A Statistical Approach to Retrieving Historical Manuscript Images without Recognition

Toni M. Rath, Victor Lavrenko and R. Manmatha*
Center for Intelligent Information Retrieval
University of Massachusetts
Amherst, MA 01002

Abstract

Handwritten historical document collections in libraries and other areas are often of interest to researchers, students or the general public. Convenient access to such corpora generally requires an index, which allows one to locate individual text units (pages, sentences, lines) that are relevant to a given query (usually provided as text). Several solutions are possible: manual annotation (very expensive), handwriting recognition (poor results) and word spotting - an image matching approach (computationally expensive).

In this work, we present a novel retrieval approach for historical document collections, which does not require recognition. We assume that word images can be described using a vocabulary of discretized word features. From a training set of labeled word images, we extract discrete feature vectors, and estimate the joint probability distribution of features and word labels. For a given feature vector (i.e. a word image), we can then calculate conditional probabilities for all labels in the training vocabulary. Experiments show that this relevance-based language model works very well with a mean average precision of 89% for 4-word queries on a subset of George Washington's manuscripts.

1. Introduction

Libraries are in the transition from offering strictly paper-based material to providing electronic versions of their collections. For simple access, multimedia information, such as audio, video or images, requires an index that allows one to retrieve data, which is relevant to a given query (usually provided as text).

At this time, historical manuscripts like the George Washington collection at the Library of Congress, are scanned page-by-page and then transcribed manually in order to build an index from the electronic transcript. This process is prohibitive for large collections, because of the extensive manual labor that is involved. Automatic approaches using handwriting recognition cannot be applied (see results in [17]), since the current technology for recognizing handwriting from images, i.e. *offline* recognition[1], has only been successful in domains with very limited lexicons and/or high redundancy, such as legal amount processing on checks and automatic mail sorting. An alternative approach called word spotting [15] involves clustering multiple instances of the same word image using image matching. Frequent clusters can then be used as index entries, because the contained images have links to the original pages. This technique is expensive - it requires $O(N^2)$ matching operations for N word images - and does not easily allow for text queries.

Here we present an approach to retrieving handwritten historical documents from a single author, using a relevance-based language model [10, 11]. Relevance models have been successfully used for both retrieval and cross-language retrieval of text documents and more recently for image annotation[9]. In their original form, these models capture the joint statistical occurrence pattern of words in two languages, which are used to describe a certain domain (e.g. a news event). By learning this dependency, one can identify texts of interest, i.e. *relevant* documents, in a foreign language by describing their content in a familiar language.

This paradigm can be used for the image domain, by describing images with words from a *feature vocabulary*, thus generating an "image description language". When the joint statistical occurrence pattern of words occurring in the image vocabulary and the image annotation vocabulary (i.e. word labels) are learned, one can perform tasks such as image retrieval using text queries, or automatic image annotation.

In this work, we model the occurrence pattern of words in two languages using the joint probability distribution

*This work was supported in part by the Center for Intelligent Information Retrieval and in part by the National Science Foundation under grant number IIS-9909073 and in part by SPAWARSYSCEN-SD grant numbers N66001-99-1-8912 and N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

[1] *Online* recognition, which records the pen position, etc. during writing, has been much more successful (see *TabletPCs*).

over the image description vocabulary and the annotation vocabulary. From a training set of annotated images of handwritten words, we learn this joint probability distribution and perform retrieval experiments with text queries on a test set. We describe word images using a vocabulary that is derived from a set of word shape features.

Our model differs from others in a number of respects. Unlike traditional handwriting recognition paradigms [12], our approach does not require perfect recognition for good retrieval. The work presented here is also related to models used for object recognition/image annotation and retrieval [6, 1, 3, 9]. However, those approaches were proposed for annotating/retrieving general-purpose colored photographs and primarily used color and texture as features. Here our focus is on word images, where such features are not available. Instead we use shape features to retrieve images. This model is not limited to handwritten document retrieval, but can be extended to many shape-related retrieval and annotation tasks in computer vision.

Using this relevance-based language model, we have conducted retrieval experiments on a set of 20 pages from the George Washington collection. The mean average precision scores we achieve lie in the range from 54% to 89% for queries using 1 to 4 words (respectively).

