Survey on Kannada Digits Recognition Using OCR Technique

Vishweshwarayya C. Hallur,
Dept of MCA
AITM, Belgaum

Avinash A. Malawade,
Dept of MCA
AITM, Belgaum

Seema G. Itagi
Dept of MCA
AITM,Belgaum

Abstract— **Day by day technology is going very fast, digital recognitions are playing wide role and providing more scope to perform research in OCR techniques. Recognition of Kannada handwritten numeral is complicated compared to English and other western numerals. However, many researchers have provided real time solution for printed Kannada numerals. Printed numeral documents recognition still offers many motivating challenges to researchers. Current research offers many solutions on Kannada handwritten documents recognition even then reasonable accuracy and concerning handwritten numeral recognition.**

***Index Terms— OCR, pre-processing, image extraction and classification*.**

I. INTRODUCTION

For the development of a high performance OCR algorithm has become essential. OCR research work has been undertaken by several researchers which aim at developing a high performance OCR algorithm.

The purpose behind an OCR is to identify and analyze a document image by dividing the page into line elements, further sub-dividing into words, and then into characters. These characters are compared with image patterns to predict the probable characters. Recognition of characters can be done either from printed documents or from hand written documents.

In particular, Kannada hand written OCR is more complicated than other related work. This is because Kannada numerals have more angles (curves). Challenges that researchers face during recognition process are due to the curve in the numerals and number of strokes and holes, sliding numerals, different writing styles so on.

The steps involved in character recognition comprise pre-processing, segmentation feature extraction and classification. There are three types of features, namely statistical, structural and hybrid can be analyzed there.

Researchers have come up with many approaches for the character recognition, however, many of them have surveyed in the paper. Apart from that, challenges and issues still prevailing in this even for future research has also been surveyed in these papers.

This paper is organized in the following manner; Preprocessing techniques are surveyed in section 2. Section 3 illustrates the various segmentation techniques available. In section 4 Feature extraction methods are explained. Various classification approaches are available explained in section 5. In section 6 the scope of our research in this area and conclusion.

II. HEADINGS AND FOOTNOTES

There are so many numbers of tasks to be completed before performing character recognition. A hand written document must be scanned and converted into a suitable format for processing. Pre-processing consisting of a few types of sub processes to clean the document image and make it appropriate to carry the recognition process accurately. The sub process which gets involved in pre-processing are illustrated bellow:

i. Binarization

ii. Noise reduction

iii. Normalization

iv. Skew correcting, thinning and slant removal

2.1 BINARIZATION

Binarizing in a method of transforming a gray scale image into a black and white image through thresholding [1][2]. Another approach, Otsu’s method may be used to perform histogram based thresholding [3][4] to get binarized image automatically. Most researchers use thresholding concepts to extract the fore ground image from back ground image [5][6][7]. Threshold value is fixed in this method by taking any value between two foreground gray code images. Histogram based thresholding approach can also be used to identify the local gray value contrast of image. This will help to extract text information from low quality documents.

2.2 NOISE REMOVAL

Digital images are having tendency to many types of noises. Noise in a document image is due to poorly photocopied pages. Median filtering [8], Wiener Filtering method [9] and morphological operations can be performed to remove noise [10]. Intensity of the character images are replaced using Median filters [11]. Whereas images are smoothened using Gaussian filters [12].

2.3 NORMALISATION

Normalization is the process of converting random sized images into a standard size. The Bicubic interpolation [3], linear sized normalization [4] and Java Image Class [13] normalization techniques could be used for standard sized images. The Roi-Extraction [14] method is used to get the single structural element from the image. In many works, input images are normalized to a size 40 * 40 after finding the bounding box of each hand written image for razing processing.

