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

corr

,

? corr

, 

corr

, 

> corr

, 

,
Possible Patches

Millions of patch location and sizes
SEARCH SPACE

Search space: $1 \times 10^6$

Search space: $1 \times 10^{12}$
Computer vision

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.

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.

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 understanding can be seen as the disentangling of symbolic information from images using models constructed with the aid of geometry, physics, statistics, and learning theory.

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.

Minnesota

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

Minnesota ( 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

Computer vision

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 perform with ease and natural perception. Understanding in this context means the transformation of visual images (the input of the retina) into descriptions of the world that can interact with other thought processes and result in appropriate action. This image understanding may be symbolic information from image data using models constructed with the aid of geometry, physics, statistics, and learning theory.[1]

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 scans. 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.[2]

Contents

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

Minnesota

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

Minnesota ([ˈmɪnəˈsoʊtə] (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 Minnesota Territory. The state has a large number of lakes, and is known by the slogan “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 second-largest in the United States. Agriculture, 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 northern forests, 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 1600s. 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 American Civil War, Minnesota’s farmers moved into the Minnesota–Dakota territory, and by the 1870s, 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.

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

Images: 0
Video: 0
Science: 0
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

Mean lamp

Mean pillow

Visual word descriptor
FEATURE REPRESENTATION OF LOCAL PATCH

Lamp

Pillow

SIFT

Mean lamp

Mean pillow

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

\[ X = \{ x_1, \ldots, x_n \} \]
**Dictionary Construction from Training Data**

Pool of SIFT descriptors

\[ X = \{ x_1, \ldots, x_n \} \]

\[ \mathbb{R}^{128} \]
**Dictionary Construction from Training Data**

Pool of SIFT descriptors

\[ X = \{ x_1, \ldots, x_n \} \]

K-means clustering
**Dictionary Construction from Training Data**

Pool of SIFT descriptors

\[ X = \{ x_1, \ldots, x_n \} \]

Dictionary \sim \text{centroids}

\[ Y = \{ y_1, \ldots, y_k \} \]

K-means clustering

\[ \mathbb{R}^{128} \]
CLUSTERED VISUAL PATCHES USING K-MEANS
Results for recognition task with 6347 images
**Visual Bag-of-Word Representation**

\[ z_i \in \mathbb{R}^k \]

\[ k: \# \text{ of visual vocabularies} \]

Term frequency

Visual vocabularies in dictionary
**Visual Bag-of-Word Representation**

$z_i \in \mathbb{R}^k$

$k$: # of visual vocabularies

Term frequency

Visual vocabularies in dictionary
**Visual Bag-of-Word Representation**

$z_i \in \mathbb{R}^k$

$k$: # of visual vocabularies

Term frequency

Visual vocabularies in dictionary
CLASSIFICATION

Max margin SVM classifier

\[ z \cdot w + b > 0 \quad \text{Positive D.} \]
\[ z \cdot w + b < 0 \quad \text{Negative D.} \]
**Accuracy Measure**

<table>
<thead>
<tr>
<th>Ground truth label</th>
<th>Prediction</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bedroom</td>
<td>Bedroom</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>Beach</td>
<td>Beach</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

Confusion matrix

- # of correct prediction on Bedroom data
  
  \[
  \frac{\text{# of correct prediction on Bedroom data}}{\text{# of Bedroom data}}
  \]

- # of incorrect prediction on Bedroom data
  
  \[
  \frac{\text{# of incorrect prediction on Bedroom data}}{\text{# of Bedroom data}}
  \]
### Accuracy Measure

<table>
<thead>
<tr>
<th>Ground truth label</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bedroom</td>
<td>0.7</td>
</tr>
<tr>
<td>Beach</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**Confusion matrix**

- Number of correct predictions on Bedroom data: 0.7
- Number of incorrect predictions on Bedroom data: 0.2
- Total number of Bedroom data: 1

**Accuracy**

\[
\text{Accuracy} = \frac{\text{# of correct predictions}}{\# \text{ of Bedroom data}} = \frac{0.7}{1} = 0.7
\]

\[
\text{Accuracy: mean of correct predictions} = \frac{(0.7 + 0.8)}{2} = 0.75
\]
COMPARISON