In the following section we discuss prior work in the field, followed by a detailed description of the relevance-based model in section 2. After explaining the features we use in our approach (section 3), we present line-retrieval results on the George Washington collection (section 4). Section 5 concludes the paper with an outlook on further work.

1.1. Previous Work

There are a number of approaches reported in the literature, which model the statistical co-occurrence patterns of image features and annotation words, in order to perform such diverse tasks as image annotation, object recognition and image retrieval. Mori et al. [13] estimate the likelihood of annotation terms appearing in a given image, by modeling the co-occurrence relationship between clustered feature vectors and annotation terms. Duygulu et al. [6] go one step further by actually annotating individual image regions (rather than producing sets of keywords for an image), which is in effect object class recognition. Their model uses the Expectation-Maximization (EM) algorithm to build a probability table that links "blobs" (clusters of image representations) to annotation terms. Barnard and Forsyth[1] extended Hofmann's Hierarchical Aspect Model for text and proposed a multi-modal approach to hierarchical clustering of images and words using EM. Blei and Jordan [3] extended their Latent Dirichlet Allocation (LDA) Model and proposed a Correlation LDA model, which relates words and images. They show only a few examples for labeling specific regions in an image, so it is difficult to tell how well this technique works.

The authors of [9] introduced the model used in this work for automatic image annotation and retrieval. With the same data and feature set, the results for image annotation were dramatically better than previous models - for example twice as good as the translation model [6]. This work extends that model to a different domain (word images in a noisy document environment), uses an improved feature representation and different attributes (shape). Shape has to be described by features that are very different from the previously utilized color and texture features. We test the model on a data set **with larger annotation vocabulary than previous experiments** and a feature vector discretization that preserves more detail than the clustering algorithms which are utilized in other approaches. In addition, our application (line retrieval) uses a new retrieval model formulation. Other authors have previously suggested document-retrieval systems that do not require recognition, but queries have to be issued in the form of examples in the image domain (e.g. see [16]). To our knowledge, our system is the first to allow retrieval without recognition using text queries.

All of the *image-to-word translation* approaches we are aware of, operate on image collections of good quality (e.g. the Corel image data base[6, 9]), which usually contain color and texture information. Color is known to be one of the most useful features for describing objects. Duygulu et al. [6], for example, use half of the entries in their feature vectors for color information. Images of handwritten words, on the other hand, do not generally contain color or texture information, and in the case of historical documents, the image quality is often greatly reduced.

The lack of other features makes shape a typical choice for offline handwriting recognition approaches. We make use of holistic word shape features that are justified by psychological studies of human reading[12], and which are widely used in the field [5, 15, 18]. However, these traditional features have varying sizes proportional to the length of the words and also tend to capture variations in word images, which are not always desirable. In order to capture the essential word shape and to get feature vectors of constant size, we use the low order DFT coefficients [7] of these features to represent a word image.

2. Model Formulation

Before explaining our model in detail, we would like to provide some intuition for it. Previous research in cross-lingual information retrieval has shown that co-occurrence probabilities of words in two languages (e.g. English and Chinese) can be effectively estimated from a parallel corpus, that is, a collection of document pairs, where each document is available in two languages. Reliable estimates can

be achieved even without any knowledge of the involved languages.

[11] describes how to capture the co-occurrence probabilities of an English word e and Chinese words c_i in the joint probability distribution $P(e, c_1 \dots c_k)$. By computing the conditional distribution

$$P(e|c_1 \dots c_k) = \frac{P(e, c_1 \dots c_k)}{P(c_1 \dots c_k)}$$

one can estimate the probability of occurrence of the term e in an English document given the occurrence of the terms c_i in a Chinese document (which talks about the same subject).

Here we apply this concept by describing images of words with an *image description language* in text. To do this, we extract features from the images and discretized them, which allows us to represent each word image in terms of a discrete vocabulary. From a set of labeled images of words we can then estimate the joint probability $P(w, f_1 \dots f_k)$, where w is a word label (the word "transcription") and the f_i are words from the image description language. Using the conditional density $P(w|f_1 \dots f_k)$ we can perform retrieval of handwritten text without recognition with high accuracy.

