Content Based Retrieval Of Kannada Document Image From Multilingual Document Image

Mr.Nithya.E
Research Scholar
Dept. Of CSE
Dayanandasagar Colg Engg,
Bangalore-78

Dr. Rameshbabu.D.R
Professor and Head
Dept. Of CSE
DSCE, Bangalore-78

Tanuja.K
IV sem, Dept. Of CSE
Dr.AIT,Bangalore-56

***Abstract* – This paper presents a system for recognition and retrieval of relevant Kannada document image from large multilingual document image collection. We can achieve the effective recognition of language, searching and retrieval from a large collection of printed multilingual document images by matching image at word – level. For the representation of words, morphological operation is performed to get the height, width and coordination of each word in the document. Upon the features extracted from the document images, language is identified. On the given query image Fast Fourier Transformation is performed to determine the phase angle which is used to matching the words and retrieves the document. System level issues for retrieval are addressed in this.**

Index Terms—
Fast Fourier Transformation, Inverse Fast Fourier Transformation, Morphological operation, Phase based image matching. Multi-lingual Document, Language Identification, Horizontal Lines,Vertical Lines, Feature Extraction.

1. INTRODUCTION

A number of collections of historical prints, writings, manuscripts and books exist in Indian languages that need search options in images. The document images of such collections cannot be recognized accurately. For Example the Famous Sanskrit poet and dramatist of India "Kalidasa" Books are translated into many languages. Such collections can be made available to large communities through electronic media. There is a need for easy and efficient access to such collections. The search procedures available for text domain can be applied, if these document images are converted into textual representations using recognizers. However, it is an infeasible solution due to the unavailability of efficient and robust OCRs for Indian languages.

Addressing this problem, this paper proposes an efficient recognition for Kannada based on the visible features of the languages (Kannada, English, Hindi and Malayalam) from multilingual document images and retrieval algorithm using phase based image matching – an image matching technique using the phase components in Fast Fourier Transformation which determines the phase angle of input image and query image that helps in matching word for the retrieval of document. This approach is faster as it does not match the image pixel by pixel.

2. LITERATURE SURVEY

2.1 Literature Survey- Script identification

From the literature survey, it has been revealed that some amount of work has been carried out in script/language identification. Peake and Tan [7] have proposed a method for automatic script and language identification from document images using multiple channel (Gabour) filters and gray level co-occurrence matrices for seven languages: Chinese, English, Greek, Korean, Malayalam, Persian and Russian. Tan [2] has developed rotation invariant texture feature extraction method for automatic script identification for six languages: Chinese, Greek, English, Russian, Persian and Malayalam. In the context of Indian languages, some amount of research work on script/language identification has been reported [8, 10, 11, and 13]. Pal and Choudhuri [8] have proposed an automatic technique of separating the text lines from 12 Indian scripts (English, Devanagari, Bangla, Gujarati, Kannada, Kashmiri, Malayalam, Oriya, Punjabi, Tamil, Telugu and Urdu) using ten triplets formed by grouping English and Devanagari with any one of the other scripts. Santanu Choudhuri, et al. [3] have proposed a method for identification of Indian languages by combining Gabour filter based technique and direction distance histogram classifier considering Hindi, English, Malayalam, Bengali, Telugu and Urdu. Basavaraj Patil and Subbareddy [9] have developed a character script class identification system for machine printed bilingual documents in

English and Kannada scripts using probabilistic neural network. Pal and Choudhuri [10] have proposed an automatic separation of Bangla, Devanagari and Roman words in multilingual multiscript Indian documents. Nagabhushan et.al. [13] havep oposed a fuzzy statistical approach to Kannada vowel recognition based on invariant moments. Pal et. al. [12] have suggested a word-wise script identification model from a document containing English, Devanagari and Telugu text. Chanda and Pal [11] have proposed an automatic technique for word-wise identification of Devanagari, English and Urdu scripts from a single document. Spitz [18] has proposed a technique for distinguishing Han and Latin based scripts on the basis of spatial relationships of features related to the character structures. Pal et al. [19] have developed a script identification technique for Indian languages by employing new features based on water reservoir principle, contour tracing, jump discontinuity, left and right profile. Ramachandra et al. [20] have proposed a method based on rotation-invariant texture features using multichannel Gabor filter for identifying six (Bengali, Kannada, Malayalam, Oriya, Telugu and Marathi) Indian languages. Hochberg et al. [21] have presented a system that automatically identifies the script form using cluster-based templates. Gopal et al. [22] have presented a scheme to identify different Indian scripts through hierarchical classification which uses features extracted from the responses of a multichannel log-Gabor filter.

