International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A Comprehensive Survey on OCR Techniques for Kannada Script

Chandrakala H T[1], Dr. Thippeswamy G[2]

[1]BNM Institute of Technology, Department of CSE, Bengaluru, India
[2]BMS Institute of Technology, Department of CSE, Bengaluru, India

Abstract: ***In modern days, there is a pervasive inclination towards digitization of text documents for the ease of their access and maintenance. Digitized documents can be preserved for the future since this form has a longer shelf life. Optical Character Recognition (OCR) system translates a digitized text document from human readable form to machine editable codes. Many commercial OCRs are available today for documents written in English, Japanese, Chinese, Arabic and a few Indian scripts. Kannada is the official language of Karnataka, which is one of the southern states of India. Development of OCR for Kannada script is an active research area currently. Kannada language consists of a large set of characters, many of which are very similar in structure. This makes the job of developing an OCR for this language several magnitudes more complicated than for a language like English. The very fact that research on developing OCRs for Kannada language is very promising and is still emerging necessitated this survey paper. The aim of this paper is to discuss in detail: the peculiarities of the Kannada script, challenges they pose for recognition, techniques reported in the literature, recognition accuracies and a comparison with other OCR systems.***

Keywords: Kannada script, Preprocessing, Feature Extraction, Classification, OCR

1. Introduction

Optical Character Recognition (OCR) is software that can recognize characters in scanned documents and makes it possible for the user to edit or search the document's content. OCR process assigns a character image to a class by using a classification algorithm based on the features extracted from the characters and the relationship among the features[26].In recent years, OCR is an emerging research area of pattern recognition which is a subfield of Document Image Analysis. OCRs are useful in a wide variety of applications like: (i) processing bank cheques without human involvement, (ii) reading aid for the blind, (iii) automatic text entry into the computer for desktop publication, library cataloguing, health care and ledgering, (iv) automatic reading of city names and addresses for postal mail, (v) document data compression, and (vi) natural language processing[23]
It is observed from the literature that many successful efforts to design OCRs with reasonable accuracies are reported for Indian scripts. Sufficient amount of work has been carried out for the development of OCR systems for Kannada script. Generally, the OCR systems designed for printed documents are not suitable for processing handwritten documents. Handwritten Character Recognition is a complex task because of various factors like variations in writing style, mood of the writer, quality of pen and paper, and size of the characters. The aim of this survey is to review works published on OCRs for Printed and Handwritten Kannada script and provide an analysis on how research on OCRs for Kannada language has evolved over the years. In this paper we have discussed the technicalities of works related to printed and handwritten Kannada character recognition. We have provided a detailed comparison of the reported methods in terms of preprocessing, feature extraction, classification and recognition accuracy. We have highlighted shortcomings of existing OCR systems.

The rest of the paper is organized as follows: Section 2 describes the characteristics of Kannada Varnamale. Section 3 gives a detailed discussion of the reported works in the fields of Handwritten Kannada Character Recognition and Printed Kannada Character Recognition Challenges in designing OCRs for Kannada scripts have been discussed in section 4 . Section 5 gives a comprehensive comparison of Kannada OCRs based on preprocessing, feature extraction, classification techniques and recognition accuracy. Conclusions are provided in section 6.

2. The Kannada Script

Karnataka's official language Kannada is written in Brahmi script. [5] It is a phonemic abugida of 49 letters. There are 15 vowels (swaras) and 34 consonants (vyanjanas) in Kannada alphabet as shown in figure 2.1

Written Kannada is composed of akshara or Kagunitha corresponding to syllables formed by combining consonants with diacritics for vowels. There are 34 x 15 = 510 possibilities of such vowel – consonant combinations. A sample listing of Kagunitha for the letters ಠ and ಗ is given in figure 2.2

Kannada script is rich in conjunct consonant clusters (Vattaksharas). There are 34 x 510 = 17340 possible combinations of conjunct consonants. From OCR design perspective, training the classifier to recognize these many combinations of characters is herculean task. A few sample combinations of Vattaksharas are shown in figure 2.3.

