Relevance Models for Automatic Image and Video Annotation & Retrieval

R. Manmatha
Center for Intelligent Information Retrieval
University of Massachusetts Amherst

With S. L. Feng, J. Jeon, V. Lavrenko and T. Rath
How do we retrieve images?

• Using CBIR systems
  – Hard to represent information needs using abstract image features; color percentages, color layout and textures.
How do we retrieve images?

• IBM QBIC system example using color

• Users prefer textual queries!
How do we retrieve images?

• Use Google image search!
  – Google uses filenames, surrounding text and ignores contents of the images.
How do we retrieve images?

• Using manual annotations
  – Libraries, Museums
  – Manual annotation is expensive.

Picture from Library of Congress
American Memory Collections

CREATED/PUBLISHED: 1940 August
NOTES: Store or cafe with soft drink signs: Coca-Cola, Orange-Crush, Royal Crown, Double Cola and Dr. Pepper.
SUBJECTS: Carbonated beverages Advertisements Restaurants United States—Mississippi—Natchez Slides--Color
CALL NUMBER: LC-USF35-115
Motivation

• How to retrieve images/videos?
  – Retrieval based on similarity search of visual features
    • Doesn’t support textural queries
    • Doesn’t capture “semantics
  – Automatically annotate images then retrieve based on the textual annotations.

Example Annotations:
Tiger, grass.
Automatic Annotation and Retrieval

- Automatically annotate unseen images/keyframes
  - A training set of annotated images/keyframes.
    - Do not know which word corresponds to which part of image.
  - Compute visterms (image features).
  - Learn a model and annotate a set of test images/keyframes.
  - Learn all annotations at the same time.

- Retrieval based on the annotation output.
  - Use query likelihood language model.
  - Rank test images/videos according to the likelihoods
Overview

• Motivation
• Image and word vocabularies.
  – Visterms and keywords.
• Cross Media Relevance Model
  – Annotation
  – Retrieval
• Continuous Relevance Model (CRM).
• Multiple Bernoulli Relevance Model and Normalized CRM Model.
• Application to Handwriting.
• Conclusions and Future Work.
Automatic Image Annotation

• Automatically annotate unseen images given a training set of images and annotations.
• Create a vocabulary of visterms.
  – Each image is generated by a set of visterms.
• Each image is described using two parallel vocabularies “visterms” and “words”
• Given a training set of visterms and image annotations, learn a model and annotate a test set of images.
• Given two parallel vocabularies, possible approaches
  – Analogous to the problem of Statistical Machine Translation (Duygulu et al, ECCV 2002)
  – Analogous to the problem of Cross-Lingual Retrieval (Approach here).
Image Vocabulary - Visterms

• Can we represent all the images with a finite set of symbols?
  – Text documents consist of words
  – Images consist of visterms

V123 V89 V988
V4552 V12336 V2
V765 V9887
Construction of Visterms

• Segmented images (e.g. Blobworld, Normalized-cuts algorithm.)
• Cluster segments. Each cluster is a visterm
Discrete Visterms

- Partition keyframe, clusters across images.
- Rectangular partition works better!
- Segmentation needs to be improved!
Visterms

• Or partition using a rectangular grid and cluster.
• Actually works better.
Now our goal is finding relationship between visterms and words.

- \( P(\text{Tiger} | V1) \), \( P( V1 | \text{Tiger} ) \), \( P( \text{Maui} | V3, V4 ) \)
Co-Occurrence Models

- Mori et al. 1999
- Create the co-occurrence table using a training set of annotated images
- Tend to annotate with high frequency words
- Context is ignored
  - Needs joint probability models

<table>
<thead>
<tr>
<th></th>
<th>w1</th>
<th>w2</th>
<th>w3</th>
<th>w4</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>V2</td>
<td>32</td>
<td>40</td>
<td>13</td>
<td>32</td>
</tr>
<tr>
<td>V3</td>
<td>13</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V4</td>
<td>65</td>
<td>43</td>
<td>12</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
P( w1 | v1 ) = \frac{12}{12+2+0+1} = 0.8
\]
\[
P( v3 | w2 ) = \frac{12}{2+40+12+43} = 0.12
\]
Translation Models

- Duygulu et al. 2002
  - Use classical IBM machine translation models to translate visterms into words
  - IBM machine translation models
    - Need a bi-lingual corpus to train the models

Examples:

- Mary did not slap the green witch
- Mary no daba una botefada a la bruja verde

- Maui People Dance
- Tiger grass sky

V2 V4 V6
V1 V34 V321 V21
Using Context in Images

• Images are defined by spatial context.
  – Isolated pixels have no meaning.
  – Context simplifies recognition/retrieval.
  – Tiger is associated with grass, tree, forest.
    • Unlikely to be associated with computers.

