Dr. M Chandwani is a Director in Institute of Engineering and Technology, D.A.V.V., Indore, Madhya Pradesh.
E-mail : chandwanim@rediffmail.com

A New Contour Based Invariant Feature Extraction Approach for the Recognition of Multi-lingual Documents

Manjunath Aradhya V N Hemantha Kumar G Shivakumara P Noushath S

Abstract

Now a day, developing a single OCR system for recognizing multi-lingual documents becomes essential to enhance the ability and performance of the existing document analysis system. Hence in this paper, we present a new technique based on contour detection and distance measure for recognizing multi-lingual characters comprising south Indian languages (Kannada, Tamil, Telugu, Malayalam, English Upper case, English Lower case, English Numerals and Persian Alphanumeric). Proposed method finds boundary for a character using contour detection and the result of contour detection is given to feature extraction scheme to obtain distinct and invariant features for identifying different characters of different languages. The method extracts invariant features by computing distance between the centroid and the pixels of contour of character image.

We compare the experimental results of proposed method with result of existing methods to evaluate the performance of the method. Based on experimental results it is realized that the proposed method gives 100% accuracy with minimum expense and time. In addition, the method is invariant to Rotation, Scaling and Translation transformations (RST).

Keywords : Contour detection, Distance Measure, Invariant features, Character recognition, OCR

0. Introduction

In some situations like border places of state or countries and places where the different people meet together, the document may contain different languages. To understand such documents, there are methods in literature called hybrid OCR system. To build hybrid OCR we need to have different OCR systems for different languages. This is time consuming and it is not economically feasible. In addition to this, the system suffers from the following drawbacks. The methods fail to segment the words from the line containing different words of different languages. The method also fails to segment the different characters of different languages present in a single word. However, this kind of document is obvious in the place of railway station where we use the reservation form to reserve the seat, advertisements and any label of the product released by company.

Hence there is a necessity of developing novel technique to meet the above requirements. Therefore in this paper we present novel concept of single OCR for recognizing the multi-lingual documents. The proposed method involves only feature extraction scheme, which identifies the language through character recognition. The advantage of this concept is that it requires less computations, time and complexity compared to hybrid OCR system where it has different schemes for recognizing different languages and segmentation algorithm for segmenting the languages from single document.

1. Related Literature

In this section, we give the related literature for recognizing the characters of different languages.

(Pal .U and Chaudhuri. B.B, 2004) have proposed a review of the OCR work done on Indian Language Scripts. They discussed different methodologies applied in OCR development in International and national scenario. However, in this paper they have not addressed the problem of Indian languages like Kannada, Tamil etc.

(Pal .U and Chaudhuri. B.B, 2002) have proposed technique to identify different script lines from multi-script documents. In this paper, they have addressed the problem of development of an automatic technique for the identification of printed Roman, Chinese, Arabic, Devangari and Bangla text lines from a single document. However, the method works only at text line level but not word level and character level. In addition method fails to identify the text lines of south Indian language documents since the structure of the ext lines is almost similar.

(Pal .U and Chaudhuri. B.B, 2001) have proposed method to identify the machine printed and hand written text lines in the single document. In this paper, they have presented a machine-printed and hand-written text classification scheme for Bangla and Devangari, the two most popular Indian scripts. However, the method works for only two languages and the method fails to identify the south Indian languages.

(Pal .U et al, 2003) have introduced water reservoir concept to segment the touching numerals. In this paper, they have developed a new technique for automation segmentation of unconstrained handwritten connected numerals. The method fails to identify the characters as the number of character increase. In addition, the method has given accuracy about 94.8%.

(Chew Lim Tan et al, 2002) have proposed a method for image document text without OCR. Documents are segmented into character objects. Image features namely, the Vertical Traverse Density (VTD) and Horizontal Traverse Density (HTD), are extracted. An n-gram based document vector is constructed for each document based on these features. The method is language independent. The method works particularly if document images are of similar fonts and resolution such as in a corpus of newspaper.

(Pal .U et al, 2000) have proposed a technique to deal with an OCR (Optical Character recognition) error detection and correction technique for a highly inflectional language, Bangla, the second-most popular language in India and fifth-most popular in world. The technique is based on morphological parsing. The method is limited to only Bangla characters.