2.4 SKEW CORRECTION, THINNING AND SLANT REMOVAL

Thinning is a pre-processor which results in single pixel width image to recognize the hand written character easily. It is applied repeatedly leaving only pixel wide linear representations of the image character. Cumulative Scalar product (CSP)[8] of windows text lock with Gabor filters has been used for thinning purpose. Morphology based thinning algorithm [15] and other thinning algorithms [1][16][17] has also been used for better symbol representation and to thin the character images. Skeletonization is the process of shedding off a pattern to as many pixels as possible without affecting the general shape of the pattern. Skew is inevitably introduced into the incoming document image during document scanning. Fourier spectrum [18], normalization [19] techniques are used for correction of the slant, angle stroke, width and vertical scaling.

III. SEGMENTATION

Segmentation process is used to split the document images into lines, words and characters. Segmentation of the handwritten document is more complex than type written documents. Histogram profiles and connected components analysis [14][20] are used for line segmentation. In this segmentation process, paragraph space has been checked for identifying paragraphs. Image's histogram is used to detect the horizontal line's width [21]. Special space detection technique has been used for word segmentation. Histogram method is used to detect the both width of the words [15] and also to convert the image to glyph [6].

The vertical histogram profile methods [3] [1] is used to find spacing within the lines to identify the word boundary. Region probe algorithm [16] is used to get individual characters from the image. Modified cross counting techniques [12], histogram profile [14] and connected component analysis are also found in the segmentation problem.

IV.FEATURE EXTRACTION

There are three classes of feature extraction namely statistical features, Structural features and the hybrid features. Quantitative measurements are used in statistical feature technique, where as structural techniques use qualitative measurements for feature extraction. In hybrid approach, these two techniques are combined and used for recognition.

4.1 STRUCTURAL TECHNIQUE

Scale invariant feature Transformation (SIFT)[14] is used to transfer the character image into a set of local features. 128 dimensions of SIFE features re identified from the character image. An image is converted to two tone image, and then converted into frame. The frame point obtained from the frame will process the vectors. The normalized feature vector (NFV) obtains the prototype from vector.

4.2 STATESTICAL TECHNIQUE

In the zone based method, pixel destiny is calculated for each zone. Then the pixel destiny is used for representing the features. The height and width of character pixels are counted using the encoding Binary variation approach. The process halts, when the top level of row and width is reached. A feature is extracted from it and a binary flag is set as per the approach.

All images are scaled to the same height and width using bilinear interpolation technique [20]. Sobel edge detection algorithm [20] is used to correct the unwanted portions. Octal graphs [22] are used to derive structural features like end point, holes, length, shape and curvature of individual stroke.

Boundaries of the images are traced using “eight-neighbor” adjustment method. The approach scans until it finds the boundary of image. Then, the Fourier description [11] is used to find the co-efficient and obtain the total number of boundaries. This number of invariant descriptors is given as important to a natural network for future classification.

4.3. HYBRID TECHNIQUE

Hough transform [23] is used to detect the horizontal and vertical lines. They have been analyzing branch and position using another algorithm. Bilinear interpolation [24] is used to extract the features such as slant and strip.

V. CLASSIFICATION

The extracted features are given as the input to the classification process. There are some approaches are used to classify the character features in the existing systems such as K-nearest Neighbor approach, fuzzy system, neural network and so on. For all these approaches, a bag of key points extracted from the feature extraction approaches are used for classification.

TABLE I
LIST OUT THE ACCURACY OBTAINED BY THE VARIOUS OCR AND THEIR APPRECIATION OBTAINED BY VARIOUS CHARECTOR RECOGNITION METHOD

