Character Level Separation and Identification of English and Gujarati Digits from Bilingual (English-Gujarati) Printed Documents

Shailesh A. Chaudhari
M.Sc.(I.T.) Programme, Veer Narmad South
Gujarat University, Surat, India,

Ravi M. Gulati
Dept., of Computer Science, Veer Narmad South
Gujarat University, Surat, India

ABSTRACT

Nowadays, it is observed that English script has interspersed within the Indian languages. So there is a need for an optical character recognition (OCR) system which can recognize these bilingual documents and store it for future use. Hence, in this paper an OCR system is proposed that can read documents containing Gujarati and English scripts (Only digits). These scripts have many features in common and hence a single system can be modelled to recognize them. Here, we have used template matching classifier. The normalized feature vector is used as a feature to classify English and Gujarati digits. The system shows a good performance for multi-font, size independent printed bilingual English- Gujarati digits. An average classification rate 98.30% is obtained for Gujarati digits and 98.88% is obtained for English digits at character level.

Keywords

Segmentation, Normalization, Vector, Template, Correlation, etc.

1. INTRODUCTION

Optical character recognition (OCR) is one of the oldest and largest sub fields of pattern recognition with a rich contribution for the recognition of printed documents. In India, there are 24 official (Indian constitution accepted) languages. Two or more of these languages may be written in one script. Twelve different scripts are used for writing these languages. In the Indian scenario, documents are often bilingual or multi-lingual in nature and therefore a multi-script OCR is a present day requirement. In addition to an Indian language, English is used as a link language in most of the important official document, reports, magazines, news papers and technical papers. Indian scripts, in general, are rich in patterns and variations. A lot of work is done in for recognition of Monolingual Scripts. Monolingual OCRs fail in bilingual contexts and there is a need to extend the operation of current monolingual systems to bilingual system. Recognition of bilingual script in an image of a document page is of primary function for a system processing bilingual document. Throughout the country, every government office uses at least two languages, English and the official language of the corresponding state.

Gujarati is the official language of the Gujarat state, however many national organizations such as Banks, use English and Gujarati. Even all the documents in the government offices of Gujarat state usually appear in these two languages. This is the major reason for choosing these two languages for experimentation. The aim of the automation of document processing is to convert the scanned paper document to the machine readable forms. In this research work the scanned paper document is experimented by taking the bilingual documents printed in English and Gujarati languages.

The present work proposes a bilingual OCR system to recognize complete set of printed Gujarati and English digits. This paper is organized in the following sections; Section 2 describes the characteristics of Gujarati and English digits. Section 3 describes the early attempts made in Indian language OCR. Section 4 explains proposed model with block diagram. Section 6 is devoted to results and discussion. Lastly in section 7 conclusions is explained.

2. CHARACTERISTICS OF GUJARATI AND ENGLISH DIGITS

Gujarati is a phonetic language and is spoken by more than 50 million people in Gujarat- a western state of India and also in surrounding states Rajasthan, Maharashtra, Madhya Pradesh etc. Though it is a very widely spoken language, limited work is found in the literature that addresses the recognition of Gujarati language. Like other languages Sanskrit, Hindi, Marathi which has been derived from Devanagari, some of the Guajarati characters are very similar in appearance. The digits in Indian languages are based on sharp curves and hardly any straight line is available. Fig.1 is a set of Gujarati digits.

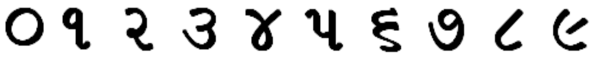


Fig. 1: Gujarati Digits 0-9.

As can be seen in Fig. 1, Gujarati digits are very peculiar in nature. Only two Gujarati digits one (1) and five (5) are having straight line, whereas other digits have curves which makes digit identification more difficult. Also Gujarati digits are often misclassified as shown in Fig. 2. As shown in Fig. 2.(a) digits zero (૦), three (૩) and seven (૭), Fig. 2.(b) digits one (૧) and six (૬), and Fig. 2.(c) digits eight (૮) and nine (૯) share similar shapes. Also identification of such Gujarati digits with different font style, font size, paper quality, etc becomes more difficult.

