Retrieval from Document Image Collections

A. Balasubramanian, Million Meshesha, and C.V. Jawahar

Centre for Visual Information Technology,
International Institute of Information Technology,
Hyderabad - 500 032, India
jawahar@iiit.ac.in

Abstract. This paper presents a system for retrieval of relevant documents from large document image collections. We achieve effective search and retrieval from a large collection of printed document images by matching image features at word-level. For representations of the words, profile-based and shape-based features are employed. A novel DTW-based partial matching scheme is employed to take care of morphologically variant words. This is useful for grouping together similar words during the indexing process. The system supports cross-lingual search using OM-Trans transliteration and a dictionary-based approach. System-level issues for retrieval (eg. scalability, effective delivery etc.) are addressed in this paper.

1 Introduction

Large digital libraries, such as Digital Library of India (DLI) [1] are emerging for archiving large collection of printed and handwritten documents. The DLI aims at digitizing all literary, artistic, and scientific works of mankind so as to create better access to traditional materials, easier preservation, and make documents freely accessible to the global society. More than 25 scanning centers all over India are working on digitization of books and manuscripts. The mega scanning center we have, has around fifty scanners, each one of them capable lof scanning approximately 5000 pages in 8 hours. As on September 2005, close to 100 thousand books with 25 million pages were digitized and made available online by DLI (*http://dli.iiit.ac.in*) as document images.

Building an effective access to these document images requires designing a mechanism for effective search and retrieval of textual data from document image collections. Document image indexing and retrieval were studied with limited scope in literature [2]. Success of these procedures mainly depends on the performance of the OCRs, which convert the document images into text. Much of the data in DLI are in Indian languages. Searching in these document image collections based on content, is not presently possible. This is because OCRs are not yet able to successfully recognize printed texts in Indian languages. We need an alternate approach to access the content of these documents [3]. A promising alternate direction is to search for relevant documents in image space without any explicit recognition. We have been motivated by the successful attempts on

H. Bunke and A.L. Spitz (Eds.): DAS 2006, LNCS 3872, pp. 1–12, 2006.

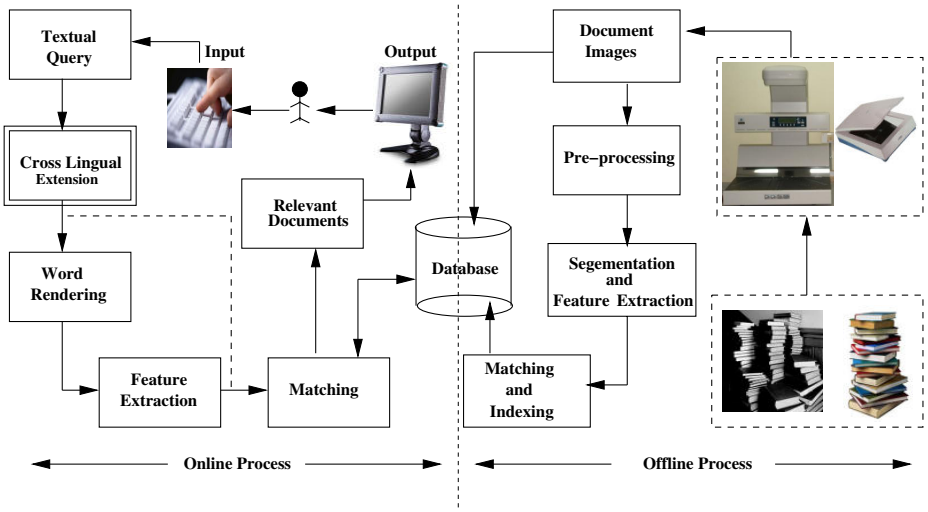


Fig. 1. Conceptual Diagram of the Searching Procedure from Multilingual Document Image Database. A Web Demo for the Above Procedure is Available Online at http://cvit.iiit.ac.in/wordsearch.html.

locating a specific word in handwritten English documents by matching image features for historical documents [4, 5].

We have already addressed algorithmic challenges for effective search in document images [6] . This paper describes the issues associated with the implementation of a scalable system for Indian language document images. A conceptual block diagram of our prototype system is shown in Figure 1. Our system accepts textual query from users. The textual query is first converted to an 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 pages from document image collections containing queried word sorted based on their relevance to the query.

