Word Spotting in Scanned Images of Printed Devanagri Documents

Pankaj,Premendra Tiwari,Siddharth Patel and Sunny Sangwan

Computer Engineering, Army Institute of Technology,
Pune,Maharashtra,India

Computer Engineering, Army Institute of Technology,
Pune,Maharashtra,India

Computer Engineering, Army Institute of Technology,
Pune,Maharashtra,India

Computer Engineering, Army Institute of Technology,
Pune,Maharashtra,India

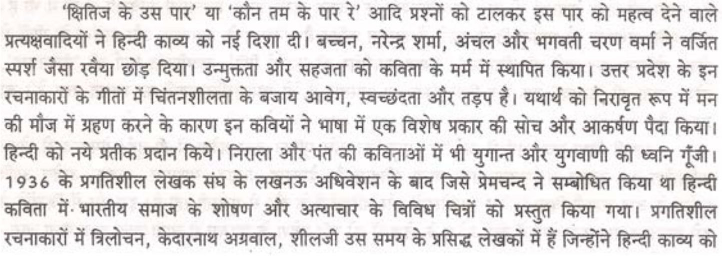
Abstract

In this paper, we propose a keyword retrieval system for locating words in printed Devanagari document images. Government has initiated various schemes to scan and digitally preserve such documents for future use. Such scanned images of documents are now available to users. Even though the documents are available in their digital format, it is still difficult to search for a single word or phrase as they are scanned images. Traditional optical character recognition techniques (OCR) and other text retrieval systems fail on such type of document images due to various types of noises. In such situations word spotting shall be a major help to users to automatically search for a particular word/phrase in millions of such document images. An attempt is made here to design and implement a word spotting technique for printed Devanagari documents. Based on the word spotting technology, a collection of document images is converted into a collection of word images by word segmentation, and a number of profile-based features are extracted to represent word images. Dynamic Time Warping(DTW) is applied for image comparisons. Image-to-image matching is done by calculating similarities between a query word image and each word image in the collection, and consequently, a ranked result is returned in descending order of the similarities. The system supports GUI. The optimal performance of the system is validated.

Keywords: *Devanagari documents,OCR,DTW,profile based features,GUI*

1. Introduction

Devanagari documents contain precious culture heritages of human beings. At present, many countries are digitizing their native language documents to preserve them as long as possible, which enables the public to access them more conveniently and faster, for instance, via Internet. In Devanagari documents., there are a large number of ancient books. Figure 1 shows a fragment of one Devanagari document image. Although the public can browse these digital images without the need of visiting the library, it is difficult to retrieve them due to the lack of indexing. Traditionally, there are two approaches for creating indexing in the field of document image retrieval (DIR). The first one is manual annotation. Since the document contains large number of pages so it is a highly expensive and tedious task for such a large collection. The second one is an automatic approach. It utilizes the optical character recognition (OCR) technology to convert images into texts and the indexing is created on the OCR'ed texts.



'क्षितिज के उस पार' या 'कौन तम के पार रे' आदि प्रश्नों को टालकर इस पार को महत्व देने वाले प्रत्यक्षवादियों ने हिन्दी काव्य को नई दिशा दी। बच्चन, नरेन्द्र शर्मा, अंचल और भगवती चरण वर्मा ने वर्जित स्पर्श जैसा रवैया छोड़ दिया। उन्मुक्तता और सहजता को कविता के मर्म में स्थापित किया। उत्तर प्रदेश के इन रचनाकारों के गीतों में चिंतनशीलता के बजाय आवेग, स्वच्छंदता और तड़प है। यथार्थ को निरावृत रूप में मन की मौज में ग्रहण करने के कारण इन कवियों ने भाषा में एक विशेष प्रकार की सोच और आकर्षण पैदा किया। हिन्दी को नये प्रतीक प्रदान किये। निराला और पंत की कविताओं में भी युगान्त और युगवाणी की ध्वनि गूँजी। 1936 के प्रगतिशील लेखक संघ के लखनऊ अधिवेशन के बाद जिसे प्रेमचन्द ने सम्बोधित किया था हिन्दी कविता में· भारतीय समाज के शोषण और अत्याचार के विविध चित्रों को प्रस्तुत किया गया। प्रगतिशील रचनाकारों में त्रिलोचन, केदारनाथ अग्रवाल, शीलजी उस समय के प्रसिद्ध लेखकों में हैं जिन्होंने हिन्दी काव्य को

