Image Classification and Object Recognition

Selim Aksoy
Department of Computer Engineering
Bilkent University
saksoy@cs.bilkent.edu.tr
Image classification

- Image (scene) classification is a fundamental problem in image understanding.

- Automatic techniques for associating scenes with semantic labels have a high potential for improving the performance of other computer vision applications such as:
  - browsing (natural grouping of images instead of clusters based only on low-level features),
  - retrieval (filtering images in archives based on content), and
  - object recognition (the probability of an unknown object/region that exhibits several local features of a ship actually being a ship can be increased if the scene context is known to be a coast with high confidence but can be decreased if no water related context is dominant in that scene).
Image classification

- The image classification problem has two critical components: representing images and learning models for semantic categories using these representations.

- Early work used low-level global features extracted from the whole image or from a fixed spatial layout.

- More recent approaches exploit local statistics in images using patches extracted by interest point detectors.

- Other configurations that use regions and their spatial relationships are also proposed.
Hierarchical image classification

Hierarchical image classification

- **Image representation:**
  - Mean and std. dev. of LUV values in 10x10 blocks for indoor/outdoor classification.
  - Edge direction histograms for city/landscape classification.
  - Histograms of HSV and LUV values for sunset/mountain/forest classification.

- **Classification:**
  - Class-conditional density estimation using vector quantization.
  - Bayesian classification.
Hierarchical image classification

<table>
<thead>
<tr>
<th>Test Data</th>
<th>Database Size</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set</td>
<td>2,541</td>
<td>94.2</td>
</tr>
<tr>
<td>Test Set 1</td>
<td>2,540</td>
<td>88.2</td>
</tr>
<tr>
<td>Test Set 2</td>
<td>1,850</td>
<td>88.7</td>
</tr>
<tr>
<td>Entire Database</td>
<td>6,931</td>
<td>90.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Data</th>
<th>EDH</th>
<th>EDCV</th>
<th>CH</th>
<th>CCV</th>
<th>EDH &amp; CH</th>
<th>EDCV &amp; CCV</th>
<th>EDCV &amp; CH</th>
<th>EDCV &amp; CCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set</td>
<td>94.7</td>
<td>97.0</td>
<td>83.7</td>
<td>83.5</td>
<td>94.8</td>
<td>95.4</td>
<td>96.4</td>
<td>96.9</td>
</tr>
<tr>
<td>Test Set</td>
<td>92.0</td>
<td>92.9</td>
<td>75.4</td>
<td>76.0</td>
<td>92.5</td>
<td>92.8</td>
<td>93.4</td>
<td>93.8</td>
</tr>
<tr>
<td>Entire Database</td>
<td>93.4</td>
<td>95.0</td>
<td>79.6</td>
<td>79.8</td>
<td>93.7</td>
<td>94.1</td>
<td>94.9</td>
<td>95.3</td>
</tr>
</tbody>
</table>
Hierarchical image classification

<table>
<thead>
<tr>
<th>Test Data</th>
<th>EDH</th>
<th>EDCV</th>
<th>CH</th>
<th>CCV</th>
<th>SPM</th>
<th>EDH &amp; CH</th>
<th>EDCV &amp; CCV</th>
<th>EDCV &amp; CH</th>
<th>EDCV &amp; CCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set</td>
<td>88.3</td>
<td>88.3</td>
<td>96.2</td>
<td>99.2</td>
<td>98.9</td>
<td>95.9</td>
<td>96.6</td>
<td>95.5</td>
<td>97.0</td>
</tr>
<tr>
<td>Test Set</td>
<td>86.3</td>
<td>89.0</td>
<td>89.7</td>
<td>93.9</td>
<td>93.9</td>
<td>90.1</td>
<td>95.4</td>
<td>90.5</td>
<td>95.1</td>
</tr>
<tr>
<td>Entire Database</td>
<td>87.4</td>
<td>88.7</td>
<td>93.0</td>
<td>96.6</td>
<td>96.4</td>
<td>93.0</td>
<td>96.0</td>
<td>93.0</td>
<td>96.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Data</th>
<th>EDH</th>
<th>EDCV</th>
<th>CH</th>
<th>CCV</th>
<th>SPM</th>
<th>EDH &amp; CH</th>
<th>EDCV &amp; CCV</th>
<th>EDCV &amp; CH</th>
<th>EDCV &amp; CCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set</td>
<td>83.4</td>
<td>78.1</td>
<td>92.0</td>
<td>98.9</td>
<td>98.4</td>
<td>94.1</td>
<td>98.4</td>
<td>93.6</td>
<td>98.4</td>
</tr>
<tr>
<td>Test Set</td>
<td>87.1</td>
<td>77.2</td>
<td>91.4</td>
<td>91.9</td>
<td>93.6</td>
<td>93.0</td>
<td>92.5</td>
<td>93.5</td>
<td>91.9</td>
</tr>
<tr>
<td>Entire Database</td>
<td>85.3</td>
<td>77.7</td>
<td>91.7</td>
<td>95.5</td>
<td>96.0</td>
<td>93.6</td>
<td>95.5</td>
<td>93.6</td>
<td>95.2</td>
</tr>
</tbody>
</table>

