RECOGNITION OF HANDWRITTEN KANNADA CHARACTERS USING HYBRID FEATURES

Saleem Pasha[1] and M.C. Padma[2]

[1]Assistant Professor, Department of Information Science and Engineering, PES College of Engineering, Mandya – 571401, Karnataka, India
[2]Professor and Head, Department of Computer Science and Engineering, PES College of Engineering, Mandya – 571401, Karnataka, India
saleempashapes@gmail.com, padmapes@gmail.com

Abstract: The challenging task in the field of Document Image Analysis is automatic recognition of handwritten characters present in a scanned document. The recognition of characters in a document is achieved by Optical Character Recognition (OCR) system. In this paper, a hybrid feature extraction technique is proposed for recognizing handwritten Kannada characters. The proposed technique uses the local and global features as hybrid features. These features are extracted from each input image. 3600 samples are used as training data set to obtain consistent feature values. K–nearest neighbor classifier is used to classify the characters based on the feature values. The proposed method is tested on a dataset of 1200 samples and at present it shows an overall accuracy of 87.33%.

Keywords: Document image analysis, Optical Character Recognition (OCR), Preprocessing, Feature extraction, Hybrid features, Local features, Global features, Classifier, K-nearest neighbor.

INTRODUCTION

Document Image Analysis refers to algorithms and techniques that are applied to images of documents to obtain a computer-readable description from pixel data. It is an emerging research area in the field of Pattern Recognition and Image Processing since many decades [1]. The Optical Character Recognition (OCR) is used torecognize the text present in a document. OCR is the process of converting scanned document images of machine printed or handwritten text into a computer editable format.

The recognition of handwritten characters of adocument is one of the challenging tasks due to diversified style of writing, mood of each individual, size of characters, quality of pen, aging of documents, quality of paper, color of ink, etc. Generally, the OCR system that is designed for printed documents cannot be used to process handwritten documents. Hence, there is a great demand to develop an OCR system that can process the documents containing handwritten text. The recognition system can be either on-line or off-line. On-line handwritten characters are obtained by using cameras or by writing the characters on a sensitive surface such as digital tablet PCs. Off-line handwritten characters are obtained from scanned images of handwritten text. In most of the realistic applications, majority of the documents received at various offices are filled by humans for specific purposes. In such scenarios, an OCR system is needed that can process the documents containing portions of printed text as well as handwritten text. OCR has been greatly developed in recent years because of the prevalence of internet and multimedia techniques. Nowadays, handwritten character recognition has received more attention in academic and production fields. Also, recognizing handwritten characters is essential for various application such as library automation, automatic postal address readers, automatic data entry from paper documents to computer, reading aid for the visually impaired.

In this paper, an attempt is made to develop a system to recognize off-line handwritten characters. Development of an efficient and robust OCR system involves several stages such as preprocessing, feature extraction and classification. However, in this paper suitable technique for extracting the hybrid features from handwritten Kannada character which could be further used in development of an OCR system is presented.

In rest of the paper, we brief out about the previous work carried out on OCR systems, give brief introduction of the Kannada script, brief out the necessary preprocessing steps, present the complete description of the proposed model and finally give the conclusion.

LITERATURE SURVEY

In this section, the literature relevant to the proposed research is highlighted.From the literature survey on character recognition, it is noticed that many OCR systems are developed for printed documents, but less work is reported for handwritten documents of Indian scripts. It is also observed from the literature that many successful attempts are reported for recognition of characters from both printed and handwritten documents for other than Indian scripts. OCR systems for printed and handwritten numerals for Indian scripts are also developed. Sufficient amount of work has been carried out for the development of OCR systems for printed documents of Kannada script.