2 Challenges in Design and Implementation of the System

Search and retrieval from large collection of document images is a challenging task, specially when there is no textual representation available. To design and implement a successful search engine in image domain, we need to address the following issues.

Search in images: Search in image space requires appropriate representational schemes and similarity measures. Success of content-based image retrieval(CBIR) schemes were limited by the diversity of the image collections. Digital libraries primarily archive text images, but of varying quality, script, style, size and font.

We need to come up with appropriate features and matching schemes, which can represent the content (text), while invariant to the popular variations.

Degradations of documents: Documents in digital libraries are extremely poor in quality. 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 of 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 etc. We need to design an appropriate representational scheme and matching algorithm to accommodate the effect of degradation.

Need for cross-lingual retrieval: Document images in digital libraries are from diverse languages. Relevant documents that users need may be available in different languages. Most educated Indians can read more than one language. Hence, we need to design a mechanism that allows users to retrieve all documents related to their queries in any of the Indian languages.

Computational speed: Searching from large collection of document images pass through many steps: image processing, feature extraction, matching and retrieval of relevant documents. Each of these steps could be computationally expensive. In a typical book, there could be around 90,000 words and processing all of them online is practically impossible. We do all computationally expensive operations during the offline indexing (Section 4) and do minimal operations during online retrieval (Section 5).

Indian languages: Indian languages pose many additional challenges [7]. Some of these are: (i) lack of standard representation for the fonts and encoding, (ii) lack of support from operating system, browsers and keyboard, and (iii) lack of language processing routines. These issues add to the complexity of the design and implementation of a document image retrieval system.

3 Representation and Matching of Word Images

Word images extracted from documents in digital libraries are of varying quality, script, font, size and style. An effective representation of the word images will have to take care of these artifacts for successful searching and retrieval. We combined two categories of features to address these effects: word profiles and structural features. Explicit definitions of these features may be seen in [6].

Word Profiles: Profiles of the word provide a coarse way of representing a word image for matching. Profiles like upper word, lower word, projection and transition profiles are used here for word representation. Upper and lower word profiles capture part of the outlining shape of a word, while projection and transition profiles capture the distribution of ink along one of the two dimensions in a word image.

Structural Features: Structural features of the words are used to match two words based on some image similarities. Statistical moments (such as mean and standard deviation) and region-based moments (such as the zeroth- and first-order moments) are employed for describing the structure of the word image. For artifacts like salt and pepper noise, structural features are found to be reasonably robust.

Some of these features provide the sequence information, while others capture the structural characteristics. Given a document image, it is preprocessed offline to threshold, skew-correct, remove noise and thereafter to segment into words. Then features are extracted for each of the segmented word. They are also normalized such that the word representations become insensitive to variations in font, style, size and various degradations popularly present in document images.

Spotting a word from handwritten images is attempted by pairwise matching of all the words [5]. However for proper search and retrieval, one needs to identify the similar words and group them based on their similarity, and evaluate the relative importance of each of these words and word clusters. Matching is used to compute dissimilarity between word images. We use a simple squared Euclidean distance while computing the dissimilarity.

For matching word images we use Dynamic Time Warping (DTW) that computes a sequence alignment score for finding the similarity of words [6]. The use of the total cost of DTW as a distance measure is helpful to cluster together word images that are related to their root word, which is discussed in Section 4.

DTW is a dynamic programming based procedure [5] to align two sequences of signals and compute a similarity measure. 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 is given by:

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

where $d(i,j) = \sum_{k=1}^{N} \left(F(i,k) - G(j,k)\right)^2$ (the cost in aligning the ith element of $\mathbf{F}$ with jth element of $\mathbf{G}$). Using the given three values $D(i, j-1), D(i-1, j)$ and $D(i-1, j-1)$ in the calculation of $D(i,j)$ realizes 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. Structural features can also be incorporated into the framework by computing them for the vertical strips. Detailed discussion of the algorithms is available in [6].