TABLE V
Classification Accuracies (in Percent) for Sunset/Forest/Mountain Classification; SPM Stands for “Spatial Color Moments”

TABLE VI
Classification Accuracies (in Percent) for Forest/Mountain Classification
Image classification using bag-of-words

Caltech data set: 13 natural scene categories.

IDIAP data set (left to right): mountain, forest, indoor, city-panorama, city-street.
Image classification using bag-of-words

Image classification using bag-of-words

A codebook obtained from 650 training examples from 13 categories. Image patches are detected by a sliding grid and random sampling of scales.
Figure 5. Internal structure of the models learnt for each category. Each row represents one category. The left panel shows the distribution of the 40 intermediate themes. The right panel shows the distribution of codewords as well as the appearance of 10 codewords selected from the top 20 most likely codewords for this category model.
Figure 6. Examples of testing images for each category. Each row is for one category. The first 3 columns on the left show 3 examples of correctly recognized images, the last column on the right shows an example of incorrectly recognized image. Superimposed on each image, we show samples of patches that belong to the most significant set of codewords given the category model.
Image classification using bag-of-words

**Figure 7.** Left Panel. Confusion table of Theme Model 1 using 100 training and 50 test examples from each category, the grid detector and patch based representation. The average performance is 64.0%. **Right Panel.** Rank statistics of the confusion table, which shows the probability of a test scene correctly belong to one of the top $N$ most probable categories. $N$ ranges from 1 to 13.
Image classification using bag-of-words

- Probabilistic Latent Semantic Analysis (PLSA) is used to learn aspect models to capture co-occurrences of visterms (visual terms).
- Bag-of-visterms representation or the aspect parameters are given as input to Support Vector Machines for classification.

## Image classification using bag-of-words

<table>
<thead>
<tr>
<th>Total class. error</th>
<th>11.1 (0.8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gr. Truth</td>
<td>Classification (%)</td>
</tr>
<tr>
<td>indoor</td>
<td>89.7</td>
</tr>
<tr>
<td>city</td>
<td>14.5</td>
</tr>
<tr>
<td>landscape</td>
<td>1.2</td>
</tr>
</tbody>
</table>

**TABLE III**

CONFUSION MATRIX FOR THE THREE-CLASS CLASSIFICATION PROBLEM, USING VOCABULARY $V_{1000}$.

Total class. error rate: 20.8 (2.1) (Baseline: 30.1 (1.1))

<table>
<thead>
<tr>
<th>Total class. error (BOV: 20.8 (2.1), Baseline: 30.1 (1.1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>m.</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>mount.</td>
</tr>
<tr>
<td>forest</td>
</tr>
<tr>
<td>indoor</td>
</tr>
<tr>
<td>city-pan.</td>
</tr>
<tr>
<td>city-str.</td>
</tr>
</tbody>
</table>

**TABLE V**

CLASSIFICATION RATE AND CONFUSION MATRIX FOR THE FIVE-CLASS, USING BOV AND VOCABULARY $V_{1000}$.