Tiny image representation + NN

(a) Image
(b) Tiny Image
<table>
<thead>
<tr>
<th>Kitchen</th>
<th>Store</th>
<th>Bedroom</th>
<th>LivingRoom</th>
<th>Office</th>
<th>Industrial</th>
<th>Suburb</th>
<th>InsideCity</th>
<th>TallBuilding</th>
<th>Street</th>
<th>Highway</th>
<th>OpenCountry</th>
<th>Coast</th>
<th>Mountain</th>
<th>Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kit</td>
<td>Sto</td>
<td>Bed</td>
<td>Liv</td>
<td>Off</td>
<td>Ind</td>
<td>Sub</td>
<td>Cty</td>
<td>Bld</td>
<td>St</td>
<td>HW</td>
<td>OC</td>
<td>Cst</td>
<td>Mnt</td>
<td>For</td>
</tr>
</tbody>
</table>

Accuracy: 0.205333

Tiny image representation + KNN (10)

(a) Image
(b) Tiny Image
Comparison

BoW + KNN (10)

Accuracy: 0.512667
### Comparison

<table>
<thead>
<tr>
<th>Kitchen</th>
<th>Sto</th>
<th>Bed</th>
<th>Liv</th>
<th>Off</th>
<th>Ind</th>
<th>Sub</th>
<th>Cty</th>
<th>Bld</th>
<th>St</th>
<th>HW</th>
<th>OC</th>
<th>Cst</th>
<th>Mnt</th>
<th>For</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Accuracy: 0.629333

BoW + SVM
**Matching**

Feature

Feature
Matching

Feature

Histogram

13 10

Feature

Histogram

12 11
Matching

Feature

Histogram

3 4

4 2

3 3

3 1

Feature

Histogram

3 4

3 3

3 2

3 2
SIMILARITY MEASURE—PYRAMID MATCHING KERNEL

Histogram intersection:
\[ \mathcal{I}(H_x^0, H_y^0) = \sum_{d=1}^{D} \min(H_x^0(d), H_y^0(d)) \]
SIMILARITY MEASURE—PYRAMID MATCHING KERNEL

Histogram intersection:
\[
I(H^0_x, H^0_y) = \sum_{d=1}^{D} \min(H^0_x(d), H^0_y(d))
\]
**Similarity Measure—Pyramid Matching Kernel**

Feature

Histogram

Histogram intersection:

\[
\mathcal{I}(H_x^0, H_y^0) = \sum_{d=1}^{D} \min(H_x^0(d), H_y^0(d))
\]

22
**Similarity Measure - Pyramid Matching Kernel**

Feature

Histogram

$H_X$

Histogram

$H_Y$

Feature

Histogram

Histogram

Histogram

Histogram intersection:

$\mathcal{I}(H_X^1, H_Y^1) = \sum_{d=1}^{D} \min(H_X^1(d), H_Y^1(d))$
**Similarity Measure - Pyramid Matching Kernel**

Histogram intersection:

\[ \mathcal{I}(H^1_x, H^1_y) = \sum_{d=1}^D \min(H^1_x(d), H^1_y(d)) \]
**Similarity Measure - Pyramid Matching Kernel**

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

22 21
**Similarity Measure - Pyramid Matching Kernel**

Kernel:
\[ \kappa^t(X, Y) = I^t + \sum_{i=0}^{t-1} \frac{1}{2^{i+1}} (I^i - I^{i+1}) \]
**SIMILARITY MEASURE—PYRAMID MATCHING KERNEL**

Feature | Histogram | Histogram | Histogram | Histogram
---|---|---|---|---

$H_1$

$H_2$
Original images

Feature histograms:
Level 3

Level 2

Level 1

Level 0

Total weight (value of pyramid match kernel): \[ I_3 + \frac{1}{2}(I_2 - I_3) + \frac{1}{4}(I_1 - I_2) + \frac{1}{8}(I_0 - I_1) \]
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)


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

<table>
<thead>
<tr>
<th>Level</th>
<th>Weak features (16)</th>
<th>Strong features (200)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single-level</td>
<td>Pyramid</td>
</tr>
<tr>
<td>0</td>
<td>15.5 ±0.9</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>31.4 ±1.2</td>
<td>32.8 ±1.3</td>
</tr>
<tr>
<td>2</td>
<td>47.2 ±1.1</td>
<td>49.3 ±1.4</td>
</tr>
<tr>
<td>3</td>
<td>52.2 ±0.8</td>
<td>54.0 ±1.1</td>
</tr>
</tbody>
</table>