Searching in Document Images

C. V. Jawahar, Million Meshesha and A. Balasubramanian
Center for Visual Information Technology
International Institute of Information Technology
Gachibowli, Hyderabad – 500 019, India
jawahar@iiit.net

Abstract

Searching in scanned documents is an important problem in Digital Libraries. If OCRs are not available, the scanned images are inaccessible. In this paper, we demonstrate a searching procedure without an intermediate textual representation. We achieve effective retrieval from document databases by matching at word-level using image features. Word profiles, structural features and transform domain representations are employed for characterising the word images. A novel partial matching approach based on dynamic time warping (DTW) is proposed to take care of word form variations. With the new partial matching procedure, morphologically variant words become similar in image space. This is specially useful for grouping together similar words for indexing purpose. We extend our formulation for cross-lingual search with the help of transliteration.

Figure 1. Regional Mega Scanning Center at IIIT, Hyderabad, Engaged in Scanning and Archival of Books and Manuscripts

1. Introduction

With storage becoming cheaper, efforts are on the way to archive every page from printed or handwritten document collections. Large libraries are emerging for storage and delivery of such digital content. Figure 1 shows a picture from our scanning center, which is focusing on archiving books and manuscripts. There are twenty scanners in this center each one of them capable of scanning approximately 5000 pages in 8 hours. Much of the books scanned in such centers are in Indian languages, where robust OCRs are not yet available for converting these scanned images into textual form. Therefore, these books can only be searched based on the meta-data of the books, and not by the content within it. Indexing and retrieval from document image collections were studied by different researchers. Success of these procedures depends on the performance of the OCRs, which convert the document images into text. For Indian, African and many other oriental languages, we need alternate methods to retrieve relevant documents from the digital libraries containing scanned images.

This paper proposes an innovative solution for searching in document image collections. Given a textual query, we search for the similar words in image space. We also take care of word form variations without a textual representation. Results of the search are presented back to the user in a ranked manner based on the relevance to the query. We extend the method to cross-lingual retrieval by transliteration among Indian languages and a table-lookup translation for other languages.

The paper is organized as follows. Section 2 provides the necessary background and review of related literature. A word image is represented with the help of a set of features that addresses common artifacts. We propose a matching procedure to compare two word-images in a relevant feature space in Section 3. Partial matching procedure is explained in Section 4 along with the details of the retrieval. Performance of the proposed scheme is reported in Section 5, with associated discussions. Finally, concluding remarks are given in Section 6.

2. Related Work

Large quantities of all significant printed and handwritten documents are getting digitized and archived as images in digital libraries. Digital Library of India (DLI) [1] and the Universal Library (UL) [2] need special mention in this regard. They are digitizing all literary, artistic, and scientific works of mankind so as to make documents freely available to the global society. This entails the popularity of document images as an information source. When the textual representation is not available for these document images, search becomes practically impossible. An excellent review of the indexing and retrieval methods for documents is available in [3]. Doermann [3] focuses on the research issues when the document is represented as text with the help of an OCR. Possibly this is more relevant for languages like English. There is an increasing focus on indexing and retrieval from poor quality documents in recent years [4, 5, 6, 7]. For Indian Languages, Chaudhury et al. [4] proposed a method to access at word level by exploiting the structural characteristics of the Indian scripts. They employ geometric feature graphs for representation, and suffix trees for indexing the printed text. Annotation of documents by hyperlinking is provided for personalised access. In this work, we go beyond accessing the documents. We search and retrieve the document images based on their relevance to the query. The successful steps employed by a standard text search engine are replicated in a novel way in document images.

Word level matching has been attempted for printed [4], offline [5] and online [6] documents. They are useful for locating similar occurrences of the query word. There have been successful attempts on locating a specific word in a handwritten text by matching image features for historical documents [5]. Their approach segments the page into words and calculates equivalence classes by matching across words. Dynamic Time Warping (DTW) technique, that aligns and compares sets of features extracted from two images is used for matching. Our work extends the word spotting approach [5, 7] for searching queried words from printed document images of newspapers and books. Dynamic time warping based word-spotting algorithm for indexing and retrieval of on-line documents is also reported [6]. Features for matching words are computed from the constituent strokes. None of these matching schemes are designed to address partial matches, which is very important for addressing word-form variations for effective search.