In [4], it is assumed that a given document should contain the text lines in one of the three languages Kannada, Hindi and English. In one of my previous papers [14], the results of detailed investigations were presented related to the study of the applicability of horizontal and vertical projections and segmentation methods to identify the language of a document considering specifically the three languages Kannada, Hindi and English. It is reasonably natural that the documents produced at the border regions of Karnataka may also be printed in the regional languages of the neighbouring states like Telugu, Tamil, Malayalam and Urdu. The system [4] was unable to identify the text words for such documents having text words in Telugu, Tamil, Malayalam, Urdu languages and hence these text words were misclassified into any one among the three languages, whichever is nearer and similar in its visual appearance. For example, Telugu is misclassified as Kannada and Tamil is misclassified as English. If the document consists of text words in other than the anticipated languagesthis algorithm fails to identify the type of the language by misclassifying the text words.

Keeping the drawback of the this method [15] in mind, we have proposed a system that would more accurately identify and separate language portions of Kannada by separating Hindi, English and Malayalam text from document images as our intension is to identify only Kannada from multilingual document images. The system identifies the Kannada with the help of knowledge base as our main aim is to focus only on Kannada.

2.2 Literature Survey- Retrieval

In this chapter, we look at the literature of indexing and retrieval techniques used for search in large image databases. The topic of interest overlaps with databases, pattern recognition, content based image retrieval, digital libraries, document image processing and information retrieval.

A number of approaches have been proposed in recent years for efficient search and retrieval of document images. There are essentially two classes of techniques to search document image collections. The first approach is to convert the images into text and then apply a search engine.

In recognition based search and retrieval techniques, the document images are passed through an optical character recognizer (OCR) to obtain text documents. The text documents are then processed by a text search engine to build the index. The text index makes the document retrieval efficient. An example is the gHMM approach of Chan et al. suggested for printed and handwritten Arabic documents. It uses gHMMs with a bi-gram letter transition model, and kernel principal component analysis (KPCA) / linear discriminant analysis (LDA) for letter discrimination. In this approach segmentation and recognition go hand in hand. The words are modeled at letter level, where the likelihood of a word given a segment is used for discriminating words. The Byblos system also uses a similar approach to recognize documents where a line is first segmented out and then divided in to image strips. Each line is then recognized using an HMM and a bi-gram letter transition model.

Taghva et al. built a search engine for documents obtained after recognition of images. Searching is done based on the results of similarity calculation between the query words and the database words. The indexed terms are divided into two groups of correctly recognized and incorrectly recognized words based on frequency calculations using a dictionary. Similar words are identified from the correct terms by applying mutual information measure. There have been attempts to retrieve complete documents (rather than searching words) by considering the information from word neighbourhood (like n-grams) to improve the search in presence of OCR errors. Word spotting is a method of searching and locating words in document images by treating a collection of documents as a collection of word images. The words are clustered and the clusters are annotated for enabling indexing and searching over the documents. It involves segmentation of each document into its corresponding lines and then into words. The word spotting approach has been extended to searching queried words from printed document images of newspapers and books. Dynamic time warping (DTW) based word-spotting

algorithm for indexing and retrieval of online documents is also reported.

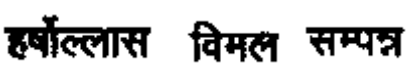
The remainder of the paper describes our current development effort in more detail. Section III describes the architecture of the research prototype we are developing. Section IV details the implementation evaluation procedures we are developing. Section V presents some experimental results. Section VI concludes the paper.