Figure 2.1: Kannada Varnamale

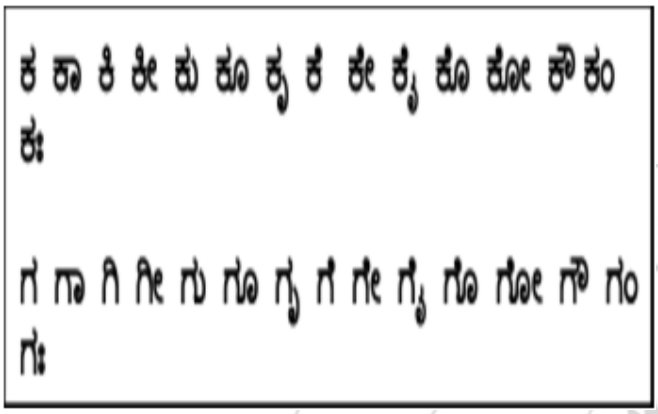

Figure 2.2: Kannada Kagunitha

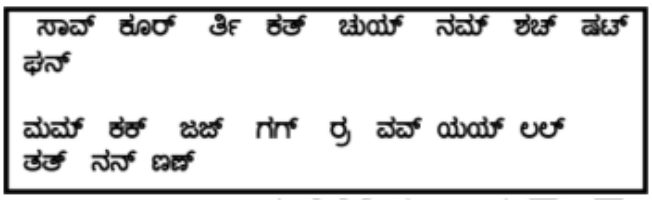

Figure 2.3: Kannada Vattaksharas

3. Related Work

3.1 Printed Kannada Character Recognition

Mamatha etal (2011) [13] have proposed an algorithm for recognition of Kannada vowels with different fonts and sizes. In their approach; they have used Discrete Curvelet Transform of different scales to create the feature vector used for recognition. Standard deviation is performed on the coefficients obtained in order to reduce the size of the feature vector. A k-NN classifier is adopted for classification and is reported to deliver accuracy of 90.17%.

An Online Kannada Text Transliteration system for mobile phones is proposed by Aravindhan etal (2010) [14]. Kannada text is captured using mobile-camera and the edges detected by using canny edge detection technique is sent to server via MMS/e-mail. The server performs feature extraction and recognition. The feature extraction consists of Cavity Analysis, Shape/Hole based Analysis, End Point contour analysis, Boundary analysis. The classification is done based on cavity features like End-point and Junction-point features using Hierarchical tree based classifier.

Karthik etal (2010)[8] propose an OCR for printed Kannada using k-means clustering. They decompose each character into components from 3-base classes' namely base consonants, modifier glyphs and subscripts. They utilize centroids obtained from k-means clustering as features. They suggest that k-means algorithm reduces the size of the training database while achieving comparable accuracy for well segmented texts. Nearest Neighbor classification is used for Recognition with a reported accuracy of 95.01%.

An OCR system for printed basic Kannada characters is presented by Kunte and Samuel (2007) [15]. Segmentation of characters is performed using connected components and projection profile techniques. They have considered Hu's invariant moments and Zernike moments for feature extraction. RBF neural networks are used as classifiers to identify and classify the characters. Their methodology is capable of handling different font sets and sizes. They have achieved a recognition accuracy of 96.8%.

In yet another work, Kunte and Samuel (2007) [16] propose a bilingual OCR system for printed Kannada and English text. Gabor filter based features are used for separating the Kannada and English words from the bilingual document. Discrete Wavelet Transform is applied to the contour points to get wavelet descriptors which serve as features for the characters which the pre-trained neural classifiers uses to recognize the character and generate the class information. Multilayer Perceptrons are being used as classifiers with an overall recognition rate being 90.5%.

