Word-Level Access to Document Image Datasets

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

Abstract

This paper presents a novel approach for retrieval of relevant documents from a large collection of printed document images, where search in corresponding textual content is practically impossible due to the unavailability of robust OCRs. We achieve effective search by matching at word-level using image features. Importance of a document for a given query is also computed at image-level.

1 Introduction

Indexing and retrieval of textual data is an important aspect in building effective access to a digital library. In literature, emphasis has been on the evaluation of the relevance of documents based on statistics of the collection, as well as on the history of user behaviour. Most of these ideas work only on textual data. Indexing and retrieval from document image collections were also studied with limited scope [1, 2]. 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, OCR technology is not yet able to successfully recognise printed text of varying quality, size and font. To have effective access to document image databases in these languages, alternate approaches are needed [2, 3]. A promising alternate direction is to search without explicit recognition of characters. There have been successful attempts on locating a specific word in a handwritten text by matching image features for historical documents [4, 5].

With storage becoming cheaper and imaging devices becoming increasingly popular, efforts are on the way to archive every page of printed or handwritten documents. Large digital libraries are emerging for this purpose [6]. Much of the data in these libraries are in languages where OCRs are not available as of now. We need a method to access the content of these documents. This is the motivation behind this work. This work aims at building a prototype system for finding out the most relevant documents for a given query word from a set of printed text document images. This needs to address the following questions.

1. Can we search for the presence of a word without converting the document image into textual form?
2. Can we index the document images based on relative importances to a query?
3. Can we take care of the morphological variations of words for effective indexing?

In this paper, we propose solutions to the above problems without converting the images into text. In section 2, a representation scheme for word images is proposed to take care of the common artifacts in printed document images. Word images are matched using these feature descriptions. Matching procedure is explained in Section 3. The proposed matching algorithm can take care of the popular morphological variations of the words. Similar words are grouped into a cluster, which allows to compute the weight of the word and relevance of the document. Involved steps are briefly explained in Section 4. Sample results for three languages –Amharic, Hindi and English are given in Section 5.

2 Representation of Word Images

Word images, particularly in newspapers and old books, are of extremely poor 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. Sample images from real-life documents, we experimented with, with these artifacts are shown in Figure 1. An effective representation of the word images will have to take care of these artifacts for successful indexing and retrieval.

We find that at least three categories of features are effective to address these effects. We have not yet attempted a systematic study of effectiveness of the individual features for specific artifacts and specific script/language. The features could be a sequential representation or a structural representation of the word [5, 7].

Word Profiles: Profiles of the word provide a coarse way of representing a word image for matching. Profiles like upper word, lower word, projection, density and transition profiles are used here for word representation. Some of them (for eg. upper and lower word profiles) capture part of the outlining shape of a word, while some others (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. Features employed in image processing literature for shape recognition are promising candidates for this. Moments, mean, distribution of black pixels, curvatures of characters etc. are employed for describing the structure of the word. For many of the artifacts like pepper and salt noise, structural features are found to be reasonably robust.

Transformed Domain Representations: A compact representation of a series of observations (like profiles) is using Fourier Transform. Fewer set of coefficients are enough to represent robustly in a transformed domain, and these coefficients are matched at a coarse level for recognition.

Some of these features provide the sequence information, while others capture the structural characteristics. Both of them are independently matched for computing word similarities. Given a document image, it is preprocessed offline to threshold, skew-correct, remove noise and thereafter to segment into words [2]. 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.

3 Matching and Grouping of Words

Spotting a word from handwritten images is attempted by pairwise matching of all the words [4, 5]. However for proper indexing and retrieval, one needs to identify the similar words and group them and evaluate the relative importance of each of these words and word clusters. For the indexing process, we propose to identify the word set by clustering them into different groups based on similarities

Figure 1: Frequent Artifacts in Images of Printed Text Documents. Example words are shown from Amharic, Telugu, Tamil, Hindi and English

between words. Distance or dissimilarity between words are computed using the features discussed in the previous section. We use a simple Euclidean distance.

In this paper, we find the similarity of words based on two components: (a) A sequence alignment score computed from a Dynamic Time Warping Procedure. (b) Structural similarity of Word images by comparing the shapes. The use of the total cost of Dynamic Time Warping(DTW) as a distance measure is helpful to cluster together word images that are related to their root word by partial match as explained in the next section.

Dynamic Time Warping is a dynamic programming based procedure [5] to align two sequences of signals and compute a similarity measure. This is a popular technique in speech recognition. 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, $d(i,j)$ is 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 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. A simple plot for matching profiles of two words is shown in Figure 2.

Large number of words in the document image database is grouped into a much smaller number of clusters. Each of these clusters are equivalent to a variation of the single word in font, size and style. Similar words are clustered together and characterised using a representative word. We follow a hierarchical clustering procedure [8] to group these words. Clusters are merged until the dissimilarity between

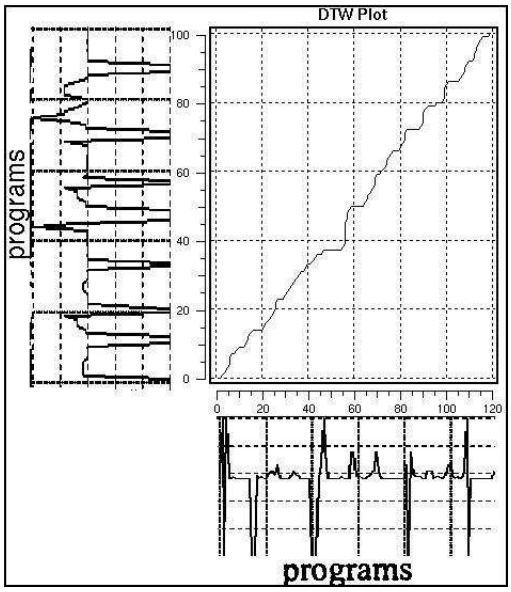

Figure 2: A sample DTW plot during matching. Two words, corresponding profiles and the optimal matching path are shown

two successive clusters became very high.