2.1. Model Estimation

Suppose we have a collection $\mathcal{C}$ of annotated manuscripts. We will model this collection as a sequence of random variables W_i, one for each word position i in $\mathcal{C}$. Each variable W_i takes on a dual representation: $W_i = \{h_i, w_i\}$, where h_i is the image of the handwritten form at position i in the collection and w_i is the corresponding transcription of the word. As we describe in the following section, we will represent the surface form h_i as a set of discrete features $f_{i,1} \dots f_{i,k}$ from some feature "vocabulary" $\mathcal{H}$. The transcription w_i is simply a word from the English vocabulary $\mathcal{V}$. Consequently, each random variable W_i takes values of the form $\{w_i, f_{i,1} \dots f_{i,k}\}$. In the remaining portions of this section we will discuss how we can estimate a probability distribution over the variables W_i.

We assume that for each position i in the collection there exists an underlying multinomial probability distribution $P_i(\cdot)$ over the union of the vocabularies $\mathcal{V}$ and $\mathcal{H}$. We further assume that actual values $\{w_i, f_{i,1} \dots f_{i,k}\}$ observed at position i represent an i.i.d. random sample drawn from $P_i(\cdot)$. In other words, the probability of a particular observation is given by:

$$P(W_i = w_i, f_{i,1} \dots f_{i,k}|I_i) = P_i(w_i|I_i) \prod_{j=1}^{k} P_i(f_{i,j}|I_i) \quad (1)$$

where I_i is the word image with representation W_i (we will omit the conditioning on I_i in the further derivations). Now suppose we are given an arbitrary observation $W = \{w, f_1 \dots f_k\}$, and would like to compute the probability of that observation appearing as a random sample somewhere in our corpus $\mathcal{C}$. Because the observation is not tied to any position, we have to estimate the probability as the expectation over every position i in our entire collection $\mathcal{C}$:

$$\begin{aligned} P(w, f_1 \dots f_k) &= E_i\left[P(W_i = w, f_1 \dots f_k)\right] \\ &= \frac{1}{|\mathcal{C}|} \sum_{i=1}^{|\mathcal{C}|} P_i(w) \prod_{j=1}^{k} P_i(f_j) \end{aligned} \quad (2)$$

Here $|\mathcal{C}|$ denotes the aggregate number of word positions in the collection. Equation (2) gives us a powerful formalism for performing automatic annotation and retrieval over handwritten documents.

2.2. Automatic Annotation and Retrieval of Manuscripts

Suppose we are given a training collection $\mathcal{C}$ of annotated manuscripts, and a target collection $\mathcal{T}$ where no annotations are provided. Given an arbitrary handwritten image h we can automatically compute its image vocabulary (≈feature) representation $f_1 \dots f_k$ and then use equation (2) to predict the words w which are likely to occur jointly with the features of h. These predictions would take the form of a conditional probability:

$$P(w|f_1 \dots f_k) = \frac{P(w, f_1 \dots f_k)}{\sum_{v \in \mathcal{V}} P(v, f_1 \dots f_k)} \quad (3)$$

This probability could be used directly to annotate new handwritten images with highly probable words. We provide a brief evaluation for this kind of annotation in section 4.2. However, if we are interested in retrieving sections of manuscripts we can make another use of equation (3).

Suppose we are given a user query $Q = q_1 \dots q_m$. We would like to retrieve sections $S \subset \mathcal{T}$ of the target collection that contain the query words. More generally, we would like to *rank* the sections S by the probability that they are relevant to Q. One of the most effective methods for ranked retrieval is based on the statistical language modeling framework [14]. In this framework, sections S of text are ranked by the probability that the query Q would be observed during i.i.d. random sampling of words from S:

$$P(Q|S) = \prod_{j=1}^{m} \hat{P}(q_j|S) \quad (4)$$

In text retrieval, estimating the probability $\hat{P}(q_j|S)$ is straightforward – we just count how many times the word q_j actually occurred in S, and then normalize and smooth the counts. When we are dealing with handwritten documents we do not know what words did or did not occur in a given section of text. However, we can use the conditional estimate provided by equation (3):

$$\hat{P}(q_j|S) = \frac{1}{|S|}\sum_{o=1}^{|S|} P(q_j|f_{o,1}\ldots f_{o,k}) \quad (5)$$

Here $|S|$ refers to the number of word-images in S, the index o goes over all positions in S, and $f_{o,1}\ldots f_{o,k}$ represent a set of features derived from the word image in position o. Combining equations (4) and (5) provides us with a complete system for handwriting retrieval.