Some considerable amount of work has been reported on handwritten documents having Kannada script. Sangame et al have proposed an OCR system for recognition of isolated handwritten Kannada vowels by extracting invariant moment features from zoned images and K-nearest neighbor classifier was used for classification [2]. Their work is limited only to vowels with an accuracy of 85.53%. Also if the same features are applied for complete character set, the recognition rate may still go down.Thungamani et al have designed an OCR system for handwritten Kannada text recognition using features extracted from Zernike moments and using support vector machine as classifier [3]. The feature extraction techniques proposed in [2] and [3] uses global features which may not yield consistent feature values and hence result in less recognition rate. Niranjan et al have designed a system for recognition of handwritten Kannada characters using fisher linear discriminant analysis [4]. In paper [4], only unconstrained handwritten characters are considered for recognition. Manjunath Aradhya et al have proposed a probabilistic neural network based approach for handwritten character recognition with an accuracy of 88.64% [5]. They have used fourier transform and principal component analysis for feature extraction, but fourier transform technique is computationally expensive for images containing large data sets.

In [6], a handwritten Kannada (numerals and vowels) and English character recognition system based on spatial features is presented. They have considered only vowels and achieved an accuracy of 90.1%. Venkatesh Narasimha Murthy et al have proposed a novel dexterous technique for fast and accurate recognition of the handwritten Kannada and Tamil characters [7]. They have considered online characters but all characters are not recognizable when the data is collected at the word level. An interesting work on handwritten Kannada Kagunita recognition was described by Leena et al using moment features [8]. They have considered only compound characters known as Kagunita to derive moment features from directional and cut images. In [9], a handwritten Kannada word recognizer with unrestricted vocabulary using statistical dynamic space warping classifier was developed. They have considered online words for recognition. In general, it is observed from the above discussed work that the OCR systems developed for handwritten documents containing Kannada script are limited to either vowels or characters or words and some works have considered online characters or words. The limitations observed in the literature have led to develop a suitable method for extracting efficient features from handwritten Kannada characters. However, in [10], we have presented a quadtree based feature extraction technique for recognizing handwritten Kannada characters and a set of 28 features were extracted as feature values. K-nearest neighbor was used as a classifier for recognition purpose. The features used in [10] are limited to particular quadrant and achieved an overall accuracy of 85.43%. In order to improve the accuracy, a hybrid feature extraction technique which contains a new set of local and global features is used in the proposed model.

KANNADA SCRIPT

Kannada is the official language of Karnataka, a south Indian state. Kannada characters were developed from Kadamba and Chalukya scripts, descendents of Brahmi. Modern Kannada has 51 base characters called Varnamale [11]. There are 16 vowels and 35 consonants and also 10 different numerals. Each of these can modify a primary consonant to form a compound character. Thus, a compound character consists of consonant–vowel and consonant–consonant–vowel combinations. The set of all such combinations together with the base characters are known as aksharas.

PREPROCESSING

Preprocessing is an important step which transforms the raw image into a processed image that is suitable for better feature extraction. Several preprocessing steps are used to achieve transformation such as skew detection and correction, binarization, noise removal, thinning [12, 13]. Some of these preprocessing steps are implemented using built in functions available in Matlab[14]. The preprocessing steps are explained below.

i. Skew detection and correction: It is performed using Fourier transform [15].
ii. Binarization: Before binarization, extra space in the selected character is removed.Binarizationis a process, which converts a gray scale image into binary image using a global thresholding approach.
iii. Noise removal: Median filtering is used for noise removal, which is a nonlinear operation often used in image processing to reduce “salt and pepper” noise. A median filter is more effective when the goal is to simultaneously reduce noise and preserve edges.
iv. Thinning: It is performed to make the image crisper by reducing the binary-valued image regions to lines that approximate the skeletons of the region. Thinning is also known as skeletonization.

Thus, the preprocessed image is prepared ready for further processing such as feature extraction.

PROPOSED MODEL

In this paper, we have proposed a novel feature extraction technique called as Hybrid feature extraction technique, which is explained below.

Feature Extraction Technique

Feature extraction technique is an integral part of any recognition system. Feature extraction is defined as the process of extracting a set of features from the individual character. Different feature extraction methods are reported in [16] for representation of the characters such as solid binary characters, character contours, thinned characters or gray-scale sub images of each individual character.

