Thanuja C, Shreedevi G R / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 4, Jul-Aug 2013, pp.1329-1335

Content Based Image Retrieval System for Kannada Query Image from Multilingual Document Image Collection

Thanuja C[1], Shreedevi G R[2]
[1, 2] (Department of Information Science, Sambhram institute of technology, Bangalore-97

ABSTRACT

In a multilingual country like India, a document may contain text words in more than one language. 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. An electronic Searching system for Kannada text, based on the content is needed to access such multilingual documents. So, it is necessary to identify different language portions of the document before the retrieval. The objective of this paper is to propose visual clues based procedure to identify Kannada text from a multilingual document which contains Hindi, English and Malayalam text portions along with Kannada.

Keywords **-** Content based image retrieval, Correlation Coefficients, Feature Extraction, Multi-lingual Document image, Script Identification.

I. INTRODUCTION

Large electronic collections of historical prints, writings, manuscripts and books exist in Indian languages that needs search options in images. The objective of language identification is to translate human identifiable documents to machine identifiable codes. For Example the heritage inscriptions are being digitized which may contain more than one language. Such collections can be made available to large communities through electronic media.

Identification of the language in a document image is of primary importance for retrieval. Language identification may seem to be an elementary and simple issue for humans in the real world, but it is difficult for a machine, primarily because different scripts (a script could be a common medium for different languages) are made up of different shaped patterns to produce different character sets. A document containing text information in more than one language is called a multilingual document. For such type of multilingual documents, it is very essential to identify the text language portion of the document, before the retrieval. Language identification is one of the vision application problems. Generally human system identifies the language in a document using some visible characteristic features such as texture, horizontal lines, vertical lines, which are visually perceivable and appeal to visual sensation. This human visual perception capability has been the motivator for the development of the proposed system. With this context, in this paper, an attempt has been made to simulate the human visual system, to identify the type of the language based on visual clues, without reading the contents of the document. There is a need for easy and efficient access to such documents.

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, the paper proposes an efficient recognition and retrieval of Kannada language from multilingual document images, based on the visual features of the text.

II. RELATED WORK

2.1 SCRIPT IDENTIFICATION

From the literature survey, it is evident that some amount of work has been carried out in script identification. Peake and Tan [1997] 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 [1998] 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 language identification has been reported [1997, 1997, 2005, and 2003]. Pal and Choudhuri [2001] 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. [2000] 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 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 [1997] have proposed an automatic separation of Bangla, Devanagari and Roman words in multilingual multiscript Indian documents. Nagabhushan et.al. 20013] havep oposed a fuzzy statistical approach to Kannada vowel recognition based on invariant moments. Pal et. al. [2005] has suggested a word-wise script identification model from a document containing English, Devanagari and Telugu text. Chanda and Pal [2005] have proposed an automatic technique for word-wise identification of Devanagari, English and Urdu scripts from a single document. Spitz [1997] 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. [2003] 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. [2002] 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. [2006] have presented a system that automatically identifies the script form using cluster-based templates. Gopal et al. 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.

Having all these 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 a knowledge base as the main aim is to focus only on Kannada text.

2.2 CONTENT BASED IMAGE RETRIEVAL

In this section, 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, and 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.

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. 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 3 demonstrates visible feature of the four languages we have considered for experiment. Section 4 gives the overall system architecture. Section 5 briefs on word level segmentation. Section 6 describes the supportive knowledge base for script identification. Section 7 details the implementation and experimental results of the developed system. And Section 8 concludes the paper.

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

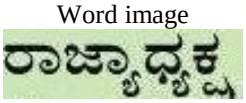
Feature extraction is an integral part of any recognition system. The aim of feature extraction is to describe the pattern by means of minimum number of features or attributes that are effective in discriminating pattern classes. The new algorithms presented in this paper are inspired by a simple observation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

3.1. FEATURES OF KANNADA TEXT

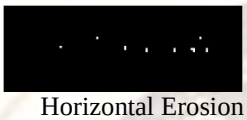
It could be seen that most of the Kannada characters have horizontal line like structures.

Kannada character set has 50 basic characters, out of which the first 14 are vowels and the remaining characters are consonants. 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 with some characters having more than one component. Typical Kannada word with vertical and horizontal lines after erosion is given below.

