Discovering objects and their location in images

Josef Sivic et al.
ICCV 2005

Presented by Elden Yu, with quite some slides borrowed

Motivation

• Is it possible to learn object classes **without supervision**?
  – Results in speech recognition highlight the importance of huge amounts of training data
  – Annotation is expensive
  – Much more unsupervised data than labeled data
• The success in text mining
  – The bag of words representation
  – Latent semantic analysis

Latent Semantic Analysis

• $D = (d_1, \ldots, d_N)$ N documents
• $W = (w_1, \ldots, w_M)$ M words
• $N_{ij} = \#(d_i, w_j)$ NxM co-occurrence term-document matrix

• A kind of Singular Value Decomposition to represent each document by its top K topics instead of by its M words

Visual object classes

Latent Semantic Analysis

• $D = (d_1, \ldots, d_N)$ N documents
• $W = (w_1, \ldots, w_M)$ M words
• $N_{ij} = \#(d_i, w_j)$ NxM co-occurrence term-document matrix

• A kind of Singular Value Decomposition to represent each document by its top K topics instead of by its M words
**Probabilistic Latent Semantic Analysis (pLSA)**

[Hofmann '99]

\[ P(w_i|d_j) = \sum_{k=1}^{K} P(z_k|d_j)P(w_i|z_k) \]

- **Maximize likelihood of data using EM.**
- **Minimize KL divergence between empirical distribution and model.**

| Model fitting: find topic vectors \( P(w|z) \) common to all documents, and mixture coefficients \( P(z|d) \) specific to each document. |

**Learning the pLSA parameters**

\[ L = \prod_{i=1}^{M} \prod_{j=1}^{N} P(w_i|d_j)^{n(w_i,d_j)} \]

Unlike LSA, pLSA does not minimize any type of squared deviation. The parameters are estimated in a probabilistically sound way.

**Visual Words**

- Vector Quantized SIFT descriptors computed in regions.
- Regions come from elliptical shape adaptation around interest point, and from the maximally stable regions of Matas et al.
- Both are elliptical regions at twice their detected scale.

**EM for pLSA**

- **E-step:** compute posterior probabilities for the latent variables

\[ P(z_k|d_j) = \frac{P(w_i|z_k)P(z_k|d_j)}{\sum_{z_k} P(w_i|z_k)P(z_k|d_j)} \]

- **M-step:** maximize the expected complete data log-likelihood

\[ P(w_i|z_k) = \frac{\sum_{d_j} n(w_i,d_j)P(z_k|d_j)}{\sum_{z_k} \sum_{d_j} n(w_i,d_j)P(z_k|d_j)} \]

\[ P(z_k|d_j) = \frac{\sum_{w_i} n(w_i,d_j)P(w_i|z_k)}{\sum_{z_k} \sum_{w_i} n(w_i,d_j)P(w_i|z_k)} \]

**Visual object discovery - overview**

- **Find words**
- **Form histograms**
- **Discover topics**

**Building a Vocabulary**

- Collection of images
- Vocabulary (grid)
- Documents
Experiment on topic discovery

- For each image $d_j$
  - Compute $P(z_k|d_j)$ over $k$, and classify the image as containing object $k$ according to the max of $P(z_k|d_j)$ over $k$
- With(1)/without(2) explicit background

<table>
<thead>
<tr>
<th>Ex</th>
<th>Categories</th>
<th>$K$</th>
<th>pLSA</th>
<th>KSM baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>4</td>
<td>4</td>
<td>98</td>
<td>79</td>
</tr>
<tr>
<td>(2)</td>
<td>4 + bg</td>
<td>5</td>
<td>78</td>
<td>50</td>
</tr>
<tr>
<td>(2')</td>
<td>4 + bg</td>
<td>6</td>
<td>74</td>
<td>48</td>
</tr>
<tr>
<td>(2')</td>
<td>4 bg</td>
<td>7</td>
<td>83</td>
<td>70</td>
</tr>
<tr>
<td>(2')</td>
<td>4 bg, bg</td>
<td>7</td>
<td>93</td>
<td>250</td>
</tr>
</tbody>
</table>

Table 1: Summary of the experiments. Column "pLSA" shows the classification accuracy measured by the average of the diagonal of the confusion matrix. Column "KSM baseline" shows the total number of misclassifications. See text for a more detailed description of the experimental results. In the case of $K'$, the two classes of background topics are assigned to one category. Evidently the baseline method performs poorly, showing the lack of the pLSA clustering.

Experiment on classifying new images

- For each image $d_j$
  - Compute $P(z_k|d_{test})$ over $k$, with the fold-in heuristic
  - Classify the image as containing object $k$ according to the max of $P(z_k|d_{test})$ over $k$
- Experiment (3) is a modification of (2)

Table 3: Confusion table for unseen test images in experiment (3)

Experiment on classifying new images

- Experiment (4) is to compare with that of Fergus, with quite some hard heuristics

<table>
<thead>
<tr>
<th>Category</th>
<th>pLSA</th>
<th>Constellation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faces</td>
<td>3.3</td>
<td>3.6</td>
</tr>
<tr>
<td>Motorbikes</td>
<td>8.0</td>
<td>6.7</td>
</tr>
<tr>
<td>Airplanes</td>
<td>1.6</td>
<td>7.5</td>
</tr>
<tr>
<td>Cars</td>
<td>7.6</td>
<td>9.7</td>
</tr>
</tbody>
</table>

- Comparable performance to constellation model
- Level of supervision: pLSA; one number (of topics)
- Constellation: 400 labels for each category
- Also an indication of the level of difficulty of the Caltech 4 dataset
Image as a mixture of topics (objects)

Use posterior over topics to classify individual visual words:

\[ P(z_k | w_i, d_j) = \frac{P(w_i | z_k) P(z_k | d_j)}{\sum_{k=1}^{K} P(w_i | z_k) P(z_k | d_j)} \]

Example (sparse-) segmentations

So far: Image as a ‘bag-of-visual-words’

Shortcomings:
- soft segmentation
- all spatial relations between visual words are lost

Use image segmentation to propose groupings of visual words