A Search Engine for Historical Manuscript Images

Toni M. Rath, R. Manmatha and Victor Lavrenko*
Center for Intelligent Information Retrieval
University of Massachusetts
Amherst, MA 01003

ABSTRACT

Many museum and library archives are digitizing their large collections of handwritten historical manuscripts to enable public access to them. These collections are only available in image formats and require expensive manual annotation work for access to them. Current handwriting recognizers have word error rates in excess of 50% and therefore cannot be used for such material. We describe two statistical models for retrieval in large collections of handwritten manuscripts given a text query. Both use a set of transcribed page images to learn a joint probability distribution between features computed from word images and their transcriptions. The models can then be used to retrieve unlabeled images of handwritten documents given a text query. We show experiments with a training set of 100 transcribed pages and a test set of 987 handwritten page images from the George Washington collection. Experiments show that the precision at 20 documents is about 0.4 to 0.5 depending on the model. To the best of our knowledge, this is the first automatic retrieval system for historical manuscripts using text queries, without manual transcription of the original corpus.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object Recognition*

General Terms

Algorithms, Measurement, Experimentation

Keywords

Handwriting retrieval, historical manuscripts, relevance models

*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 number 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.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.

1. INTRODUCTION

This paper describes the construction of the first known automatic retrieval system for handwritten historical manuscripts. The system does not involve recognition, but instead uses a training set of transcribed manuscripts to automatically retrieve a test set of un-transcribed handwritten page images. The retrieval system is trained using an annotated set of 100 pages of George Washington's manuscripts and is used to query a dataset of 987 page images from the same collection.

Many libraries, museums and other organizations possess large quantities of handwritten document collections. Examples include the George Washington collection at the Library of Congress and Isaac Newton's manuscripts. Access to such material often requires that people travel to the library. Since the documents are fragile and valuable, access to originals is usually restricted to a few researchers. Many organizations intend to digitize such material and put it on the web to make it available to a wider audience. However, the information is usually in image format and the collections are large (often the correspondence of a single person may be on the order of 100,000 pages) which makes it difficult and inefficient to access. The current approach is to use metadata or indices, which are created manually in a tedious, labor intensive and expensive process. This makes automatic approaches to searching and accessing this material very attractive.

The obvious approach to this problem is to use handwriting recognition followed by a text search engine. While handwriting recognition has been successful in constrained domains like bank check processing or automatic mail sorting, the error rates for large vocabulary documents are high. Typical word error rates for large vocabulary handwriting recognition exceed 50% [10, 7] and may be even higher for historical documents which are often of poor quality [4].[1]

Here we look at an alternate approach to searching handwritten manuscripts using text queries without recognition. Given a set of transcribed pages, we have words in two different vocabularies - a vocabulary for describing word images and a vocabulary of annotation terms. The problem of retrieving a manuscript page in response to a text query is then similar to the problem of cross-lingual retrieval. We can therefore adapt models from cross-lingual retrieval to solve this problem. Our approach is to learn a statistical relevance

[1]The Tablet PC and PDAs use *online* handwriting recognition which is an easier problem - in online recognition additional information such as pen stroke, position and velocity information is available.

(based language) model by training on a transcribed set of (word for word) pages. Specifically, the relevance model allows us to automatically annotate word images in the test set with words in the lexicon and associated probabilities. A language model can then be used to retrieve word images given a text query. An initial test of this approach to do line based retrieval is described in [12] using a set of 19 training pages and 1 test page.

Given the small nature of the previous test set, there are a number of issues involved in scaling this to a more realistic data set. The specific contributions of this paper include experiments on a larger more realistic data set containing 100 training pages and 987 page test set (roughly 300,000 word images, 8GB of uncompressed data). We describe how to improve the performance of this model by reordering the results. We also describe a better performing second model called the direct retrieval model for this problem which involves no prior annotation of the test word images. Instead, a query relevance model and an image relevance model are computed and the Kullback-Leibler divergence between these is used to rank page images. Experiments using this model are performed for the same set of 987 test pages. Both these models have been used to build the first known retrieval system for handwritten (historical) manuscripts which does not involve humans transcribing the entire corpus.

The remainder of this article is organized as follows. After reviewing related work in the next section, we provide an overview of the system (section 3) and explain our word image representation in section 4 and the retrieval model in section 5. We describe the data collection effort in 6. The retrieval performance of the proposed model is evaluated on a large dataset in section 7. Section 8 concludes the paper with an outlook on future work.