Word image



Vertical Erosion



Horizontal Erosion

3.2 FEATURES OF HINDI TEXT

In Hindi, many characters have a horizontal line at the upper part. This line is called sirorekha in Devanagari. 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. Typical Hindi word with vertical and horizontal lines after erosion is given below.

Word image

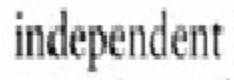
Vertical Erosion

Horizontal Erosion

3.3. FEATURES OF ENGLISH TEXT

It has been found that a distinct characteristic of most of the English characters is the existence of vertical line-like structures and uniform sized characters with each characters having only one component (except "i" and "j" in lower-case). Typical English word with vertical and horizontal lines after erosion is given below.

Word image



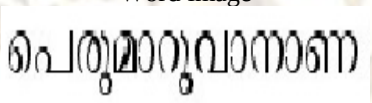
Vertical Erosion

Horizontal Erosion

3.4. FEATURES OF MALAYALAM TEXT

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. Typical Malayalam word with vertical and horizontal lines after erosion is given below.

Word image



Vertical Erosion

Horizontal Erosion

Table-1: Presence and absence of discriminating features of Kannada, Hindi, English Malayalam text words.

I. LANGUAGE FEATURE → / ↓	**Horizontal lines**	**Vertical lines**
Kannada	Yes	No
Hindi	Yes	Yes
English	Yes	Yes
Malayalam	Yes	Yes

(Yes means presence and No means absence of that feature)

IV. SYSTEM ARCHITECTURE

Figure 1 is the architecture of the system which accepts a textual query from users. The textual query is first converted to an image by

rendering, features are extracted from images and then recognition of Kannada language and a search is carried out for retrieval of relevant multilingual documents.

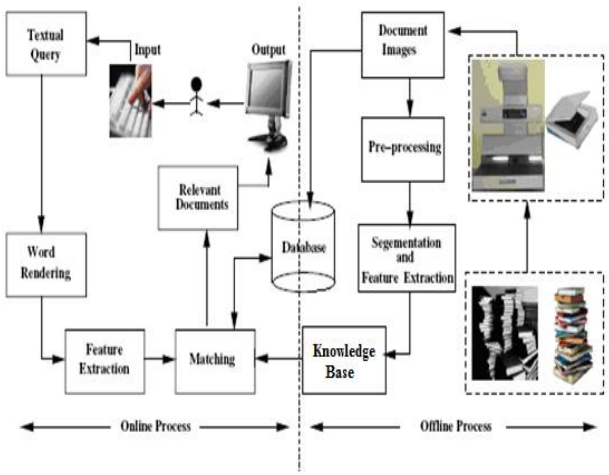


Fig 1. System Architecture

V. SEGMENTATION

The result of pre-processing and dilation is the connected components which makes the characters of the words to connect as a single group of pixels. These single groups of pixels are treated as a single word and segmented. Document's word images marked for segmentation is shown in Fig. 2.

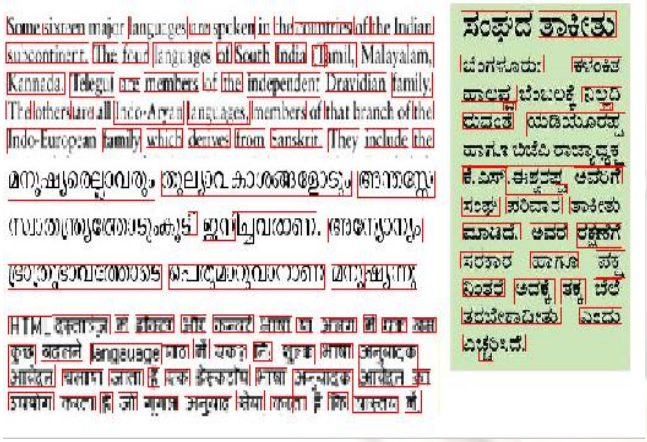


Fig 2. Segmentation

VI. SUPPORTIVE KNOWLEDGE BASE FOR SCRIPT IDENTIFICATION

Knowledge base plays an important role in Recognition of any pattern and knowledge base is a repository of Derived information .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 vertical line and horizontal line density of segmented word images of the four languages- Kannada, Hindi, English and Malayalam- are practically computed using sufficient data set. Erosion is used to obtain the features of the languages. Based on the experimental results, a supportive knowledge base is constructed considering the density of the vertical and horizontal lines of each language text words. The density of two visual features for each word image for the four languages are practically computed through extensive experimentation and stored in the knowledge base for later use during decision-making. The given below is a summary plot of the values from the knowledgebase.