Fig 1. A fragment of one Devanagari document image

So far as we know, there are very few OCR software for Devanagari and they yield clumsy conversion. The goal of word spotting is to find all word images in the collection that are similar to a given query word image by image matching. There are various image matching methods for measuring similarities between word images. Several image matching methods were compared with each other, and the dynamic time warping (DTW) algorithm was the best one. Consequently, DTW has been widely used for image matching in word spotting .However, the word spotting technology has several drawbacks, such as that the query word image needs to be selected from document images by users and that the accuracy of word segmentation might be very poor on poorly scanned document images, especially for historical printed. Moreover, image matching methods, such as DTW, are often time consuming, which makes it impossible for online image-to-image matching during the retrieval.

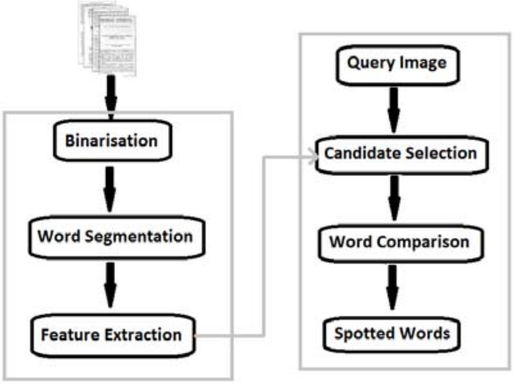


Fig. 2 The architecture of the proposed system

2. Characteristics of Devanagari Languages

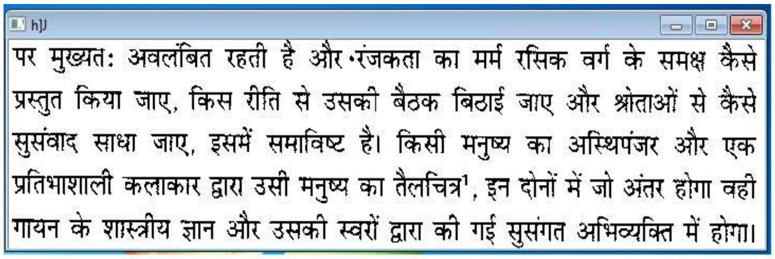
The characteristics are as follows:

Devanagari Needs No Spelling Arrangement:
Spelling has been eliminated in Devanagari because every sound had been correctly analyzed and placed into its phonetic classification and the consonants and the vowels which have different functions have been assigned definite mode of behavior.

Series Modulation by Vowel Signs:
Devanagari and all script of the Brahmi family have distinctive graphemes for all vowels in order to join with the consonants representative signs. These have been evolved along with the script.

Differentiation for the Pronunciation of Vowels:
As pronunciation was to be very accurately managed the Indian Grammarians made differentiation between the signs for the short and long sounds of the same vowel. It will be seen that the vowels which are short, flourish to the left and their longer signs to the right. This is noticeable in Devanagari, Tamil, Malayalam and other scripts of South Indians who were very careful in preserving their traditions.

The Vertical Line:
In earlier Brahmi script the vertical line is absent; it shows the addition of the "Aa matra" to a consonant so that it could be fully pronounced and written.



पर मुख्यत: अवलंबित रहती है और ·रंजकता का मर्म रसिक वर्ग के समक्ष कैसे प्रस्तुत किया जाए, किस रीति से उसकी बैठक बिठाई जाए और श्रोताओं से कैसे सुसंवाद साधा जाए, इसमें समाविष्ट है। किसी मनुष्य का अस्थिपंजर और एक प्रतिभाशाली कलाकार द्वारा उसी मनुष्य का तैलचित्र[1], इन दोनों में जो अंतर होगा वही गायन के शास्त्रीय ज्ञान और उसकी स्वरों द्वारा की गई सुसंगत अभिव्यक्ति में होगा।

Fig. 3 Binarized image

The Top Line:
The top line is an integral and essential feature of Devanagari orthography. The continuous top line is a later development. In old manuscripts, copper plates and inscriptions the top line was limited to the character only and did not join the next letter. The top line knits letters into distinct words. A gap indicates separation of one word from another.