<table>
<thead>
<tr>
<th>Total class. error</th>
<th>11.9(1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gr. Truth</td>
<td>Classification (%)</td>
</tr>
<tr>
<td>indoor</td>
<td>86.6</td>
</tr>
<tr>
<td>city</td>
<td>14.8</td>
</tr>
<tr>
<td>land.</td>
<td>1.3</td>
</tr>
</tbody>
</table>

**TABLE VIII**

CLASSIFICATION ERROR AND CONFUSION MATRIX FOR THE THREE-CLASS PROBLEM USING PLSA, WITH $V_{1000}$ AND 60 ASPECTS.

<table>
<thead>
<tr>
<th>Total error rate (BOV: 20.8 (2.1), Baseline: 30.1 (1.1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>m.</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>mountain</td>
</tr>
<tr>
<td>forest</td>
</tr>
<tr>
<td>indoor</td>
</tr>
<tr>
<td>city-pan.</td>
</tr>
<tr>
<td>city-str.</td>
</tr>
</tbody>
</table>
Fig. 7. The 10 most probable images from the D1 data set for seven aspects (out of 20) learned on the D3 data set.
Image classification using bag-of-regions

  - Region segmentation
  - Region clustering $\rightarrow$ region codebook
  - Above-below spatial relationships $\rightarrow$ region pairs
  - Statistical region selection: identify region types that
    - are frequently found in a particular class of scenes but rarely exist in other classes, and
    - consistently occur together in the same class of scenes.
  - Bayesian scene classification using
    - bag of individual regions,
    - bag of region pairs.
Image classification using bag-of-regions

Examples for region clusters.
Each row represents a different cluster.
Image classification using bag-of-regions

Table 3. Confusion matrix for the bag of individual regions representation after region selection.

<table>
<thead>
<tr>
<th></th>
<th>Assigned</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coast</td>
<td>forest</td>
<td>highway</td>
<td>insidecity</td>
<td>mountain</td>
<td>opencountry</td>
<td>street</td>
<td>Total</td>
<td>% Agree</td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>38</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>50</td>
<td>76.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>36</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>50</td>
<td>72.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>32</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>50</td>
<td>64.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>12</td>
<td>22</td>
<td>2</td>
<td>0</td>
<td>10</td>
<td>50</td>
<td>44.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>32</td>
<td>6</td>
<td>2</td>
<td>50</td>
<td>64.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>14</td>
<td>14</td>
<td>1</td>
<td>50</td>
<td>28.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>27</td>
<td>50</td>
<td>54.00</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>52</td>
<td>63</td>
<td>36</td>
<td>60</td>
<td>34</td>
<td>47</td>
<td>350</td>
<td>57.43</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Confusion matrix for the bag of region pairs representation after region selection.

<table>
<thead>
<tr>
<th></th>
<th>Assigned</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>coast</td>
<td>forest</td>
<td>highway</td>
<td>insidecity</td>
<td>mountain</td>
<td>opencountry</td>
<td>street</td>
<td>Total</td>
<td>% Agree</td>
<td></td>
</tr>
<tr>
<td>True</td>
<td>42</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>50</td>
<td>84.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>38</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>50</td>
<td>76.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>31</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>50</td>
<td>62.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td>12</td>
<td>19</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>50</td>
<td>38.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>3</td>
<td>1</td>
<td>50</td>
<td>80.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>25</td>
<td>0</td>
<td>50</td>
<td>50.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>12</td>
<td>2</td>
<td>3</td>
<td>22</td>
<td>50</td>
<td>44.00</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>54</td>
<td>52</td>
<td>40</td>
<td>61</td>
<td>42</td>
<td>43</td>
<td>350</td>
<td>62.00</td>
<td></td>
</tr>
</tbody>
</table>
Image classification using bag-of-regions

Examples for correctly classified scenes.

Examples for wrongly classified scenes.
Image classification using factor graphs


Figure 1. (a) A beach scene. (b) Its manually-labeled materials. The true configuration includes sky above water, water above sand, and sky above sand. (c) The underlying graph showing detector results and spatial relations.
Recognizing and Learning Object Categories

Li Fei-Fei, UIUC
Rob Fergus, MIT
Antonio Torralba, MIT
Agenda

- Introduction
- Bag of words models
- Part-based models
- Discriminative methods
- Segmentation and recognition
- Conclusions
object

Perception Key (əˈbʒekt, -jəkt')

n.