4 Offline Indexing

The simple matching procedure described in Section 3 may be efficient for spotting or locating a selected word-image. However the indexing process for a good

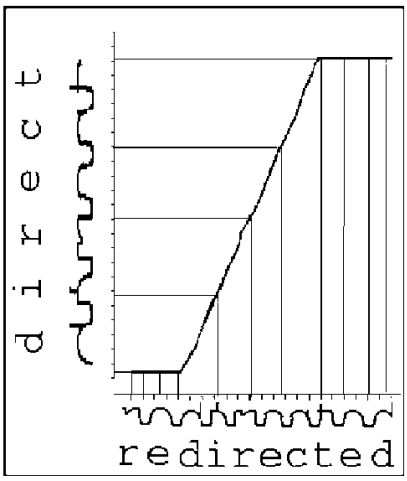


Fig. 2. Plot Demonstrating Matching of Two Words 'direct' and 'redirected' using Dynamic Time Warping and the Optimal Matching Path. Similar Word Form Variations are Present in Indian Languages.

search engine is more involved than the simple word-level matches. A word usually appears in various forms. Variation of word forms may obey the language rules. Text search engines use this information while indexing. However for text-image indexing process, 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 six 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. With this we find that a large set of words get grouped into one cluster. We expect to extend this for more general variations of words.

The optimal warping path is generated by backtracking the DTW minimal score in the matching space. As shown in Figure 2, extracted features (using upper word profile) of the two words 'direct' and 'redirected' are aligned using DTW 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 additional cost is identified and removed while backtracking.

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 2, the path deviates from the diagonal line at the two extreme ends.

This happened 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 and hence subtracted from the total matching cost so as to compute the net cost. Such word form variations are very popular in most languages.

For the indexing process, we propose to identify the word set by clustering them into different groups based on their similarities. This requires processing the page to be indexed for detection of relevant words in it. Many interesting measures are proposed for this. We propose the following steps for effective retrieval at image level.

Detection of Common Stop Words: Once similar words are clustered, we analyze the clusters for their relevance. A very simple measure of the uniformity of the presence of similar words across the documents is computed. This acts as an inverse document frequency. If a word is common in most of the documents, this word is less meaningful to characterize any of the document.

Document Relevance Measurement: Given a query, a word image is generated and the cluster corresponding to this word is identified. If a cluster is annotated, matching query word is fast and direct. For other clusters, query word image and prototype of the cluster are compared in the image domain. In each cluster, documents with highest occurrence of similar words are ranked and listed.

Clustering: Large number of words in the document image database are grouped into a much smaller number of clusters. Each of these clusters are equivalent to a variation of the single word in morphology, font, size, style and quality. Similar words are clustered together and characterized using a representative word. We follow a hierarchical clustering procedure [8] to group these words. Clusters are merged until the dissimilarity between two successive clusters become very high. This method also provides scope for incremental clustering and indexing.

Annotation: After the clustering process has been completed offline, we have a set of similar words in each cluster. These clusters are annotated by their root word to ease searching and retrieval. Suppose a cluster contains words such as 'programmer', 'programmers', 'programming', 'programs' and 'program'. Then, we annotate the cluster with the root word "program". Likewise all clusters are manually annotated. If the annotation is not available, we identify an image-representative for the cluster. However, presence of image-prototype can slow down the search process. During searching, cluster prototypes are accessed and checked for their similarity with the query word. This makes sure that search in image domain is as fast as search in text domain.

5 Online Retrieval

A prototype web-based system for searching in document images, is also developed. This is presently available at http://cvit.iiit.ac.in/wordsearch.html. The system has many basic features as discussed below.

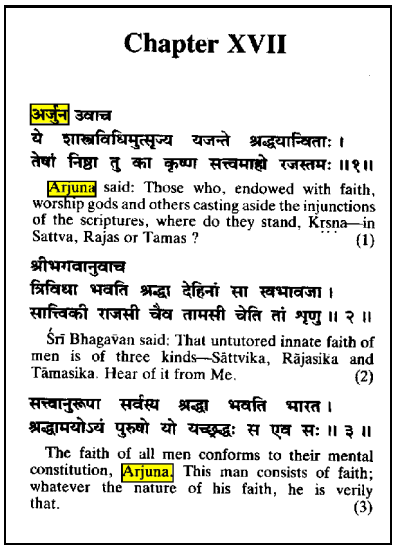
Web-based GUI: The web interface allows the user to type in Roman text and simultaneously view the text in one of the Indian languages of his choice. Users can also have the option to search with cross lingual retrieval and can specify the kind of retrieval they want to use. Many retrieval combinations are also provided in the advanced search options such as case insensitivity, boolean searching using !, &, | *and parenthesis*, displaying up to 50 search results per page, and various others. There is on the fly character transliteration available. The user can first choose a particular language (such as Hindi, Telugu, etc.) and then see the text in the corresponding language as he keeps typing the query in Roman (OM-Trans).