• Relevance models are a nice way of introducing context (without having to do so explicitly).
  – Do this by computing the joint probability of images and words.
  – Relevance Models originally introduced for text retrieval and cross-lingual retrieval
  – A formal way of introducing query expansion.
Cross Media Relevance Models

- Goal: Estimating Relevance Model – the joint distribution of words and visterms.

- Find probability of observing word w and visterms $b_i$ $P(w,b_1,\ldots,b_m)$ together.

- To annotate image with visterms
  - Grass, tiger, tree, road
  - $P(w|b_{\text{grass}},b_{\text{tiger}},b_{\text{tree}},b_{\text{road}})$
  - If top two probabilities are for words
    - grass, tiger,
    - Then annotate image with grass, tiger,
Relevance Models for Images

- Every image is represented both in terms of words \( \{w_i\} \) and visterms \( \{b_j\} \).
- For each image \( I \) there is a probability distribution \( P(.|I) \) which is a joint distribution of words and visterms—relevance model.
- How does one estimate \( P(.|I) \)?
  - The observed representation \( \{b_1 \ldots b_m\} \) results from \( m \) random samples from \( P(.|I) \).

- Annotation involves sampling \( n \) words \( \{w_1 \ldots w_n\} \) from the relevance model \( P(.|I) \).
- Notion: Tiger is associated with certain other visual categories grass, tree, forest…
Relevance Models

- Annotation \( P(w \mid I) \approx P(w \mid b_1 \ldots b_m) \)

- Or

\[
P(w \mid b_1 \ldots b_m) = \frac{P(w, b_1 \ldots b_m)}{\sum_w P(w, b_1 \ldots b_m)}
\]

Training

• Joint distribution computed as an expectation over the training set $J$

$$P(w, b_1...b_m) = \sum_{J} P(J)P(w, b_1,..., b_m \mid J)$$

• Since the events are independent

$$P(w, b_1...b_m) = \sum_{J} P(J)P(w \mid J)\prod_{i=1}^{m} P(b_i \mid J)$$
Estimation

• Use smoothed maximum likelihood estimates

\[
P(w \mid J) = (1 - \alpha_j) \frac{(w, J)}{|J|} + \alpha_j \frac{(w, T)}{|T|}
\]

\[
P(b \mid J) = (1 - \beta_j) \frac{(b, J)}{|J|} + \beta_j \frac{(b, T)}{|T|}
\]

– #(w,J) occurrences of w in image J’s caption.
– #(w,T) occurrences of w in all captions in training set T.
– |J| no. of visterms and words in image J.
– |T| size of training set.
Annotation

- Compute $P(w|I)$ for different $w$.
- Probabilistic Annotation: (PACMRM)
  - Annotate the image with every possible $w$ in the vocabulary with associated probabilities.
  - Useful for retrieval but not for people.

- Fixed Length Annotation: (FACMRM)
  - For people, take the top (say 3 or 4) words for every image and annotate images with them.

<table>
<thead>
<tr>
<th>Word</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>water</td>
<td>0.2301</td>
</tr>
<tr>
<td>people</td>
<td>0.2277</td>
</tr>
<tr>
<td>pool</td>
<td>0.2239</td>
</tr>
<tr>
<td>swimmers</td>
<td>0.2228</td>
</tr>
<tr>
<td>sky</td>
<td>0.0058</td>
</tr>
<tr>
<td>tree</td>
<td>0.0056</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Retrieval

• Language Modeling Approach:
• Given a Query Q, the probability of drawing Q from image I is

\[ P(Q \mid I) = \prod_{j=1}^{k} P(w_j \mid I) \]