1. Something one or more of the senses, especially sight or touch; a visible or tangible thing.
2. A focus of attention, thinking, thought, or action: an object of compassion.
3. The purpose or goal of a specific action or effort: the object of a work of art.
   a. A noun, pronoun, or noun phrase that receives or is affected by the action of a verb within a sentence.
   b. A noun or substantive governed by a preposition.
5. Philosophy. Something intelligible or perceptible by the mind.
6. Computer Science. A discrete item that can be selected and maneuvered, such as an onscreen graphic. In object-oriented programming, objects include data and the procedures necessary to operate on that data.
How many object categories are there?

~10,000 to 30,000

Biederman 1987
So what does object recognition involve?
Verification: is that a bus?
Detection: are there cars?
Identification: is that a picture of Mao?
Object categorization

- sky
- building
- flag
- banner
- face
- street lamp
- wall
- bus
- cars
Scene and context categorization

- outdoor
- city
- traffic
- ...

[Image of a cityscape with a famous building and buses]
Challenges 1: view point variation

Michelangelo 1475-1564
Challenges 2: illumination

slide credit: S. Ullman
Challenges 3: occlusion

Magritte, 1957
Challenges 4: scale
Challenges 5: deformation

Xu, Beihong 1943
Challenges 6: background clutter

Klimt, 1913
History: single object recognition
Challenges 7: intra-class variation
History: early object categorization
\~10,000 to 30,000
Scenes, Objects, and Parts

Object categorization: the statistical viewpoint

Bayes rule:

\[
p(\text{zebra} \mid \text{image}) \quad \text{vs.} \quad p(\text{no zebra} \mid \text{image})
\]

\[
\frac{p(\text{zebra} \mid \text{image})}{p(\text{no zebra} \mid \text{image})} = \frac{p(\text{image} \mid \text{zebra})}{p(\text{image} \mid \text{no zebra})} \cdot \frac{p(\text{zebra})}{p(\text{no zebra})}
\]

posterior ratio
likelihood ratio
prior ratio
Object categorization: the statistical viewpoint

\[
\frac{p(\text{zebra} \mid \text{image})}{p(\text{no zebra} \mid \text{image})} = \frac{p(\text{image} \mid \text{zebra})}{p(\text{image} \mid \text{no zebra})} \cdot \frac{p(\text{zebra})}{p(\text{no zebra})}
\]

- **Posterior ratio**
- **Likelihood ratio**
- **Prior ratio**

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

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

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

<table>
<thead>
<tr>
<th>$p(image \mid zebra)$</th>
<th>$p(image \mid no \ zebra)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Middle</td>
</tr>
<tr>
<td>High</td>
<td>Middle $\rightarrow$ Low</td>
</tr>
</tbody>
</table>
Three main issues

- Representation
  - How to represent an object category

- Learning
  - How to form the classifier, given training data

- Recognition
  - How the classifier is to be used on novel data
Representation

- Generative / discriminative / hybrid
Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
- Invariances
  - View point
  - Illumination
  - Occlusion
  - Scale
  - Deformation
  - Clutter
  - etc.
Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
- Invariances
- Part-based or global w/sub-window
Representation

- Generative / discriminative / hybrid
- Appearance only or location and appearance
- Invariances
- Parts or global w/sub-window
- Use set of features or each pixel in image
Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning.
Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)

- Methods of training: generative vs. discriminative
Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)
- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)
- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels

Contains a motorbike
Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning

What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)

Level of supervision
- Manual segmentation; bounding box; image labels; noisy labels

Batch/incremental (on category and image level; user-feedback)
Learning

- Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning.

- What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)

- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels

- Batch/incremental (on category and image level; user-feedback)

- Training images:
  - Issue of overfitting
  - Negative images for discriminative methods
Unclear how to model categories, so we learn what distinguishes them rather than manually specify the difference -- hence current interest in machine learning)

What are you maximizing? Likelihood (Gen.) or performances on train/validation set (Disc.)