2.3. Estimation Details

In this section we provide the estimation details necessary for a successful implementation of our model. In order to use equation (2) we need estimates for the multinomial models $P_i(\cdot)$ that underly every position i in the training collection $\mathcal{C}$. We estimate these probabilities via smoothed relative frequencies:

$$\begin{aligned}\hat{P}_i(x) = \; & \frac{\lambda}{1+k}\delta(x \in \{w_i, f_{i,1}\ldots f_{i,k}\}) \\ & + \frac{(1-\lambda)}{(1+k)|\mathcal{C}|}\sum_{l\in\mathcal{C}}\delta(x \in \{w_l, f_{l,1}\ldots f_{l,k}\}) \quad (6)\end{aligned}$$

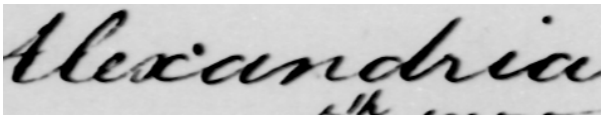
where $\delta(x \in \{w, f_1\ldots f_k\})$ is a set membership function, equal to one if and only if x is either w or one of the feature vocabulary terms $f_1\ldots f_k$. Parameter λ controls the degree of smoothing on the frequency estimate and can be tuned empirically.

3. Word Image Features

The mathematical formulation of our retrieval approach requires that word images are represented in terms of a feature vocabulary with discrete entries. This is achieved in a four-step process (see Figure 1 for an illustration):

1. extract single-valued/scalar features (e.g. image width) and variable-length features (e.g. projection profile) from the word image.
2. compute a fixed-length description of the variable-length features by using low-order Fourier coefficients.
3. combine scalar features and Fourier coefficients into a fixed-length feature vector.
4. discretize each feature dimension using a binning scheme, and produce one vocabulary term per bin.

In the following sections, the word image features are described, followed by an explanation of the vocabulary generation (i.e. feature discretization) process. These steps require image normalization which removes some of the variability that is present even in single-author handwriting. Figure 2 shows the results of background cleaning and slant/skew/baseline-correction on a typical input image.



(a) original image, as segmented from document,

(b) after cleaning and normalization.

Figure 2: Image cleaning and normalization.

3.1. Scalar Features

Each of the features described here, can be expressed by a scalar (a single number). Part of them have been used previously (see e.g. [15]) to quickly determine coarse similarity between word images. For a given image with tight bounding box (no extra space around word) we extract:

1. the height h,
2. the width w,
3. the aspect ratio w/h,
4. the area $w \cdot h$, and
5. an estimate for the number of descenders in the word, i.e. strokes below the baseline (e.g. lower part of 'p').

3.2. Variable-Length Features

The variable-length features we use give a much more detailed view of a word's shape than single-valued features can. All of the time series features below have been successfully used in a whole-word matching approach [15]. Each feature results from recording a single number *per image column* in the word image, thus creating a time series of the same length as the width of the image.

We generate three time series:

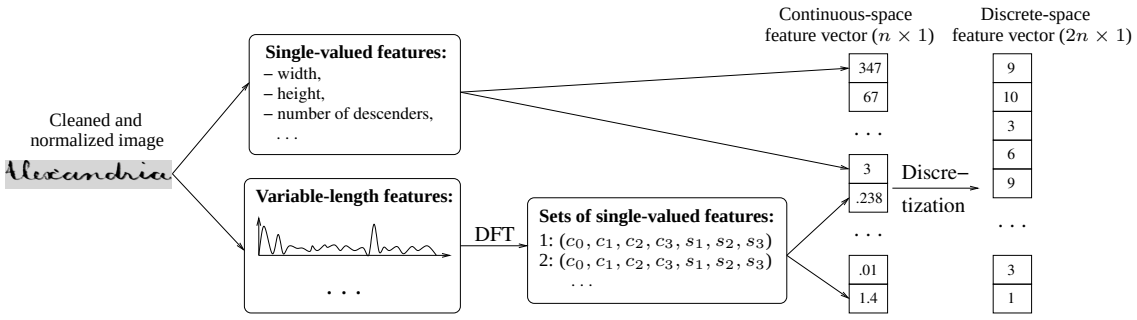
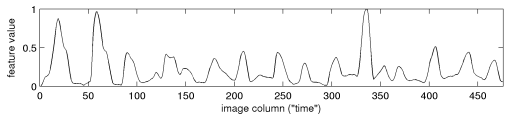
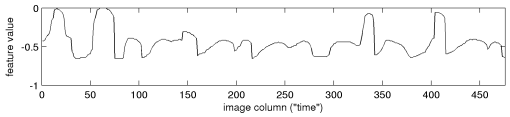


Figure 1: Feature generation process.