Kunte and Samuel (2007) [17] have designed an OCR for a complete set of printed Kannada characters, consisting of more than 600 characters. They apply wavelet transform to the character contours and extract wavelet coefficients which are used as features for further recognition. They reduce the large set of characters by categorizing them into main characters and subscript characters. Then they perform two stage classifications by employing Multilayer Perceptrons. The first stage feeds the character into either main classifier or a subscript classifier depending upon whether it is a main character or subscript character. The second stage of classification selects an appropriate group classifier to further classify a consonant or its composite character within its group. With this approach they have achieved a recognition rate of 91%.

Ashwin and Sastry (2002)[23] proposed the first ever font and size independent OCR system for printed Kannada documents. The individual characters are segmented from the document using projection profile technique and these characters were further subdivided into three zones signifying the top, middle and bottom region of the character. Exploiting the rounded appearance of Kannada characters, they extract the distribution of ON pixels in each zone in radial and angular direction. This forms a 48 dimensional feature vector which is fed to the classification stage. The final recognition is achieved by employing a number of 2-class classifiers based on Support Vector Machine (SVM) method. An overall recognition accuracy of 86.11% is reported.

Sagar etal (2008) [3] have used a database approach to design an OCR for printed Kannada text. After segmenting the text up to character level some metadata about the character, like ASCII value name, width length and total number of ON pixel in the character image are computed. These computed values are compared with the pre-defined values of similar characters already stored in the database. There will be exactly one unique match for each character which is retrieved from the database as same font and size is used for all characters hence giving 100% accuracy.

The same authors have improvised their work by adding a post processing stage [4]. The post processing stage performs spell checking of the recognized Kannada characters to eliminate typographic errors. It also uses a Ternary Search Tree (TST) data structure to store all the valid Kannada suggestions as Dictionary. This makes the database search faster improving the time efficiency of the system.

3.2 Handwritten Kannada Character Recognition

Shashikala and Dhandra (2015) [21] have tried to exploit the curved nature of Kannada characters in their work. They have employed Second Generation Discrete Curvelet transform for feature extraction from Handwritten Kannada vowels and consonants. Recognition has been implemented using KNN classifier with 2-fold cross validation and accuracy of 90.57% is reported. The system is reported to be robust to confusions among similar characters.

Pasha and Padma (2013) [19] have proposed a hybrid feature extraction technique for Handwritten Kannada vowels and characters. This technique extracts local features from the characters using quad tree method and also extracts certain global features like Image density, Aspect ratio, Euler number, Width and point features. The local and global features are combined to form hybrid features useful for classification. KNN classifier is used for recognition and the reported accuracy is 87.33%.

In yet another work, Padma and Pasha (2014) [20] have exploited some distinct features of Kannada characters like lines and curves in various directions. They have employed a Quad tree approach for feature extraction which divides each character into 4 quadrants and extracts 7 line and curve features making a feature vector of 28 features per character. A KNN classifier is trained using this feature set to recognize Handwritten Kannada vowels and characters. The recognition accuracy achieved is 85.43%.

Deepak and Ramakrishnan (2012) [6] have proposed a method for feature extraction and classification of Kannada as well as English characters extracted from scene or natural images. Features are extracted in the form of Discrete Cosine Transform coefficients and are represented using Angular Radial Transform region based descriptors. Nearest Neighbor classifier is used for classification. They have evaluated their method on the complete test set of Chars74k dataset for English and Kannada scripts consisting of handwritten and synthesized characters, as well as characters extracted from camera captured images. Their reported accuracy for Kannada characters is 33.3%.

The impact of grid based approach in offline handwritten Kannada word recognition has been studied by Patel and Sanjay (2014) [10].Their method divides each word into four grids and computes Eigen Vectors of each grid using the subspace learning method Principal Component Analysis. These extracted features are used for classification based on Euclidean distance measure technique and the maximum reported accuracy is 68.57 %.