3. VISUAL DISCRIMINATING FEATURES OF KANNADA, HINDI, ENGLISH AND MALAYALAM TEXT WORDS

Feature extraction is an integral part of any recognition system. The aim of feature extraction is to describe the pattern by means of minimum number of features or attributes that are effective in discriminating pattern classes [13]. The new algorithms presented in this paper are inspired by a simple observation that every script/language defines a finite set of text patterns, each having a distinct visual appearance [1]. The character shape descriptors take into account any feature that appears to be distinct for the language [1] and hence every language could be identified based on its visual discriminating features. Presence and absence of the four discriminating features of Kannada, Hindi and English text words are given in Table-1.

3.1. Visual discriminating features of Hindi language

In Hindi language, many characters have a horizontal line at the upper part. This line is called sirorekha in Devanagari [8]. However, we shall call it as head-line. It could be seen that, when two or more characters sit side by side to form a word, the character head-line segments mostly join one another in a word resulting in only one component within each text word and generates one continuous head-line for each text word. Since the characters are connected through their head-line portions, a Hindi word appears as a single component and hence it cannot be segmented further into blocks, which could be used as a visual discriminating feature to recognize Hindi language. We can also observe that most of the Hindi characters have vertical line like structures. It could be seen that since two or more characters are connected together through their head-line portions, the width of the block is much larger than the height of the text line. Some typical Hindi words are given below:



3.2. Visual discriminating features of English language

It has been found that a distinct characteristic of most of the English characters is the existence of vertical line-like structures [8] and uniform sized characters with each characters having only one component (except “i” and “j” in lower-case).

3.3. Visual discriminating features of Kannada language

It could be seen that most of the Kannada characters have horizontal line like structures. Kannada character set has 50 basic characters, out of which the first 14 are vowels and the remaining characters are consonants [11]. A consonant combined with a vowel forms a modified compound character resulting in more than one component and is much larger in size than the corresponding basic character. It could be seen that a document in Kannada language is made up of collection of basic and compound characters resulting in equal and unequal sized characters [11] with some characters having more than one component, which could be expected to support in identifying the text words of Kannada language. Some typical Kannada words are given below:

3.3. Visual discriminating features of Malayalam language

In Malayalam language, many characters have a horizontal line. This could be used as a visual discriminating feature to recognize Malayalam language. We can also observe that most of the Malayalam characters have vertical line like structures. Some typical Malayalam words are given below:

Table-3.1 Presence and absence of discriminating features of Kannada, Hindi, English Malayalam text words.

(Yes means presence and No means absence of that feature. F1: Horizontal lines; F2: Vertical lines)

Discriminating Features / Text words	F1	F2
Kannada	Yes	No
Hindi	Yes	Yes
English	Yes	Yes
Malayalam	Yes	Yes

4. SUPPORTIVE KNOWLEDGE BASE FOR SCRIPT IDENTIFIATION

Knowledge base plays an important role in Recognition of any pattern and knowledge base is a repository of Derived information [14]. A supportive knowledge base is constructed for each specific class of patterns, which further helps during decision making to arrive at a conclusion. In the present method, the percentage of the presence of the four features for each text of the four languages Kannada, Hindi, English and Malayalam are practically computed using sufficient data set. Based on the experimental results, a supportive knowledge base is constructed to store the percentage of the presence of the visual features. The technique of obtaining the visual features from the input image through experimentation is explained below:

Feature 1-Horizontal lines: In the binary image of each text line, if there are continuous one's in a row greater than the horizontal threshold value (Horizontal threshold value is calculated for each text line. Horizontal threshold value = 75% of the X-height of that text line), then such continuous one's are retained resulting in a horizontal line and if there are no continuous one's greater than the horizontal threshold value, then such one's are changed to zeroes. A component has a horizontal line-like structure, if a black run length (sequence of continuous one's) of that component is greater than the horizontal threshold value of that text line.

