STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720 (Online)
Volume 6, Issue 1
www.stmjournals.com

Historical Kannada Handwritten Character Recognition using K-Nearest Neighbour Technique

Parashuram Bannigidad, Chandrashekar Gudada*
Department of Computer Science, School of Mathematics and Computing Science,
Rani Channamma University, Belagavi, Karanataka, India

Abstract

Most of the historical Kannada handwritten documents are preserved in the manuscript preservation center and archaeological department, and these documents are generally degraded in nature and it is very difficult to read and understand the contents in it. Hence, it is very much essential to digitize the historical Kannada handwritten document and recognize its originality of dynasty. The main objective of this paper is to digitize and recognize the historical Kannada handwritten documents by applying bonding box segmentation method and extracting the geometrical shape features, classification is performed by using K-nearest neighbor classifier. The average classification accuracy of the historical Kannada handwritten character from the different dynasties based on their age-type is: Kadamba 97.83%, Badami chalukya 97.78%, Kalyana chalukya 97.92%, Hoysala 97.87%, Vijayanagara 100%, Mysore wodeyars 97.87% and Aadhunika Kannada 97.96%. The results are compared with manual results obtained by the epigraphists and language experts, which demonstrate the efficacy of the proposed method.

Keywords: *Restoration, segmentation, Kannada; K-NN, character recognition, handwritten script, historical documents, document image analysis*

****Author for Correspondence*** E-mail: chandrugudada@gmail.com

INTRODUCTION

Kannada is the most prominently speaking and official language in the state of Karnataka in India, Most of the Indian dynasties were ruled in this area, and Kannada language is one of the oldest among 20 languages of the Dravidian group. Kannada script evolved over a period of 1500 years and variations occurred in the scripts used by different dynasties. The evolution of Kannada character has undergone many inventions and changes during its usage throughout the centuries of dynasties. The first dynastic rule of Karnataka is recognized as *Kadamba script* and can be seen in the scripts of 5th century AD. The Adi Ganga script was used by the Ganga dynasty during 4th to 6th AD and has been classified as *Adi Ganga script*. This script was almost resemblance to the Kadamba script. The script used by the Badami Chalukya is called *Badami Chalukya script* and it can be seen in the records of 6th–7th century AD. The script used by the Rastrakuta rulers is called *Rastrakuta script* and it is available in the records of 8th–10th century AD. The script used by the Kalyana Chalukya rulers is called *Kalyana Chalukya script*, which is found in the records of 10th–12th century AD. Kalyana Chalukya Script is one of the most decorative forms of ancient Kannada script in the dynasties. The Hoysala kings used the *Kalyana chaluyka script* in a more decorative and cursive manner. The present day Kannada script is almost similar script to the Kalyana Chalukya and *Hoysala script*. Soap stones were used during this period to write cursive nature of those scripts. The Vijayanagara rulers have given less importance to the writing material of their records during their rule between 14th and 16th centuries, as most of the records were written on rough granite stone. Hence, the script of Vijayanagara is not decorative or uniform. The Mysore Wodeyars are the last dynasties, which ruled in 18th–19th century AD; in this period, the *Vijayanagar scripts* were evolved and it was almost similar to the *Aadhunika Kannada scripts*.