S A Angadi and S H Angadi (2015) [18] have proposed a recognition method for Hand written Kannada vowels and consonants based on structural features. They have extracted structural and topological features like eccentricity, orientation, area and perimeter from the characters. SVM classifier is used for classification with recognition accuracy reported to be 89.84%vfor vowels and 85.14% for consonants.

Gururaj etal (2012) [7] have designed a bilingual OCR engine for Kannada and English character recognition based on zone features. Morphological opening, closing are used for Binarization of the input images. The individual characters are segmented by dilation and connected component techniques. Then each character is divided into 64 zones and pixel density is computed for each zone, thereby generating 64 features. These features are fed to SVM classifier for recognition. An average percentage of recognition accuracy of 83.02% is obtained for mixture input of both Kannada and English characters.

An OCR for ancient Kannada script of Ashoka and Hoysala periods is presented by A Sowmya and G Hemantha Kumar (2015) [1]. They have employed Nearest Neighbor Clustering algorithm for segmentation of the input image. Statistical features such as Mean, Variance, Standard Deviation, Kurtosis, Skewness, Homogeneity, Contrast, Correlation, Energy and coarseness are extracted to form a training set and for later comparison. Mamdani Fuzzy Classifier is used for recognition and the ancient Kannada is transliterated and displayed in modern Kannada form.

An online handwritten Kannada character recognition system was presented by Vishwas etal (2012) [24] using Direction based Stroke Density (DSD) principle. The input was fed to this system through a digital pen or mouse movement. The system extracted features like direction of the stroke, density of the stroke and number of clicks for those characters. These features were normalized to fall in a narrow range and fed into a Kohonen Neural Network classifier, which calculates a winner neuron with respect to the output weights. The winning neuron is subsequently mapped to the corresponding character to get the result with an accuracy of 94.4%

4. Challenges for Kannada OCR

1. Kannada character set is very vast containing 17340 possible character combinations\newline. If each of this character is considered as a separate class then designing a classifier for recognition of these many classes is very complex.

2. There are many characters in Kannada which are very similar to each other in structure. Hence it is very tedious to train the classifier to recognize these confusing characters accurately.
3. The size of the different characters and words in Kannada is not uniform. This poses difficulty in their segmentation.
4. Unavailability of robust feature set for Kannada script is also a major difficulty for OCR design.
5. If only the image information (shape and structure) is used for character recognition, the OCR is prone to give incorrect results due to the structural complexity of Kannada script. In order to improve the accuracy rate of recognition, it is required to employ post-processing operation. Post processing makes use of language knowledge to correct the recognition result.

5. Comparison

5.1 Printed Character Recognition

5.2 Handwritten Character Recognition

Authors	*Pre processing*	*Input*	*Features*	*Classifier*	*Post processing*	*Reported Accuracy*
Pasha, Padma(2013)	Median Filtering, Global Threshold, Fourier Transform	Handwritten Kannada characters	Local and global (Hybrid) features	KNN	None	87.33%
S A Angadi etal (2015)	Thinning, Resize	Handwritten Kannada characters	Structural features	SVM	None	85.14%
Shashikala etal(2015)	Otsu's global thresholding Median filter	Handwritten Kannada characters	Discrete Curvelet Transform with wrapping technique	KNN	None	90.57%
Padma, Pasha (2014)	Median Filtering, Global Threshold, Fourier Transform	Handwritten Kannada characters	Quad tree based features	KNN	None	85.43%
Deepak Kumar etal (2012)	Min max Binarization	Handwritten Kannada characters from scene images	Discrete Cosine Transform Angular Radial Transform	Nearest Neighbor Classifier	None	33.3%
M S Patel etal (2014)	Noise Removal	Handwritten Kannada words	Grid based approach PCA	Euclidean distance measure	None	68.57%
Gururaj etal(2012)	Morphological Opening and closing	Handwritten Kannada and English characters	Zoning and pixel density	SVM	None	73.3%
A Soumya etal(2015)	Nearest Neighbor Clustering	Handwritten Ancient kannada Epigraphs	Statistical features	Mamdani Fuzzy Classifier	None	Not available
Vishwaas M etal (2012)	Collection of x-y coordinates	Online Handwritten Kannada characters	Direction based Stroke Density	Kohonen Neural Network	None	94.4%