Feature 2- Vertical lines: In the binary image of each text line, if there are continuous one's in a column greater than the vertical threshold value (Vertical threshold value is computed for each text line. Vertical threshold value = X-height of that text line) then such continuous one's are retained resulting in a vertical line and if there are no continuous one's greater than the vertical threshold value, then such one's are changed to zeroes. A component has a vertical line-like structure, if a black run length (sequence of continuous one's) of that component is greater than vertical threshold of the text line.

The percentage of the spatial occurrence of two visual features for each of the four languages are practically computed through extensive experimentation and stored in the knowledge base for later use during decision-making.

5. SYSTEM ARCHITECTURE

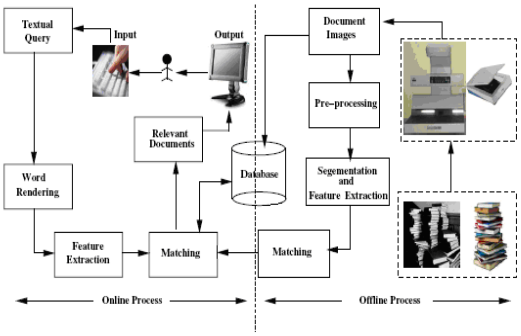


Fig (5.1) System Architecture

This system accepts a textual query from users. The textual query is first converted to an image by rendering, features are extracted from these images and then recognition of Kannada language and a search is carried out for retrieval of relevant multilingual documents. Results of the search are pages from document image collections containing the retrieved words sorted based on their relevance to the query. This work mainly aims at addressing some of the issues involved in effective and efficient retrieval in document images with effective representations of the word images.

6. IMPLEMENTATION

An efficient mechanism for retrieval of a Kannada word from a large multilingual document image collection is presented in this thesis. This involves three phases: first phase includes pre processing, which is preparing the source image for recognition of Kannada language and searching the query word, second phase includes generating the query image from the query word, and third phase includes matching of images to find the query word in the source image.

In first phase, consider a source image, which is generally in form of RGB. First, the source image, which is in the RGB form, is converted to the grey scale image. Then, this grey scale image is, in turn, converted to the binary image, i.e. image will be in 0's and 1's form, with 0 representing black and 1 representing white. This process of conversion helps in performing morphological operation. Morphological operation is considered as repeated dilations of an image, called marker image, until the contour of the marker image fits under a second image, called the mask image. Morphological operation is performed to initiate the dilation, which helps in differentiating two words delimited by a space. Then, fill the holes to find any picture in the image and remove the big area, which might be, say, a photograph. Then identify the Kannada language by extracting vertical and horizontal line features from the multilingual image documents from all the four

languages. Then, record the coordinates, height and width of each word in image document.

Second phase includes generating the query image from the query word. First, read the text of English word. Then, find the corresponding Kannada letter image from the database. Next, align the thus got letter images, which form a Kannada word. Convert the query image to binary image. This helps in comparison of input image with the query image.

Third phase includes matching of images to find the query word in the source image. The method used is Phase – based image matching [1]. This phase has input as two images, source image and query image, which are converted into binary form. First, Fast Fourier transform is performed on both the images and phase angle of both the images are determined. Then, subtract the phase angle of first image with that of the second image. Inverse Fast Fourier transform is performed on thus obtained phase difference. If the highest value of IFFT result is more than the threshold then words are matching, else the words are not matching.

6. EXPERIMENTAL RESULTS

Figure (5.1) shows the input to the system and figure (5.2) and figure (5.3) shows the output. Since we are using database approach for the character recognition, in this approach for each character we need to have details like Character name, Character BMP image. This takes lot of space as well as lot of computation involved in recognizing the character.