2. RELATED WORK

While the problem of print recognition is essentially solved for standard fonts, the problem of handwriting recognition is still an open one. This seems surprising since they are essentially both images of characters. Techniques applied to print recognition cannot be assumed to apply to handwriting recognition without investigating them in the context of handwriting recognition. This is also true of other modalities. For example, although speech and music are both audio waves, techniques applied to speech retrieval cannot be assumed to music retrieval without specifically investigating them for music retrieval. Similar observations may be applied to recognizing handwriting and photographic images. Unfortunately, techniques in one area are not necessarily useful in another area.

The challenges in the historical manuscript domain are numerous: unlike the heavily researched domains of check processing and mail sorting, the vocabulary of historical documents is essentially unlimited. Even on high-quality document images, current state-of-the-art recognizers have word error rates in excess of 50% on large-vocabulary text [10] and these errors may be even larger for historical documents [4]. The age of the manuscripts causes problems such as stained paper, ink bleed and similar effects. Figure 1 is one of the nicer examples in our collection, showing typical problems such as faded ink and dirt marks. Further noise is also added by the image acquisition process; for example, the collection of George Washington's manuscripts we work with has been scanned from microfilm, rather than from the originals. Scanning artifacts, such as black borders around a page require careful processing in order to extract the text.

Word spotting is a different approach to indexing historical handwritten manuscripts and involves creating an index (similar to a book index) by matching word images with each other [8, 13]. A number of different matching techniques for this problem were investigated by Rath & Manmatha [13], including dynamic time warping of 1D features and shape context matching. While reasonable matching can be achieved, the techniques are very computationally expensive and it is currently impossible to build a system for even a small number of pages within a reasonable amount of time.

Our work here is more closely related to ideas in information retrieval, cross-lingual retrieval and automatic image annotation and retrieval. For example, in recent work Duygulu et al. [3] approached the problem of image annotation as similar to that of machine translation, where the task is to translate/map from an *image language* to an annotation language (e.g. English). Barnard & Forsyth used Hofmann's hierarchical aspect model [1] and Blei & Jordan used a latent Dirichlet allocation model [2] for this problem. Closest to this work in spirit is the work by Jeon et al. [5]. They adapt a relevance based language models for cross-lingual information retrieval [6]. The key idea is to describe image content with an *image description vocabulary*, and to model the occurrence of terms from the image and annotation vocabulary with a joint probability distribution. All of the above models were developed for general-purpose photograph datasets.

As mentioned above, although handwriting and general images (e.g. of tigers, grass and so on) are superficially similar in the sense of both being images, they are actually very different. Handwritten word images provide different challenges. Since words look much more alike than general images, it is much harder to distinguish between distinct word images while ensuring that similar word images are grouped together. Even humans have difficulties when reading some of the manuscripts when no context is available.

The feature sets for the above image models consist predominantly of color and texture, which have no discriminative power for handwritten documents. Instead we use features which characterize the shape of the word images [7]. The image vocabularies for photograph annotation models are created by first segmenting the image, then clustering the features computed over each region into "blobs" [3, 5]. Our datasets are much larger than those used for the general purpose photographs and hence clustering is not a viable option. Instead we use a binning technique to create a discrete vocabulary from the features. Another important distinction is that the image vocabulary in [3, 5] assumes that each region is described by one vocabulary word and that multiple regions in the image may map to the same annotation word. The vocabulary of features for handwriting instead assumes that a word image is generated by a number of different items from the feature vocabulary and that all these items, therefore, map to one feature word.

3. SYSTEM OVERVIEW

Our current system for the retrieval of handwritten manuscripts consists of a number of components. The following is a brief overview of the most important parts of this system and the necessary processing steps. First we collected a