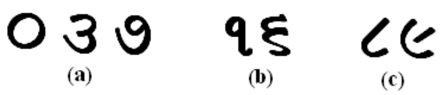


Fig. 2: Confusing Gujarati Digits.

English is a universal and international language and is spoken all around the world. It is a very widely spoken language and lot of work is found in the literature that addresses the recognition of English language. Most of the alphabets in English language have vertical bar but, digits in

English languages are based on sharp curves and hardly any straight line is available. Fig.3 is a set of English digits.

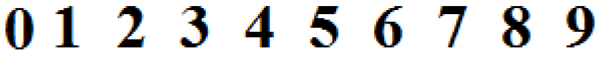


Fig. 3: English Digits 0-9.

As can be seen in Fig. 3, only two English digits viz. one (1) and five (4) have straight line, which makes digit identification a little more difficult. Also English digits are often misclassified as shown in Fig. 4. As shown in Fig. 4.(a) digits five (5) and six (6), Fig. 4.(b) digits one (1), four (4) and seven (7), and Fig. 4.(c) digits three (3), eight (8), nine (9) and zero (0) share similar shapes.

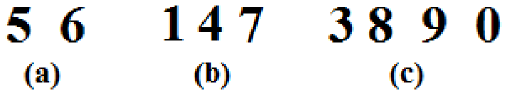


Fig. 4: Confusing English Digits.

Also, as can be seen in Fig.1 and Fig.3, two Gujarati digits zero (૦) and three (૩) and two English digits zero (0) and three (3) having almost similar shape.

3. LITERATURE REVIEW

All existing language identification techniques, fall into either local geometric or a global statistical approach. The local approaches analyze a list of connected components (like line, word and character) in the document images to identify the type of the language. In contrast, global approaches employ analysis of regions comprising at least two lines and hence do not require fine segmentation. Script separation for Indian language documents is extensively investigated by Pal and Chaudhuri [1] [2], mainly using structural features. They conducted script separation studies very effectively for many pairs of languages most of the time at line level and rarely at paragraph or word level. They have obtained about 97.33% accuracy

Spitz and Tan [3-5] have some contributions in the document script identification but rarely attempted to process multi-script, multi-lingual documents. It may be due to the reason that such multi-script, multi-lingual documents appear only in Indian Society. B. B. Chaudhuri and U.Pal [6] have discussed an OCR system to read two Indian language scripts, Bangla and Devnagari (Hindi).

Pal, Sinha and Chaudhari [7] proposed a generalized scheme for script line identification in printed multiscript documents that can classify as many as 12 Indian scripts. Features chosen in the proposed method are headlines, horizontal projection profile, water reservoir-based features, left and right profiles, and feature based on jump discontinuity, which refers to the maximum horizontal distance between two consecutive border pixels in a character pattern. Experimental results show an average script line identification accuracy of 97.52 %.

Patil and Subbareddy [8] was developed; a neural network-based architecture for identification of printed Latin, Devnagari, and Kannada scripts. It consists of a feature extractor followed by a modular neural network. The modular neural network structure consists of three independently trained feed forward neural networks, one for each of the three scripts under consideration. It was seen that such a system can classify English and Kannada with 100 percent accuracy, while the rate is slightly lower (97 %) in recognizing Devnagari.

A zone-based feature extraction algorithm scheme proposed by Rajeshekararadhya and Ranjan [9] for the recognition of off-line handwritten digits of four popular Indian scripts. The nearest neighbour feed forward back propagation neural network and support vector machine classifiers are used for subsequent classification and recognition purposes. They obtained a recognition rate of 98.65 % for Kannada digits, 96.1 % for Tamil digits, 98.6 % for Telugu digits and 96.5 % for Malayalam digits using the support vector machine.

An algorithm for language identification of Kannada, Hindi and English text lines from printed documents is proposed by Padma, Vijaya and Nagabhushan [10]. The approach is based on the analysis of the top and bottom profiles of individual text lines and hence does not require any character or word segmentation. Experimental results demonstrate that relatively simple technique can reach a high accuracy level for identifying the text lines of Kannada, Hindi and English languages. The performance has turned out to be 95.4% rate.