6. Conclusion

Development of OCR for Kannada script is an active research area recently. Recognition of Kannada text is a complicated task due to the structural complexity, large character set and many similar shaped character classes contained in Kannada script. This survey paper is organized around the published works related to recognition of Printed and Handwritten Kannada characters. Various works on OCR for Kannada script have been reviewed in detail in this paper. An effort has been made to throw light on the various research trends in modern OCR systems, compare the techniques adopted and the datasets used in the design of OCR systems. The paper also discusses the challenges specific to Kannada OCR, thus indicating some open problems in this field of research. It was observed that addition of a post processing stage followed by OCR can improve the recognition accuracy.

References

[1] A Soumya and G Hemantha Kumar 2015 Feature Extraction and recognition of Ancient Kannada Epigraphs. Computational Intelligence in Data Mining-Volume 3, Springer, 978-81-322-2202-6-42
[2] A Soumya and G Hemantha Kumar 2014 Recognition of Ancient Kannada Epigraphs using Fuzzy-Based Approach. International Conference on Contemporary Computing and Informatics, IEEE, 978-1-4799-6629-5
[3] B M Sagar, Shobha G and Ramakanth Kumar P 2008 Complete Kannada Optical Character recognition with Syntactical analysis of the script. International Conference on Computing, Communication and Networking IEEE 978-1-4244-3595-1/08
[4] B M Sagar, Shobha G and Ramakanth Kumar P 2008 OCR for printed Kannada text to Machine editable format using Database approach. 9th WSEAS International Conference on AUTOMATION AND INFORMATION 978-960-6766-77-0
[5] Chandrakala H T 2013 A Kannada Document Image Retrieval system based on Correlation Method. International Journal of Computer Applications, 0975-8887.Volume 77-No.3
[6] Deepak Kumar and A G Ramakrishnan 2012 Recognition of Kannada characters extracted from scene images. Proceeding of the workshop on Document Analysis and Recognition Pages 15-21 ACM, New York, USA 978-1-4503-1797-9
[7] Gururaj Mukarambi, B V Dhandra and Mallikarjun Hangarge 2012 A Zone based Character Recognition Engine for Kannada and English Scripts. Elsevier Procedia Engineering 38 3292-3299
[8] Karthik Sheshadri, Pavan Kumar T Ambekar, Deeksha Padma Prasad and Ramakanth P Kumar 2010 An OCR system for printed Kannada using k-means clustering. International Conference on Advances in Computing IEEE 978-1-4244-5697-0/10