(a) projection profile time series,



(b) upper word profile time series.

Figure 3: Two of the three utilized time series features. Both features were directly extracted from image 2(b).

1. Projection Profile: each time series value is the sum of the pixel intensities in the corresponding image column (see Figure 3(a) for an example).

2. Upper Word Profile: each value is the distance from the top of the word's bounding box to the first "ink" pixel in the corresponding image column (see Figure 3(b)).

3. Lower Word Profile: same as upper word profile, but distance is measured from bottom of image bounding box.

All of these features are normalized so that their maximum range range is $[0..1]$. This ensures that features are comparable across words of different heights. The quality of these features strongly depends on good image normalization. For example, slant can affect the visibility of parts of words in terms of the word profile features (e.g. the 'l' leaning over the 'e' in Figure 2(a)).

While these time series features capture the shape of a word in great detail, they vary in length, and thus cannot be used in our framework, which requires fixed-length feature vectors. A time series can be adequately approximated by the lower-order coefficients of its Discrete Fourier Transform (DFT) [7]. The DFT representation also takes into account that images can have different lengths, since one period of the DFT basis functions is equal to the number of sample points.

We perform the DFT on the time series $\mathbf{s} = s_0 \ldots s_{n-1}$ to get its frequency-space representation $\mathbf{S} = S_0 ... S_{n-1}$:

$$S_k = \sum_{l=0}^{n-1} s_l \cdot \mathrm{e}^{-2\pi i l k/n}, \qquad 0 \le k \le n-1. \tag{7}$$

From the DFT representation we extract the first 4 real (cosine) components and 3 imaginary (sine) components[2] for use as single-valued features. Figure 4 shows a reproduction of the time series in Figure 3(a) using these features. For our purposes, this approximation suffices, since the goal is not to represent the original signal in detail, but rather to capture the global word shape with a small number of descriptors.

[2] For real-valued signals, the first imaginary component of the DFT is always 0.

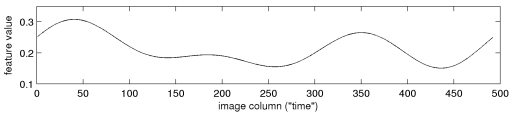


Figure 4: Projection profile time series from Figure 3(a), reconstructed using 4 lowest-order DFT coefficients.

3.3. Discretizing Features / Vocabulary

With all features combined, we have a continuous-space vector with $5 + 3 \cdot (4 + 3) = 26$ entries. Our relevance model requires us to represent all feature vectors in terms of a fixed-size "feature vocabulary". This can be achieved by discretizing each entry of the feature vector and creating one vocabulary term per discretization step. Then the vocabulary representation of a feature vector is comprised of the terms that correspond to the discretization steps of each vector entry.

We chose a discretization strategy that divides the observed range of each feature dimension in the training set into 10 parts (bins) of equal size. Since similar feature values could end up in neighboring bins if they fall into the region where two bins meet, we use a second set of 9 bins with shifted bin centers. Figure 5 illustrates this idea. This discretization process uses two vocabulary terms (e.g. `feature12_binset1_bin5` and `feature12_binset2_bin4`) to represent a feature vector entry. Per word image, this results in a representation that uses $26 \cdot 2 = 52$ feature vocabulary terms.

4. Experimental Evaluation

In this section we discuss the experiments we carried out to evaluate the proposed retrieval model. We will discuss two types of evaluation. First, we briefly look at the predictive capability of the annotation as outlined in section 2. We train a model on a small set of annotated manuscripts and evaluate how well the model was able to annotate each word in a held-out portion of the dataset. Then we turn to evaluating the model in the context of ranked retrieval.

The data set we used in training and evaluating our approach consists of 20 manually annotated pages from George Washington's handwritten letters. Segmenting this collection yielded a total of 4773 images, from which the majority contain exactly one word. An estimated 5-10% of the images contain segmentation errors of varying degrees: parts of words that have faded tend to get missed by the segmentation, and occasionally images contain 2 or more words or only a word fragment.