 © STM Journals 2019. All Rights Reserved

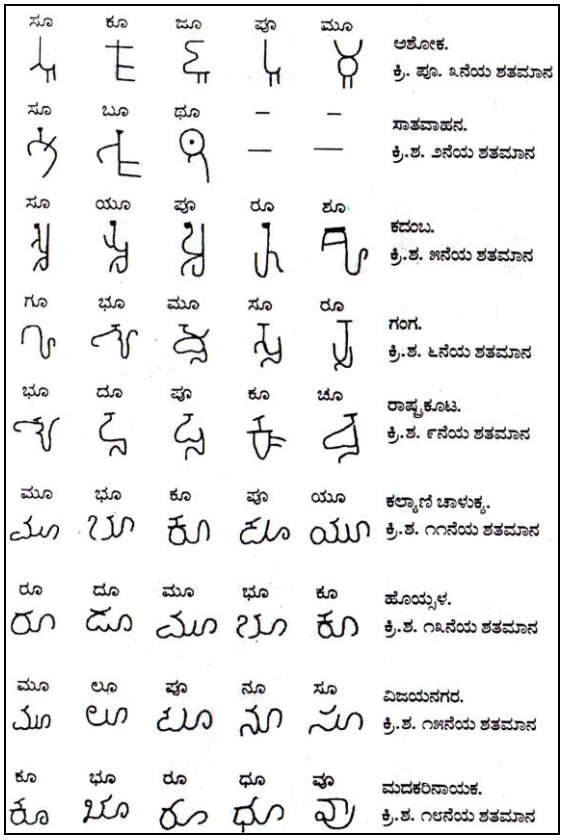

Fig. 1: *The Evolution of Sample Historical Kannada Handwritten Scripts of Different Dynasties.*

The evolutions of sample historical Kannada handwritten scripts of different dynasties are shown in the Figure 1.

The holistic word recognition for handwritten historical documents has been proposed by Lavrenko *et al.* [1]. Handwritten text recognition for historical documents in the transcriptorium project was investigated by Sanchez *et al.* [2]. A feature extraction technique based on character geometry for character recognition was presented by Dileep *et al.* [3]. Handwritten text recognition for historical documents was done by Romero *et al.* [4]. Classical Mongolian words recognition in historical document was investigated by Gao *et al*. [5]. Performance analysis of random forests with SVM and KNN classifiers are used for ancient Kannada scripts was carried out by Soumya *et al.* [6]. Curvelet transform based approach for prediction of epigraphical scripts era was suggested by Gangamma *et al.* [7]. SVM classifier based predication of era of an epigraphical script was carried out by Soumya *et al.* [8]. Recognition of historical records using gabor and zonal features were carried out by Soumya *et al.* [9]. Stone inscripted Kannada character matching using SIFTS has been investigated by Mohana *et al.* [10]. The recent achievements in offline handwriting recognition system were proposed by Verma *et al*. [11]. Age-type identification and recognition of historical Kannada handwritten document images using HOG feature descriptors and GLCM feature was presented by Parashuram *et al*. [12, 13]. Parashuram *et al.* have proposed an image enhancement method for degraded historical Kannada handwritten document images and ensured the quality of images [14–16].

PROPOSED METHOD

The objective of the present study is to recognize the historical Kannada handwritten characters by applying bonding box segmentation method and extracting geometric shape features using K-nearest neighbor technique. The purpose of the proposed method is to recognize and identify the characters in the script of the dynasty, whether it belongs to Kadamba or Badami Chalukya or Kalyana Chalukya or Hoysala or Vijayanagara or Mysore Wodeyars or Aadhunika Kannada? Out of many geometric features used in the literature [3], it has been observed that, there are six geometric features, namely, area, eccentricity, extent, orientation, perimeter and circularity are provides better results, and they are defined below:

- **Area:** It is the actual number of pixels in the region.
- **Eccentricity:** It is the ratio of the length of the highest chord of the shape to the longest chord perpendicular to it.

$$\frac{Major\_axislength}{Minor\_axislength}$$

- **Extent:** It specifies the ratio of pixels in the region to pixels in the total bounding box, computed as the area divided by the area of the bounding box.
- **Orientation:** It is the angle between the x-axis and the major axis of the ellipse that has the same second-moments as the region.
- **Perimeter:** It is the distance around the boundary of the region.

- **Circularity:** $\frac{perimeter^2}{4\pi * area}$

The above geometric features are determined in two different sets i.e., F1 dataset and F2 dataset. These datasets are described as:
F1= {area, eccentricity, extent, orientation and perimeter}
F2= {eccentricity, extent and circularity}

The proposed method is divided into two phases, i.e., Training phase and testing phase and these are described in the form of algorithm is as follows:

Proposed Algorithm

Training Phase

Algorithm: Segmentation and extraction of geometric features from historical Kannada handwritten script.