• Or using the probabilistic annotation.

\[ P(Q \mid I) = \prod_{j=1}^{k} \sum_{J} P(J) P(w_j \mid J) \prod_{i=1}^{m} P(b_i \mid J) \]

• Rank images according to this probability.
Continuous Relevance Model

- Describe every image using two parallel vocabularies.
  - Visterms – from visual features.
  - Words – from annotations.
- Similar to Cross-Lingual Retrieval Problem (or Machine Translation).
- Learn the joint distribution of words and visterms.
- Annotate test images using this distribution.
Continuous Visterms

• Partition keyframe, features extracted from each rectangle

\{ r_1, r_2, \ldots, r_n \}

• Each feature vector contains visual information about color and texture (e.g. region color average, standard derivation and Gabor energy)
• Annotation:

\[ P(w \mid r_1...r_m) = \frac{P(w, r_1...r_m)}{\sum_w P(w, r_1...r_m)} \]

• Goal: to learn the joint distribution of a set of words and visterms \( P(w, r_1...r_m) \)

• Compute a mixture over all training samples:

\[ P(w, r_1...r_m) = \sum_J P(J)P(w \mid J) \prod_{i=1}^{m} P(r_i \mid J) \]

• “V. Lavrenko, R. Manmatha and J. Jeon”, *A Model for Learning the Semantics of Pictures, NIPS’03*
Continuous Relevance Model

- A generative model
- Words $w_j$ generated by an i.i.d. sample from a multinomial
- Features $r_i$ generated by a multi-variate (Gaussian) density
Estimation in CRM

- P(J) uniform.
- Kernel density estimate.

\[
P(r \mid J) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{\|r - r_i\|}{\beta}\right)
\]

- Smoothed likelihood estimates for word probabilities

\[
P(w \mid J) = (1 - \alpha_J) \frac{\#(w, J)}{|J|} + \alpha_J \frac{\#(w, T)}{|T|}
\]
Annotation Models

- **Co-occurrence Model (Mori et al):** Compute the co-occurrence of blobs and words.
  - Mean precision of **0.07**
- **Translation Model (Duygulu, Barnard, de Freitas and Forsyth):** Treat it as a problem of translating from the vocabulary of blobs to that of words. (Also try labeling regions). No ranked retrieval.
  - Mean precision of **0.14**
- **Cross Media Relevance Model (Jeon, Lavrenko, Manmatha):** Use a relevance (based language) model. Discrete model.
  - Mean precision of **0.33**
- **Correlation Latent Dirichlet Allocation (Blei and Jordan):** Model generates words and regions based on a latent factor. (Also try labeling regions).
  - Direct comparison on same dataset not available. Comparable results?
- **Continuous Relevance Model (Lavrenko, Manmatha, Jeon):** Relevance Model with continuous features.
  - Mean precision of **0.6**
Annotation Models

- Barnard et al, JMLR paper.
- Carbonetto, De Freitas, Barnard, Markov Random Fields
Problems/Improvements

• Are the two assumptions of CRM appropriate?
  – Segmentation
    • The quality of segmentation strongly affects the overall annotation performance.
    • Extremely expensive for large-scale dataset.
  – Multinomial Distribution of Words
    • Unsuitable when annotation length varies widely.
    • Spreads probability between words.
    • Multinomial distributions imply that annotations focus on presence rather than prominence.
Grid vs Segmentation

• Segmentation vs Rectangular Partition.
• Results - Rectangular Partition better than segmentation!
  – Model learned over many images. Segmentation over one image.
  – Similar result for a different model by Carbonetto and de Freitas.
Multinomial Model Appropriate?

• Keywords – at most one instance of each keyword per keyframe.

• Annotations vary widely in length from frame to frame
One example: $P(\text{face} \mid J)$?

*Face, female_face, outdoors, tree, news_subject*

- **Multinomial:** $P(\text{face} \mid J) = 1/5$
  
- **Bernoulli:** $P(\text{face} \mid J) = 1$

*face*

- **Normalized:** $P(\text{face} \mid J) = 1/5$
  
- **$P(\text{face} \mid J) = 1$**
Models

• Continuous Relevance Models.
  – Problems with varying length annotations.