Delivery of Images: In order to facilitate users access to the retrieved document images, there is a need to control image size and quality. When a book is typically scanned at a resolution of 600 dpi, the original scanned size of a single page is around 12MB as a PNG file. Viewing such page is too slow and needs network resources. It is wise to make these images available in a compressed form. We compress the above image to a size ranging between 30 to 40 KB in TIFF format, by reducing the size of the image. This makes sure that the delivery of images are faster over the Internet. TIFF image format helps us in general for achieving the trade-off between image size and quality. It keeps the quality of the image during the compression process over JPG and BMP formats.

Speculative Downloading: Our system also supports speculative downloading, where some related pages with the currently retrieved page are prefetched for quick viewing during searching and retrieval as per users query. This mechanism is helpful especially when the user is viewing a collection, page by page, with the assumption that he might view the next page also. Speculative downloading is a background process.

Dynamic Coloring: When a user searches for relevant pages to a given query, our system searches and displays the result with dynamic coloring of all the words in the page that are similar to the queried word. This helps users to easily evaluate relevance of the retrieved page to their need. We adopt false coloring mechanism such that each word in a query carries a unique color in a document image. All this coloring happens at runtime (Figure 3) at image level.

Scalability: DLI is a one million book scanning project. Hence it archives huge collection of document images. Searching in this situation raises the question of scalability. The current prototype system searches in three books, that are a mixture of English and other Indian Languages (Hindi and Telugu). Each book on the average consists of 350 pages, and each page with 300 words. This brings the total number of words to 360,000. This is relatively a small number. The system should aid in searching the huge one million book collection and thus the scalability issues come to forth. Indexing this large collection takes immense time. For us indexing is an offline activity. Searching and retrieval is the only online process. That is why the system manages to run fast in the above sample database. Because it only checks keywords of the index to search for similar words with the query. Even with an increase in the size of document images we do not expect much increase in the number of clusters. Because, words are



Chapter XVII

अर्जुन उवाच
ये शास्त्रविधिमुत्सृज्य यजन्ते श्रद्धयान्विताः ।
तेषां निष्ठा तु का कृष्ण सत्त्वमाहो रजस्तमः ॥१॥

Arjuna said: Those who, endowed with faith, worship gods and others casting aside the injunctions of the scriptures, where do they stand, Krsna—in Sattva, Rajas or Tamas ? (1)

श्रीभगवानुवाच
त्रिविधा भवति श्रद्धा देहिनां सा स्वभावजा ।
सात्त्विकी राजसी चैव तामसी चेति तां शृणु ॥ २ ॥

Śrī Bhagavān said: That untutored innate faith of men is of three kinds—Sāttvika, Rājasika and Tāmasika. Hear of it from Me. (2)

सत्त्वानुरूपा सर्वस्य श्रद्धा भवति भारत ।
श्रद्धामयोऽयं पुरुषो यो यच्छ्रद्धः स एव सः ॥ ३ ॥

The faith of all men conforms to their mental constitution, Arjuna. This man consists of faith; whatever the nature of his faith, he is verily that. (3)

Fig. 3. Search Result with Dynamic Coloring for Query Word 'Arjuna' seen Both in English and Devanagari

