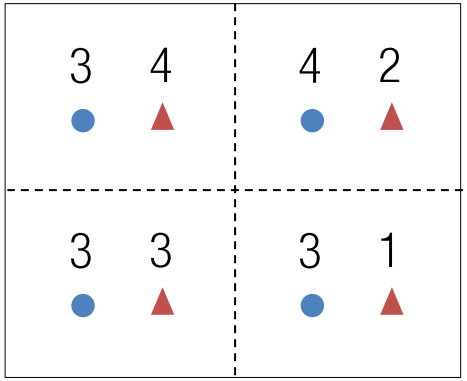
MATCHING



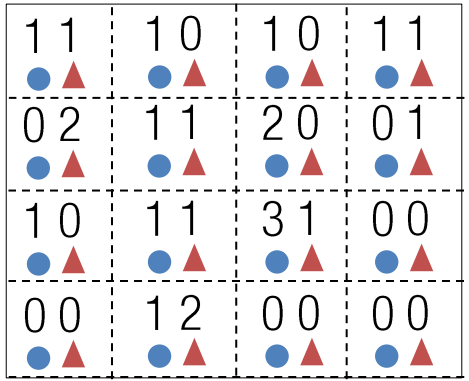
Feature



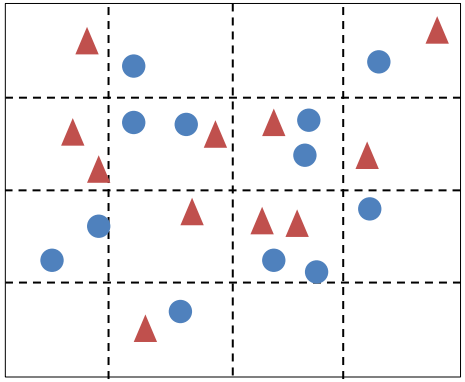
Histogram



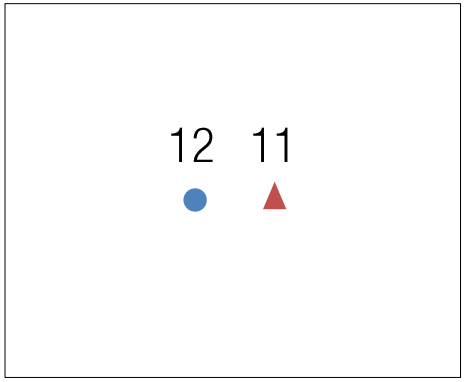
Histogram



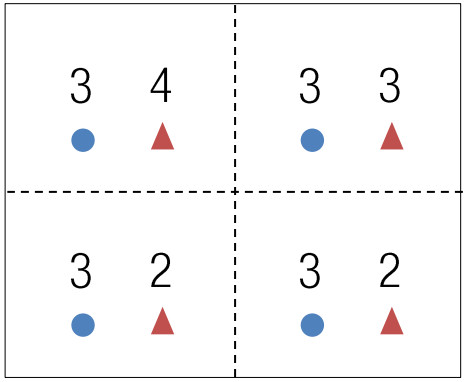
Histogram



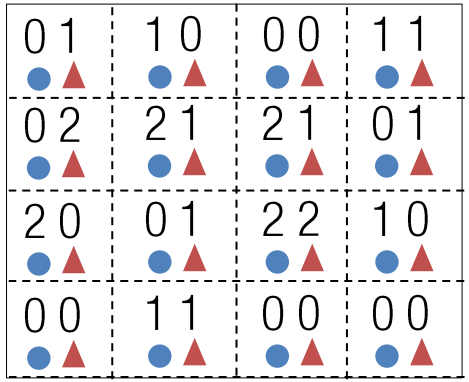
Feature



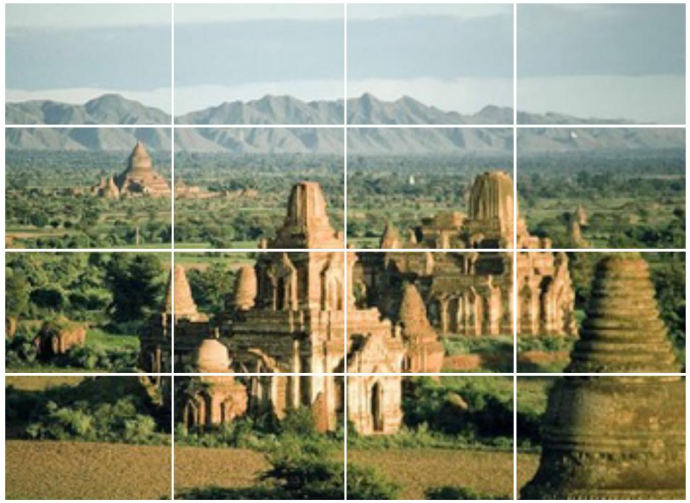
Histogram

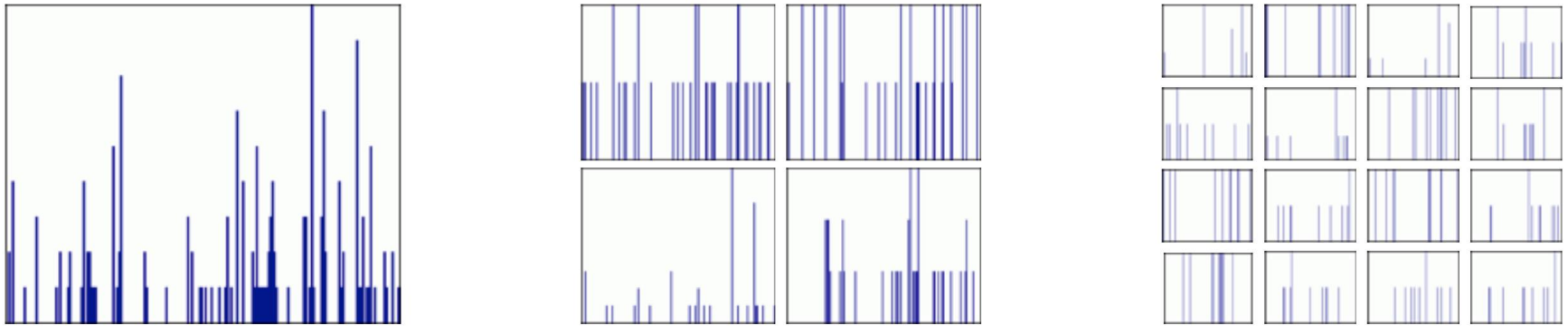


Histogram

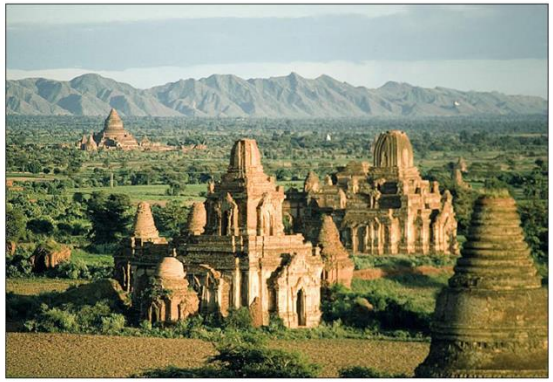


Histogram

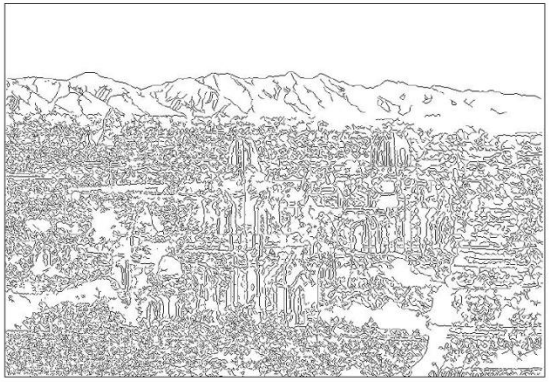