Processing text documents is a well studied area for indexing and retrieval. Text documents undergo many preprocessing steps before getting indexed [3, 8]. During preprocessing, tokenization [8] is done for text documents. For document images this is equivalent to the word detection. Tokenization is followed by normalization. During normalization, suffixes are stripped off using the stemming process and stop words are filtered out for the representation of the document. A representation for the documents can be obtained by applying a serious of text processing techniques that extract relationships between terms, and their relative weights. For measuring the importance of the documents, these weights are used. Most systems weigh the term frequencies by the inverse document frequencies [8]. This is based on the realisation that words, which are common to all the documents, are not really important for the search. Once the documents are indexed, the resulting index vectors are used for searching and retrieval.

By observing the steps followed for text search, we find that there are three important components in developing an effective searching scheme for textual data at image level. They are: (a) a representation capable of taking care of common artifacts in printed document images, (b) a matching procedure to mimic the text processing algorithms employed in search engines, (c) an indexing scheme to compute the relevance of a document to a given query. All these three together help us to provide an effective search in a document image dataset.

3. Feature Extraction and Matching

Generic content-based image retrieval systems use colour, shape or/and texture features for characterising the content. In the case of document images, features can be more specific to the domain as they contain image-description of the textual content in it. Word images, particularly from newspapers and books, are of extremely poor quality. Common problems in such document database will have to be analysed before identifying the relevant features. Popular artifacts in printed document images include (a) Excessive dusty noise, (b) Large ink-blobs joining disjoint characters or components, (c) Vertical cuts due to folding of the paper, (d) Cuts in arbitrary direction due to paper quality or foreign material, (e) Degradation of printed text due to the poor quality of paper and ink, (f) Floating ink from facing pages. An effective representation of the word images will have to take care of these artifacts for successful indexing and retrieval. We found that three categories of features are needed to address these artifacts: word profiles, structural features and transform domain representations. The features could be a sequential representation of the word shape or a structural representation of the word image [9].

Profiles of the word provide a coarse way of representing a word image for matching. We employ profiles such as (a) upper word, (b) lower word, (c) projection, (d) density and (e) ink-to-background transition. Upper and lower word profiles capture the part of the outlining shape of a word, while projection and transition profiles capture the distribu-

tion of ink along one of the two dimensions in a word image. Structural features of the words are also used to match two words based on image similarities. Normalised moments, such as first-order moments: (f) M_{00}, (g) M_{01}; central moment: (h) CM_{01}; and statistical moments: (i) mean, (j) standard deviation, (k) skew are employed in this work for describing the structure of the word. For artifacts like pepper and salt noise, structural features are found to be very useful. A compact representation of a series of observations (such as profiles) can be derived using Fourier Transform. Fewer set of coefficients are enough to represent a word robustly in a transformed domain, and these coefficients are compared at a coarse level for matching. We use five predominant Fourier coefficients ((l) - (p)) for the word representation. All together we employ 16 features (a-p) describing various properties of the word. We have observed that these representations work well for popular fonts, but only with limited success for fancy fonts. This representation is found to be sufficient for printed document images.

Document images are preprocessed offline to threshold, skew-correct, remove the noise and thereafter to segment into words. Then the features are extracted for individual words. They are also normalized such that the word representations become insensitive to variations in size, font and various degradations popularly present in the text documents.

For proper search, we need to identify the similar words. Distance or dissimilarity between words is computed using the features discussed above. In this paper, we find the similarity of words based on two components: (a) A sequence alignment score computed using a Dynamic Time Warping (DTW) procedure. (b) Structural similarity of word images by comparing the shapes.

Dynamic Time Warping is a dynamic programming based procedure [5] to align two sequences. This can also provide a similarity measure. This is a popular matching procedure in speech analysis and recognition [10].

Let the word images (say their profiles) are represented as a sequence of vectors $\mathcal{F} = \mathbf{F_1}, \mathbf{F_2}, \ldots \mathbf{F_M}$ and $\mathcal{G} = \mathbf{G_1}, \mathbf{G_2}, \ldots, \mathbf{G_N}$. The DTW-cost between these two sequences is $D(M, N)$, which is calculated using dynamic programming as:

$$D(i,j) = \min \begin{cases} D(i-1, j-1) \\ D(i, j-1) \\ D(i-1, j) \end{cases} + d(i,j)$$

where, $d(i,j)$ is the local distance cost in aligning the ith element of $\mathbf{F}$ with jth element of $\mathbf{G}$ and is computed using a simple squared Euclidean distance:

$$d(i,j) = \sum_k \left(F_i^k - G_j^k\right)^2$$

where F_i^k is the k^{th} feature of $\mathbf{F_i}$ and G_i^k is the k^{th} feature of $\mathbf{G_i}$.