[9] K Indira and S Sethu Selvi 2009. Kannada Character Recognition System : A Review InterJRI Science and Technology, Vol. 1
[10] M S Patel and Sanjay Linga Reddy 2014 An impact of Grid based Approach in Offline handwritten Kannada Word Recognition. International Conference on Contemporary Computing and Informatics IEEE 987-1-4799-6629-5/14
[11] M S Patel and Sanjay Linga Reddy 2014 An impact of Grid based approach in Offline Handwritten Kannada Word Recognition. International Conference on Contemporary Computing and Informatics 978-1-4799-6629-5
[12] Madhavaraj A, A G Ramakrishnan, ShivaKumar H R and Nagaraj Bhat 2014 Improved recognition of aged Kannada documents by effective segmentation of merged characters. International workshop on multilingual OCR IEEE 978-1-4799-4665-5/14
[13] Mamatha H R, Sucharitha S and Srikanta Murthy K 2011 Multi-font and Multi-size Kannada Charcter Recognition based on Curvelets and Standard Deviation. International Journal of Computer Applications(0975-8887) Volume 35-No.11
[14] R Aravindhan, G N Rathna and Vipin Gupta 2010 An Online Kannada Text Transliteration System for Mobile Phones. International Conference on Communications and Mobile Computing IEEE 978-0-7695-3989-8/10
[15] R Sanjeev Kunte and R D Sudhaker Samuel 2007 A bilingual machine interface OCR for printed Kannada and English text employing wavelet features. International Conference on Information Technology IEEE 0-7695-3068-0/07
[16] R Sanjeev Kunte and R D Sudhaker Samuel 2007 A simple and efficient Optical Character Recognition system for basic symbols in printed Kannada text. Sadhana 32: 521-533
[17] R Sanjeev Kunte and R D Sudhaker Samuel 2007 An OCR system for printed Kannada text using two-stage Multi-network classification approach employing Wavelet features. International Conference on Computational Intelligence and Multimedia Applications, IEEE, 0-7695-3050-8.
[18] S A Angadi and S H Angadi 2015 Structural features for recognition of handwritten character based on SVM. International Journal of Computer Science, Engineering and Information Technology(IJCSEIT) Vol.5, No.2
[19] Saleem Pasha and M C Padma 2013 Recognition of Handwritten Kannada characters using Hybrid features. International Conference on Emerging Technologies IEEE 978-1-84919-842-4
[20] Saleem Pasha and M C Padma 2014 Quadtree Based Feature Extraction Technique for Recognizing Handwritten Kannada Characters. International Conference on Emerging Research in Electronics, Computer Science and Technology Springer 978-81-322-1157-0_74
[21] Shashikala Parameshwarappa and B V Dhandra 2015 Handwritten Kannada Character Recognition using Curvelet Transform. International Journal of Computer Applications (0975-8887)
[22] Soumen Bag and Gaurav Harit 2013 A survey on Optical Character Recognition for Bangla and Devanagari scripts. Sadhana 38: 133-168
[23] T V Ashwin and P S Sastry 2002 A font and size independent OCR system for printed Kannada documents using support vector machines. Sadhana 27: 35-58
[24] Vishwas M, Arjun M M and Dinesh R 2012 Handwritten Kannada Character Recognition based on Kohonen Neural Network. International Conference on Recent Advances in Computing and Software Systems, IEEE, 978-1-4673-0255-5
[25] A S Kavitha, P Shivakumara, G H Kumar and Tong Lu 2016 Text Segmentation in Degraded Historical Document Images. Egyptian Informatics Journal, Elsevier, 1110-8665
[26] Rangachar Kasturi, Lawrence O Gorman and Venu Govindaraju 2002 Document Image Analysis : A Primer Sadhana Vol 27, Part 1, pp 3-22
[27] V N Manjunatha Aradhya, G Hemantha Kumar, S Noushath and P Shivakumara 2006 Fisher Linear Discriminant Analysis based technique useful for Efficient Character Recognition. IEEE, 1-4244-0612-9

Author Profile

Chandrakala H T received her MTech degree in CSE from Visvesvaraya Technological University in 2012. Currently she is a pursuing her PhD in the field of Digital Image Processing under Visvesvaraya Technological University. She has 6 years of Teaching experience and 2 years of Research experience. She is now working as Assistant Professor in the Department of CSE, BNM Institute of Technology, Bengaluru, India.

Dr. Thippeswamy G received his ME in CSE from Bangalore University in 1997 and PhD in Digital Image Processing from Mangalore University in 2012. He has 21 years of Teaching experience and 6 years of Research experience. He participated in DSI-RFBR sponsored research project entitled "Mathematical models and morphological analysis based algorithm for image compression and classification in computer visual system" and presented his research progress at Lomonosov Moscow state University, Moscow, Russia. He is now working as Professor and HOD in the Department of CSE, BMS Institute of Technology, Bengaluru, India.