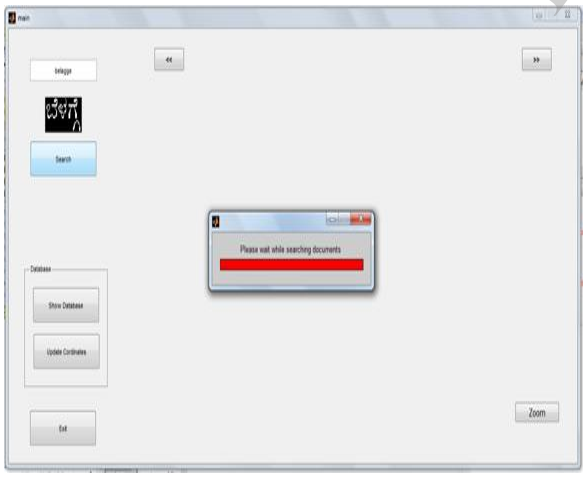


Fig (5.1) Input to the system

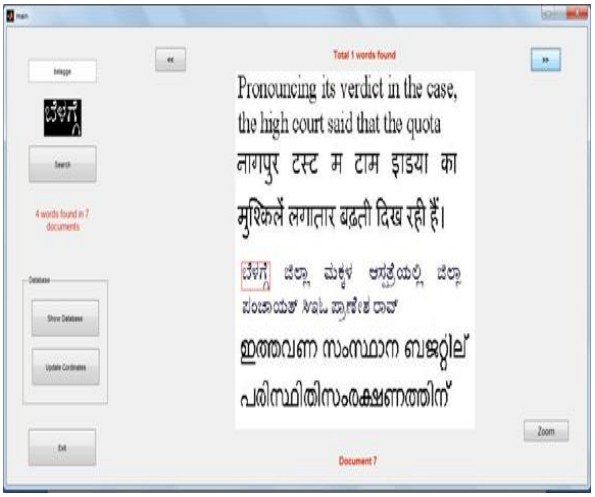


Fig (5.2) Result of search

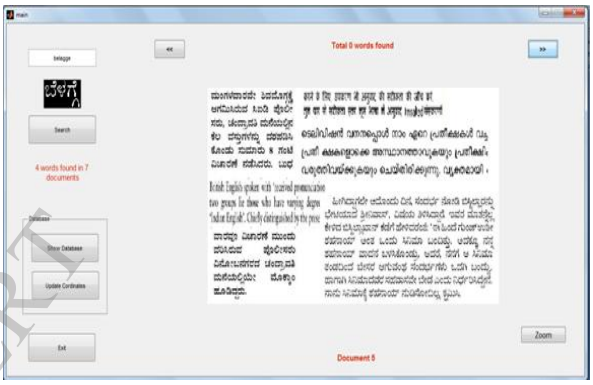


Fig (5.3) Result of search

CONCLUSION

In this paper we have presented recognition and retrieval in large document image collections. The recognition based on visual features and matching technique based on phase - based image matching, for search in large collections of document word images is applied to obtain good performance. The approaches used for word spotting so far, dynamic time warping and/or nearest neighbour search tend to be slow for large collection of books. Direct matching of pixels in images is inefficient due to the complexity of matching and thus impractical for large databases. This problem is solved by directly storing word image representations.

Some of the possible directions in which the future work can be carried out are as below. The effect of combination of different fonts in a single collection can be one possible direction for exploring the feasibility of the proposed technique and improving it.

Reference:

1. P.Naghabhushan, Radhika M Pai, "Modified Region Decomposition Method and Optimal Depth Decision Tree in the Recognition of non-uniform sized characters – An Experimentation with Kannada Characters", *Journal of Pattern Recognition Letters*, 20, 1467-1475, (1999).

2. T.N.Tan, "Rotation Invariant Texture Features and their use in Automatic Script Identification", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(7), 751- 756, (1998).

3. Santanu Choudhury, Gaurav Harit, Shekar Madnani, R.B. Shet, "Identification of Scripts of Indian Languages by Combining Trainable Classifiers", ICVGIP 2000, Dec., 20-22, Bangalore, India.

4. M.C.Padma, P.Nagabhushan, "Horizontal and Vertical linear edge features as useful clues in the discrimination of multiligual (Kannada, Hindi and English) machine printed documents", *Proc. National Workshop on Computer Vision, Graphics and Image Processing (WVGIP)*, Madhurai, 204-209, (2002).

5. U.Pal, B.B.Choudhuri, "OCR in Bangla:an Indo-Bangladeshi language", *IEEE, no.2*, 1051-4651, (1994).

6. U.Pal, B.B.Choudhuri, "An OCR system to read two Indian language scripts:Bangla and Devanagari(Hindi)", *Proc. 4th ICDAR*, Uhn, 18-20, (1997).