(Nagabhushan. P and Radhika M. Pai, 1999) have proposed modified region decomposition method and optimal depth tree in the recognition of non-uniform sized characters. However, the method is limited to Kannada characters only. In addition, the method is found to be computationally expensive.

(Masayoshi Okamoto and Kazuhiko Yamamoto, 1999) have proposed an on-line character recognition method that simultaneously uses both directional features, otherwise Known as off-line features, and direction-change features, which designed as on-line features. The method works for online character recognition. The method fails for Kannada characters.

(Anil K. Jain et al, 1995) have given survey on Feature extraction methods for character recognition. They have given an overview of feature extraction methods for off-line recognition of segmented (isolated) characters. We have found that no algorithms are reported in the paper to recognize the characters of south Indian languages.

(Rejean Plamondon and Sargur N. Srihari, 2000) have given survey on online and offline hand written recognition. They have described the nature of handwritten languages, how it is transduced into electronic data, and the basic concept behind written language recognition algorithms. The method works for only English characters.

(Hemantha Kumar et al, 2004) proposed a method based on construction of concentric rings for the recognition of Malayalam characters. This method draws concentric circles on the character and extracts feature values such as number of black and white pixels from each of the rings drawn. The method is invariant to rotation but variant to scaling. Further, the method requires more than 6 features to recognize the characters of different languages. Hence the method becomes computationally expensive.

(Hemantha Kumar et al, 2003) proposed a method based on distance measure and directional codes for the recognition of alphanumeric characters. The method works based on Euclidean distance measure and City-block distance measure for both thick and thinned characters. However, the method is invariant to rotation of 45-degree inclinations but it is variant to scaling. Further, the method is said to computationally expensive since the method involves two features.

Further (Hemantha Kumar et al, 2004) proposed method based on Polar Transformation. This method maps the spatial coordinate of the image to polar coordinates of polar domain. The features are extracted by counting Number of black and white pixels in each ring. The method is invariant to RST Transformation but it has less accuracy.

From the above discussion, it is revealed that to the best of our knowledge no single algorithm is reported for the recognition of characters of multi-languages.

In this paper, we introduced novel concept of single OCR system based on contour detection and distance measure to recognize the multi-lingual documents. The proposed method has three stages. In first stage, we develop an algorithm to obtain contour (boundary) of a character image. New invariant feature extraction scheme for identifying the different contours of different languages results in second stage. The contour is obtained by performing the mask designed. The features are extracted for each contour by computing distance between the centroid and black pixels of the contour. The square of number of black pixels in the contour divided by the sum of all distances gives invariant feature for each contour. In the last stage, we have designed database using linear search tree and binary search tree to study the performance of the proposed method.

The rest of the paper is organized as follows: Section 3 discusses the proposed methodologies. Section 4 presents the experimental results. Section 5 and 6 presents Comparative study and Conclusions respectively.

2. Proposed Methodology

This section presents a technique for developing single recognition system for multi-lingual documents. Proposed methodology is divided into two sub sections. Algorithm for boundary detection using contour is presented in section 3.1. A new feature extraction scheme is introduced in section 3.2 to obtain invariant features for recognizing multi-lingual documents. In this work, we have considered the following data set for designing the single OCR for six languages including English upper case letters, lower case letters and numerals (ref. Fig 1 – Fig 8). In this method, we have assumed that the characters are in isolated form.