Employing $D(i, j-1), D(i-1, j)$ and $D(i-1, j-1)$ in the calculation of $D(i,j)$ enforces a local continuity constraint, which ensures no samples are left out in time warping. Score for matching the two sequences $\mathcal{F}$ and $\mathcal{G}$ is considered as $D(M, N)$, where M and N are the lengths of the two sequences.

We also imposed a global constraint using Sakoe-Chiba band to ensure the maximum steepness or flatness of the DTW path [10]. As shown in Figure 2 (a), Sakoa-Chiba band restricts the matching space (the space in gray colour). This limits DTW computation time to a great extent. Matching between two words therefore takes place only within this space. The warping path for the words 'direct' and 'redirected' (i.e. the path at the middle) shows this fact. In this way, the Sakoe-Chiba band constraint ensures this path stays close to the diagonal of the matrix which contains the $D(i,j)$. As a result, warpings that align a small portion in one sequence to a large portion in the other are avoided.

Thus given two word images, we are able to compute the distance (or dissimilarity) between words with the help of a dynamic time warping procedure. The dissimilarity score is computed using a comprehensive set of features, which is selected, specially, to take care of popular artifacts in printed text documents. Our representation is more comprehensive and robust compared to the other word matching algorithms.

4. Partial Matching and Retrieval

The simple matching procedure described above may be efficient for spotting or matching word-images. However the indexing process for a good search engine is more involved than the simple word-level matching. A word, with similar meanings, usually appears in various forms. This requires a method to conflate such different morphological variants of a word to a common stem/root [8]. Stemming removes word variations such that words with the same underlying stem become similar. In the text domain, variation of word forms may obey the language rules. Text search engines use this information while indexing. However for text-image indexing process, which we are interested in, this information is not directly usable.

We take care of simple, but very popular, word form variations taking place at the beginning and end. For this, once sequences are matched, we backtrack the optimal cost path. During the backtracking phase, if the dissimilarity in words is concentrated at the end, or in the beginning, they are deemphasized. For instance, for a query "direct", the matching scores of the words "directed" and "redirected" are only the matching of the 6 characters, 'd-i-r-e-c-t', of both words. Once an optimal sub-path is identified, a normalized cost corresponding to this segment is considered as the matching score for the pair of words.

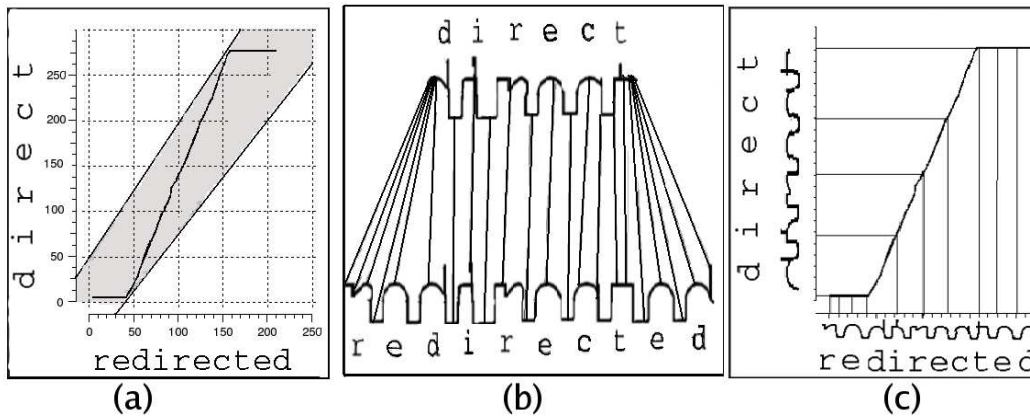

Figure 2. Plots Demonstrating the Dynamic Time Warping Procedure for Matching Words (a) Sakao-Chiba Band Constraint (b) Alignment of Two Words: 'Direct' and 'redirected' (c) Profiles of Two Words and the Optimal Matching Path

The optimal warping path is generated by backtracking the DTW minimal score in the matching space. First, as shown in Figure 2 (b), extracted features (here we use upper word profile for demonstration) of the two words 'direct' and 'redirected' are aligned using dynamic time warping algorithm. It is observed that features of these words are matched in such a way that elements of 're' at the beginning as well as 'ed' at the end of the word 'redirected' get matched with characters 'd' and 't' of the word "direct". This is identified and removed while backtracking. A simple plot for matching profiles of the two words is shown in Figure 2 (c).