• Bernoulli Models
  – Good annotation performance, poor retrieval.

• Normalized CRM
  – Annotation performance identical to Bernoulli Model.
  – Good retrieval.
Multiple Bernoulli Model (MBRM)

- For both CRM and MBRM

\[ P(w, r) = \sum_{J} P(w | J) \prod_{r \in r} P(r | J) \]

- Sample \( P(w | J) \) using a Bernoulli model

\[ P(w, r) = \sum_{J} P(J) \prod_{v \in w} P(v | J) \prod_{v \notin w} (1 - P(v | J)) \prod_{i=1}^{m} (P(r_i | J)) \]

- Compare with CRM model

\[ P(w, r) = \sum_{J} P(J) P(w | J) \prod_{i=1}^{m} (P(r_i | J)) \]

Bernoulli Model

• In theory we need to compute for every subset \( w \)

• Trick (proof long).
  – Assume vocabulary = \((w_1, w_2, w_3)\).
  – Then if we assume that the joint probability of \( w_1 \)
    and \( r \) is the probability of the union of all sets
    containing \( w_1 \).

  – Maybe used to simplify equation so that each \( w_i \)
    can be independently estimated
    using (Beta prior)
    \[
    P(w \mid J) = \frac{\mu \delta_{w,J} + N_w}{\mu + N}
    \]
Annotation Examples:

- Compute $P(w|J)$ for different $w$.
- Probabilistic Annotation:
  - Annotate the frame with every possible $w$ in the vocabulary with associated probabilities.
  - Useful for retrieval.

<table>
<thead>
<tr>
<th>CRM</th>
<th>NCRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>food</td>
<td>graphics_and_text</td>
</tr>
<tr>
<td>outdoors</td>
<td>text_overlay</td>
</tr>
<tr>
<td>monologue</td>
<td>graphics_and_text</td>
</tr>
<tr>
<td>outdoors</td>
<td>non_studio_setting</td>
</tr>
<tr>
<td>graphics_and_text</td>
<td>people_event</td>
</tr>
<tr>
<td>Text_overlay</td>
<td>face</td>
</tr>
<tr>
<td>non_studio_setting</td>
<td>male_face</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Results of Automatic Image Annotation

On Corel data set:

<table>
<thead>
<tr>
<th>Models</th>
<th>Translation</th>
<th>CRM</th>
<th>CRM-Rectangles</th>
<th>MBRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>#words with recall ≥ 0</td>
<td>49</td>
<td>107</td>
<td>119</td>
<td>122</td>
</tr>
<tr>
<td>Results on 49 best words, as in[4, 6]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean per-word Recall</td>
<td>0.34</td>
<td>0.70</td>
<td>0.75</td>
<td>0.78</td>
</tr>
<tr>
<td>Mean per-word Precision</td>
<td>0.20</td>
<td>0.59</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td>Results on all 260 words</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean per-word Recall</td>
<td>0.04</td>
<td>0.19</td>
<td>0.23</td>
<td>0.25</td>
</tr>
<tr>
<td>Mean per-word Precision</td>
<td>0.06</td>
<td>0.16</td>
<td>0.22</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Rectangles – Partition better than Segmentation! MBRM best.
Results of Automatic Image Annotation

- On Trec Video Dataset: A subset of NIST’s Video Trec dataset, consisting of 5200 key frames (1730 for test), with each key frame partitioned into 35 rectangles.

<table>
<thead>
<tr>
<th>Models</th>
<th>CRM-Rectangles</th>
<th>MBRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>#words with recall ≥ 0</td>
<td>79</td>
<td>83</td>
</tr>
</tbody>
</table>

Results on all 110 words.

<table>
<thead>
<tr>
<th></th>
<th>CRM-Rectangles</th>
<th>MBRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean per-word Recall</td>
<td>0.23</td>
<td>0.26</td>
</tr>
<tr>
<td>Mean per-word Precision</td>
<td>0.23</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Results on all words with recall ≥ 0

<table>
<thead>
<tr>
<th></th>
<th>CRM-Rectangles</th>
<th>MBRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean per-word Recall</td>
<td>0.32</td>
<td>0.34</td>
</tr>
<tr>
<td>Mean per-word Precision</td>
<td>0.32</td>
<td>0.35</td>
</tr>
</tbody>
</table>
Bernoulli Model

• Multiple-Bernoulli Relevance Model (MBRM)
  – Good Annotation performance.
  – With Bernoulli retrieval model
    • Poor retrieval performance.
  – With Multinominal Retrieval Model
    • Good retrieval performance but model not clean.