4.1. Evaluation Methodology

Our dataset comprises 4773 total word occurrences arranged on 657 lines. Because of the relatively small size of the dataset, all of our experiments use a 10-fold randomized cross-validation, where each time the data is split into a 90% training and 10% testing sets. Splitting was performed on a line level, since we chose lines to be our retrieval unit. Prior to any experiments, the manual annotations were reduced to the root form using the Krovetz morphological analyzer. This is a standard practice in Information Retrieval, it allows one to search for semantically similar variants of the same word. For our annotation experiments we use every word of the 4773-word vocabulary that occurs in both the training and the testing set. For retrieval experiments, we remove all function words, such as "of", "the", "and", etc. Furthermore, to simulate real queries users might pose to our system, we tested all possible combinations of 2, 3 and 4 words that occurred on the same line in the testing, but not necessarily in the training set. Function words were excluded from all of these combinations.

We use the standard evaluation methodology of Information Retrieval. In response to a given query, our model produces a ranking of all lines in the testing set. Out of these lines we consider only the ones that contain all query words to be relevant. The remaining lines are assumed to be non-relevant. Then for each line in the ranked list we compute *recall* and *precision*. Recall is defined as the number of relevant lines above (and including) the current line, divided by the total number of relevant lines for the current query. Similarly, precision is defined as number of above relevant lines divided by the rank of the current line. Recall is a measure of what percent of relevant lines we found, and precision suggests how many non-relevant lines we had to look at to achieve that recall. In our evaluation we use plots of precision vs. recall, averaged over all queries and all cross-validation repeats. We also report Mean Average Precision, which is an average of precision values at all recall points.

4.2. Discussion of Results

Figure 6 shows the performance of our model on the task of assigning word labels to handwritten images. We carried out two types of evaluation. In **position-level** evaluation, we generated a probability distribution $P(w|f_{i,1}\ldots f_{i,k})$ for

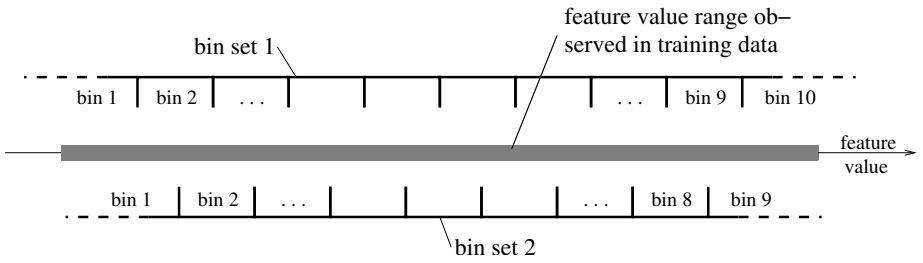


Figure 5: Binning scheme used in discretizing feature values (shown for one feature dimension).

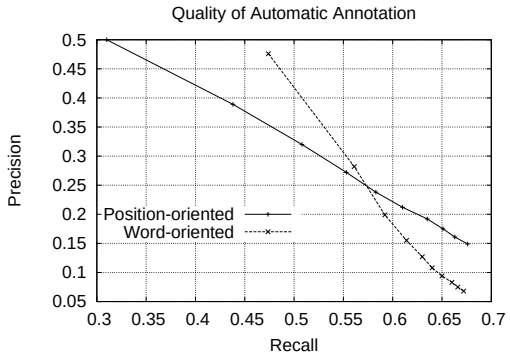


Figure 6: Performance on annotating word images with words.

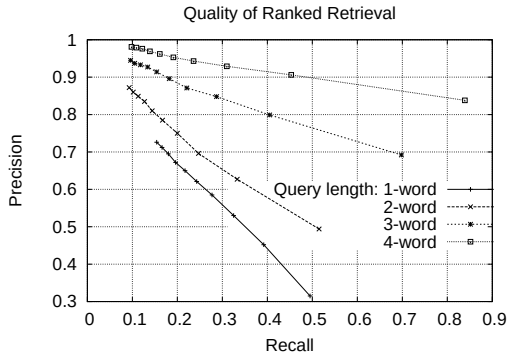


Figure 7: Performance on ranked retrieval with different query sizes.

every position i in the testing set. Then we looked for the rank of the correct word w in that distribution and averaged the resulting recall and precision over all positions. Since we did not exclude function words at this stage, position-level evaluation is strongly biased toward very common words such as "of", "the" etc. These words are generally not very interesting, so we carried out a **word-level** evaluation. Here for a given word w we look at the ranked list of all the positions i in the testing set, sorted in the decreasing order of $P(w|f_{i,1}\ldots f_{i,k})$. This is similar to running w as a query and retrieving all *positions* in which it could possibly occur. Recall and precision were calculated as discussed in the previous section.