Level of supervision
- Manual segmentation; bounding box; image labels; noisy labels

Batch/incremental (on category and image level; user-feedback)

Training images:
- Issue of overfitting
- Negative images for discriminative methods

Priors
Recognition

- Scale / orientation range to search over
- Speed
Object Class Recognition using Images of Abstract Regions

Yi Li, Jeff A. Bilmes, and Linda G. Shapiro
Department of Computer Science and Engineering
Department of Electrical Engineering
University of Washington
Problem Statement

**Given:** Some images and their corresponding descriptions

{trees, grass, cherry trees} {cheetah, trunk} {mountains, sky} {beach, sky, trees, water}

**To solve:** What object classes are present in new images
Abstract Regions

Original Images | Color Regions | Texture Regions | Line Clusters

[Images of original images, color regions, texture regions, and line clusters]
Object Model Learning (Ideal)

region attributes $\rightarrow$ object

Learned Models

Sky =
Tree =
Water =
Boat =
Our Scenario: Abstract Regions

Multiple segmentations whose regions are not labeled; a list of labels is provided for each training image.

labels
{sky, building}

various different segmentations

region attributes from several different types of regions
Object Model Learning

Assumptions

- The feature distribution of each object within a region is a Gaussian;

- Each image is a set of regions; each region can be modeled as a mixture of multivariate Gaussian distributions.
Model Initial Estimation

- Estimate the initial model of an object using all the region features from all images that contain the object
Expectation-Maximization

Initial Model for “trees”

Initial Model for “sky”

Final Model for “trees”

Final Model for “sky”

EM
1. Initialization Step (Example)

Image & description

\[ I_1 \rightarrow \begin{cases} O_1 \\ O_2 \end{cases}, \quad I_2 \rightarrow \begin{cases} O_1 \\ O_3 \end{cases}, \quad I_3 \rightarrow \begin{cases} O_2 \\ O_3 \end{cases} \]

\[ N_{O_1}^{(0)} = \begin{cases} \text{Red: } W=0.5 \\ \text{Green: } W=0.5 \end{cases}, \quad N_{O_2}^{(0)} = \begin{cases} \text{Green: } W=0.5 \\ \text{Blue: } W=0.5 \end{cases}, \quad N_{O_3}^{(0)} = \begin{cases} \text{Blue: } W=0.5 \end{cases} \]
2. Iteration Step (Example)

\[ I_1 \rightarrow \begin{cases} O_1 \\ O_2 \end{cases} \]

\[ I_2 \rightarrow \begin{cases} O_1 \\ O_3 \end{cases} \]

\[ I_3 \rightarrow \begin{cases} O_2 \\ O_3 \end{cases} \]

\[ E-Step \]

\[ M-Step \]

\[ N_{O_1}^{(p)} \]

\[ N_{O_2}^{(p)} \]

\[ N_{O_3}^{(p)} \]

\[ N_{O_1}^{(p+1)} \]

\[ N_{O_2}^{(p+1)} \]

\[ N_{O_3}^{(p+1)} \]
Recognition

Test Image

Color Regions

Object Model
Database

To calculate $p(\text{tree} \mid \text{image})$

$$p(\text{tree} \mid \text{image}) = f \begin{pmatrix} p(\text{tree} \mid \text{image}) \\ p(\text{tree} \mid \text{image}) \\ p(\text{tree} \mid \text{image}) \\ p(\text{tree} \mid \text{image}) \end{pmatrix}$$

$$p(\text{o} \mid F_i^a) = f \left( \prod_{r^a \in F_i^a} p(\text{o} \mid r^a) \right)$$
Combining different abstract regions

- Treat the different types of regions independently and combine at the time of classification.

\[ p(o \mid \{F_i^a\}) = \prod_a p(o \mid F_i^a) \]

- Form intersections of the different types of regions, creating smaller regions that have both color and texture properties for classification.
Experiments (on 860 images)

- 18 keywords: mountains (30), orangutan (37), track (40), tree trunk (43), football field (43), beach (45), prairie grass (53), cherry tree (53), snow (54), zebra (56), polar bear (56), lion (71), water (76), chimpanzee (79), cheetah (112), sky (259), grass (272), tree (361).