Normalized CRM

• Alternative: Normalized CRM
  – Pad annotations to a fixed length (using nulls) and use multinomial
  – Annotation performance same as Bernoulli Model (probabilities same up to a constant factor).
  – Multinomial retrieval model gives good performance.
Retrieval Results on video set

<table>
<thead>
<tr>
<th>Query length</th>
<th>1 word</th>
<th>2 words</th>
<th>3 words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of queries</td>
<td>107</td>
<td>431</td>
<td>402</td>
</tr>
<tr>
<td>Relevant images</td>
<td>6649</td>
<td>12553</td>
<td>11023</td>
</tr>
<tr>
<td>Precision at 5 retrieved key frames</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRM</td>
<td>0.36</td>
<td>0.33</td>
<td>0.42</td>
</tr>
<tr>
<td>normalized CRM</td>
<td>0.49</td>
<td>0.47</td>
<td>0.58</td>
</tr>
<tr>
<td>Mean Average Precision</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRM</td>
<td>0.26</td>
<td>0.19</td>
<td>0.25</td>
</tr>
<tr>
<td>normalized CRM</td>
<td>0.30</td>
<td>0.26</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Data: 6 hrs of Trec Video – 5200 keyframes.
Retrieval example

Query: Bill_Clinton
Retrieval Examples

Query: Basketball

CRM

Normalized-CRM

Query: Outdoors, Sky, Transportation

CRM

Normalized-CRM
Application to Retrieval of Handwritten Historical Manuscripts.
Handwritten Manuscripts.

• Retrieve historical documents written by a single author.

• Examples
  – George Washington’s papers
  – Isaac Newton’s manuscripts.
  – Joseph Grinnell’s (biologist) field notes at the Museum of Vertebrate Zoology in Berkeley.
  – Old Diaries of New England farmers for climatological studies.

• Existing state of the art.
  – Hand transcribe manuscripts --> text search engine.
  – Expensive and tedious.
  – Handwriting Recognition WER > 50%
To Captain Robert Stewart, at Winchester.

You are hereby required to take charge of the Recruits sent to Winchester by Captain Gist, whose Son you must order to proceed immediately and join his Father.

Captain Gist this day received one hundred pounds to recruit with, and the same Orders that were given to the other Officers on the 3rd. Instant. Ic.

Alexandria: December 8th, 1755.
Joseph Grinnell’s Field Notes
(courtesy Museum of Vertebrate Zoology, Berkeley)

Grinnell-1915
Yosemite Valley
May 31

Western Tanager (about 8; one I watched carrying nest material to site at end of out-merging branch of yellow pine; fully 25 feet up);

Yellow Warbler (fully 20, nearly all in deciduous trees along streams or sloughs);

Hermit Warbler (at least 6 singing males, in yellow pines, incense cedars and black oaks);

Western Bluebird.
Collection

• Have roughly 140,000 scanned pages of George Washington’s manuscripts from the Library of Congress.
• Scanned in 8 bit gray level at 300dpi.
• 300 GB lossless compressed.
• Scanned from microfilm
  – Quality not as good as scanning from original.
    • For example, boundary artifacts, noise etc.
  – Probably done for reasons of cost, fragility of manuscripts and security.
Preprocessing Example

Original

Cleaned and Cropped

Slant Corrected
Image Vocabulary

• Create a vocabulary of word image features by:
  – Segmenting image into words.
  – Compute features over each word.
  – Discretize features into bins to create a discrete vocabulary.
    • Use two overlapping discretizations.
Word Image Features

• Scalar features:
  – describe words in terms of a single number
  – examples: word length, word height, aspect ratio, …

• 1D profile features:
  – describe words with a series of numbers
  – example: projection profile
Feature Normalization

- 1D profile features vary in length from word to word
- But: model requires fixed-length description of images.
- Solution: use low order Fourier coefficients