7. G.S. Peake, T.N.Tan, "Script and Language Identification from Document Images", *Proc. Eighth British Mach. Vision Conference.*, 2, 230-233, (1997).

8. U.Pal, B.B.Choudhuri, "Script Line Separation From Indian Multi-Script Documents", *Proc. 5th International Conference on Document Analysis and Recognition(IEEE Comput. Soc. Press)*, 406-409, (1999).

9. S.Basvaraj Patil, N.V.Subba Reddy, "Character script class identification system using probabilistic neural network for multi-script multi lingual document processing", *Proc. National Conference on Document Analysis and Recognition*, Mandya, Karnataka, 1-8,

10. U.Pal B.B.Choudhuri, "Automatic Separation of Words in Multi Lingual multi Script Indian Documents", Proc. 4th *International Conference on Document Analysis and Recognition*, 576-579, (1997).

11. S.Chanda, U.Pal, "English, Devanagari and Urdu Text Identification", *Proc. International Conference on Document Analysis and Recognition*, 538-545, (2005).

12. U.Pal, S.Sinha, B.B.Choudhuri, "Word-wise script identification from a document containing English, Devanagari and Telugu text", *Proc. 2nd National Conference on Document Analysis and Recognition*, Karnataka, India, 213-220, (2003).

13. P.Nagabhushan, S.A.Angadi, B.S.Anami, "A Fuzzy Statistical Approachto Kannada Vowel Recognition based on Invariant Moments", *proc. 2nd National Conference, NCDAR*, Mandya, 275-285, (2003).

14. M.C.Padma, P.Nagabhushan, "Study of the Applicability of Horizontal and Vertical Projections and Segmentation in Language Identification of Kannada, Hindi and English Documents", *Proc. National Conference NCCIT*, Kilakarai, Tamilnadu, 93-102, (2001).

15. M.C.Padma, P.Nagabhushan, "Identification and separation of text words of Kannada, Hindi and English languages through discriminating features", *Proc. 2nd National Conference on Document Analysis and Recognition*, Mandya, Karnataka, 252-260, (2003).

16. U.Pal, B.B.Choudhuri, "Automatic Identification of English, Chinese, Arabic, Devanagari and Bangla Script Line", *Proc. 6th International Conference on Document Analysis and Recognition*, 790-794, (2001).

17. R.C.Gonzalez, R.E.Woods, *Digital Image Processing* Pearson Education Publications, India, 2002.

18. A.L.Spitz, "Determination of the Script and language Content of Document Images", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, 235-245, 1997.

19. U.Pal, S.Sinha, B.B.Choudhuri, "Multi-Script Line Identification from Indian Documents", *Proc. 7th International Conference on Document Analysis and Recognition (ICDAR 2003)* vol. 2, 880-884, 2003.

20. Ramachandra Manthalkar and P.K. Biswas, "An Automatic Script Identification Scheme for Indian Languages", NCC, 2002.

21.J.Hochberg, P.Kelly, T.Thomas, L.Kerns, "Automatic Script Identification from Document Images using Cluster –based Templates", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 176-181, 1997. Gopal Datt Joshi, Saurabh Garg, Jayanthi Sivaswamy, "Script Identification from Indian Documents", *DAS 2006, LNCS 3872*, 255-267, 2006.

22. Koichi Ito, Ayumi Morita, Takafumi Aoki Tatsuo Higuchi, Hiroshi Nakajima, and Koji Kobayashi, A Fingerprint Recognition Algorithm Using Phase-Based Image Matching for Low-Quality Fingerprints

23. Ashwin T V 2000 *A font and size independent OCR for printed Kannada using SVM*. M E Project Report, Dept. Electrical Engg., Indian Institute of Science, Bangalore

24. A. Balasubramanian, Million Meshesha, and C.V. Jawahar Retrieval from Document Image Collections Centre for Visual Information Technology, International Institute of Information Technology, Hyderabad - 500 032, India