S.No	Title	Accuracy	Appreciations
1[25]	Handwritten numeral/Mixed numeral recognition of south Indian scripts: The zone based feature extraction method	96.10%	2000 training samples and 2000 testing samples were used in experiments
2[26]	Spatial features for multi font/ multi-size kannada numerals and vowels recognition	98.45%	Training samples = 550, test samples= 550 and number of features = 13
3[27]	A single Euler number feature for multi-size, multi-font kannada numeral recognition	99%	A total of 1500 numeral images with different font sizes are tested
4[28]	Multi font/ Multi-size Kannada numeral recognition based on structural features	100%	1150 samples of numerals, 20 font styles and 16 to 50 font sizes
5[29]	Printed and handwritten kannada numerals recognition using directional stroke and directional density with KNN	98.40%	A total 5000 numeral images(4000 handwritten numeral images and 1000 printed numeral images)
6[30]	A script independent approach for handwritten Bilingual kannada and telugu digit recognition	95.50% for KNN classifier, 96.22% for SVM classifiers	Training samples=500,test samples = 500 and number of features =64
7[31]	Zone based features for hand written printed mixed kannada digits recognition	97.32% for handwritten and 98.30% for printed kannada numerals	Training samples=550,test samples = 550 and number of features =64
8[32]	Printed and handwritten mixed kannada numerals recognition using SVM	97.76%	Experimented on 5000 numeral images consisting hadwritten and printed numerals each of size 2500 pixels
9[33]	Zone based feature extraction and statistical classification technique for kannada handwritten numeral recognition	99% with traing samples and 98% with testing samples	Data collected from 125 different writers has 1250 training and testing digits and 2500 numerals

10[34]	Optical character recognition(OCR) for kannada numerals using left bottom 1/4th segment minimum feature extraction	98%	--
11[35]	Hand written numeral recognition of kannada script	95% for NNC classifiers, 92.85% for BPNN classifiers, 96.05% for SVM classifiers	Training samples=2000 and testing samples= 2000
12[35]	Kannada and English numeral recognition system	95.25%	The hand written and printed kannada and English numerals are tested for classification on 4000 sample images
13[37]	Kannada, Telagu and Devanagari handwritten numeral recognition with probabilistic neural network: A script independent approach	96.8% for Kannada, 97.20% for overall accuracy including kannada, Telagu and Devanagari numerals	2550 image samples were taken for recognition result
14[38]	Printed and handwritten kannada numeral recognition using crack codes and Fourier descriptors plate	99.76% for printed numerals and 95.22% for handwritten numerals	2500 printed multi font printed numeral image samples for testing. And 3150 handwritten kannada numeral image samples were taken for testing
15[39]	Kannada, Telagu and Devanagari handwritten numeral recognition with probabilistic neural network: A novel approach	99.40% for kannada, 99.60% for Telagu, and 98.40% for Devanagari numerals	Training samples=2000, testing samples=500 and number of features=13 with radial value 0.05
16[40]	Multilevel classifiers in recognition of handwritten kannada numerals	98%	1600 sample sizes and 44 feature sizes

VI. CONCLUSION

Maximum research work exists in the survey for Kannada Handwritten numeral recognition. However, there is no standard solution to identify all kannada numerals with reasonable accuracy. Different approaches has been used in each phase of recognition process, where as each approach provides solution only for few numeral sets. Challenges still prevails in the recognition of normal as abnormal writing, slanting numerals, similar shaped numerals, curves and so on during recognition process.

The following key challenges can be further explored in my future research work.

- Curves in the Kannada numeral
- Significant variation in writing styles.
- Difficulties faced in viewing angles, shadows and unique fonts.

ACKNOLEDGMENT

Authors are grateful to Dr.R.S.Hegadi, Associate Professor Solapur University, Solapur for his valuable guidance and also grateful to Dr. H. N. Nagaraja, Director and Principal of Angadi Institute of technology and Management, Belgaum for his constant support for research works.

REFERENCES

[1] Shanthi N and Duraiswami K, “Performance comparison of different image size for recognizing unconstrained handwritten Tamil character using SVM”, Journal of Computer Science vol-3(9): page (3) 760-764, 2007
[2] Jagadeesh Kumar R, Prabhakar R and Suresh R.M, “Off-line cursive handwritten Tamil characters recognition”, International Conferences on Security Technology, page(s): 159-164, 2008
[3] Shanthi N and Duraiswami K, “A novel SVM based handwritten Tamil character recognition system”, Springer, Pattern Analysis &Applications, Vol-13, No.2, 173-180, 2010
[4] Ramanathan R, Ponmathavan S, Thaneshwaran L, Arun S. Nair and Valliappan N, “Tamil font recognition using Gabor and support vector machines”, International Conference on Advances in Computing, Control & Telecommunication Technologies, page(s):613-615, 2009