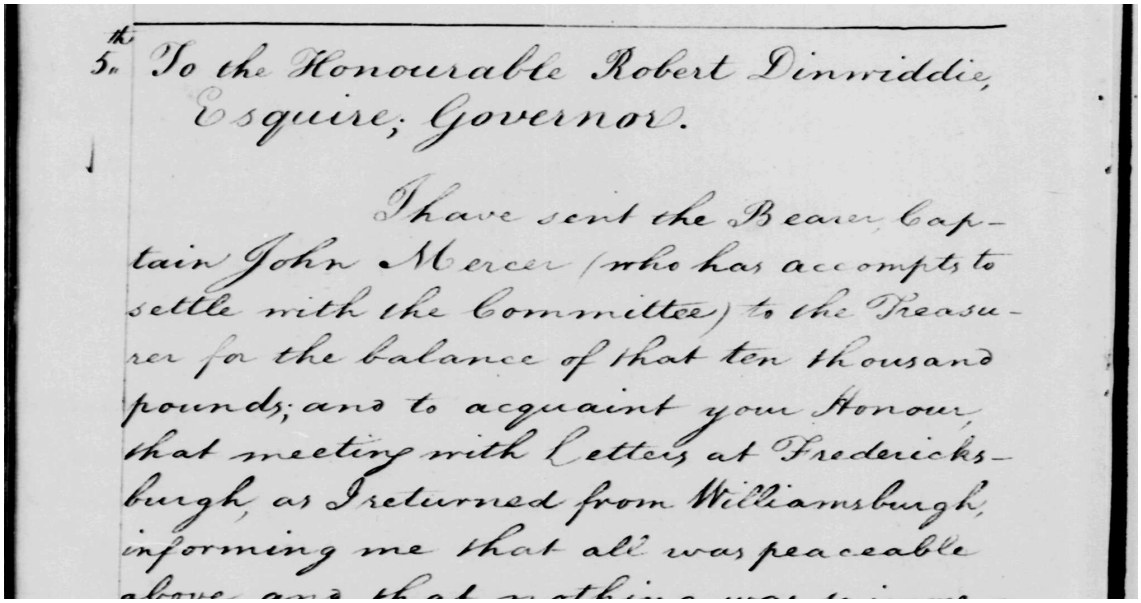


Figure 1: Part of a scanned document from the George Washington collection.

training set consisting of 100 pages with 24665 labeled word images. This allows us to learn the association of word images with the respective annotations. The testing set is a collection of 987 page images.

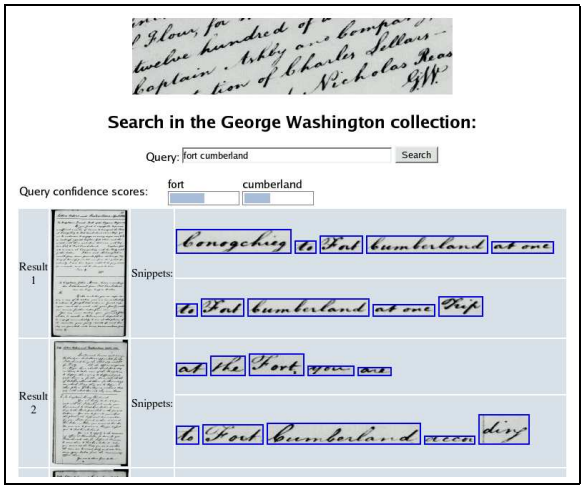


Figure 2: Screen shot of the web-based retrieval system interface (page image retrieval).

When the probabilistic annotation model is used, each word image in the testing set is annotated with every term in the annotation vocabulary and a corresponding probability. For page retrieval, these annotation probability distributions are averaged over all images that occur in a page, thus creating a language model of the page. During retrieval, these language models are used to rank word images or pages using query likelihood. The estimated language models are stored in an inverted list for quick access during querying. With this preprocessing, typical query times are less than one second.

In the case of direct retrieval, a query is used to estimate a distribution over the feature vocabulary that one would expect to observe jointly with the query. By comparing this distribution with a distribution of the feature vocabulary of each word image using Kullback-Leibler divergence, one may rank all word images in the testing set at query time. This real-time retrieval aspect of the model makes querying slower: typical query times on a 550MHz machine are about 40 to 45 seconds but less off-line processing is required.

We have implemented demonstration systems using both models that can be accessed at `http://ciir.cs.umass.edu/research/wordspotting`. Figure 2 shows a screen shot of the page retrieval engine: thumbnails of the retrieved pages are shown on the left-hand side, with snippets from the original page on the right-hand side. The snippets consist of the word images that are estimated to be the best match for the given query terms, with extra words around to provide some context. This is intended to help the user judge the relevance of an entire page by glancing at the snippets. Query *confidence scores*, are also displayed to give feedback about the result quality the user can expect from the provided query terms. These scores are computed from the number of training examples that are available for the query terms the user entered. Query terms, for which many training examples are available, generally yield better results.

4. WORD IMAGE REPRESENTATION

We use holistic shape features to represent images of handwritten words (see detailed description in [7]). The decision to use whole-word features is motivated by the large amount of noise in historical documents. Holistic features are computed over an entire word image, without having to break it down into smaller units (e.g. characters), thus avoiding the well-known segmentation problem, which is one of the main obstacles for handwriting recognition. By treating words as a unit, it is possible to avoid the problem of character segmentation entirely, which would be especially problematic in the highly noisy historical manuscript domain. The set of holistic word shape features used in our retrieval system

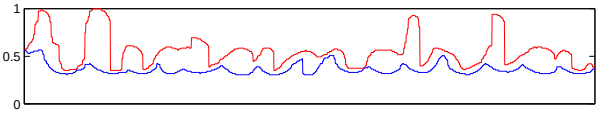
consists of two parts:

1. Simple shape features: these are very simple descriptors, such as the width and height of a word image. We use a total of 5 such features.