Before feature extraction, all the images are resized to a constant resolution of 128×128, since scanned images of Kannada characters may have different resolutions. In order to extract the discriminating information in the characters, features can be classified into two categories.

i. Local features: These features contain structural elements like loops, lines, crossing point, branches, joints, curves, strokes, etc. Due to presence of structural elements, local features are called as structural features.
ii. Global features: These features may be topological (zoning, connectivity, projection profiles, number of holes, etc) or statistical (invariant moments, histogram and distance, etc.).

Figure 1 shows the proposed model for feature extraction technique. Initially preprocessed image is taken as an input to feature extraction. During feature extraction, local and global features are extracted from the image and these features are combined to form hybrid features and stored in a feature vector.

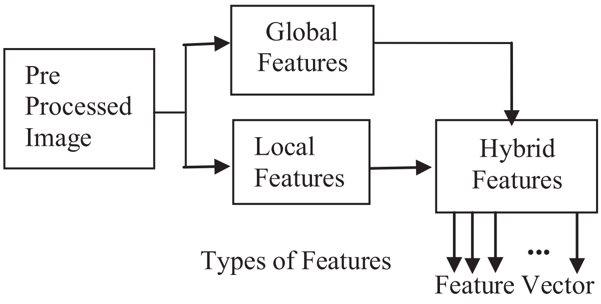


Fig. 1 Feature Extraction Process.

Local Feature Extraction

The aim of the feature extraction method is to select good feature set, which is reasonably invariant with respect to shape variations caused by various writing styles. Before extracting local features, the preprocessed image is divided into four quadrants. We use quadtree concept to partition the input image, upto first level. A quadtree is a tree data structure in which each internal node has exactly four sub nodes [17]. Zone is an alternative name for the quadrant. In the proposed method, let the four zones be Z1, Z2, Z3 and Z4 as shown in the figure 2. A set of 5 features (F1 to F5) are extracted from each zone, which are explained below. Finally 20 local features are extracted from four zones.

Kannada characters have distinct characteristic features due to presence of different segments such as horizontal lines, vertical lines, and curves. The presence of such segments inspired us to extract these features from individual characters.In paper [10], we have used quadtree based technique for extracting seven segments as features. These segments may also be called as strokes. But in this paper, we have reduced these seven features to five features. The reason for using these five features is that, these are quite enough to extract entire character information. Among these five strokes, the first four strokes are horizontal stroke, vertical stroke, primary diagonal stroke and secondary diagonal stroke. For these four strokes, the number of presence of the stroke type is computed.The length of each stroke type is also computed which is the total number of pixels present in that stroke. Through experimentation, it is decided to use a stroke size of five pixels to identify a particular stroke type. Fifth stroke is called zonal density, which is computed for each zone. These strokes are explained below.

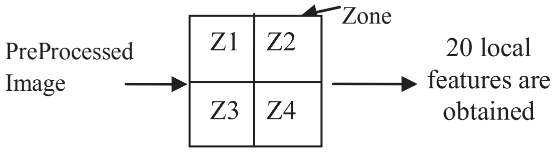


Fig. 2 Process of local feature extraction.

i. Horizontal stroke: The occurrence of the pixels in a segment will be in a single row with the column number being changed in steps of one, either from left to right or right to left. The feature named F1 is computed using the equation (1),
F1=Number of horizontal strokes × Total length of horizontal strokes (1)
ii. Vertical stroke: The occurrence of the pixels in a segment will be in a single column with the row number being changed in steps of one, either from top to bottom or from bottom to top. The feature named F2 is computed using the equation (2),
F2=Number of vertical strokes × Total length of vertical strokes (2)
iii. Primary diagonal stroke:This stroke starts with a column say *i* and moves towards left that is from *i*-1 to 1 till the last pixel of that stroke and in each step the row number is increased by one. The featurenamedF3 is computed using the equation (3),
F3=No. of primary diagonal strokes ×Total length of primary diagonal strokes (3)
iv. Secondary diagonal stroke: This stroke starts with a column say *i* and moves towards right that is from *i*+1 to n till the last pixel of that stroke (n is the last column) and in each step the row number is increased by one. The feature named F4 is computed using the equation (4),
F4=No. of secondary diagonal strokes ×Total length of secondary diagonal strokes (4)
v. Zonal density: This feature is computed for each zone with respect to skeleton of the image. It is needed because characters have different densities at

particular zones. This feature named F5 is computed using the equation(5),

$$F5 = \frac{\text{Number of black pixels in each zone}}{\text{Total number of pixels in entire image}} \qquad (5)$$