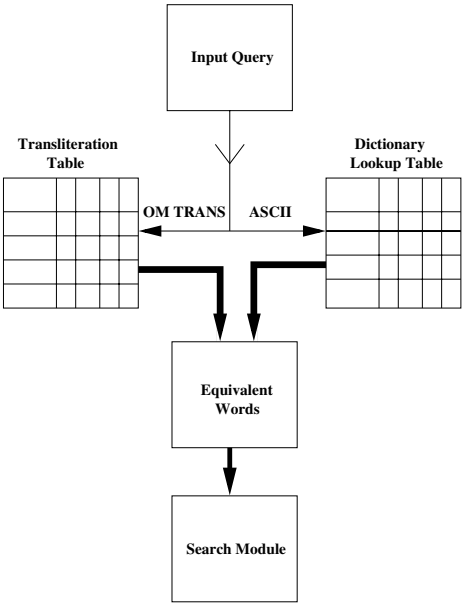
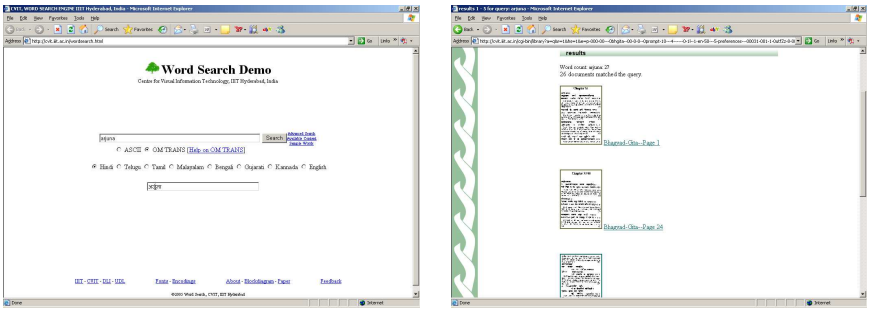


Fig. 4. A Conceptual Diagram that Shows Document Searching in Multiple Languages using Transliteration and Dictionary Based Approach

Alphabet	a	aa	i	ii	u	uu	....
Hindi	अ	आ	इ	ई	उ	ऊ	....
Telugu	అ	ఆ	ఇ	ఈ	ఉ	ఊ	....
:	:	:	:	:	:	:	
:	:	:	:	:	:	:	
:	:	:	:	:	:	:	

Fig. 5. Sample Entries of the Transliteration Map Built for Cross-lingual Retrieval in English, Devanagari and Telugu



(a) (b)

Fig. 6. Screenshots of Implementation Results for Cross-lingual Search. (a) An Interface where Users Enter their Query, (b) View of Result of the Search as Thumbnails

limited in every language and they are only the morphological variants of the root word. The system handles addition of new books without re-indexing every time. This saves much time and creating new indexes will be a smooth process. However, we need to deal with delivery of the document images. An increase in the number of pages viewed may slow the transfer process. A good compression technique needs to be applied.

Cross-lingual Search: Our system can also search for cross-lingual documents for a given query. As shown in Figure 4 we achieve this in two ways: transliteration and dictionary-based approaches. Figure 4 is an expanded view of the cross-lingual block diagram presented in Figure 1.

Since Indian scripts share a common alphabet (derived from *Brahmi*), we can transliterate the words across languages. This helps us to search in multiple languages at the same time. We use OM-Transliteration scheme [9]. In OM-Trans scheme, there is a Roman equivalent for all the basic Indian language characters.

Figure 5 shows a sample transliteration map built for this purpose. For Example, "Bhim " can be typed in its Roman equivalent using OM-trans as "Bhima". Then the transliteration table is looked up for searching in Hindi (Devanagari script) and Telugu pages. Their cumulative result is finally displayed back to the user. A screenshot showing implementation result is presented in Figure 6.

Fig. 7. Results: (a) Sample Word Images Retrieved for the Queries Given in Special Boxes. Examples are from English 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.

We also have a dictionary-based translation for cross-lingual retrieval. In this approach, every English word has an equivalent word in the corresponding Indian and other oriental languages. If a user queries for the word 'India', the dictionary lookup points to 'भारत' in Hindi for searching relevant documents across languages. This table is extended also for other Indian languages. The result of the search are documents that contain the query word 'India' in all the languages.

We tried searching in scanned documents from the book 'Bhagavat-Gita'. Pages from this book contain English and Devanagari text. These pages are of poor quality. We search for the occurrences of the word 'arjuna'. It fetched pages which contain 'Arjuna' in both English and Devanagari. Sample results are shown in Figure 7 (b). In this respect, we need to exploit the available technology at WordNet Project [10] and Universal Language Dictionary Project [11]. WordNet is a lexical database that has been widely adopted in artificial intelligence and computational linguistics for a variety of practical applications such as information retrieval, information extraction, summarization, etc. The Universal Language Dictionary is an attempt to create a list of concepts along with words to express those concepts in several "natural" and "artificial" (constructed) languages.