Conjunct Formation of Consonants:
Devanagari has evolved as a specialized script for the highly developed languages like Sanskrit. Some reformers recommended formation of conjunct letters with halant sign even in case of letters with full vertical lines.

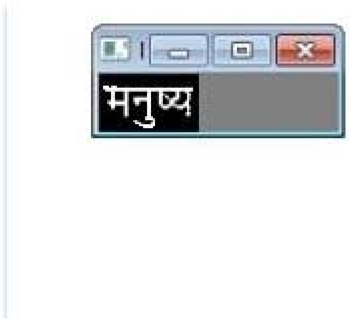
Diacritical and Special Marks:
Devanagari script is said to be capable of expressing many varieties of sounds. Various signs for variety of long and short vowels were made in the language.

Direction from Left to Right:
Perhaps all ancient scripts were first written from the right to left. Only the Hebrew, the Arabic, the Persian and the Urdu continued with old way, but Brahmi the mother script of Devanagari changed the direction of writing from left to right. In this respect the Devanagari is somewhat nearer to the Roman script.

Independent Grapheme exists for Pure Vowels and Consonants:
Like the Roman script, Devanagari too has independent vowels. Every basic vowel has a different phonetic origin; each should reasonably have independent grapheme. These vowels are given representative forms and these are used for modulation of consonants which are supposed be

pronounced with the help of vowels into a modulating series.

Basic Arrangement of Phonemes is Phonetic:
The Brahmi script has evolved from the studies in phonemics in which the Dravidians had specialized and hence is completely phonetic. There are some pronunciations imported from foreign languages and they are indicated by a dot at the bottom of the nearer phoneme. Half open vowels imported from European tongues are indicated by a half-moon sign over nearer phoneme.



मनुष्य

Fig. 4 Query word image

3. Overview of the keyword retrieval system

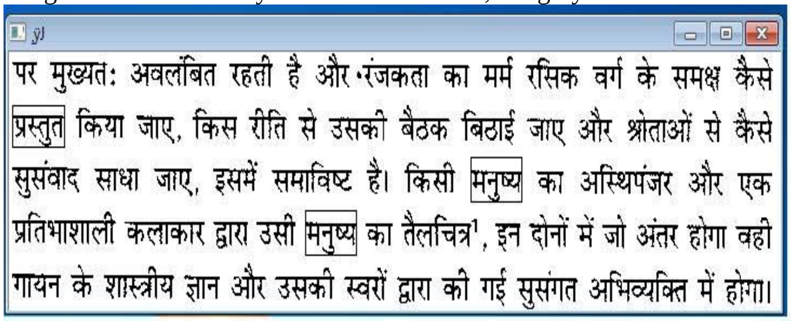
The proposed keyword retrieval system utilizes the word spotting technology to index word images and to retrieve them by image-to-image matching. The whole architecture of this system is shown in Fig. 2. Firstly, the scanned images are converted into binary images. Then, word images can be obtained by connected components analysis on each binary image. Next, several profile features are extracted to represent each word image. The coordinate information of word images are saved into a database. User provides a query word image by selecting an instance of it from the document image itself. DTW is applied to get the distance between query word image and each word image in the database. And then, the ranked results will be displayed by sorting similarities between the query word image and each word image in the database.

4. Preprocessing

The scanned document images are saved in colored PNG format with 600 dpi. The collection of the document images needs to be converted into a collection of word images. Firstly, these color images are converted into binary images then word images can be obtained by connected components analysis on each binary image. The detailed procedures of binarization and word segmentation are described as follows.

a. Binarisation.

The scanned color images are converted into gray level images and smoothed by Median filter. Then, the gray level images are processed by the well-known global thresholding method i.e, Otsu algorithm. to convert them into binary document images.



पर मुख्यत: अवलंबित रहती है और •रंजकता का मर्म रसिक वर्ग के समक्ष कैसे प्रस्तुत किया जाए, किस रीति से उसकी बैठक बिठाई जाए और श्रोताओं से कैसे सुसंवाद साधा जाए, इसमें समाविष्ट है। किसी मनुष्य का अस्थिपंजर और एक प्रतिभाशाली कलाकार द्वारा उसी मनुष्य का तैलचित्र[1], इन दोनों में जो अंतर होगा वही गायन के शास्त्रीय ज्ञान और उसकी स्वरों द्वारा की गई सुसंगत अभिव्यक्ति में होगा।

Fig. 5 Spotted words

Otsu's method is used to automatically perform clustering-based image thresholding.

$$\sigma_w^2(t) = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t) \tag{1}$$

Weights ω_i are the probabilities of the two classes separated by a threshold t and σ_i^2 are variances of these classes. Figure 3 shows the binarized image of a sample document.

