Zoning Feature for Script Identification and Character Recognition

Shailesh A. Chaudhari, Assistant Professor,
Dept. of ICT, Veer Narmad South Gujarat University, Surat, Gujarat, India
(E-mail: sachaudhari@vnsgu.ac.in)

Abstract *–* Script identification is an essential pre-processing step in design of multi-script Optical Character Recognition (OCR) systems. Most of the reported work in the field of OCR is for mono script. This paper presents zoning feature for character level script identification from offline printed bilingual Gujarati-English text. The database consists of 4,760 basic Gujarati characters and 7,280 English (Uppercase + Lowercase) characters with a variety in terms of size, font type and style. The input character image is divided into predefined non-overlapping zones to extract the features. The proposed method goes through two phases: Classification and Recognition. In the first phase classification, script is identified by KNN and SVM classifiers, then in second phase based on identified script, appropriate script specific OCR is employed for character recognition. The same Zoning features are used in recognition phase. Results from the experiments show that, presented method is robust for character level printed bilingual script identification and character recognition.

***Keywords* --** bilingual, Zoning, classification, recognition.

I. INTRODUCTION

OCR is an ever emerging research area since it was initially introduced in software industry. Many OCR systems for printed mono script text are available, but the work is still emerging stage for multi-script text. In the digital era, the world is becoming automated and multilingual with the advancement of digital technologies. Digital solution provides easy storage, access and retrieval of information, stored digitally. During last few decades, document processing systems such as OCR have set a benchmark of success. Nowadays, due to OCR, we can easily process documents either in off line or on line mode. Unfortunately, OCR is restricted to some kwon script and hence fails with multi-script documents. Therefore, in designing of multi-script OCR system, script identification is essential. Here, we focus on character level off line script identification. It can be noted that, script identification, especially in Indian documents, becomes more difficult because of variation in font style, size and inconsistent gap in words and characters.

II. RELATED WORK

Development of a multi-script OCR system for Indian languages is more challenging than mono script OCR development. This is because of the large variation of character sets in each Indian script. This causes difficulty in digitizing multi-script documents.

Spitz presented a global approach based technique, early in the history of script identification to distinguish Asian (Chinese, Japanese and Korean) and European (Roman) [1][2]. This technique examines presence and location of upward concavities of characters in the text images. Elgammal and Ismail were separated English and Arabic language at text line level by horizontal projection [3]. In this technique, researchers examine the horizontal projection profiles of both Arabic and English text. They observed that Arabic text have one major peak with respect to the baseline, whereas, English text have two major peaks with respect to baseline.

Chaudhuri and Pal contributed major work of script identification for Indian Languages [4][5]. First, they developed a technique to distinguish the Roman, Bangla and Devnagari scripts at text line level. They used horizontal line at the top of the text (Headline) as an important feature. Using this feature they distinguished Bangla and Devnagari script from Roman script.

Dhndra et al. proposed a morphological reconstruction based method [6] for script identification at word level. They used morphological erosion and opening of words in four directions horizontal, vertical, right, and left diagonal using line structuring elements and also the hole filling was performed for those character which contains loop.

Kunte and Samuel developed a bilingual OCR system for printed Kannada and English [7]. They used Gabor function based features to identify the script at word level using pre-trained neural classifier. Then, to separate Kannada and English characters, Wavelet descriptors were used as features with neural network for classification. Dhandra and Hengarge presented OCR system based on discriminating features such as aspect ratio, strokes densities, eccentricity, etc. Using these features they separated printed Kannada numeral words from bilingual and trilingual documents representing Kannada, Devnagari, Tamil, Odiya and Malayalam scripts [8]. Chaudhari and Gulati reported initial work for bilingual printed Gujarati-English documents [9][10][11]. They used statistical feature for script identification at line level in printed bilingual

Gujarati-English documents. Then, they proposed a system for word level script identification using Gabor feature and SVM classifier.

Singh et al. presented a texture feature based technique for script identification at page level [12]. Texture features were extracted from document pages based gray-level co-occurrence matrix. Then features like energy, entropy, inertia, contrast, local homogeneity, cluster shade, cluster prominence and information measure of correlation were calculated. Finally, scripts like Bangla, Devanagari, Telugu and Roman were identified using Multi Layer Perceptron (MLP) classifier.