From the experimentation, it is decided to use a 3×3 mask that is applied on the image to determine the presence of a particular stroke such as horizontal stroke, vertical stroke, primary diagonal stroke and secondary diagonal stroke. The masks used for detecting these strokes are shown in figure 3.

0	0	0
1	1	1
0	0	0

a) Horizontal stroke

0	1	0
0	1	0
0	1	0

b) Vertical stroke

0	0	1
0	1	0
1	0	0

(c) Primary diagonal

1	0	0
0	1	0
0	0	1

d) Secondary diagonal

Fig. 3 Masks used for Stroke detection.

Global Feature Extraction

In addition to the local features, few global properties are analyzed from Kannada characters based on their distinguishing appearance to obtain global features. Global features used in the proposed method are explained below.

i. Width feature: The preprocessed image is divided into four equal partitions in vertical direction. The number of black pixels in each partition is determined and denoted by n1, n2, n3, n4, respectively. Actual number of pixels along the width of each partition is denoted as 'X'. Then, the width feature 'W' is calculated using the equation (6).

$$\text{Width feature, } W = X \times \left(\frac{\text{Nmax}-\text{Nmin}}{\text{Nmax}}\right) \qquad (6)$$

where X = 32, N_{max}=max{n_i}, N_{min}=min{n_i}, *i*=1,…,4 denotes four partitions.

ii. Image density: It is defined as the ratio of number of black pixels to the total number of pixels in the entire image.

iii. Point feature: A point feature represents a geographic location of interest. It is identified by a point symbol, and it may have feature information stored along with the position information. In point feature,number of end points in the entire image is extracted.

iv. Average Aspect Ratio: Aspect ratio is calculated for each zone as local feature and is defined as the ratio of maximum number of black pixels along the width to the maximum number of pixels along the height of each zone. Average aspect ratio(AAR) is defined as the average of the aspect ratio of the four zones. It is computed using the equation (7),

$$\text{AAR}=\frac{1}{4}\sum_{i=1}^{4} \frac{\text{Pixels along the width}}{\text{Pixels along the height}} \qquad (7)$$

Where *i*=1,…,4 denote the number of zones.

v. Euler Number: It is defined as the number of objects in the region minus number of holes in those objects. Hole is a set of pixels that form a loop.

Hence 5 global features are extracted from entire image.

Combined Features

We have a set of 20 features obtained as local features and 5 features as global features. These local and global features are combined to form hybrid features and stored in a feature vector. This vector contains a total of 25 features and used as an input to recognize the character.

Algorithm for hybrid feature extraction technique

```
Algorithm:Hybrid feature extraction technique
Input: Images containing handwritten Kannada character
Output: 25 Features from the entire image
Algorithm Begins
    Step 1: Local features: The input image is divided into four zonesZ1, Z2, Z3 and Z4. From each zone, a set of 5features are extracted. From four zones, we get a total of 20features.
    Step 2: Global features: A set of 5 features are extracted from the entire image.
    Step 3: The 20 features from step 1 and 5 features from step 2 are combined to form hybrid features that contains a total of 25 features, which is used for recognizing thehandwritten Kannada characters.
Algorithm Ends
```

Classification

An effective technique for classification problemis nearest neighbor classifier in which the pattern classes exhibit a reasonably limited degree of variability. Nearest neighbor classifier is an instance-based learning or lazy learning, where the function is only approximated locally and all computation is deferred until classification. In this classifier, we use a K value which is a positive integer. Based on the value of K, the object is assigned to the class of its nearest neighbor. In this paper, K-nearest neighbor classifier is used for the recognition of handwritten Kannada characters with different values of K=1,2,3. Better accuracy is achieved for recognition of handwritten Kannada characters using K-nearest neighbor classifier with the value of K=3. K-nearest Neighbor classifier is used for classifying total samples of 1200 handwritten Kannadacharacters.