2. Fourier coefficients of profile features: detailed descriptions of a word's shape can be obtained with profile features, such as the upper and lower profiles (see Figure 3 for an illustration). Since these features are vectors that vary in length, depending on the width of a word image, they cannot be compared easily across word images. For this reason, we compute fixed-length representations of the profile features by using the first 7 coefficients from a Discrete Fourier Transform of each profile. Using the lower-order coefficients ensures that most of the energy of the original signal is preserved, while blocking out unwanted detail that is due to handwriting variations. From the projection, upper and lower profiles we obtain $3 \cdot 7 = 21$ features.

Together, each word image is represented by a 26-dimensional *continuous-space* feature vector.

(a) Cleaned and normalized word image,



(b) resulting upper and lower profile features displayed together.

Figure 3: Two of the three shape profile features.

Since we cast the document image retrieval problem as a cross-language problem, we need to represent each word image in terms of a *discrete* feature vocabulary (an "image language"). To do this, we divide the range of observed values in each feature dimension into 10 bins of equal size, and associate a unique feature vocabulary term with each bin. The representation of a particular feature vector is then given by the terms, which correspond to the bins that the feature vector values fall into. This process is repeated (see [12] for details) with 9 bins for each dimension, resulting in a representation consisting of $2 \cdot 26 = 52$ feature terms per word image, out of a feature vocabulary of size $(10+9) \cdot 26 = 494$. This is different from the discretization strategy used in [5], which assigned one feature term per image region cluster. In our case, this would correspond to creating clusters of word images, essentially making a classification decision. Our discretization technique preserves a greater level of detail which is required for handwritten images.

5. MODEL FORMULATION

The two models we discuss here both use some kind of "translation" or mapping. We assume that we have two different vocabularies. The first is the vocabulary of (English) words used in the lexicon. The second is a feature vocabulary for the word images. It is obtained by first computing a set of features over each word image and then discretizing them as explained in the previous section. The assumption is that each word image is described by a fixed dictionary of k ($k = 52$) discretized features from this feature vocabulary. In the first model, which we call the "Probabilistic Annotation" model, we take all the word images in the dataset and probabilistically map them to words, i.e. we assign to each word image in the test set all the words in the lexicon and associated probabilities. The model for mapping is learned using a training set of transcribed annotations. Given a text query, retrieval can be done with these probabilistic annotations in a language model based approach using query-likelihood ranking.

In the second model, which we call the "Direct Retrieval" model, we take each text query and compute the probability of generating a member of the feature vocabulary. We do this for every element in the feature vocabulary. In other words, we map the query to a distribution $P(f|Q)$ over the feature vocabulary. We can also obtain a distribution over the feature vocabulary for each word image I_i i.e. $P(f|I_i)$. By comparing these two distributions using the Kullback-Leibler divergence (this is a specific instance of the risk minimization framework in text retrieval) we can rank the word images. We now discuss these formulations in more detail.

5.1 Probabilistic Annotation Model

Assume we have a training set $\mathcal{T}$ of word images from a set of manuscript images which have been transcribed and aligned word for word. Each word is represented as an image and as an ASCII word. We will model this annotated training set using a set of random variables W_i where the i is an index for the word position in the training set $\mathcal{T}$.[2] Our aim is to treat this as similar to the cross-lingual retrieval problem and hence we can use the dual representation $W_i = \{h_i, w_i\}$, where h_i is the representation of the handwritten form at position i in the collection and w_i is the corresponding transcription of the word. We compute a set of discrete features $f_{i,1} \ldots f_{i,k}$ over each word and assume that the features characterize each word. The features come from the feature vocabulary $\mathcal{H}$ while the words w_i come from an English vocabulary $\mathcal{V}$. Hence, each W_i is of the form $\{w_i, f_{i,1} \ldots f_{i,k}\}$. All we need to do is to estimate a probability distribution over W_i.

Imagine an urn which contains all the possible features and words that are associated with word image I_i at position i. The features $f_1 \ldots f_k$ we observe can be assumed to have been obtained by taking k random samples. It follows from the urn model that the probabilities of observing $w, f_1 \ldots f_k$ are mutually independent once we pick a word image I_i with representation W_i.