It can be observed that for word variants the DTW path deviates from the diagonal line in the horizontal or vertical direction from the beginning or end of the path, which results in an increase in the matching cost. In the example figure (Figure 2 (c)), the path deviates from the diagonal line at the two extreme ends. This happens during matching the two words, that is, the root word (direct) and its variant (redirected). Profiles of the extra characters ('re' and 'ed') have minimal contribution to the matching score. With the reduction in the dissimilarity score of a variant word (against its root word), we find that a large set of words get grouped together. Table 1 shows dissimilarity of a set of example words with the word 'direct'. Note that the first five words have very low dissimilarity, while any arbitrary word of similar length can have very high dissimilarity. With our partial matching technique, morphologically similar words have smaller dissimilarity than arbitrary words. Hence during searching, say for the example word 'direct' our system not only searches strictly for the occurrence of word 'direct' but also for all occurrence of its variants that are coming in the predefined range of acceptance. Detection of similar words for indexing is very important to increase recall and precision of the retrieval system. The use of the net cost of the new partial matching algorithm as a distance measure is helpful to group together large number of similar words from the document image database. This can make it easier to determine the relevance of documents to a given query. Because once similar words are grouped together, we analyse each group for its relevance and determine documents with highest occurrence of similar words. This is also important from the users perspective as they are searching for relevant documents to their queries.

Word	Dissimilarity
direct	0.000000
directed	0.011650
redirect	0.019092
indirect	0.023002
redirected	0.051173
detail	0.171924
secure	0.185319
belief	0.194533
happen	0.249376
knowledge	0.291347

Table 1. Dissimilarity Cost of a Set of Example Words, Both Variants and Arbitrary Words, with the Word 'direct'

Stop words are words that are common in all the documents. These words are less meaningful to characterize any of the document. Flagging such words from the list of word representations can have a significant impact on the search process. Some of the identified stop words include: a, by, is, to, the, of, in, and, are and class. They are taken care of while searching.

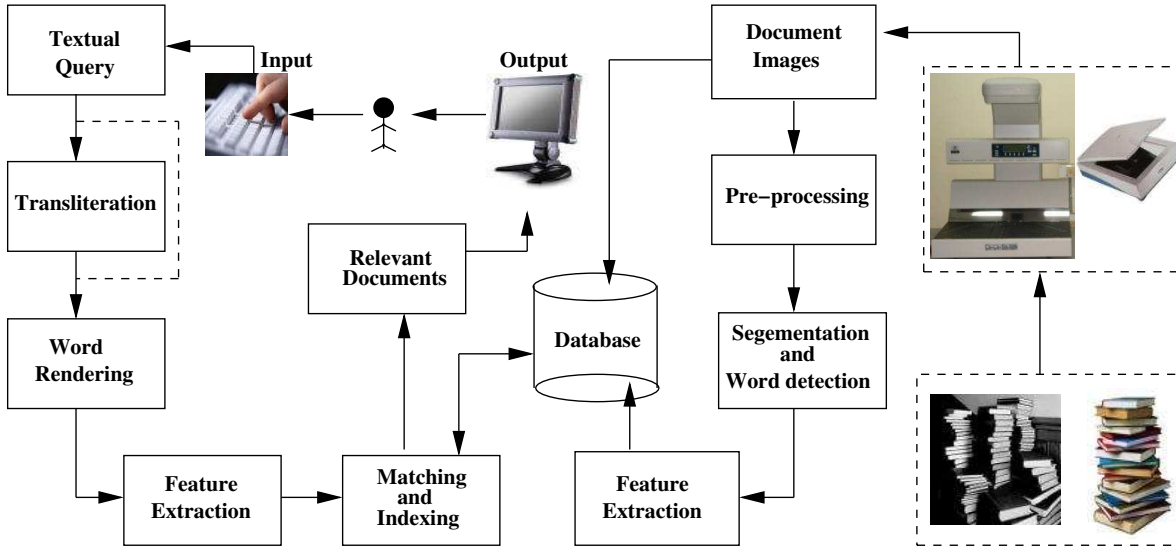

Figure 3. Conceptual Diagram of the Searching Procedure

Language	Data Set	Test*	Prec.	Recall
English	2507	15	95.89	97.69
Hindi	3354	14	92.67	93.71
Amharic	2547	14	94.51	96.63

[1]*Number of words used for testing

Table 2. Performance of the Proposed Approach on Three Data Sets in English, Hindi and Amharic. Percentages of Precision and Recall are Reported for Some Test Words

5. Results and Discussion

We have a prototype system for searching in document images. A conceptual block diagram is shown in Figure 3. Books are scanned and processed offline to index the document pages. System accepts textual query from users. The textual query is first converted to image by rendering, features are extracted from these images and then search is carried out for retrieval of relevant documents. Results of the search are document images containing queried word sorted based on their relevance to the query. To facilitate searching, scanned document images are preprocessed, features of representative word images are extracted and then indexed.