Bindu Philip and R. D. Sudhaker Samuel [11] addressed the problem of bilingual character recognition using Gabor features for script identification and Dominant Singular Values as features for classification. The proposed algorithm has been tested successfully and an overall recognition rate of 96.5% is achieved.

Antani and Agnihotri [12] describe the classification of a subset of printed or digitized Gujarati characters. It has low recognition rate of 67 %. A research prototype of Gujarati OCR has been designed and implemented [13]. The accuracy of recognition is low but can be improved upon by modifying distance measures used and tweaking the code. It recognizes each word in the input document image and outputs UNICODE text equivalent to it. The overall system was tested on various images from various sources.

A novel hybrid feature extraction technique is suggested by Desai [14] which is constituted by a structural approach and statistical approach of feature extraction. Image is subdivided and then the pixel information is used as a structural approach whereas the aspect ratio of the number is considered as a statistical approach. For classification kNN classifier has been used. This model gives overall accuracy of 96.99% for the handwritten Gujarati digits.

4. PROPOSED MODEL

We have followed the system model as shown in Fig.5, for accomplishing the task of recognising English-Gujarati mixed digits. The model is divided in basic image processing stages like pre-processing, segmentation, and feature extraction and classification. For English-Gujarati mixed digits, we have used standard procedure for digitization that give digitized text image.

The first stage is document pre-processing, the second stage is segmentation, and the third stage is feature extraction and classification technique. Pre-Processing includes Reading image file, Binarization, and noise removal. In binarization grey-scale image file or color image file is converted into Binary file (Black & White) using UTSU's global thresholding technique [15]. During noise removal objects containing less than 10 pixels considered as noise and is removed. After Pre-processing, the segmentation is performed in which, first the line segmentation is performed and then using connected component analysis character segmentation is performed. A bounding box is created for individual connected component to extract a single digit by finding minimum, maximum black pixel for both horizontal and vertical direction. Then bounding box is cropped for feature extraction. The bounding box is of varying size due to varying font size or shape.