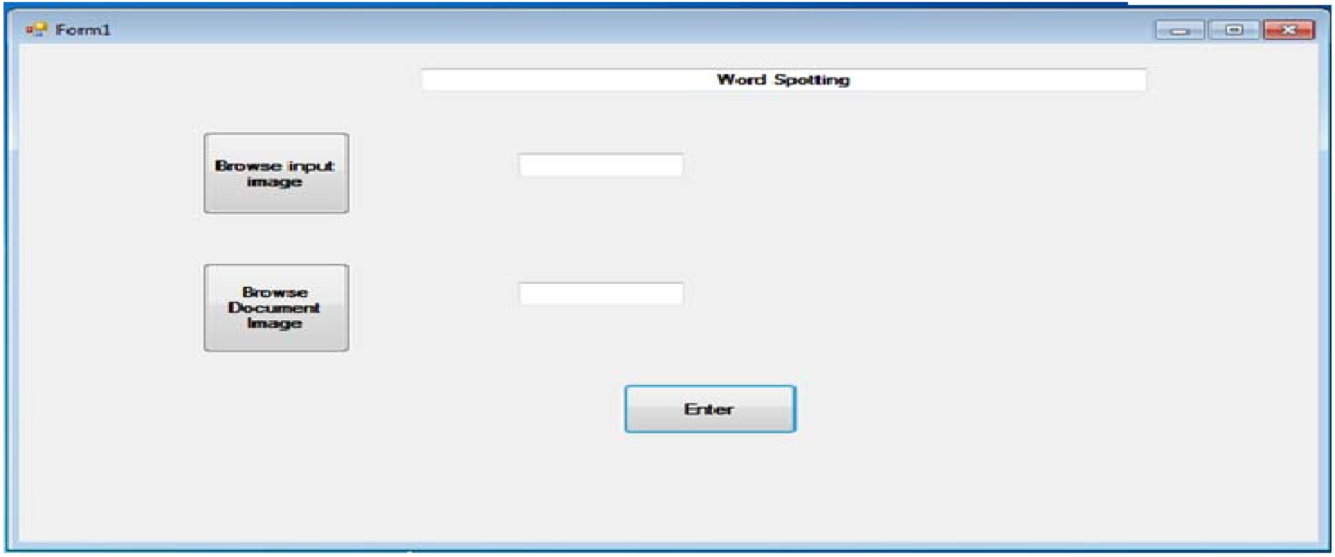
b. Extracting Word Images.

Printed documents have very well defined gaps in between text lines and in between words. These can be used to separate lines first and consequently extract the words. Horizontal profiling is a technique in which the image is traversed row wise and for each row the count of black pixels are stored. Values less than a predefined threshold(which may be zero also)gives us the white line separating two text lines. The same method can be used to extract words from a text line line giving us the coordinates of words.

c. Word Image Representation

In word spotting, profile-based features are widely used to represent historical handwritten [4,5,6] or printed [7,8] word images. The profile-based features are popularized by Rath and Manmatha [4], which were adopted to represent the word images of the Washington's manuscripts. A feature value is associated with each column of the word image, and the word image along with the writing direction is represented as a feature vector. The size of the feature vector equals to the width of the word image. Inspired by the previous work, five profile features have been selected for representing the word images in our technique.

Fig 6 Screen shot of the system



1. Horizontal projection profile: It can capture the distribution of foreground pixels along with writing direction. It is computed by summing the foreground pixels in each image row.

2. Left profile: The left profile of a word image is made of the distances from the left boundary to the nearest foreground pixels in each image row.

3. Right profile: The right profile of a word image is made of the distances from the right boundary to the nearest foreground pixels in each image row.