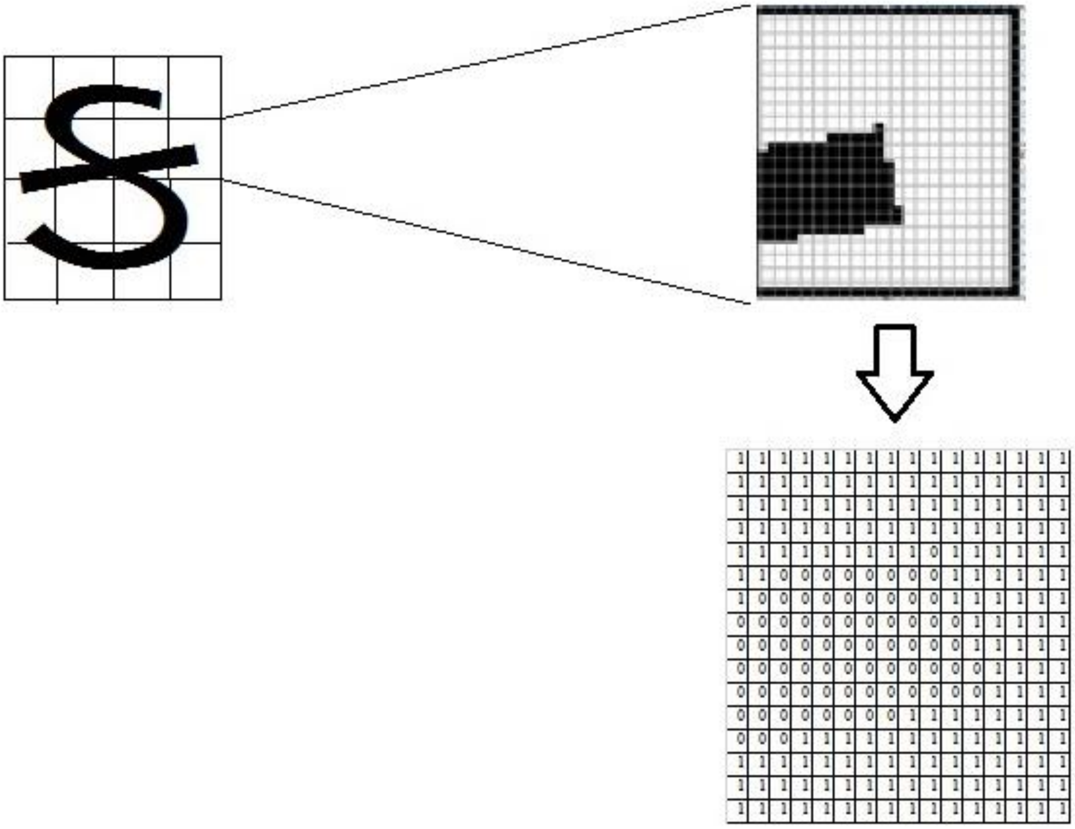
Kartar et al. presented a zoning feature with four types of projection histogram- horizontal, vertical, left diagonal and right diagonal for handwritten Gurmukhi character recognition [13].

III. FEATURE EXTRACTION

Zoning is a very popular and extensive technique for the extraction of features as it can manage up with variability and differences of character patterns. Here, the input character image is first resized to size of 32*32 pixels and then divided into 16 equal zones or blocks each of size 8*8 pixels [13][14]. The features are extracted by counting the number of black pixels in each zone. This procedure is repeated sequentially for all 16 zones which are stored in the form of signature array for each character. Thus, for each Gujarati or English character we get a signature array of length 16 calculated from each zone using following equation.

$$D(i) = \frac{\text{Number of foreground pixels in zone i}}{\text{Total number of pixels in zone i}}$$

This method has been widely adopted in order to obtain valuable information on the local characteristics of the character pattern. In general, let “A” be a pattern image, Z be a zoning method then Zm can be thought out to be as a partition of “A” into M sub images (M being an integer greater than 1). These sub-images are the named zones, i.e; Zm = {z1, z2,…,zM}, where each one provides local information on patterns. In the early days, methods based on zoning were extensively put into action for recognition as well as analysis of printed characters. In this case, the pattern image under consideration is partitioned into zones and the relevance of the information carried out from each zone is evaluated, in the context of human recognition processes.



1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1
1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Figure 1 Zoning and Binary representation of Gujarati Character 'Ka'

The most important aspect of optical character recognition scheme is the selection of good feature set, which is reasonably invariant with respect to shape variations. The zoning method is used to compute the percentage of black pixel in each zone. The rectangle area of the character is divided into several overlapping, or non-overlapping regions and the densities of black pixels within these regions are computed and used as features as shown within Fig.1. The major advantage of this approach exploits from its robustness to small variation, ease of implementation and good recognition rate. Zone-based feature extraction method provides good result even when certain pre processing steps like filtering, smoothing and slant removing are not considered. The detailed description of Zoning Algorithm is as under:

A. *Zoning Algorithm*

```
 Input: Binary image (size 32*32)
block_id = 1
 for i from 1 to 4
        for i from 1 to 4
             read 8*8 pixels in matrix format
                     block[block_id] = read matrix
             block_id = block_id + 1
 for block_id from 1 to 16
             read block[block_id]
        count = number of black pixels in the block
 signature_array[block_id] = count
 Result : signature array of the image
```

The input character image is divided into N×M zones. From each zone features are extracted to form the feature vector by computing pixel density in each zone. The goal of zoning is to obtain the local characteristics instead of global characteristics. We created 16 (4*4) zones of 8*8 size each out of our 32*32 normalized samples by horizontal and vertical division. We obtained the density of each zone by dividing the number of foreground pixels in each zone by total number of pixels in each zone i.e. 64. Thus, we obtained 16 zoning density features (Fig.1). Finally, ’’n’’ such features will be obtained for classification and recognition.