More formally, we assume that at each position i (i.e. image I_i), the features and words are sampled from an underlying multinomial probability distribution $P(\cdot|I_i)$ over the union of the vocabularies $\mathcal{V}$ and $\mathcal{H}$ and that the actual observed values $\{w, f_1 \ldots f_k\}$ represent an i.i.d. random sample drawn from $P(\cdot|I_i)$. Then, the probability of a particular observation is given by:

[2]Our models do not make use of word position information, although this could certainly provide improved performance in the future. Here we only use i as an index for the words in our collection.

$$P(W_i = w, f_1 \ldots f_k | I_i) = P(w|I_i) \prod_{j=1}^{k} P(f_j|I_i) \quad (1)$$

Given an arbitrary observation $W = \{w, f_1 \ldots f_k\}$, we would like to compute the probability of that observation appearing as a random sample somewhere in our training set $\mathcal{T}$. We thus need to estimate the probability as the expectation over every position i in our entire collection $\mathcal{T}$:

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

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

Given a test collection $\mathcal{C}$ which has no transcriptions, we can use the above equation to probabilistically annotate every word image in the test collection. Given a word image, we first compute its image vocabulary (≈feature) representation $f_1 \ldots f_k$ and then use equation (2) to predict the words w which are likely to occur jointly with the features of h using:

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

We can now use the probabilities obtained from equation (3) to create a search engine for handwritten manuscripts using a language model retrieval approach.

5.1.1 Annotation and Retrieval of Manuscripts

Given a text query $Q = q_1 \ldots q_m$, we would like to retrieve pages $Pg \subset \mathcal{C}$ of the test collection that contain the query words. One of the most effective methods for ranked retrieval is based on the statistical language modeling framework [11]. In this framework, we rank a page Pg by the probability that the query Q would be observed during i.i.d. random sampling of words from Pg. That is,

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

In text retrieval, simple frequency counts may be used to estimate the probability $\hat{P}(q_j|Pg)$. This is somewhat more difficult with handwritten documents since we do not have these frequencies. We can instead use equation (3) as an estimator. That is,

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

Here $|Pg|$ refers to the number of word-images in Pg, the index o goes over all positions in Pg, and $f_{o,1} \ldots f_{o,k}$ represent a set of features derived from the word image in position o. We average the per-term annotation distributions in equation (5), because this makes the per-page annotation distribution converge to the Maximum Likelihood probability estimates of the term frequencies on page Pg for perfect estimates $P(q_j|f_{o,1} \ldots f_{o,k})$. Combining equations (4) and (5) provides us with a complete system for handwriting retrieval.

In order to use equation (2) we need estimates for the multinomial models $P(\cdot|I_i)$ that underly every position i in the training collection $\mathcal{T}$. We estimate these probabilities via smoothed relative frequencies obtained from the word image I_i and the entire training collection $\mathcal{T}$:

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

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 $f_1 \ldots f_k$. The parameter λ is tuned empirically. It controls the degree of smoothing on the frequency estimates obtained from I_i only, and from all images in the training collection $\mathcal{T}$.

5.2 Direct Retrieval

The approach adopted above assumes that we find a mapping from word image features to words (i.e. each word image is probabilistically converted to a word). However, we can go in the other direction and find a mapping for the query in terms of the word image features. Given a text query Q, we assume that the query is a random sample from a relevance model $P(\cdot|Q)$.

The probability of observing an element of the feature vocabulary f is given by:

$$P(f|Q) = \frac{P(f, Q)}{P(Q)} \quad (7)$$

As before, we can estimate the joint probability as an expectation over the training set. That is,

$$P(f, Q) = \sum_{W \in \mathcal{T}} P(W) P(f|W) P(Q|W) \quad (8)$$

where W is the word image and $\mathcal{T}$ the training set. The probabilities $P(f|W)$ and $P(Q|W)$ can be estimated using (6). The prior probabilities of a word image $P(W)$ are assumed to be uniform. We have to assume in this case that the query Q is a single word.

5.3 Reordering

Words which have never been seen in training are problematic. If a query word has never been seen in training we can detect it. The web interface simply handles such query words by informing the user that the word does not occur in the training set.

The probabilistic annotation approach suffers a different problem. Annotations are done before querying. Even if a given word image has never been seen in training, it is still annotated with the entire lexicon and associated probabilities. However, since this word image has never been seen in training, its ASCII representation is not part of the lexicon, which creates a problem. For example, assume that the word image representation of "Mandela" has not been seen in training. Assume that in feature space, the absolute distance of "Mandela" to the features representing other words is very large. While the absolute distance may be quite large, say two of these word features are closer than the

others. "Mandela" will then be incorrectly annotated with large probabilities with these words skewing the ranked list. The problem clearly occurs because our model does not look at absolute feature distance.

One solution to this problem is to take the top n ranked pages and reorder the results by absolute feature distance. Experiments (see experimental section) show that this approach actually improves results and solves a problem created by the lack of training examples.

6. DATA COLLECTION

No transcribed dataset of handwritten historical manuscripts is currently available in a form suitable for experiments in recognition or retrieval and so we had to create our own dataset. This was a very labor intensive process. In fact, to our knowledge there are only a handful of small handwriting databases and these are usually specially created for the handwriting recognition task by having people write them under constrained conditions [10]. They do not accurately represent how actual historical documents are written, preserved, scanned or archived.