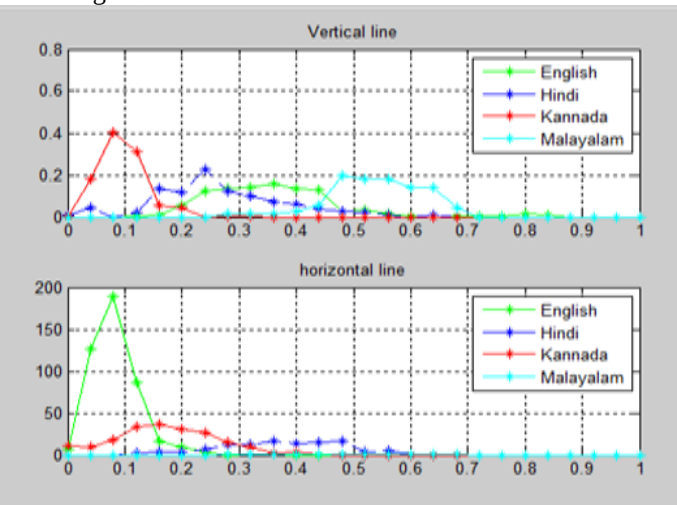


Fig 3. Plot generated from Knowledgebase

VII. IMPLEMENTATION

System accepts a textual query from users. Results of the search are pages from document image collections with the query word being highlighted.

An efficient mechanism for retrieval of a Kannada word from a large multilingual document image collection is presented in this paper. This involves i) Pre-Processing ii) Query image formulation and iii) Matching and retrieval.

7.1 PRE-PROCESSING

Pre-Processing involves preparing the source image for recognition of Kannada language. The source image is converted to the binary image. This process of conversion helps in performing morphological operation. Morphological operation is considered as repeated dilations of an image. Dilation helps in differentiating two words delimited by a space. 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 of each word in image document.

7.2 QUERY IMAGE

Query image has to be formulated from the query word given as input to the system. English text entered, is translated to Kannada and converted as image by rendering. Convert this query image to binary image. This helps in comparison of input image with the query image.

7.3 MATCHING AND RETRIEVAL

This is the stage where the documents matching with the search criteria are retrieved. Correlation coefficient of images is used for matching query word with source image. If the matching score of query and source image is more than the threshold then words are matching, and the word will be highlighted in the document.

7.4. CORRELATION MATCHING

Correlation coefficient between two variables (image matrix)) is defined as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables.

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and may be referred to as the population correlation coefficient or the population Pearson correlation coefficient. The formula for ρ is:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Pearson's correlation coefficient when applied to a sample is commonly represented by the letter r and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient. We can obtain a formula for r by substituting estimates of the covariances and variances based on a sample into the formula above. That formula for *r* is:

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

An equivalent expression gives the correlation coefficient as the mean of the products of the standard scores. Based on a sample of paired data (X_i, Y_i), the sample correlation coefficient is

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right)$$

where

$$\frac{X_i-\bar{X}}{s_X}, \bar{X}, \text{ and } s_X$$

are the standard score, sample mean, and sample standard deviation, respectively.

The absolute value of both the sample and population Pearson correlation coefficients are less than or equal to 1. Correlations equal to 1 or -1 correspond to data points lying exactly on a line, or to a bivariate distribution entirely supported on a line. A key mathematical property of the correlation coefficient is that it is invariant (up to a sign) to separate changes in location and scale in the two variables. That is, we may transform *X* to *a* + *bX* and transform *Y* to *c* + *dY*, where *a*, *b*, *c*, and *d* are constants, without changing the correlation coefficient (this fact holds for both the population and sample Pearson correlation coefficients). Note that more general linear transformations do change the correlation.

The Pearson correlation can be expressed in terms of uncentered moments. Since $\mu_X = E(X)$, $\sigma_X^2 = E[(X - E(X))^2] = E(X^2) - E^2(X)$ and likewise for *Y*, and since

$$E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y),$$

The correlation can also be written as

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - (E(X))^2}\sqrt{E(Y^2) - (E(Y))^2}}.$$

Alternative formulae for the sample Pearson correlation coefficient are also available:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

The correlation coefficient ranges from −1 to 1. A value of 1 implies that a linear equation describes the relationship between *X* and *Y* perfectly, with all data points lying on a line for which Y increases as *X* increases. A value of −1 implies that all data points lie on a line for which *Y* decreases as *X* increases. A value of 0 implies that there is no linear correlation between the variables.