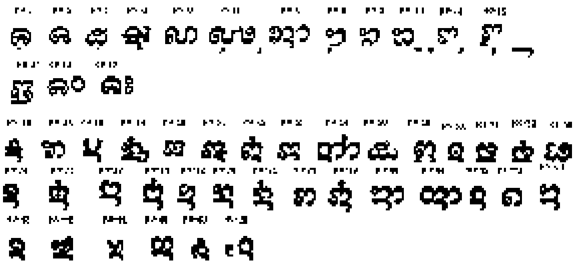


Fig. 1 Fifty Alphabets of Kannada Language

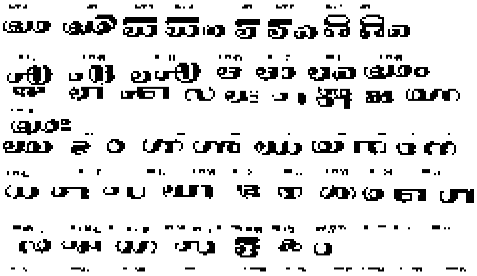


Fig. 2 Fifty-Two Alphabets of Malayalam Language

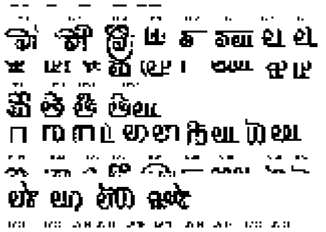


Fig. 3 Thirty-Five Alphabets of Tamil Language

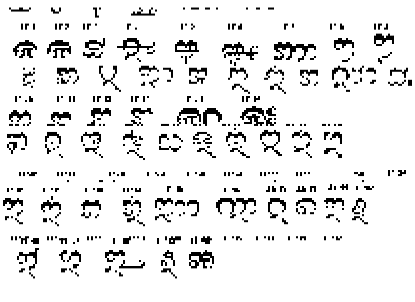


Fig. 4 Fifty Alphabets of Telugu Language

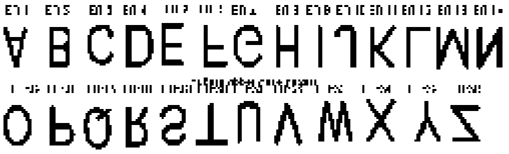


Fig. 5 Twenty-Six Alphabets of English Upper Case Letters

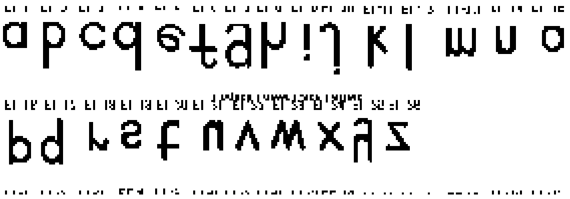


Fig. 6 Twenty-Six Alphabets of English Lower Case Letters

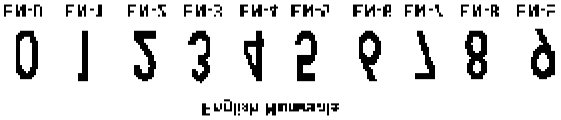


Fig. 7 Ten Numerals of English Language

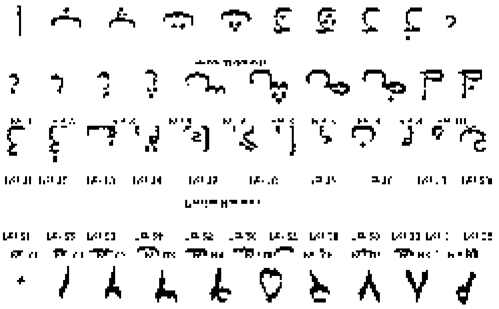
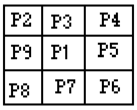


Fig. 8 Forty-Two Letters of Persian Language

2.1 Contour Detection (CD)

In this sub section, we present a new preprocessing contour detection technique to obtain boundary for a character. This technique reduces the number of black pixels of width of the character boundary. Moreover it works even width of boundary of character varies. In order to obtain boundary for a character we have designed 3X3 mask, which is given in Fig. 9, where P_2, P_3, P_4, P_5, P_6, P_7, P_8 and P_9 are the eight neighbor pixels of P_1 pixel. The technique assumes that the 1 (white) represents background color and 0 (Black) represents foreground color.



P2	P3	P4
P9	P1	P5
P8	P7	P6

Fig 9 Mask used to find contour

The technique deletes border pixel when P_2, P_3, P_4, P_5, P_6, P_7, P_8 and P_9 are 0. That is the center pixel (P_1) changes into 1 when the above condition is satisfied. This condition is satisfied not at the inside and outer boundary of the character image. Hence it removes pixel, which are present inside the character image. This procedure is repeated until no further changes in the boundary of a character. If the condition is not satisfied then the mask move to next pixel of the character image. The result of the technique is shown in Fig. 10.