DFT & Reconstruction.
Feature Discretization

- Have: feature vector with continuous-space entries.
- Need: description in terms of discrete alphabet.
- Solution: discretize feature vector along each dimension (bins of equal size)
  - Two different discretizations shifted by half a bin used.
  - 10 bins + 9 bins.
Other 1D Features

Original image

Alexandria

Upper Word Profiles

Lower Word Profile
Features

- Height $h$, width $w$, aspect ratio $w/h$, area $w\cdot h$
- No. of descenders
- Projection Profile – compute 4 real and 3 imaginary Fourier components.
- Upper Word Profile – again 7 Fourier components
- Lower Word Profile – also 7 Fourier components

- Total $5 + 3 \cdot 7 = 26$ features.
Word Image Vocabulary

- Discrete Image Vocabulary
- Size - 26 (features) * 19 (bins)
- Each of these is a visterm.
Retrieval

- Language Model Based.
- Retrieval of lines (or documents).
- Combine relevance models of all word images in a line:

\[
P(w | \text{line}) = \frac{1}{|\text{line}|} \sum_{I \in \text{line}} P(w | f_{I,1}, \ldots, f_{I,m})
\]
Retrieval

• Given a Query Q, the probability of drawing Q from line L (formed from I images) is

\[
P(Q \mid \text{line}) = \prod_{j=1}^{k} \left( \frac{1}{\text{line}} \sum_{I \in \text{line}} P(w_j \mid f_{I,1}, \ldots, f_{I,m}) \right)
\]

• Or using the probabilistic annotation.

\[
P(Q \mid \text{line}) = \prod_{j=1}^{k} \left( \frac{1}{\text{line}} \sum_{I \in \text{line}} \sum_{J} P(J)P(w_j \mid J) \prod_{i=1}^{m} P(f_{I,i} \mid J) \right)
\]

• Rank lines according to this probability.
Annotation

• Compute $P(w|I)$ for different $w$.
• Probabilistic Annotation: (PACMRM)
  – Annotate the word image with every possible $w$ in the vocabulary with associated probabilities.
  – Simulated Example.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>that</td>
<td>0.70</td>
</tr>
<tr>
<td>the</td>
<td>0.25</td>
</tr>
<tr>
<td>them</td>
<td>0.05</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Estimation and Experiments

- $P_i(w_j)$ and $P_i(f)$ are estimated from the training image at position $i$ using Maximum Likelihood
- Estimates are smoothed with Maximum Likelihood estimates over entire collection
- Line retrieval with 1- to 4-word queries on 20 pages (4773 word images, 657 lines)
- Used 90% for training, 10% for retrieval

<table>
<thead>
<tr>
<th>Num. query words</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Avg. Prec.</td>
<td>54%</td>
<td>63%</td>
<td>78%</td>
<td>89%</td>
</tr>
</tbody>
</table>
Retrieval Example

• Sample result, obtained with query “sergeant wilper fort cumberland”

rank:

1

room Sergeant Wilper has received from Fort Cumberland and this

2

at Fort Cumberland

3

Sergeant Wilper taking the receipt for the same for
• Performed retrieval on ~1000 page images
• Manually judged relevance at top 20 ranks
• 26 1-word queries
  – Retrieval of word images: Precision between 46% and 40%
  – Retrieval of page images: Precision between 58% and 52%
Beyond Handwritten Documents

• Retrieval techniques are applicable to general shape data sets
• Successful retrieval of general shapes possible with features similar to those used for handwritten words
• MPEG7: 70 categories, 87% mean avg. prec.
COIL-100 Shape Set

- 45 shape categories, extracted from color pictures of household objects on a turntable.

- Mean avg. prec. vs. # training examples

Extracted from
Conclusion

• Continuous features work better.
• Rectangular partition works better.
• Need to take into account annotation length
  – Normalized CRM significantly outperforms CRM for both annotation and retrieval.
• Application to handwriting very promising.
Future Work

• Larger Datasets
  – Workshop video trec dataset of 60 hrs of video.

• Other models
  – Have tried maxent and inference net models. Need to try more models.

• Speedup

• Try other models for handwriting.