The George Washington collection at the Library of Congress contains approximately 150,000 pages. The images were digitized from microfilm (for cost and other reasons) at 300dpi, 8bit grayscale from these pages. Scanning from microfilm introduces considerable noise and the quality of the scans varies. Washington employed secretaries, making this a multi-writer collection. This circumstance can be handled by acquiring training data that covers all observed writing styles.

The training set of 100 page images from the George Washington collection at the Library of Congress was first segmented automatically into words using the algorithm described in [9]. A java tool was created and the segmentations were hand corrected and also manually annotated by a set of undergraduates over a couple of months. We obtained 24665 words with individual labels. The whole process was made more difficult by the fact that handwritten historical manuscripts are really difficult to read. The vocabulary size of the training set (the number of distinct annotation terms) was 3087, a problem size that is considered large-vocabulary in the handwriting recognition field. Given the tedious and long nature of the task, the annotations are not perfect. While some errors were corrected, others remain. For example, the word "instruction" was sometimes misspelled as "instuction". While this creates a problem for the model (and lowers the average precision of the results) we decided to leave this in place in recognition of the fact that any similar data collection effort is likely to face the same problem. In some instances the manuscripts contain hyphenated words at the end of a line and we decided to transcribe them in the same manner.

The manual annotations were stemmed to the same root form using the Krovetz morphological analyzer allowing us to search for semantically similar variations of the same word.

7. EXPERIMENTAL EVALUATION

Our retrieval experiments were performed on a collection of 987 page images of George Washington's manuscripts. This is about 8 GB of raw image data (compare to TREC 1, 2 and 3 which together account for 4 GB). Processing this data requires much more space since the intermediate stages use floating point numbers rather than 8 bits to represent a pixel. The page images were also automatically segmented into words using a state of the art segmenter, yielding 234,754 word images. Since the segmentation is about 75% accurate, a number of words are incorrectly segmented. Given the size of the corpus it is impractical to correct these segmentations manually. We therefore use the actual (errorful) segmentations in our experiments. An incomplete partial "errorful" transcription of some of the manuscripts is available. However, these transcriptions often span multiple page images, thus making it impossible to accurately align a page image with its transcription in any straightforward manner.

Many of the processing stages may be computed off-line. A single pass of all the stages - segmentation, pre-processing, feature and model computation (for the probabilistic annotation scheme) - takes about 10 days to process 987 images using six 450 Mhz processors (the computations are parallelizeable). This assumes no mistakes have been made. Once computed, querying using the probabilistic annotation scheme is fast (less than 1 sec per query). The model computation for the direct retrieval technique is done at query time (on a single processor) and it takes about 45 seconds.

An interface allows querying to be done using text queries.

Query	# Ex.	Query	# Ex.
1755	79	Alexandria	39
arrived	29	clothes	11
Colonel	23	Cumberland	17
deliver	16	deserter	10
disobedience	2	fort	51
king	7	letter	113
lieutenant	18	march	11
order	143	provisions	13
receipt	9	received	52
recruit	40	regiment	31
Sergeant	16	vessel	7
Virginia	24	Washington	65
Williamsburgh	7	Winchester	33

Table 1: List of 1-word queries and the number of corresponding training examples.

Query	#Oc.	Query	#Oc.
10th decr.	3	1797. dear	4
1797 sir	3	20th decr.	3
ad quod	3	best endeavours	2
captain hogg	3	duly enlisted	4
esquire dr.	3	great esteem	8
g. washington	28	instructions. november	7
late letter	2	lieutenant george	3
mr. jno.	2	public meetings	2
royal notice.	2	safe hand	3
servt. g.	4	shall find	3
shall refer	2	shall suffer	3
stephen. sir	2	ten thousand	2
twelve hundred	3	utmost endeavours	4

Table 2: List of 2-word queries and the corresponding number of occurrences in the training set.

The probabilistic annotation model can handle multi-word queries while the direct retrieval approach is limited to 1 word queries at this time.