[5] Sigappi A.N, Palanivel S and Ramalingam V, “Handwritten document retrieval system for Tamil language”, Int. J of Computer Application, vol-31, 2011

[6] Suresh Kumar C and Ravichandran T, “Handwritten Tamil character recognition using RCS algorithms”, Int. J. of Computer Applications,(0975-8887) volume-8-no.8, October 2010

[7] Bremananth R and Prakash A, “Tamil numerals identification”, International Conference on Advances in Recent Technologies in Communication and Computing, page(s):620-622, 2009

[9] Stuti Asthana, Farha Haneef and Rakesh K Bhujade, “Handwritten multiscript numeral recognition using artificial neural networks”, Int. J. of Soft Computing and Engineering ISSN:2231-2307, volume-1, Issue-1, March 2011

[10] Sutha J and RamaRaj N, “Neural network based offline Tamil handwritten character recognition system”, International Conference on Computational Intelligence and Multimedia vol: 2, pages: 446-450, 2007

[11] Rajashekararadhya S.V and Vanaja Ranjan P, “Efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south Indian scripts”. Int. J. of Theoretical and Applied Information Technology, pages: 1171-1181, 2008

[12] Paulpandian T and Ganpathy V, “Translation and scale invariant recognition of handwritten Tamil characters using hierarchical neural networks”, Circuits and Systems, IEEE Int. Sym., vol.4, 2439-2441, 1993

[13] Rajashekararadhya S.V, Vanaja Ranjan P and Manjunath Aradhya V.N “Isolated handwritten Kannada and Tamil numeral recognition a novel approach”, First International Conference on Emerging Trends in Engineering and Technology, page(s): 1192-1195, 2008

[14] Subashini A and Kodikara N.D, “A novel SIFE based codebook generation for handwritten Tamil character recognition”, 6th IEEE Int. Conf. on Industrial and Information Systems (ICIIS), page(s):261-264, 2011

[15] Venkatesh J and Suresh Kumar C, “Tamil handwritten character recognition using Kohonon’s self organizing map”, Int. J. of Computer Science and Network Security, Vol.9 No.12, Dec 2009

[16] Jagadeesh Kumar R and Prabhakar R, “Accuracy augmentation of Tamil OCR using algorithm fusion”, Int. J. of Computer Science and Network Security, VOL.8 No5, May 2008

[17] Bhattacharya U, Ghosh S.K and Parui, “A two stage recognition scheme for handwritten Tamil characters”, Ninth International Conference on Document Analysis and Recognition, Vol: 1 page(s):511-515, 2007

[18] Suresh R.M, “Printed and handwritten Tamil characters recognition using Fuzzy technique”, Pro. Of the Int. Multi Conference of Engineers and Computer Scientists, vol 1, 19-2, March, 2008

[19] Sarveswaran K and Ratnaweera, “An adaptive technique for handwritten Tamil character recognition”, International Conference on Intelligent and Advanced Systems, page(s):151-156, 2007

[20] Indra Gandhi R and Iyakutti K, “An attempt to recognize handwritten Tamil character using Kohonen SOM”, Int. J. of Advance d Networking and Applications, Volume: 01 Issue: 03 ages: 188-192, 2009

[21] Banumathi P and Nasira G.M, “Handwritten Tamil character recognition using artificial neural networks”, International Conference on Process Automation, Control and Computing (PACC), page(s): 1-5, 2011

[22] Jagadeesh Kumar R and Prabhakar R, “An improved handwritten Tamil character recognition system using octal graph”, Int. J. of Computer Science, ISSN 1549-3636, Vol 4 (7): 509-516, 2008

[23] Akshay Apte and Harshad Gado, “Tamil character recognition using structural features”, 2010

[24] Hewavitharana S and Fernando H.C, “A two stage classification approach to Tamil handwritten recognition”, Tamil Internet, California, USA, 2002