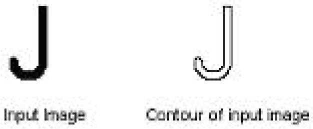


Fig 10 Result of contour detection

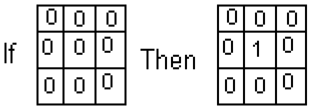
Algorithm: Contour Detection (CD)

Input: Character image

Output: Contour Image

Method Begins

Step 1: For each pixel of the character image, employ the following rule.



If

0	0	0
0	0	0
0	0	0

Then

0	0	0
0	1	0
0	0	0

Step 2: Repeat the procedure until no further changes in the

Method ends

2.2 Invariant Features (IF) Extraction

For any variations in the character boundary, the CD algorithm gives single pixel boundary of the character. The result of the CD taken as input for feature extraction approach is presented in this section. The method computes centroid of the character by finding X_{min}, X_{max} and Y_{min}, Y_{max} coordinates of the contour. Next method estimates the distance (D) between the centroid (C) and the pixel of the contour. Further, the method finds Sum of all Distances (SD) and number of black pixels (N). The ratio square of N to SD is feature which is invariant to image transformation such as rotation, Scaling and Translation (RST). The steps involved in algorithm are given in Fig. 11.

Contour of input image

showing distances to be calculated by contour image from contour centriod

Fig 11 Feature extraction procedure

Algorithm: Invariant Feature (IF) Extraction

Input: Contour Image

Output: Invariant Features

Method Begins

Step 1: Find out the centriod (C) of the contour image.

Step 1.1: Sum (Sx) of all X coordinates is $Sx \,?\, \overset{n}{\underset{i?1}{?}} Xi$ where n is the number of black pixels.

Step 1.2: Sum (Sy) of all Y coordinates is $Sy \,?\, \overset{n}{\underset{i?1}{?}} Yi$

Step 1.3: X coordinate of the C is = $Cx \,?\, \frac{Sx}{N}$, where N is the total number of black pixels

Step 1.4: Y coordinate of the C is $Cy \,?\, \frac{Sy}{N}$.

Step 2: Find out the D from C to every black pixel using Euclidean Distance (ED).

$EDi \,?\, \sqrt{(Cx \,?\, Xi)^2 \,?\, (Cy \,?\, Yi)^2}$ For i = 1 to n, where Xi and Yi are the coordinates of black pixels.

Step 3: $SD \,?\, \overset{n}{\underset{i?1}{?}} EDi$

Step 5: $IF \,?\, \frac{N^2}{SD}$

Method ends

3. Experimental Results

In this section, we present the experimental results for evaluating the efficiency of the proposed contour based method. We have considered accuracy and computations as decision parameters to establish superiority of the proposed method. Accuracy of the method depends on number of characters recognized correctly out of 291 characters data set. Computations depend on the dominant operation involved in the method. In this method the dominant operation is pixel searching. In order to know the accuracy and recognition rate we have designed database using linear search tree and binary search tree [Jean. Paul. Tremblay and Paul. G. Sorenson, 1988] we have given comparative study in next section. Further, we have also shown that the proposed method is invariant to RST. 1.4 G Hz processor system is used for the experimentation purpose in this work. The values of accuracy and computations of the proposed method are tabulated in Table 1.

From Table 1 it is clear that the proposed method gives 100% accuracy and it takes 58201 computations for all 291 characters. We have experimented the proposed method to show that the method is invariant to rotations of any degree. In Fig. 12, we have given some samples of rotations of character image to tell that the method gives same feature values for different rotations. The corresponding features are tabulated in Table 2. From Table 2 and Fig. 13 it is clear that the proposed method is invariant to rotation transformation. Similarly for scaling also we have experimented the proposed method for different resolutions, which is given in Fig. 14. The corresponding values are tabulated in Table 3. From Table 3 and Fig. 15 it is observed that the method is invariant to scaling after 100dpi. In addition to this, we also concluded that the performance of the method degrades when the low-resolution image is given.