6 Discussion

We have a prototype system for retrieval of document images. This system is integrated with 'Greenstone search engine' for digital libraries [12]. Greenstone is a suite of software for building and distributing digital library collections via the Internet. Given a textual query, we convert it to image by rendering. Features are extracted from these images and then search is carried out for retrieval of relevant documents in image space. We extend the search to cross-lingual retrieval by transliteration among Indian languages and a table-lookup translation for other languages. Results of the search are presented to the user in a ranked manner based on their relevance to the query word.

Table 1. Performance of the Proposed Approach on Two Data Sets in English, and Hindi. Percentages of Precision and Recall are Reported for Some Test Words.

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

[1]*Number of words used for testing

We evaluated the performance of the system on data sets from languages such as English and Hindi. Pages of Hindi and English are taken from digital library of India collections. The system is extensively tested on all these data sets. Sample words retrieved are shown in Figure 7 (a).

We measure the speed of the system so as to see its practicality. The system takes 0.16 seconds to search and retrieve relevant documents from image databases and 0.34 seconds to transfer that page for viewing by users over the intranet. In comparison, Greenstone text search takes 0.13 seconds to search and retrieve relevant documents from image databases and 0.31 seconds to transfer that page for viewing by users over the intranet. The speed of our system is almost comparable with the Greenstone text search. This shows the effectiveness of the system. The strategy we followed is to perform text processing and indexing offline. The search then takes place on the representative words indexed. Compressing the image (to a size of few KB) also help us a lot during the transfer of the document image for viewing.

Quantitative performance of the matching scheme is computed on sample document image databases of size more than 2500 words. Around 15 query words are used for testing. During selection of query words, priority is given to words with many variants. We computed recall and precision for these query words, as shown in Table 1. Percentage of relevant words which are retrieved from the entire collection is represented as recall, where as, percentage of retrieved words which are relevant is represented as precision. It is found that a high precision and recall (close to 95%) is registered for all the languages. High recall and precision is registered in our experiment. This may be because of the limited dataset we experimented with, that are similar in font, style and size. We are working towards a comprehensive test on real-life large datasets. Our existing partial matching module controls morphological word variants. We plan to make the module more general so that it addresses many more variations of words encountered in real-life documents. We are also working on avoiding the manual annotation and still retaining the same performance.

7 Conclusions

In this paper, we have proposed a search system for retrieval of relevant documents from large collection of document images. This method of search will be important in using large digitized manuscript data sets in Indian languages.

We have focused on computing information retrieval measures from word images without explicitly recognizing these images. The system is capable of searching across languages for retrieving relevant documents from multilingual document image database. Preliminary experiments show that the results are promising. We are currently working on a comprehensive test on large collection of document images.

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

References

1. Digital Library of India. (at: `http://www.dli.gov.in`)
2. Doermann, D.: The Indexing and Retrieval of Document Images: A Survey. Computer Vision and Image Understanding (CVIU) **70** (1998) 287–298
3. Chaudhury, S., Sethi, G., Vyas, A., Harit, G.: Devising Interactive Access Techniques for Indian Language Document Images. In: Proc. of the Seventh International Conference on Document Analysis and Recognition (ICDAR). (2003) 885–889
4. Rath, T., Manmatha, R.: Features for Word Spotting in Historical Manuscripts. In: Proc. of the Seventh International Conference on Document Analysis and Recognition (ICDAR). (2003) 218–222
5. Rath, T., Manmatha, R.: Word Image Matching Using Dynamic Time Warping. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) **2** (2003) 521–527
6. C. V. Jawahar, Million Meshesha, A. Balasubramanian: Searching in Document Images. Proc. of the 4th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP) (2004) 622–627
7. Department of Information Technology: Technology Development for Indian Languages. (at: `http://tdil.mit.gov.in`)
8. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Willey & Sons, New York (2001)
9. Indian Language Transliteration. (at: `http://www.cs.cmu.edu/~madhavi/Om/`)
10. Vossen, P., Fellbaum, C.: The Global WordNet Association. (at: `http://www.globalwordnet.org`)
11. Universal Language Dictionary Project. at: `http://ogden.basic-english.org` (2003)
12. Greenstone Digital Library Software. (at: `http://www.greenstone.org`)