IV. CLASSIFICATION AND RECOGNITION

The core task of classification is to use the feature vectors provided by feature extraction algorithm to classify and assign the object/pattern to a category. To perceive the behaviour of proposed algorithm, a comprehensive study has been made through experimental tests that were conducted on bi-script database using n-NN and SVM classifier.

k-NN and SVM both are supervised learning algorithm. The k-NN uses Euclidean distance to measure the distance between the test sample and the *k* trained neighbors. After careful examination of the k nearest neighbors, a simple majority of these k-nearest neighbors is to be predicted for the input image character.

SVM is a binary classifier which classifies dataset using by inspecting optimal hyper plane. The power of SVM lies in its ability to transform data to a high dimensional space where the data can be separated using a hyper plane. SVM is a well-developed technique to create optimal hyper plane which distinct two classes by maximizing the distance or margin between two classes.

A. *Classification of Printed Basic Gujarati and English (Uppercase + Lowercase) Characters*

A comprehensive study has been carried out through conduct of experiments to analyse the performance of proposed method for printed basic Gujarati and lowercase English characters. The collected data set contains 4,760 Gujarati Characters and 7,280 English (Uppercase + Lowercase) Characters. A Zone based pixel density feature vector of size 16 was created by dividing the input character image into non-overlapping 4X4=(16) regions.

To evaluate global script identification accracy, 10-fold cross validation is used. An average percentage of classification accuracy of 99.63%, and 99.54% using kNN, and SVM_RBF classifiers respectively obtained and is presented in Table 1.

TABLE 1. Classification Accuracy of Printed basic Gujarati and English Uppercase and Lower case Characters using DCT feature

Data set	kNN	SVM_RBF
fold1	100.00	100.00
fold2	99.76	99.76
fold3	99.76	99.76
fold4	99.88	98.45
fold5	100.00	100.00
fold6	100.00	100.00
fold7	99.88	99.88
fold8	97.02	97.50
fold9	100.00	100.00
fold10	100.00	100.00
Max	**100.00**	**100.00**
Min	**97.02**	**97.50**
Avg	**99.63**	**99.54**

From the Table 1, it is clear that the average classification accuracy with kNN is 99.63%, which is nearly 100%. The important thing to notice here is that, this accuracy is achieved when Gujarati basic character is mixed with both English Uppercase and English Lowercase characters with different font types and different font sizes. However, efforts are on to increase the classification accuracy level for 100% by adding the dominant features to the present feature vector.

B. Recognition of Printed Basic Gujarati Character

A comprehensive study has been carried out with experiments to analyse the performance of proposed method to recognize printed basic Gujarati characters with kNN and SVM classifiers. A Zone based pixel density feature vector of size 16 was created by dividing the input character image into non-overlapping 4X4 = (16) regions. The collected data set contains 4,760 Gujarati Characters.

To test the performance of recognition algorithm, 10-fold cross validation used. An average percentage of recognition accuracy of 96.60%, and 92.34% using kNN, and SVM_RBF classifiers respectively obtained and is presented in Table 2.

TABLE 2. Recognition Accuracy of printed basic Gujarati characters using DCT feature

Data set	kNN	SVM_RBF
fold1	96.85	93.91
fold2	98.74	95.38
fold3	88.87	87.40
fold4	99.16	94.75
fold5	95.80	91.18
fold6	99.79	96.01
fold7	98.53	93.91
fold8	89.71	79.21
fold9	99.58	96.22
fold10	98.95	95.38
Max	**99.79**	**96.22**
Min	**88.87**	**79.21**
Avg	**96.60**	**92.34**

From the Table 2, it is clear that the average recognition accuracy with kNN is 96.60% for printed Gujarati basic characters. The important thing to note here is that, this accuracy is achieved with different font types and different font sizes of printed Gujarati characters. However, efforts are on to increase the recognition accuracy level for 100% by adding the dominant features to the present feature vector.

C. Recognition of Printed Uppercase and Lowercase English Characters

Extensive experiments were conducted to study the performance of proposed grid based pixel density feature for the recognition of printed uppercase and lowercase English characters. The collected data set contained 7,280 English Uppercase and Lowercase Characters. To prepare the database, for English script commonly used 14 fonts with 10 different font sizes were used for both Uppercase and Lowercase characters. A Zone based pixel density feature vector of size 16 was created by dividing the input character image into non-overlapping 4X4= (16) regions.