Table 1 Accuracy and computations of proposed method for 291 characters.

Contour Based Method	
Accuracy	Computation
100%	58201

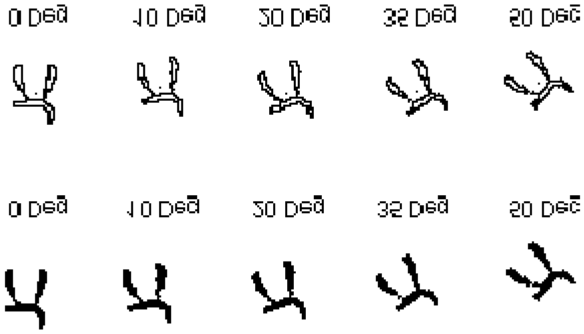


Fig 12. Different rotations of character image

Table 2 Features for different rotations

Degree	F
0	14.69
10	14.67
20	14.70
35	14.628
50	14.721

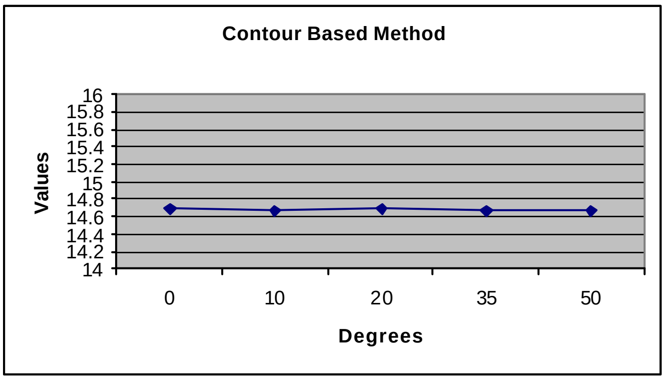


Fig 13 Graph Features v/s degrees

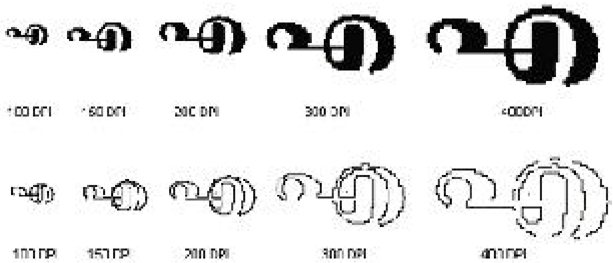


Fig 14 Different resolutions of character image

Table 3 Features for different dpi

Resolution	F
100	20.51
150	22.56
200	22.01
300	22.10
400	22.013

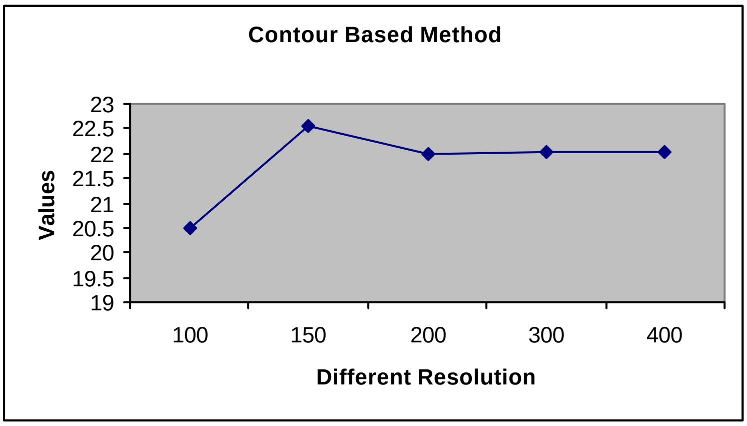


Fig 15 Graph Features v/s dpi

4. Comparative Study

In this section, we have given comparative study of proposed method with [Hemantha Kumar et al., 2004, Hemantha Kumar et al., 2004, Hemantha Kumar et al., 2003] method to evaluate the performance of the proposed method in terms of accuracy in recognition and number of computations. In order to compare the methods we have chosen parameters such as Time required to recognize the 291 characters using Linear Search Tree (TLST), Time using Binary Search Tree (TBST), computation involved in searching expected features in the database of 291 characters. In addition to this we have also consider the invariance property as parameter apart from the accuracy in recognition and number of computations.