4 Indexing and Retrieval

Addressing Word Form Variations The simple matching procedure described above may be efficient for spotting or searching a selected word or word-image. However the indexing process for a good 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 a simple, but very popular word form variation at the end. For this, once sequences are matched, we backtrack the optimal cost path, if the difference in words are concentrated at the end. Therefore the matching score of "characters" and "characterisation" are only the matching of the first 9 characters of both the words. Once an optimal sub-path is identified, 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.

DTW score for matching with and without this modification is shown in Table 1. This score along with the structural matching score allows us to take care of morphological variations. We expect to extend this for more general variations of words.

Detection of relevant words for indexing a page is very important for effective retrieval. Many interesting measures are proposed for this. We propose some simple measures in the similar direction to do a similar job at image level.

Word	Conventional	Modified
Program	0.023241	0.016076
Programmer	0.181001	0.116316
Programmers	0.197802	0.127585
Programming	0.229562	0.130781
diagram	0.162764	0.133201
खरीदा	0.013389	0.009658
खरीदो	0.109000	0.094300
खरीदी	0.123543	0.106600
खरीदना	0.127856	0.117610
खराब	0.126389	0.120170

Table 1: Word form variations and matching scores computed using the proposed approach

Detection of Common Stop Words Once similar words are clustered, we analyse 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 all the documents, this word is less meaningful to characterise any of the document.

Word Frequencies for Retrieval Given a query, a word image is generated and the cluster corresponding to this word is identified. In each cluster, documents with highest occurrence of similar words are ranked and listed.

5 Results and Discussion

We experimented the procedure on sample sets of pages from many languages. Summary of the experiments show that the procedure is found to perform well for all these languages. We are working on improving the performance of the procedure by integrating the language specific features in script and word morphology.

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

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

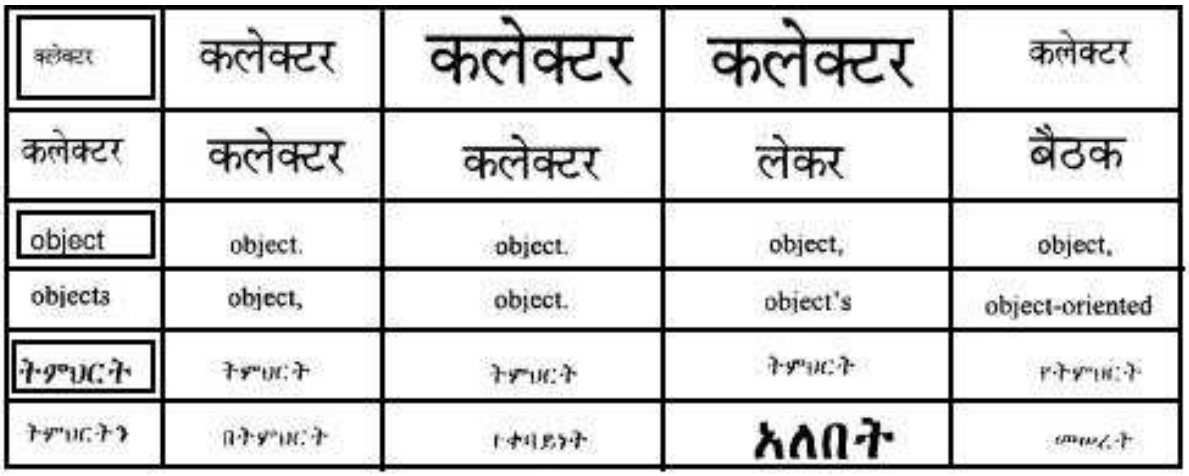

Figure 3: Some sample word images retrieved for the queries given in special boxes

We conducted detailed experiments on English, one Indian Language (Hindi) and an African Language(Amharic). Amharic is the official language of Ethiopia. Almost 95 percent of the languages spoken in Africa use Latin-based and Arabic scripts while, Amharic is the only official language with its own indigenous writing system(FIDEL) in Africa. We have built a corpus of Amharic newspaper images by scanning pages from the newspaper "Addis Zemen". This is used as a test-bed. Other data set consists of scanned pages of Hindi printed text in the lab and scanned document images from digital library of India. The proposed method is extensively tested on all these data sets. Some of the sample words retrieved are shown in Figure 3. The word in double-box is the query word and the words in that and next row are the retrieved words.

Performance of the word level access is computed on the document image databases of size approximately 2500 words. Around 15 words are used for testing. For these words precision and recall are manually tabulated in Table 2. Percentage of correct retrieval is represented as Precision, while recall is computed by the percentage of the correct words retrieved through this process. It is found that both precision and recall are close to 95% for all the languages.

6 Conclusions

In this paper, we have proposed a framework for word-level access to a collection of document images. These methods of access 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 recognising these images. Language specific information is not used in the retrieval process. We are working towards building a crosslingual retrieval for multilingual document image database.

References

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

[2] Mudit Agrawal, M. N. S. S. K. Pavan Kumar, and C. V. Jawahar, "Indexing and retrieval of devanagari text from printed documents," in *Proc of NCDAR 2003*, pp. 244–251, 2003.

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

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

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

[6] http://ulib.org http://www.dli.gov.in

[7] B. B. Chaudhuri and D. D. Majumder, *Two-Tone Image Processing and Recognition*. New Delhi: Wiley Eastern Limited, 1993.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: John Willey & Sons, 2001.