Evaluation is a difficult problem since queries and relevance judgements are not available for this task. Table 1 shows the selected 26 1-word queries and their counts (since all index terms are stemmed, this count includes all surface forms). The queries were a mixture of proper names, places as well as other nouns and also included a number in the form of a year. The 26 1-word queries were picked to be reasonably frequent words in the training set. Since the dataset is a collection of letters by George Washington over a period of time, the frequencies in the training set and test set differ and we do not know how often the same words occur in the test set (it is quite possible that some of them may not occur at all in the test set). The top 20 pages were manually judged for relevance by a group of graduate students.[3] This turned out to be an extremely tedious task given how difficult it is for humans (even motivated ones) to read handwritten historical documents. Using far more resources, TREC only judges text documents obtained by pooling a set of search engines: the top 100 documents are judged from each search engine for each query. We also tried to select 2-word queries (for the probabilistic annotation approach) by randomly selecting 26 queries from the 300 most frequent bigrams that did not contain any stop words. Bigrams that contained hyphenations were also discarded. Relevance judgements were difficult to do and it was noted that many 2-word queries did not occur in the top 20 documents in the test set. While it is possible that some of these are further down in the ranked list, it is more likely that the combinations do not occur in the test set because of the nature of the collection (letters over time). As Table 2 shows, most of the combinations occur only 2, 3 or 4 times in the training set. Since these words come from Washington's correspondence, many of these combinations may not occur at all in the test set (obvious examples include "20th decr", "10th decr"). Consequently, the results on 2-word queries probably underestimate the performance of the model and should be repeated with a more adequate query set. Without transcriptions of the test set, there is no easy solution to this problem. Note that all evaluations are performed using interpolated scores at ranks 1 to 20, averaged over all queries.

7.1 Word Image Retrieval

Here we evaluate the performance of the direct retrieval and probabilistic annotation models on word image retrieval. That is, the models were used to rank all word images in the entire test collection using the 1-word queries from Table 1. A ranked image was considered relevant if it has the same stem as the query. Figure 4 shows the interpolated precision scores obtained with the probabilistic annotation and direct retrieval model.

In the case of the probabilistic annotation model, a post-processing step was also performed in which the top 20 returned images were reordered ("reordered prob. annotation"): our feature vocabulary contains 494 entries, so we can also represent an image by a vector of that length, with exactly 52 1's, while the remaining entries are set to 0. The reordering was performed using the average dot product of the retrieved image at a particular rank and all training images for the given query. This serves as a measure of closeness between the retrieved images and the training examples for the given query. Figure 4 shows that this yields a much better ordering than the original probabilistic annotation, even better than the direct retrieval model for high ranks.

[3]It is not practical to manually annotate a dataset of 1000 pages (we estimate this would take about one man year). As ground truth data is not available, recall cannot be calculated.

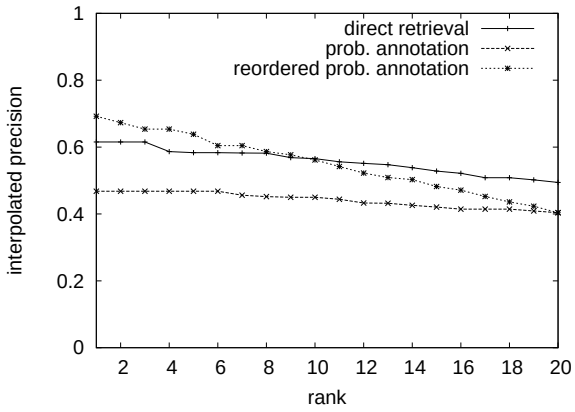


Figure 4: Retrieval of word images. Interpolated precision for the top 20 returned word images with the direct retrieval and probabilistic annotation models.

The performance of all three retrieval algorithms is quite good. However, the recall level remains unclear: the direct retrieval method did not retrieve any instances of *deserter* and *disobedience*, while the probabilistic annotation model found one *disobedience*. The low turnout of these and other terms (e.g. *vessel*) may be caused either by insufficient training examples (see Table 1) or the lack of relevant images in the testing collection.

7.2 Page Image Retrieval

Here we evaluate the performance of whole page retrieval. This evaluation can only be performed for the probabilistic annotation model, because the direct retrieval model allows us only to estimate feature distributions for *individual word images*, not page images. The retrieval was performed using query likelihood for the queries in Tables 1 and 2, using the language models estimated with the probabilistic annotation model.

Figure 5 shows the interpolated precision scores for the top 20 retrieved page images using 1-word queries. Again, the performance is quite good, even higher than in the single-word retrieval without reordering. Figure 6 shows the precision scores obtained with 2-word queries. Since 13 of the queries did not yield a single relevant page, the evaluation was also performed with those queries discarded ("nonrel. removed"). The results seem low, but we believe that a more thorough evaluation with ground truth data would yield better results. After all, the line retrieval experiments in [12] with 2-word queries suggest that mean average precision scores of about 63% are more realistic.