From Table 4, it is noticed that the proposed method gives 100% accuracy and takes less computations (ref. Fig. 16 and Fig. 17) when compared to Polar Transformation Method (PTM) and Ring Projection Method (RPM) since the proposed method involves contour detection, which reduces the number of black pixels. However, Distance measure Method (DM) takes less computation compared to proposed method since DM involves only eight directions. The proposed method is competitive with respect to time parameter (ref. Fig. 18 and Fig 19) compared to RPM, PTM and DM in case of both linear and binary search trees. This is because the PTM involves feature vector containing 15 features, RPM involves feature vector containing 6 features and DM involves feature vector containing 2 features. As the parameter computations is concerned, the number of computations required to search feature using linear search tree and binary search tree, the proposed method gives better results than existing methods (Ref. Fig. 20 and Fig. 21). This is because of the feature vector having number of features.

Table 4 Values of the parameters based on experimental results

Name of the Methods	**Linear Search**		**Binary Search**		**Accuracy in**	**Computations recognition**
	Time	Computa-tions to search a FV	Time	Computa-tions to search a FV		
PTM	5.05 m	120539475	4.16 m	52890275	67%	81480
DM	2.07	24726852	1.53	6547500	100%	37910
RPM	4.54	74180556	3.54	19642500	100%	4431930
Contour Based Method (proposed method)	2.20 m	12363426	1.60 m	3273750	100%	58201