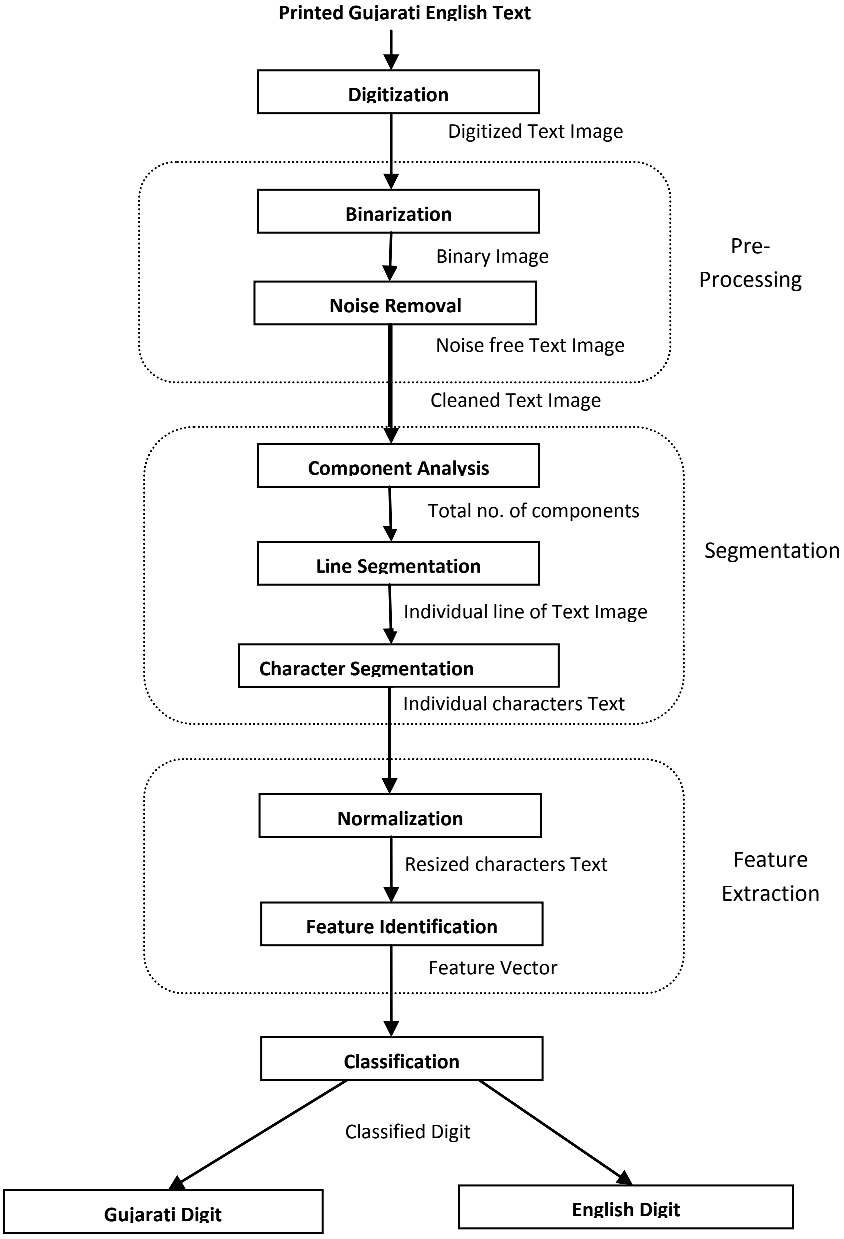


Fig. 5: Block Diagram of proposed system

Normalization of the digits is essential because of the different font style and different font size which result in several variations in shapes and sizes. Therefore to bring about uniformity in the input digits, all the segmented components should be made of the same size. For this reason segmented bounding box are fit into a standard size window of 32 X 40. Every measure has to be taken to preserve the exact aspect ratio of the input digit. The size of the window is selected due to the fact that the height of the digit is more than the width of the digit. The image of all the segmented characters are normalized (rescaled) into a common height and width producing a grid of 32 X 40 pixel-size (shaped-zones) as shown in Fig.6.

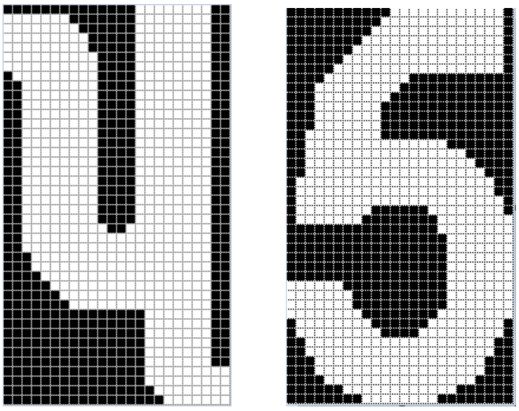


Fig. 6: 2-D matrix of Gujarati digit 5(left) and English digit 5(right).

The pixel density is calculated as binary patterns and therefore a vector is created by dividing grid into 32 columns of 1x40 size and all these 1x40 columns are combined to create a vector of 1280 pixels. However, due to varying nature of font-family, there was dissimilarity between the feature vectors of the same class.

Template matching, or matrix matching, is one of the most common classification methods. In template matching, individual image pixels are used as features. Classification is performed by comparing an input digit image with a set of templates (or prototypes) for both Gujarati and English digits. For each digit correlation coefficient is computed and saved. The correlation coefficient is defined as:

$$r = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{(\sum_m \sum_n (A_{mn} - \bar{A})^2))(\sum_m \sum_n (B_{mn} - \bar{B})^2)}}$$

where $-1.0 \leq r \leq +1.0$.

In this equation, A is the template image and $\bar{A}$ is the average value of the template. B is the segmented input image vector and $\bar{B}$ is the average value of the input image vector. Here, m and n indicates the row and column index of the respective vector.