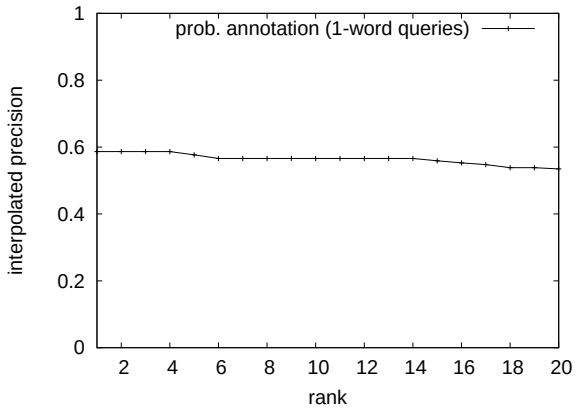


Figure 5: Retrieval of page images with 1-word queries. Interpolated precision for the top 20 returned ranks with the probabilistic annotation model.

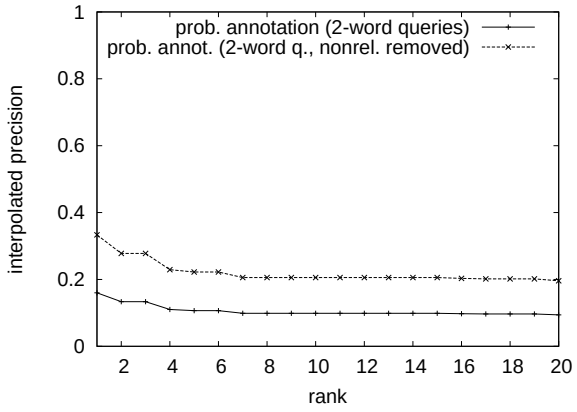


Figure 6: Retrieval of page images with 2-word queries. Interpolated precision for the top 20 returned ranks with the probabilistic annotation model.

8. CONCLUSIONS

This paper describes the first system ever built to retrieve historic handwritten manuscripts. Our results show that retrieval can be done even when recognition of handwriting remains a challenging task. In particular, adapting information retrieval models to this task has great promise. While our experiments show that adapting statistical relevance models produces good results, much remains to be done. Better models are needed. Large datasets can be handled either by using a cluster of processors, or by improving the efficiency of both the feature processing and the retrieval model stages. Finally, data collection and evaluation remain major challenges and automatic efforts to improve both are needed. In particular, the lack of training data requires attention. We are currently investigating synthetic training data as a possible solution.

9. ACKNOWLEDGMENTS

The Library of Congress kindly provided scanned images of the original manuscripts of George Washington. We would like to thank the reviewers for their valuable comments, which helped to improve the final version of this paper.

10. REFERENCES

[1] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proc. of the Int'l Conf. on Computer Vision*, volume 2, pages 408–415, Vancouver, Canada, July 9-12 2001.

[2] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proc. of the 26th Annual Int'l ACM SIGIR Conf.*, pages 127–134, Toronto, Canada, July 28-August 1 2003.

[3] 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*, volume 4, pages 97–112, Copenhagen, Denmark, May 27-June 2 2002.

[4] V. Govindaraju. Presentation. In *IEEE Workshop on Document Image Analysis for Libraries*, Palo Alto, CA, January 23-24 2004.

[5] 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.*, pages 119–126, Toronto, Canada, July 28-August 1 2003.

[6] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *Proc. of the 25th Annual Int'l SIGIR Conf.*, pages 175–182, Tampere, Finland, August 11-15 2002.

[7] V. Lavrenko, T. M. Rath, and R. Manmatha. Holistic word recognition for handwritten historical documents. In *Proc. of the Int'l Workshop on Document Image Analysis for Libraries*, pages 278–287, Palo Alto, CA, January 23-24 2004.

[8] R. Manmatha and W. B. Croft. Word spotting: Indexing handwritten manuscripts. In M. Maybury, editor, *Intelligent Multi-media Information Retrieval*, pages 43–64. AAAI/MIT Press, 1997.

[9] R. Manmatha and N. Srimal. Scale space technique for word segmentation in handwritten manuscripts. In *Proc. of the 2nd Int'l Conf. on Scale-Space Theories in Computer Vision*, pages 22–33, Corfu, Greece, September 26-27 1999.

[10] U.-V. Marti and H. Bunke. Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. *Int'l Journal of Pattern Recognition and Artifical Intelligence*, 15(1):65–90, 2001.

[11] J. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. of the 21st Annual Int'l ACM SIGIR Conf.*, pages 275–281, Melbourne, Australia, August 24-28 1998.

[12] T. M. Rath, V. Lavrenko, and R. Manmatha. Retrieving historical manuscripts using shape. Technical report, Center for Intelligent Information Retrieval, Univ. of Massachusetts Amherst, 2003.

[13] T. M. Rath and R. Manmatha. Word image matching using dynamic time warping. In *Proc. of the Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 521–527, Madison, WI, June 18-20 2003.