- A set of cross-validation experiments (80% as training set and the other 20% as test set)

- The poorest results are on object classes “tree,” “grass,” and “water,” each of which has a high variance; a single Gaussian model is insufficient.
**ROC Charts**

Independent Treatment of Color and Texture

Using Intersections of Color and Texture Regions
Sample Results

cheetah
Sample Results (Cont.)

grass
Sample Results (Cont.)

cherry tree
Sample Results (Cont.)

lion
Summary

- Designed a set of abstract region features: color, texture, structure, . . .

- Developed a new semi-supervised EM-like algorithm to recognize object classes in color photographic images of outdoor scenes; tested on 860 images.

- Compared two different methods of combining different types of abstract regions. The intersection method had a higher performance.
Groundtruth Data Set

- UW Ground truth database (1224 images)
- 31 elementary object categories: river (30), beach (31), bridge (33), track (35), pole (38), football field (41), frozen lake (42), lantern (42), husky stadium (44), hill (49), cherry tree (54), car (60), boat (67), stone (70), ground (81), flower (85), lake (86), sidewalk (88), street (96), snow (98), cloud (119), rock (122), house (175), bush (178), mountain (231), water (290), building (316), grass (322), people (344), tree (589), sky (659)
- 20 high-level concepts: Asian city, Australia, Barcelona, campus, Cannon Beach, Columbia Gorge, European city, Geneva, Green Lake, Greenland, Indonesia, indoor, Iran, Italy, Japan, park, San Juans, spring flowers, Swiss mountains, and Yellowstone.
## Groundtruth Data Set:
### ROC Scores

<table>
<thead>
<tr>
<th>Item</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>street</td>
<td>60.4</td>
</tr>
<tr>
<td>people</td>
<td>68.0</td>
</tr>
<tr>
<td>rock</td>
<td>73.5</td>
</tr>
<tr>
<td>sky</td>
<td>74.1</td>
</tr>
<tr>
<td>ground</td>
<td>74.3</td>
</tr>
<tr>
<td>river</td>
<td>74.7</td>
</tr>
<tr>
<td>grass</td>
<td>74.9</td>
</tr>
<tr>
<td>building</td>
<td>75.4</td>
</tr>
<tr>
<td>cloud</td>
<td>75.4</td>
</tr>
<tr>
<td>boat</td>
<td>76.8</td>
</tr>
<tr>
<td>lantern</td>
<td>78.1</td>
</tr>
<tr>
<td>australia</td>
<td>79.7</td>
</tr>
<tr>
<td>house</td>
<td>80.1</td>
</tr>
<tr>
<td>tree</td>
<td>80.8</td>
</tr>
<tr>
<td>bush</td>
<td>81.0</td>
</tr>
<tr>
<td>flower</td>
<td>81.1</td>
</tr>
<tr>
<td>iran</td>
<td>82.2</td>
</tr>
<tr>
<td>bridge</td>
<td>82.7</td>
</tr>
<tr>
<td>car</td>
<td>82.9</td>
</tr>
<tr>
<td>pole</td>
<td>83.3</td>
</tr>
<tr>
<td>yellowstone</td>
<td>83.7</td>
</tr>
<tr>
<td>water</td>
<td>83.9</td>
</tr>
<tr>
<td>indonesia</td>
<td>84.3</td>
</tr>
<tr>
<td>sidewalk</td>
<td>85.7</td>
</tr>
<tr>
<td>asian city</td>
<td>86.7</td>
</tr>
<tr>
<td>european city</td>
<td>87.0</td>
</tr>
<tr>
<td>stone</td>
<td>87.1</td>
</tr>
<tr>
<td>hill</td>
<td>87.4</td>
</tr>
<tr>
<td>mountain</td>
<td>88.3</td>
</tr>
<tr>
<td>beach</td>
<td>89.0</td>
</tr>
<tr>
<td>snow</td>
<td>92.0</td>
</tr>
<tr>
<td>lake</td>
<td>92.8</td>
</tr>
<tr>
<td>frozen lake</td>
<td>92.8</td>
</tr>
<tr>
<td>japan</td>
<td>92.9</td>
</tr>
<tr>
<td>campus</td>
<td>92.9</td>
</tr>
<tr>
<td>barcelona</td>
<td>92.9</td>
</tr>
<tr>
<td>geneva</td>
<td>93.3</td>
</tr>
<tr>
<td>park</td>
<td>94.0</td>
</tr>
<tr>
<td>spring flowers</td>
<td>94.4</td>
</tr>
<tr>
<td>columbia gorge</td>
<td>94.5</td>
</tr>
<tr>
<td>green lake</td>
<td>94.9</td>
</tr>
<tr>
<td>italy</td>
<td>95.1</td>
</tr>
<tr>
<td>swiss moutains</td>
<td>95.7</td>
</tr>
<tr>
<td>sanjuans</td>
<td>96.5</td>
</tr>
<tr>
<td>cherry tree</td>
<td>96.9</td>
</tr>
<tr>
<td>indoor</td>
<td>97.0</td>
</tr>
<tr>
<td>greenland</td>
<td>98.7</td>
</tr>
<tr>
<td>cannon beach</td>
<td>99.2</td>
</tr>
<tr>
<td>track</td>
<td>99.6</td>
</tr>
<tr>
<td>football field</td>
<td>99.8</td>
</tr>
<tr>
<td>husky stadium</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Groundtruth Data Set: Top Results