Each comparison results in a similarity measure between the cropped input digit image and the template. One measure increases the amount of similarity when a pixel in the observed character is identical to the same pixel in the template image. If the pixels differ, the measure of similarity may be decreased. After all templates have been compared with the observed digit image, the digit's identity is assigned as the identity of the most similar template.

We implemented the system in a research mode, without optimising for performance. We pre-processed all textual symbols in a document not only digits. We have taken connected component analysis based approach. We applied our separation and identification scheme on more then 200 image samples. The images were scanned at 200 to 300 dpi, from different literature, newspaper, magazine, books, etc. This scheme does not depend on the size of digits. Also, this approach is insensitive to font and style variation. Since there is no work reported for script recognition in bilingual document at character level, to the best of our knowledge, the results of this work could not be compared and hence we believe that this approach to identify digits from a bilingual document will give the best result.

5. RESULTS AND FINDINGS

Here, we have taken twenty templates (0-9 for Gujarati digits and 0-9 for English digits) to obtain the performance of proposed system. We have collected more than 200 image samples of scanned bilingual English-Gujarati documents at 200 to 300 dpi and experimented with proposed System. The test set used in this experiment was getting a good set of digits for classification.

The digits used for the experiment were enclosed in a bounding box and normalized to a fixed sized of template window. Different font families represent the same digit differently and the correlation between similar digits varies from font to font. This preliminary research helped us to focus our attention on these matters so that issues for building robust digit recognition can be studied. This segmentation based approach proved to be efficient for multi-font, size independent digits. The multi-font aspect for Gujarati and English alphabet characters is under investigation.

We have made our experiments on bilingual English-Gujarati documents having various qualities of papers in nature. For both types of digit, Gujarati and English, we achieved more than 98% accuracy. The following tables show the result of comparing the performance of each English digit (0-9) and each Gujarati digit (0-9) with different sizes of test datasets. It also shows the average accuracy of the English digits (0-9) and Gujarati digits (0-9).

The table 1 shows the result achieved for English digits against the testing of the system.

Table 1. Accuracy of English Digits (0-9)

English Digit	Total English Test Digits	Missed	Accuracy (%)
0	160	0	100
1	178	1	99.44
2	159	0	100
3	97	3	96.91
4	97	1	98.97
5	139	4	97.12
6	91	1	98.90
7	108	1	99.07
8	101	1	99.01
9	166	1	99.40
Average Accuracy of English Digits (0-9)			**98.88**

The table 2 shows the result achieved for Gujarati digits against the testing of the system.

Table 2. Accuracy of Gujarati digits (0-9)

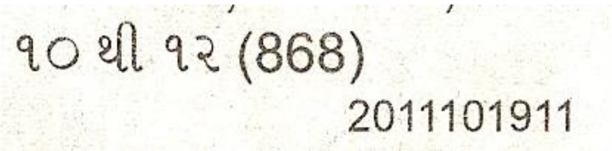
Gujarati Digit	Total Gujarati Test Digits	Missed	Accuracy (%)

૦	164	0	100
૧	131	2	98.47
૨	99	3	96.97
૩	79	0	100
૪	70	5	92.86
૫	85	2	97.65
૬	51	1	98.04
૭	67	0	100
૮	83	0	100
૯	99	1	98.99
Average Accuracy of Gujarati Digits (0-9)			**98.30**

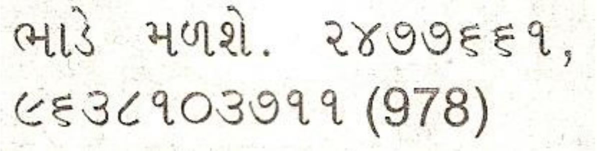
6. CONCLUSION

The proposed system used hybrid statistical and structural feature and template matching classifier for classification of pattern of bilingual English-Gujarati digit. Here the proposed system works effectively and reliably on each digit. To conclude, a bilingual English-Gujarati digit recogniser and converter is designed to investigate the process of automatic digit recognition and conversion. We obtained a classification rate 98.30% for Gujarati digit and 98.88% for English digit. From the experiment, it is clear that the performance of the proposed system greatly reduce the cost of storing scanned image document in computer's hard disk. The present work can be extended for English-Gujarati alphabets and few other symbols for complete bilingual English-Gujarati printed document. It can also be extended for skewed /tilted bilingual English-Gujarati documents.