To assess the performance of recognition algorithm, 10-fold cross validation is used. An average percentage of recognition accuracy of 96.72%, 95.10% and 95.14% using kNN, SVM_Polynomial, and SVM_RBF classifier respectively obtained and is presented in Table 3.

TABLE 3. Recognition Accuracy of printed Uppercase and Lowercase English characters using DCT feature

Data set	kNN	SVM_RBF
fold1	92.99	98.63
fold2	98.21	96.43
fold3	97.94	95.33
fold4	95.88	95.20
fold5	97.94	98.21
fold6	99.04	97.80
fold7	96.30	94.37
fold8	97.25	91.07
fold9	98.63	92.72
fold10	93.00	91.62
Max	**99.04**	**98.63**
Min	**92.99**	**91.07**
Avg	**96.72**	**95.14**

From the Table 3, it is clear that the average recognition accuracy is 96.72% for printed English Uppercase and Lowercase characters. The important thing to note here is that, this accuracy is achieved with different font types and different font sizes of printed Gujarati characters. However, efforts are on to increase the recognition accuracy level to 100% by adding the dominant features to the present feature vector.

V. CONCLUSION

A feature extraction technique based on Zoning from character image is presented for script identification and character recognition. It is assumed that the text is in well printed form. The proposed method is robust with respect to font size, font type and style. The worst case recognition accuracy rate still needs to be improved. The script identification and recognition accuracy is very promising as it is one of the untouched research areas of script identification in OCR. The present work will be further extended for other printed bilingual Indian documents as well as hand written bilingual Gujarati - English text.

REFERENCES

[1] Spitz.A.L, Determination Of The Script And Language Content Of Document Images, IEEE Tran. On Pattern Analysis And Machine Intelligence, 1994.

[2] Spitz.A.L, Determination Of The Script And Language Content Of Document Images, IEEE Tran. On Pattern Analysis And Machine Intelligence, 1997 Vol. 19, pp.234-245.

[3] Elgammmal.A.M and Ismail.M.A, Techniques for Language Identification for Hybrid Arabic-English Document Images, Proc. Sixth Int'l Conf. Document Analysis and Recognition, 2001, pp. 1100-1104.

[4] Chaudhuri.B.B and Pal.U, An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi), In Proc. 4th ICDAR, Uhn, 1997, pp18-20.

[5] U.Pal and B.B. Chaudhuri, Script Line Separation from Indian Multi-Script Documents, Proc. Int'l Conf. Document Analysis and Recognition, 1999, pp. 406-409.

[6] B.V. Dhandra, P. Nagabhushan, M. Hangarge, R. Hegadi, and V.S. Malemath, Script Identification Based on Morphological Reconstruction in Document Images, Proc. IEEE Int'l Conf. Pattern Recognition, 2006, vol. 2, pp. 950-953.

[7] R Sanjeev Kunte and R D Sudhaker Samuel, A Bilingual Machine-Interface OCR for Printed Kannada and English Text Employing Wavelet Features, 10th International Conference on Information Technology IEEE, 2007, 0-7695-3068-0/07.

[8] B.V.Dhandra and Mallikarjun Hangarge, On Separation of English Numerals from Multilingual Document Images, Journal of Multimedia, 2007, VOL. 2, NO. 6.

[9] S. Chaudhari and R. Gulati, Script Identification from bilingual Gujarati-English Documents, International Journal of Computer Applications (IJCA), 2014, Vol. 93.

[10] S. Chaudhari and R. Gulati, Script Identification Using Gabor Feature and SVM Classifier, Elsevier Procedia Computer Science, 2016, Vol. 79, pp. 85–92.

[11] S. Chaudhari and R. Gulati , A Comparative Analysis of Feature Extraction Techniques and Classifiers Inaccuracies for Bilingual Printed Documents (Gujarati-English), International Journal of Applied Information Systems (IJAIS), 2016, Vol. 1.

[12] Singh PK, Dalal SK, Sarkar R, Nasipuri M, Page-level script identification from multi-script handwritten documents. In:3rd international conference on computer, communication, control and information technology, Hooghly, 2015, pp 1–6.

[13] Kartar Singh Siddharth , Mahesh Jangid, Renu Dhir, Rajneesh Rani, "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features", International Journal on Computer Science and Engineering (0975-3397), Vol. 3 No. 6 June 2011.

Mr. Shailesh A. Chaudhari hold M.Phil. form Veer Narmad South Gujarat University and master degre in computer Science and Applications (MCA) from same university. Mr. Chaudhari has several years work experience in the areas of teaching, research and programmimg. He has several research publications in well-koown international journals and conferences. He has also been engaged to create linkage between industry and academia. He associate with CSI and ISTE.