From the graphs in Figure 6 we observe that our model performs quite well in annotation. For position-level annotation, we achieve 50% precision at rank 1, which means that for a given position i, half the time the word w with the highest conditional probability $P(w|f_{i,1}\ldots f_{i,k})$ is the correct one. Word-oriented evaluation also has close to 50% precision at rank 1, meaning that for a given word w the highest-ranked position i contains that word almost half the time. Mean Average Precision values are 54% and 52% for position-oriented and word-oriented evaluations respectively.

Now we turn our attention to using our model for the task of retrieving relevant portions of manuscripts. As discussed before, we created four sets of queries: 1, 2, 3 and 4 words in length, and will test them on retrieving line segments. Our experiments involve a total of 1950 single-word queries, 1939 word pairs, 1870 3-word and 1558 4-word queries over 657 lines. Figure 7 shows the recall-precision graphs. It is very encouraging to see that our model performs extremely well in this evaluation, reaching over 90% mean precision at rank 1. This is an exceptionally good result, showing that our model is nearly flawless when even such short queries are used. Mean average precision values were 54%, 63%, 78% and 89% for 1-, 2-, 3- and 4-word queries respectively. Figures 8, 9 and 10 show three retrieval results (two good and one bad) with variable-length queries. We have implemented a demo web-interface for our retrieval system, which can be found at **URL here!**.

5. Summary and Conclusion

We have presented a relevance-based language model for the retrieval of handwritten documents. Our model estimates the joint probability of occurrence of word annota-

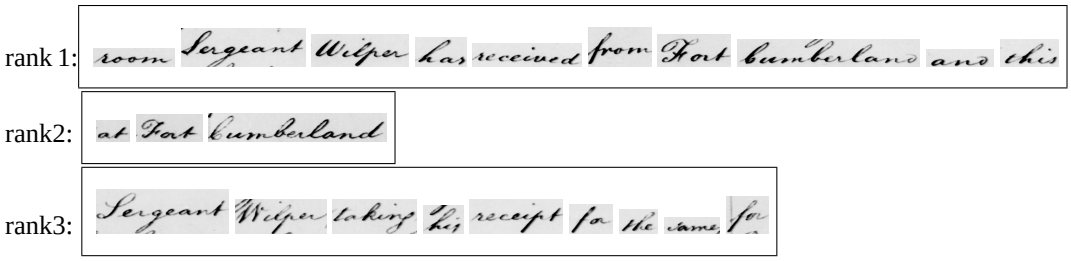


Figure 8: Retrieval result for the 4-word query “sergeant wilper fort cumberland” (one relevant line in collection).

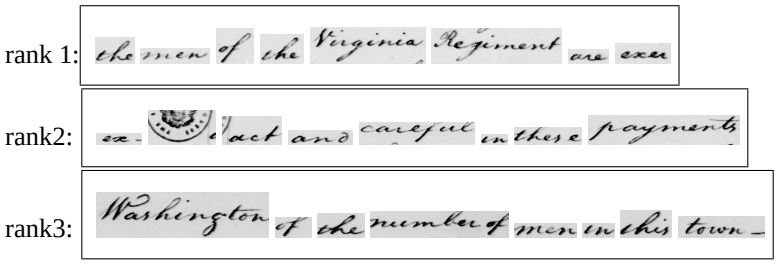


Figure 9: Retrieval result for the 3-word query “men virginia regiment” (one relevant line in collection).

Figure 10: Retrieval result for the 1-word query “sergeant” (three relevant lines in collection).

tions and feature vocabulary terms in order to perform probabilistic annotation of whole words and retrieval of lines of handwritten text. Our approach is the first to use shape-based features, and we presented appropriate shape representation, discretization and retrieval techniques. The results for line retrieval indicate performance at a level that is practical for real-world applications.