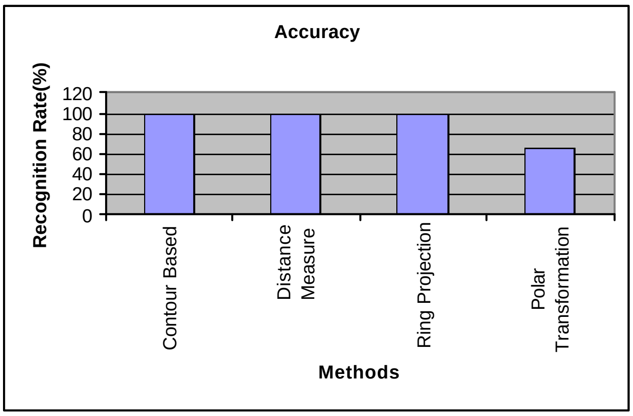


Fig 16 Graph for Accuracy v/s methods

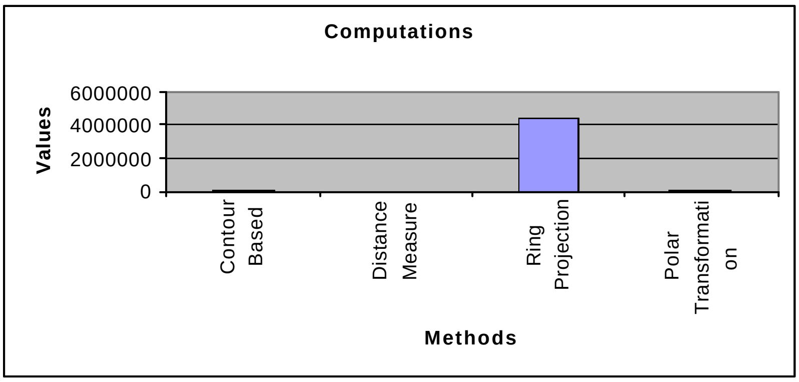


Fig 17 Graph for computations v/s methods

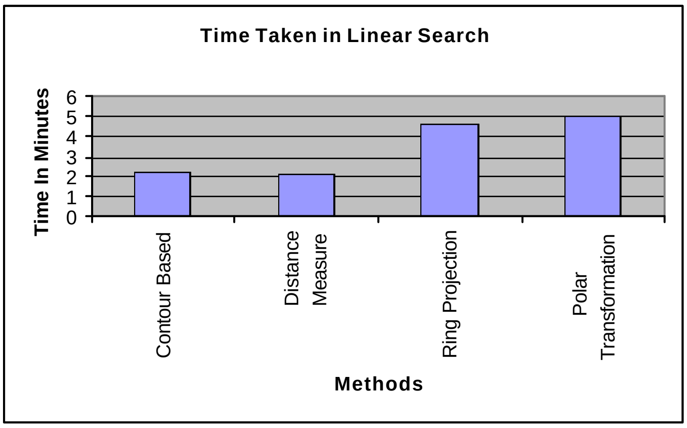


Fig 18 Graph for Time LST v/s methods

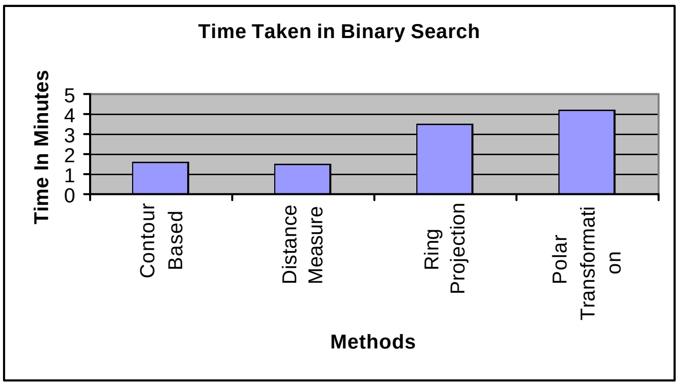


Fig 19Graph for Time BST v/s methods

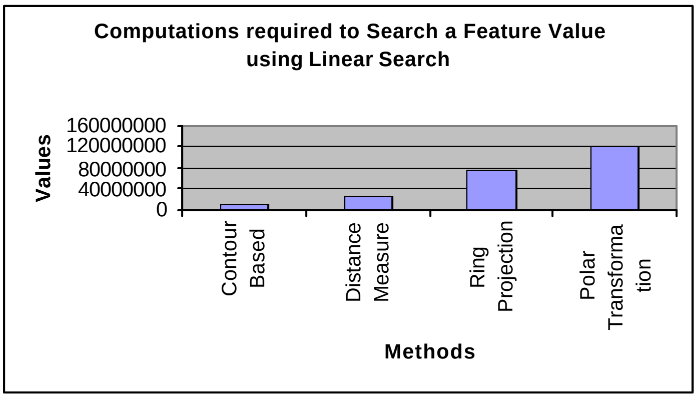


Fig 20 Graph for computations in linear Search tree v/s methods

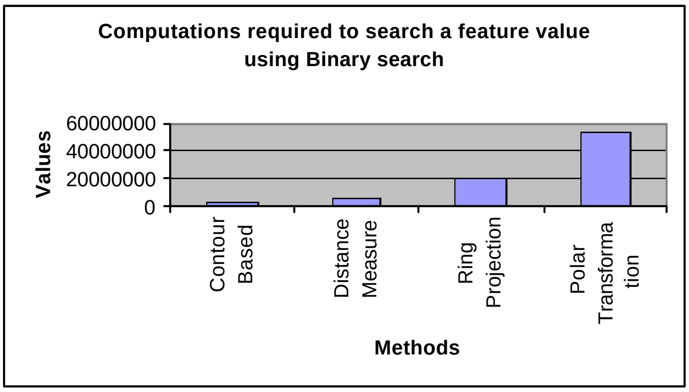


Fig 21 Graph for computations in Binary Search tree v/s methods

Table 5 Overall performances of the proposed and existing methods

Parameters	**Name of the Methods**
Accuracy	Contour based, Distance Measure, Ring Projection
Computations	Distance Measure
Number of computations involved in linear Search for a Feature Value (LS)	Contour based method
Number of computations involved in Binary Search for a Feature Value (LS)	Contour based method
RST Transformation	Contour based and Polar Transformation method

5. Conclusion

We have presented a new contour based method for the recognition of multi-lingual documents. The Proposed method is compared with the methods based on PTM [Hemantha Kumar .G et al., 2004], DM [Hemantha Kumar G et al., 2003] and RPM [Hemantha Kumar et al., 2004]. We have shown that the proposed method is better compared to other methods in terms of accuracy, computations, time required for searching a character and invariance property (ref. Table 5). However, the performance of the proposed method degrades for low-resolution images. This is an attempt to develop single OCR for all these languages. Further, the method is extended to some other languages also. This would be our future work.