We evaluated the performance of the prototype system on data sets containing documents from English, Hindi and Amharic. Amharic is the official language of Ethiopia. We have built a corpus of Amharic document images by scanning pages from the newspaper "Addis Zemen". Hindi and English pages are taken from digital library of India collections. The proposed method is extensively tested on all these data sets. Some of the sample words retrieved are shown in Figure 4 (a). The text in double-box is the query word and the images next to this word are the retrieved ones. Performance of the word level access is computed on the document image databases of size approximately 2500 words. Around 15 query words are used for testing. During selection of query words, priority is given to words with many variants. For these words precision and recall are tabulated in Table 2. Percentage of retrieved words which are relevant, is represented as precision, while recall is computed as the percentage of the relevant words which are retrieved from the entire collection. It is found that both precision and recall are close to 95% for all the languages. High precision and recall is registered in our experiment. This may be because of the limited dataset (with similar fonts, style and size) we experimented with. We are working a comprehensive test on real large datasets. Preliminary experiments shows that the results are promising. We are also working on improving the system architecture in Figure 3 to make it scalable.

Cross-lingual Search Our searching scheme can also do cross-lingual retrieval. Since Indian scripts share a common alphabet, we can transliterate the words across languages. This helps us to search in multiple languages at the same time. We also have a dictionary-based translation for cross-lingual retrieval. We tried searching in scanned documents from 'Bhagavat-Gita'. Pages from this book contain English and Devanagari text. These pages are of poor quality. We tried searching for the occurrences 'arjuna'. It fetched pages which contain 'Arjuna' in both English and Devangari. Sample results are shown in Figure 4 (b). The system also retrieved words containing variants of the word 'arjuna'.

Figure 4. Results: (a) Some Sample Word Images Retrieved for the Queries Given in Special Boxes. Examples are from English, Amharic and Hindi Languages. The Proposed Approach Takes Care of Variations in Word-form, Size, Font and Style Successfully. (b) Example Result for Cross-lingual Search from Bhagavat Gita Pages. Similar Words are Retrieved Both in Devanagari and English.

6. Comments and Conclusions

In this paper, we have proposed a framework for content-based access to a collection of document images. A novel matching procedure is proposed to take care of a class of word form variations in image matching. A sample search engine employing the features described in this paper will be demonstrated with a set of sample books in the near future. We are presently working on: (a) Selection of relevant features for specific scripts/languages. (b) Developing a sophisticated data structure and system architecture for better efficiency and real-time retrieval. We presently use a simple plain file system for the indexing and storage. (c) Learning generative models for the degradation of printed and handwritten text. This will be helpful to build improved templates for matching and retrieval. (d) Improving the indexing process by mimicking the text-search engines. (e) Building a ground truth-ed word level data set in multiple languages for evaluation of the performance on a large data set. This will act as a test bed for feature selection as well as precision - recall computations.

Acknowledgment

This work was partially supported by the MCIT, Govt. of India for Digital Libraries Activities.

References

[1] Digital Library of India, "http://www.dli.ernet.in,"

[2] The Universal Library, "http://www.ulib.org,"

[3] D. Doermann, "The Indexing and Retrieval of Document Images: A Survey," *Computer Vision and Image Understanding*, vol. 70, no. 3, pp. 287–298, 1998.

[4] S. Chaudhury, G. Sethi, A. Vyas, and G. Harit, "Devising interactive access techniques for Indian language document images," in *Proc. International Conference on Document Analysis and Recognition*, pp. 885–889, 2003.

[5] T. Rath and R. Manmatha, "Features for word spotting in historical manuscripts," in *Proc. International Conference on Document Analysis and Recognition*, pp. 218–222, 2003.

[6] A.K. Jain and Anoop M. Namboodiri, "Indexing and retrieval of on-line handwritten documents," in *Proc. International Conference on Document Analysis and Recognition*, pp. 655–659, 2003.

[7] T. Rath and R. Manmatha, "Word image matching using dynamic time warping," in *Proceeding of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 521–527, 2003.

[8] R. Korfhage, *Information Storage and Retrieval*. New York: John Willey, 1997.

[9] O. D. Trier, A. K. Jain, and T. Taxt, "Feature Extraction Methods for Character Recognition: A Survey," *Pattern Recognition*, vol. 29, pp. 641–662, 1996.

[10] H. Sakoe and S. Chiba, "Dynamic Programming Optimization for Spoken Word Recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 26, pp. 623–625, 1980.