4. Vertical projection profile: It is computed by summing the foreground pixels in each image column.

Pruning is used to quickly determine whether a pair of images is either dissimilar or likely to match each other. In [1], pruning of word pairs based on the area and aspect ratio of their bounding boxes was performed. The idea is to require word images, which will later be compared, to have similar *pruning statistics* (e.g. area of bounding box).Thus irrelevant comparisons will not be carried out.

The DTW algorithm is the best approach [2,7] for matching word images represented by the profile based features. Dynamic time warping (DTW) is a time series alignment algorithm. It aims at aligning two sequences of feature vectors by warping the time axis iteratively until an optimal match (according to a suitable metrics) between the two sequences is found. The DTW-distance between two time series

$x_1, x_2 ...... x_m$

$y_1, y_2 ...... y_m$

is D(M,N), which we calculate in a dynamic programming approach using

$$D(i, j) = \min \begin{Bmatrix} D(i, j-1) \\ D(i-1, j) \\ D(i-1, j-1) \end{Bmatrix} + d(x_i, y_i) \qquad (2)$$

The particular choice of recurrence equation and “local” distance function d(·, ·) varies with the application. 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 smooth time warping.

Backtracking along the minimum cost index pairs (i,j) starting from (M,N) yields the DTW warping path.We use the *Sakoe-Chiba* band constraint [3] to ensure this path stays close to the diagonal of the matrix which contains the D(I,j). This way, pathological warpings that align a small portion in one sequence to a large portion in the other are avoided. A more detailed discussion of continuity constraints can be found in [4].Figure 4 shows the query word image taken from document image itself. Figure 5 shows the spotted words in the document image.

5. Implementation

We have developed a keyword retrieval system for the Devanagri document images upon the Microsoft .NET Framework. Feature vectors are stored in a generic array and other information including coordinates, width and height of word images are stored in the STL(standard template library) list. Opencv is integrated with Visual Studio 2010 software to get the inbuilt image processing utilities. System is programmed in C++. The CPU is Intel core i5 and the main memory size is 4GB. Figure 6 shows a screen shot of the system.

6. Conclusion

This paper described a keyword retrieval system for locating words in the Devanagari document images by word spotting. Firstly we created indexing of the word images. Here, each index term was a fixed-length feature vector. Five profile features were used to represent each word image. A ranked result of word images was returned through calculating similarities between the query word image and each word image of the collection. However, few word images were skipped because of the distortion due to improper scanning. So, our future work will be focused on the methods and technologies for improving the spotting efficiency. Query word image was selected from the document image itself but true potential of the system can be evaluated only when it is taken from other source or generated. Future work is inclined towards this aspect.

Acknowledgments

This project is supported by faculty member Dr. Jayadevan R ,Computer Engineering ,Army Institute of Technology,Pune.

References

[1] Manmatha, R., Han, C., Riseman, E.M., Croft, W.B.: Indexing handwriting using word matching. In: Proceedings of 1st ACMA keyword retrieval system 45 International Conference on Digital Libraries (ICDL), pp. 151–159 (1996).

[2] Rath, T.M., Manmatha, R.:Word spotting for historical documents.IJDAR **9**(2), 139–152 (2007).

[3] Rath, T.M., Manmatha, R.: Word image matching using dynamic time warping. In: Proceedings of 28th InternationalConference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 521–527 (2003).

[4] Rath, T.M., Manmatha, R.: Features for word spotting in historical manuscripts. In: Proceedings of 7th International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 218–222 (2003).

[5] Rabaev, I., Biller, O., El-Sana, J., Kedem, K., Dinstein, I.: Case study in Hebrew character searching. In: Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR), pp. 1080–1084 (2011).

[6] Aghbari, Z., Brook, S.: HAN manuscripts: a holistic paradigm for classifying and retrieving historical Arabic handwritten documents.Expert Syst. Appl. **36**(8), 10942–10951 (2009).

[7] Konidaris, T., Gatos, B., Ntzios, K., et al.: Keyword-guided word spotting in historical printed documents using synthetic data and user feedback. IJDAR **9**, 167–177 (2007).

[8] Abidi, A., Siddiqi, I., Khurshid, K.: Towards searchable digital Urdu libraries—a word spotting based retrieval approach. In: Proceedings of the 11th InternationalConference on Document Analysis and Recognition (ICDAR), pp. 1344–1348 (2011).