6. Acknowledgment

The authors acknowledge the support extended by Dr. D. S Ramakrishna, Principal, APS College of Engineering, Somanahalli, Bangalore-82.

7. References

1. Anil. K Jain, Oivind Due Trier, and Torfinn Taxt, Feature Extraction Methods for Character Recognition – A Survey, Pattern Recognition, Vol.29, No.4, pp641-662, 1996.

2. Chew Lim Tan, Weihua Huang, Zhaohui Yu and Yi Xu, Imaged Document Text Retrieval Without OCR, IEEE transaction on Pattern Analysis and Machine Intelligence, Vol.24, No.6, 2002.

3. Hemantha Kumar. G, Shivakumara. P, Noushath. S, and Manjunath Aradhya. V.N, A New Invariant Algorithm for Recognition of Alphabets of Multi-Lingual Documents, Proceedings of 6 Th International Conference on Cognitive Systems – ICCS 2004, Centre for Research in Cognitive Systems, New Delhi, India, December 14-15, 2004(Accepted).

4. Hemantha Kumar. G, Shivakumara. P, Noushath. S, and Manjunath Aradhya. V.N, A Novel Feature Extraction Scheme for Malayalam Character Recognition, Journal of the Society of Statistics, Computer and Applications, Vol. 2, No. 1, 2004(New Series), pp 101-113.

5. Hemantha Kumar. G, Shivakumara. P, Noushath. S, and Manjunath Aradhya. V.N, Feature Extraction for Alphanumeric Symbols Recognition: An Approach Based on Distance Measures, Proceedings of Ist Indian International Conference on Artificial Intelligence (IICAI-03), Hyderabad, India, December 18-20 2003.

6. Jean. Paul Tremblay and Paul. G. Sorenson, An Introduction to data Structures with Applications, Mc Graw –Hill Book Company, 1988.

7. Masayoshi Okamoto and Kazuhiko Yamamoto, On-line handwriting character recognition using direction-change features that consider imaginary strokes, The Journal of Pattern Recognition Society, Vol.32, pp1115-1128, 1999.

8. Nagabhushan. P and 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, The Journal of Pattern Recognition Society, Vol.20, pp1467-1475, 1999.

9. Pal. U and Chaudhuri. B.B, Identification of different script lines from multi-script documents, Image and Vision Computing, Vol.20, pp945-954, 2002.

10. Pal. U and Chaudhuri. B.B, Machine-printed and Hand-written text lines identification, Pattern Recognition Letters, Vol.22, pp431-441, 2001.

11. Pal. U, Belaid. A and Choisy. Ch, Touching numeral segmentation using water reservoir concept, Pattern Recognition Letters, Vol.24, pp261-272, 2003.

12. Pal. U, Chaudhuri .B.B, Indian script character recognition: a survey, Pattern Recognition, Vol.37, pp 1887-1899, 2004.

13. Pal. U, Kundu. P. K, and Chaudhuri B. B, OCR Error Correction of an Inflectional Indian Language using Morphological Parsing, Journal of Information Science and Engineering, Vol.16, pp903-922, 2000.

14. Rejean Plamondon and Sargur N.Srihari, On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey, IEEE transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.1, Jan-2000.

About Authors

Manjunath Aradhya V N, Department of Studies in Computer Science, Manasagangothri, University of Mysore, Mysore-6.
E-mail : mukesh_mysore@rediffmail.com

Hemantha Kumar G, Department of Studies in Computer Science, Manasagangothri, University of Mysore, Mysore-6.
E-mail : mukesh_mysore@rediffmail.com

Shivakumara P, Assistant Professor, Department of Computer Science and Engineering, Acharya Patasala College of Engineering, Kanakapura Road, Somanahalli, Bangalore – 62.
E-mail : hudempsk@yahoo.com

Noushath S, Department of Studies in Computer Science, Manasagangothri, University of Mysore, Mysore - 6.
E-mail : mukesh_mysore@rediffmail.com