Methods	Sample size	Feature Extraction Technique	Classifier Technique	Accuracy	Remarks
Method 1 [2]	1625	Invariant moment features	K–Nearest Neighbor	85.53%	Tested on only vowels.
Method 2 [5]	5000	Fourier transform and Principal component analysis	Probabilistic Neural Network	68.89%	Tested for both vowels and consonants. This method is computationally expensive.
Method 3 [6]	1400	Directional spatial features	K–Nearest Neighbor	90.1%	Method is tested only on vowels.
Method 4[10]	4800	Quadtree based technique	K–Nearest Neighbor	85.43%	Tested for both vowels and consonants. This method is computationally expensive.
Proposed Method	4800	Hybrid (local and global features) technique	K–Nearest Neighbor	87.33%	Tested for both vowels and consonants. Better accuracy is obtained.

Table 1: Recognition Accuracy

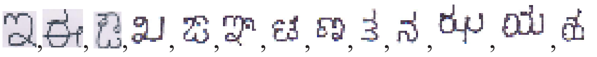
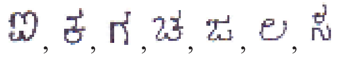
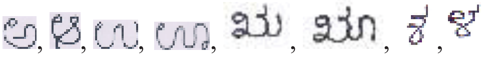
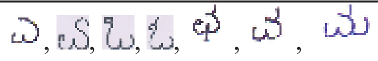
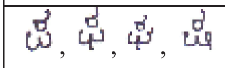
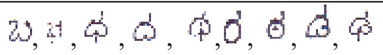
Training samples=3600, Test samples=1200, Number of features=25			
Handwritten Kannada Characters	No. of Training samples	No. of Testing samples	Accuracy in %
	75	25	92.00
	75	25	92.00
	75	25	88.00
	75	25	88.00
	75	25	84.00
	75	25	80.00
Average Accuracy			87.33

Table 2. Comparison of different methods

Experimental Results and Discussions

Handwritten character recognition is an active topic in OCR application and pattern classification.The handwritten Kannada characters are distinct due to font size and style. These characters are created with no restriction on the pen, paper, ink color, ink flow, size, etc.The data set has been created for the experimentation, since there is no standard data set for handwritten Kannada characters. At present, we have considered neatly handwritten Kannada characters for the experimentation purpose. The proposed model is implemented using Matlab in Windows7 platform. We have collected samples of handwritten Kannada characters from differ ent writers belonging to different age groups. The sampled images were scanned with a resolution of 300dpi. For extracting the features, a total of 3600 training sam ples are used. For recognition purpose, a total of 1200 testing samples are used. The recognition of handwritten Kannada characters is achieved using K-nearest neighbor classifier with different values of K=1,2,3. The performance of K–nearest neighbor classifier was better when the value of K=3 is used. Recognition accuracy for different characters isgiven in Table 1.

Some of the Kannada characters are misclassified as other characters. Misclassification occurs due to close resemblance of one character over the other. One such case is occurrence of the feature ‘dot’ may change the meaning of the character, like a dot in the character ‘ra’ may change it to ‘ta’ and a dot in the character ‘dha’ may change it to ’tha’. To handle this case, feature‘dot’is identified by size and position and a corresponding flag is set and recognized symbol is assigned a label based on the status of this flag. As, we are considering handwritten characters, there may be several other cases of misclassification.Because of these misclassifiedcharacters, we have reduction in the accuracy and achieved an overall accuracy of 87.33%. The proposed method is compared with the existing methods [2,5, 6, 10] and it is given in Table 2.