Feature extraction

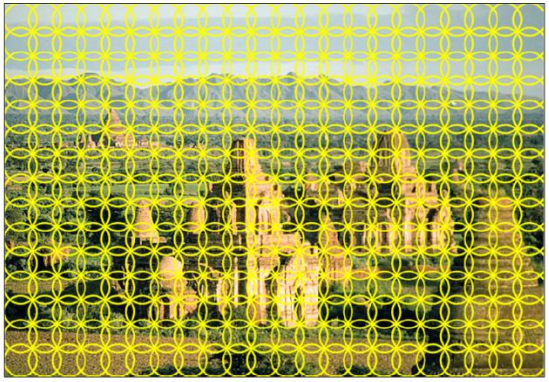


Weak features



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

Strong features

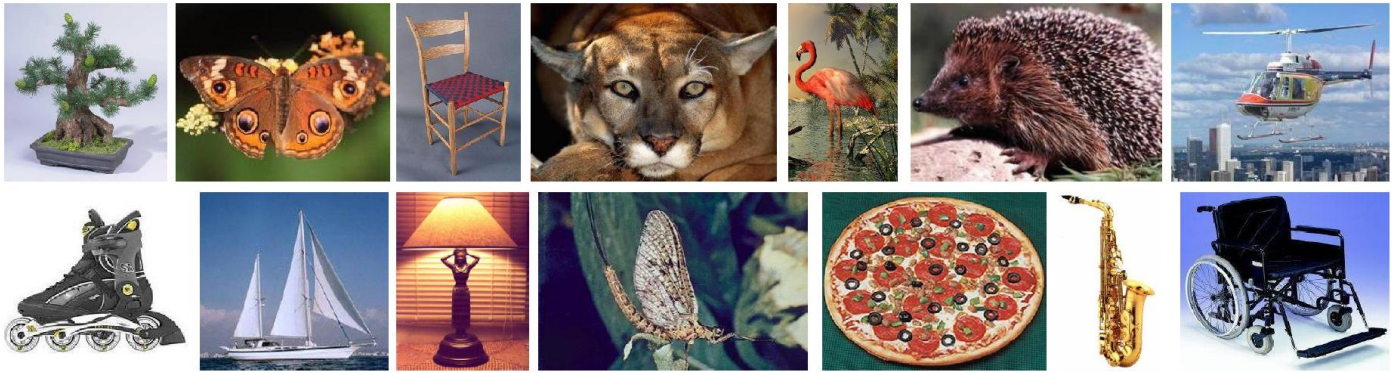


SIFT descriptors of 16x16 patches sampled on a regular grid, quantized to form visual vocabulary (size 200, 400)

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 ±0.9		41.2 ±1.2	
1	31.4 ±1.2	32.8 ±1.3	55.9 ±0.9	57.0 ±0.8
2	47.2 ±1.1	49.3 ±1.4	63.6 ±0.9	**64.6** ±0.8
3	52.2 ±0.8	**54.0** ±1.1	60.3 ±0.9	64.6 ±0.7