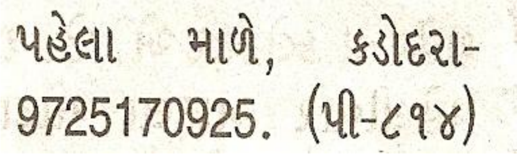
Following are the some of the document image samples tested by the proposed system.



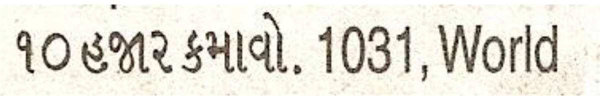
૧૦ થી ૧૨ (868)
2011101911



ભાડે મળશે. ૨૪૭૭૬૬૧,
૯૬૩૮૧૦૩૭૧૧ (978)



પહેલા માળે, કડોદરા-
9725170925. (પી-૮૧૪)



૧૦હજાર કમાવો. 1031, World

7. REFERENCES

[1] U. Pal and B. B. Chaudhuri, 1999, “Script line separation from Indian multi-script documents”, In Proc. Int. Conf. Document Analysis and Recognition (ICDAR).

[2] U. Pal and B.B.Chaudhuri, 2001, “Automatic identification of English, Chinese, Arabic, Devanagari and Bangla script line”, In Sixth International Conference on Document Analysis and Recognition (ICDAR '01).

[3] A. L. Spitz, 1997, "Determination of the Script and Language content of Document Images", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 3.

[4] T. N. Tan, 1998, "Rotation Invariant Texture Features and their use in Automatic script Identification", IEEE Transactions on PAMI, Vol.20, No.7.

[5] B. B. Chaudhuri and U. Pal, 1999, “Automatic separation of machine printed and handwritten text lines", 5th Ineternational Conference on Document Analysis and Recognition, Vol.1.

[6] U.Pal and B.B. Chauduri, 1997, "Automatic seperation of words in multi-lingual multi- script Indian documents", 4th International Conference on Document and Recognition, Vol.2.

[7] U. Pal, S. Sinha, and B.B. Chaudhuri, 2003, “Multi-Script Line Identification from Indian Documents,” Proc. Int'l Conf. Document Analysis and Recognition.

[8] S.B. Patil and N.V. Subbareddy, 2002, “Neural Network Based System for Script Identification in Indian Documents,” Sadhana, vol. 27, no. 1.

[9] S. V. Rajashekararadhya, Dr P. Vanaja Ranjan, 2009, “Handwritten Numeral/Mixed Numerals Recognition Of South-Indian Scripts: The Zonebased Feature Extraction Method” Journal of Theoretical and Applied Information Technology, Vol 7. No 1.

[10] M.C. Padma, P. A. Vijaya, P. Nagabhushan,2009, “Language Identification from an Indian Multilingual Document Using Profile Features”, International Conference on Computer and Automation Engineering, IEEE, 978-0-7695-3569-2.

[11] Bindu Philip and R. D. Sudhaker Samuel, 2009, “A Novel Bilingual OCR for Printed Malayalam-English Text based on Gabor Features and Dominant Singular Values”, International Conference on Digital Image Processing, 978-0-7695-3565-4/09, IEEE.

[12] S Antani and L Agnihotri, 1999, “Gujarati Character Recognition”, Proceedings of the International Conference on Document Analysis and Recognition, (ICDAR-99), Bangalore, India.

[13] Prof S K Shah, A Sharma, 2006, "Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching", IE(I) Journal−ET, Vol.2.

[14] Apurva A. Desai, 2010, “Handwritten Gujarati Numeral Optical Character Recognition using Hybrid Feature Extraction Technique ”, Int'l conf. IP, Vision, and Pattern Recognition, IPVC.

[15] N. Otsu, 1979, “ A threshold selection method from gray level histogram “, IEEE Trans. Syst. Man Cyb.