CONCLUSION

In this paper, a hybrid feature extraction technique for recognizing handwritten Kannada characters is presented. Hybrid feature extraction involves local and global feature extraction. The method is simple and found to be effective as the features are extracted from the partitioned image as local features and from the entire image as global features. The experimental results illustrate the performance of the proposed method. Comparative analysis also shows that the features used in the proposed method gives better result compared to the existing methods. In this paper, we have considered neatly handwritten Kannada characters for the experimentation purpose. In future, we consider handwritten Kannada characters with more complexity and extend our work on compound characters.

REFERENCES

[1] RangacharKasturi, Lawrence O’ Gorman, VenuGovindaraju (2002): Document image analysis: A primer, Sadhana, Vol. 27(1), pp. 3–22.

[2] Sangame S.K., Ramteke R.J., Rajkumar Benne (2009): Recognition of isolated handwritten Kannada vowels, Advances in computational research, Vol.1(2), pp. 52–55.

[3] Thungamani M., Dr. Ramakhanth Kumar P., KeshavaPrasanna, Shravani Krishna Rau (2011): Off-line handwritten Kannada text recognition using support vector machines and Zernike moments, International Journal of Computer Science and Network Security, Vol.11(7), pp.128–135.

[4] Niranjan S.K., Vijaya Kumar, Hemantha Kumar G., ManjunathAradhya V.N. (2008): FLD based unconstrained handwritten Kannada character recognition, SecondInternational conference on future generation communication and networking symposia, pp. 7–10.

[5] ManjunathAradhya V.N., Niranjan S.K., Hemanth Kumar G. (2010): Probabilistic neural network based approach for handwritten character recognition,IJCCT, Vol. 1, pp. 9–13.

[6] Dhandra B.V., MallikarjunHangarge, GururajMukarambi (2010): Spatial features for handwritten Kannada and English character recognition,IJCA, special issue on Recent trends in image processing and pattern recognition, pp. 146–151.

[7] VenkateshNarasimha Murthy, AngaraiGanesanRamakrishnan (2011): Choice of classifiers in hierarchical recognition of online handwritten Kannada and Tamil aksharas, Journal of Universal Computer Science,Vol. 17(1), pp. 94–106.

[8] Leena R.Ragha, Sasikumar M. (2011): Feature analysis for handwritten Kannada kagunita recognition, International Journal of Computer Theory and Engineering, Vol.3(1), pp. 94–103.

[9] RiturajKunwar, Shashikiran K., Ramakrishnan A.G. (2010): Online handwritten Kannada word recognizer with unrestricted vocabulary, 12th International conference on frontiers in handwriting recognition, pp.611–616.

[10]Padma M.C, Saleem Pasha (2013): Quadtree Based Feature Extraction Technique for Recognizing Handwritten Kannada Characters, proceedings of International conference in Springer, International conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT–12), Lecture Notes in Electrical Engineering, Vol. 248, Under Press.

[11]Indira K., SethuSelvi S. (2009): Kannada character recognition system: A review,InterJRI science and Technology, Vol. 1(2), pp. 31–42.

[12]Jomy John, Pramod K.V., KannanBalakrishnan (2011): Handwritten character recognition of south Indian scripts: A review, National conference on Indian language computing, kochi, pp. 1–6.

[13]Abdul Rahiman M., Rajasree M.S. (2009): A detailed study and analysis of OCR research in south Indian scripts, International conference on advances in recent technologies in communication and computing, pp.31–38.

[14]Rafael C. Gonzalez, Richard E. Woods, Steven L. Eddins (2011): Digital Image Processing using Matlab, 2nd edition, McGraw Hill.

[15]Postl W.(1986): Detection of liner oblique structure and skew scan in digitized documents, Proceeding of International conference on Pattern Recognition, pp. 687–689.

[16]Trier O.D., Jain A.K., Taxt J.(1996): Feature extraction methods for character recognition: A survey, Pattern recognition Vol. 29(4), pp. 641–662.

[17]Raphael Finkel and J.L. Bentley**:** Quad Trees (1974): A Data Structure for Retrieval on Composite Keys**,**ActaInformatica, Springer-Verlag,Vol. 4(1), pp.1–9.