Future work will include a retrieval system for a larger collection, with page retrieval. Extending the collection could require more features in order to discriminate better between similar words. Lastly, we would also like to work on new retrieval models.

Acknowledgments

We would like to thank the Library of Congress for providing the scanned images of the George Washington collection.

References

[1] K. Barnard and D. Forsyth: *Learning the Semantics of Words and Pictures*. In: Proc. of the Int’l Conf. on Computer Vision, vol. 2, Vancouver, Canada, July 9-12, 2001, pp. 408-415.

[2] A. Berger, and J. Lafferty: *Information Retrieval as Statistical Translation*. In: Proc. of the 22nd Annual Int’l SIGIR Conf., 1999, pp. 222-229.

[3] D. M. Blei, and M. I. Jordan: *Modeling Annotated Data*. Technical Report UCB//CSD-02-1202, 2002.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan: *Latent Dirichlet Allocation*. Journal of Machine Learning Research **3** (2003) 993-1022.

[5] C.-H.Chen: *Lexicon-Driven Word Recognition*. In: Proc. of the 3rd Int'l Conf. on Document Analysis and Recognition 1995, Montréal, Canada, August 14-16, 1995, pp. 919-922.

[6] P. Duygulu, K. Barnard, N. de Freitas and D. Forsyth: *Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary*. In: Proc. of the 7th European Conf. on Computer Vision, Copenhagen, Denmark, May 27-June 2, 2002, vol. 4, pp. 97-112.

[7] C. Faloutsos: *Multimedia IR: Indexing and Searching*. In: Modern Information Retrieval, R. Baeza-Yates and B. Ribeiro-Neto; Addison-Wesley, Reading, MA, 1999.

[8] D. Hiemstra: *Using Language Models for Information Retrieval*. Ph.D. dissertation, University of Twente, Enschede, The Netherlands, 2001.

[9] J. Jeon, V. Lavrenko and R. Manmatha: *Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models*. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf., Toronto, Canada, July 28-August 1, 2003.

[10] V. Lavrenko and W. B. Croft: *Relevance-Based Language Models*. In: Proc. of the 24th Annual Int'l SIGIR Conf., New Orleans, LA, September 9-13, 2001, pp. 120-127.

[11] V. Lavrenko, M. Choquette and W. B. Croft: *Cross-Lingual Relevance Models*. In: Proc. of the 25th Annual Int'l SIGIR Conf., Tampere, Finland, August 11-15, 2002, pp. 175-182.

[12] S. Madhvanath and V. Govindaraju: *The Role of Holistic Paradigms in Handwritten Word Recognition*. Trans. on Pattern Analysis and Machine Intelligence **23**:2 (2001) 149-164.

[13] Y. Mori, H. Takahashi and R. Oka: *Image-to-Word Transformation Based on Dividing and Vector Quantizing Images with Words*. In: 1st Int'l Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM), Orlando, FL, October 30, 1999.

[14] J.M. Ponte and W.B. Croft: *A Language Modeling Approach to Information Retrieval*. In: Proc. of the 21st Annual Int'l SIGIR Conf., Melbourne, Australia, August 24-28, 1998, pp. 275-281.

[15] T. M. Rath, R. Manmatha: *Word Image Matching Using Dynamic Time Warping*. In: Proc. of the Conf. on Computer Vision and Pattern Recognition, Madison, WI, June 18-20, 2003, vol. 2, pp. 521-527.

[16] C. L. Tan, W. Huang and Y. Xu: *Imaged Document Text Retrieval without OCR*. Trans. on Pattern Analysis and Machine Intelligence **24**:6 (2002) 838-844.

[17] C. I. Tomai, B. Zhang and V. Govindaraju: *Transcript Mapping for Historic Handwritten Document Images*. In: Proc. of the 8th Int'l Workshop on Frontiers in Handwriting Recognition 2002, Niagara-on-the-Lake, ON, August 6-8, 2002, pp. 413-418.

[18] Ø. D. Trier, A. K. Jain and T. Taxt: *Feature Extraction Methods for Character Recognition - A Survey*. Pattern Recognition **29**:4 (1996) 641-662.

[19] C. Zhai: *Risk Minimization and Language Modeling in Text Retrieval*. Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, 2002.