[25] S.V Rajashekaradhya and Dr. P. Vanaja Ranjan, “Handwritten numeral/Mixed numeral recognition of south Indian scripts: The zone based feature extraction method”, Journal of Theoretical and Applied Information Technology, page(s)63-79, Vol:7.No.1,2009

[26] B.V Dhandra, Mallikarjun Hangarge ang Gururaj Mukarambi, “Spatial features for multi font/font size Kannada numerals and vowels recognition”

[27] B.V Dhandra, R.G Benne and Mallikarjun Hangarge, “A single euler number feature for multi-font multi-size Kannada numeral recognition”, Recent Trends in Information Technology(RTIT-2009), pp101-106

[28] B.V Dhandra, R.G Benne and Mallikarjun Hangarge, “Multi-font multi-size Kannada numerals recognition based on structural features”, Emerging Trends in information Technology, page(s)193-199, 2007

[29] B.V Dhandra, R.G Benne and Mallikarjun Hangarge, “Printed and handwritten Kannada numerals recognition using directional stroke and directional density with KNN”, International Journal of Machine Intelligence (IJMI), pp121-125, Volume: 3, Issue: 3, 2011

[30] B.V Dhandra, Gururaj Mukarambi and Mallikarjun Hangarge, “A script independent approach for handwritten bilingual Kannada and Telugu digits recognition”, International Journal of Machine Intelligence (IJMI), pp155-159, Volume: 3, Issue: 3, 2011

[31] B.V Dhandra, Gururaj Mukarambi and Mallikarjun Hangarge, “Zone based features for handwritten and printed mixed Kannada digits recognition”, International Conference on VLSI, Communication and Instrumentation (ICVCI), pp5-9, 2011

[32] G.G Rajput, Rajeswari Horakeri, Sidramappa Chandrakant, “Printed and handwritten mixed Kannada numerals recognition using SVM”, International Journal on Computer Science and Engineering (IJCSE), vol: 2, No.5, pp1622-1626, 2010

[33] Ashoka H.N, Manjaiah D.H and Rabindranath Bera,”Zone based feature extraction and statistical classification technique for Kannada handwritten numeral recognition”, International Journal on Computer Science and Engineering (IJCSE), Vol: 3, No.10, pp476-482, 2012

[34] K.S Prassana Kumar,” Optical character recognition (OCR) for Kannada numerals using left bottom 1/4th segment minimum feature extraction”, Int. Journal of Computer Technology and Application, Vol: 3(1), pp 221-225, 2012

[35] S.V Rajashekararadhya and P. Vanaja Ranjan, “Handwritten numeral recognition of Kannada script”, Proceedings of the International Workshop on Machine Intelligence Research, pp 80-86, 2009

[36] B.V Dhandra, Gururaj Mukarambi and Mallikarjun Hangarge, “Kannada and English numeral recognition system”, International Journal of Computer Applications (0975-8887), Vol: 26, No.9, pp 17-22, 2011

[37] B.V Dhandra, R.G Benne and Mallikarjun Hangarge, “Kannada, Telugu and Devanagiri handwritten numeral recognition with probabilistic neural network: A script independent approach”, International Journal of Computer Applications (0975-8887), Vol: 26, No.9, pp 11-16, 2011

[38] G.G Rajput, Rajeswari Horakeri, Sidramappa Chandrakant, “Printed and handwritten Kannada numeral recognition using crack codes and Fourier descriptors plate”, IJCA Special issue on Recent Trends in Image Processing and Pattern Recognition, pp 53-58, 2010

[39] B.V Dhandra, R.G Benne and Mallikarjun Hangarge, “Kannada, Telugu and Devanagiri handwritten numeral recognition with probabilistic neural network: A novel approach”, IJCA Special issue on Recent Trends in Image Processing and Pattern Recognition, pp 83-88, 2010

[40] Dinesh Acharya U, N.V Subba Reddy and Krishnamoorthi Makkithaya,” Multilevel classifiers in recognition of handwritten Kannada numerals”, World Academy of Science, Engineering and Technology, pp 278-283, 2008

I.