- **Asian city**
  - Images related to Asian cities

- **Cannon beach**
  - Images related to Cannon beach

- **Italy**
  - Images related to Italy

- **park**
  - Images related to parks
Groundtruth Data Set:
Top Results

sky

spring flowers

tree

water
VACE Test Image Set (828 images and 10 object classes): from Boeing, VIVID, and NGA videos
Experiments: ROC Curves

<table>
<thead>
<tr>
<th>Class</th>
<th>True Positive Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>field</td>
<td>77.5</td>
</tr>
<tr>
<td>tree</td>
<td>80.6</td>
</tr>
<tr>
<td>car</td>
<td>82.3</td>
</tr>
<tr>
<td>people</td>
<td>83.9</td>
</tr>
<tr>
<td>house</td>
<td>84.9</td>
</tr>
<tr>
<td>paved road</td>
<td>87.5</td>
</tr>
<tr>
<td>forest</td>
<td>87.6</td>
</tr>
<tr>
<td>dirt road</td>
<td>89.5</td>
</tr>
<tr>
<td>airplane</td>
<td>91.1</td>
</tr>
<tr>
<td>runway</td>
<td>94.4</td>
</tr>
</tbody>
</table>
Objects detected in frames

**Objects detected in frames**

- **forest** (94.37) **house** (64.09)
  - car (46.5)
  - dirt road (23.44)
  - paved road (4.77)
  - tree (2.29)
  - airplane (1.47)
  - runway (0.03)
  - field (0.02)
  - people (0)

- **runway** (99.98) **field** (98.66) **car** (96.24)
  - people (10.04)
  - airplane (2.74)
  - paved road (2.39)
  - forest (0.82)
  - house (0.48)
  - dirt road (0.41)
  - tree (0)

- **car** (94.3) **dirt road** (91.7) **field** (16.17)
  - tree (14.23)
  - paved road (5.34)
  - airplane (5.17)
  - people (3.91)
  - forest (0.53)
  - house (0.47)
  - runway (0.41)

- **runway** (100) **car** (99.23) **field** (98.07)
  - dirt road (92.1)
  - house (85.24)
  - tree (19.43)
  - paved road (5.77)
  - airplane (3.56)
  - forest (2.85)
  - people (0.07)

- **runway** (99.98) **car** (99.84) **field** (99.27)
  - paved road (18.28)
  - people (13.13)
  - tree (8.71)
  - airplane (7.94)
  - forest (1.67)
  - house (0.14)
  - dirt road (0.08)

- **car** (97.92) **forest** (94.2) **paved road** (85)
  - dirt road (72.94)
  - tree (68.84)
  - airplane (39.13)
  - house (33.17)
  - people (12.97)
  - field (2.38)
  - runway (0.04)