More generally, note that $(X_i - X)(Y_i - Y)$ is positive if and only if X_i and Y_i lie on the same side of their respective means. Thus the correlation coefficient is positive if X_i and Y_i tend to be simultaneously greater than, or simultaneously less than, their respective means.

For uncentered data, the correlation coefficient corresponds with the cosine of the angle φ between both possible regression lines $y=g_x(x)$ and $x=g_y(y)$. For centered data (i.e., data which have been shifted by the sample mean so as to have an average of zero), the correlation coefficient can also be viewed as the cosine of the angle θ between the two vectors of samples drawn from the two random variables (see Fig.4).

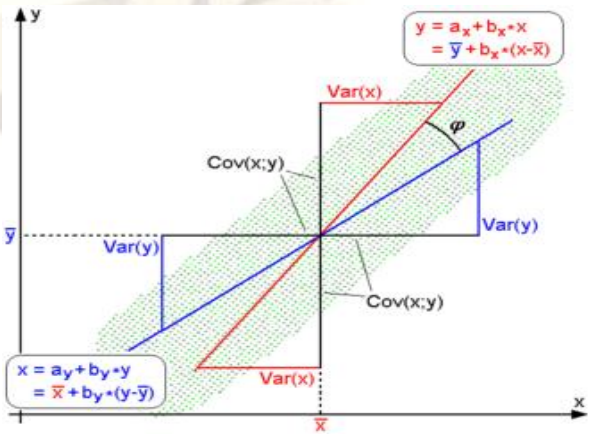


Fig 4. Regression lines for $y=g_x(x)$ [red] and $x=g_y(y)$ [blue]

Both the uncentered (non-Pearson-compliant) and centered correlation coefficients can be determined for a dataset.

The experimental results of the proposed system are given below. Figure 5 shows script identification based on the knowledge base. Only the Kannada text has been marked for searching. After script identification, the query word will be searched only in this highlighted area, which is Kannada text.

Fig 5. Script identification

Figure 6 shows the final result of the search. The Query word "ARJI" is being highlighted in the document, and this document will be given to the user as a result.

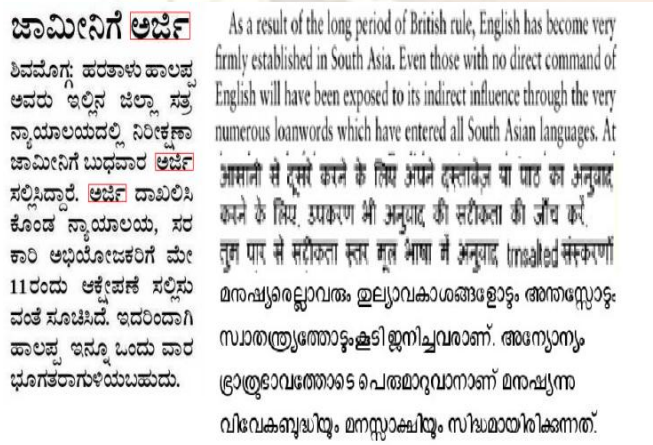


Fig 6. Result of search

VIII. CONCLUSION

In this paper, we have presented word wise identification models to identify Kannada text words from Indian multilingual machine printed documents which also contains Hindi, English and Malayalam text words . The proposed method is developed based on the visual discriminating features, which serve as useful visual clues for language identification. The methods help to accurately identify and separate language portions of Kannada from Hindi, English and Malayalam. The experimental results show that the method effectively identifies and separates the Kannada language portions of the document, which further helps in document image retrieval.

The system can be enhanced in number of directions. One can work on combination of different fonts in a single documents collection. Searching and retrieval in documents with more number of languages which has same kind of features is challenging. The system can also be enhanced to character level identification.

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] 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

[3] 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.

[4] 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

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

[6] 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.

[7] 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

[8] 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).

[9] 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).

[10] 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).

[11] 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).

[12] R.C.Gonzalez, R.E.Woods, Digital Image Processing Pearson Education Publications, India, 2002.

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

[14] 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,

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

[16] 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.

[17] 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).

[18] 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).

[19] 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).

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

[21] 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).

[22] 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.

[23] 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).

[24] 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).