Step 1. Input the historical Kannada handwritten document (RGB color) image.
Step 2. Convert the given input image into binarized image by using graytrash() function.
Step 3. Obtain the individual character from the binarized image by applying bonding box segmentation method on step 2.
Step 4. Normalize the individual characterized image on step 3 of size 64×64.
Step 5. Apply morphological skeleton operation on step 4 to obtain skeletonized image.
Step 6. For each skeletonized image, compute geometric shape features; i.e., area, eccentricity, extent, orientation, perimeter and circularity and are stored in two separate sets, i.e., F1 and F2. (F1= area, eccentricity, extent, orientation and perimeter; F2= eccentricity, extent and circularity).
Step 7. Repeat steps from 4 to 6 for all the training images of different age-types (dynasties), namely, Kadamba, Badami chalukya, Kalyana chalukya, Hoysala, Vijayanagara, Mysore Wodeyars and Aadhunika Kannada, and store them as knowledge base.

Testing Phase

Algorithm: Recognition and classification of historical Kannada handwritten characters based on their age-type (dynasties):

Step 1. Input the historical Kannada handwritten document (RGB color) image.
Step 2. Convert the given input image into binarized image by using graytrash() function.
Step 3. Obtain the individual character from the binarized image by applying bonding box segmentation method on step 2.
Step 4. Normalize the individual characterized image on step 3 of size 64×64.
Step 5. Apply morphological skeleton operation on step 4 to obtain skeletonized image.
Step 6. For each skeletonized image, compute geometric shape features; area, eccentricity, extent, orientation, perimeter and circularity and are stored in two different sets, i.e. F1 and F2.
Step 7. Repeat steps from 4 to 6 for all the testing images of different age-types of historical Kannada handwritten document images.
Step 8. Apply K-nearest neighbor classifier to recognize and classify historical Kannada handwritten characters based on their geometric shape features extracted from various dynasties; namely, Kadamba, Badami chalukya, Kalyana chalukya, Hoysala, Vijayanagara, Mysore Wodeyars and Aadhunika Kannada. K=1 is minimum distance classifier.

K-NN Classifier

The K-nearest neighbor (K-NN) classification is performed by using a reference data set (training set) which contains both the input (feature set) and the target variables (known characters) and then by comparing the unknown (test data) which contains only the input variables (features) to that reference set. The distance of the unknown characters to the K-nearest neighbors determines its class assignment by either averaging the class numbers of the K-nearest reference points or by obtaining a majority vote from them.

RESULTS AND DISCUSSION

For the purpose of experimentation, the historical Kannada handwritten documents are collected from the book by Manjunath and Devarajaswamy [17]. These Kannada

handwritten scripts are captured through Samsung PL21, 14 Mega pixels Digital Camera at 4320×3240 resolution in JPEG format. The implementation is done on Intel Core i5, @ 2.40 GHz system using MATLAB R2015b. The complete recognition process is divided into two phases, i.e., Training phase and Testing phase. During the training phase, the historical Kannada handwritten RGB document is considered as a input image (Fig. 2(a)), convert this input image into binarized image by using graytrash() function, then apply bonding box segmentation method and is shown in Fig. 2(b) to the binarized image to obtain individual characters and store them as individual image as in Fig. 2(c), next apply normalization technique to normalize the size of the individual character of size 64×64, as shown in Fig. 2(d), then perform morphological skeleton operations on Fig. 2(d) to get the correct structure of the character and extract geometric shape features and store them in two different sets, i.e., F1 and F2. These feature sets are determined based on the experimentation for the purpose of improvement in the accuracy of the results. The details of feature sets F1 and F2 are described in the next part of the paper.

The total number of characters used in the experimentation in two different datasets, i.e., dataset 1 containing 3300 characters and dataset 2 containing 1380 characters from different historical Kannada handwritten documents [18, 19]. These datasets containing different historical Kannada handwritten characters used from various dynasties for recognition are as shown in the Table 1. The classification accuracy computed by extracting F1 (5 features) and F2 (3 features) datasets. The results of dataset 1 using K-NN classifier are shown in Table 2. The accuracy has been computed by extracting F1 (5 features) and F2 (3 features) sets on dataset 2 using K-NN classifier as shown in Table 3. The average classification accuracy of the different dynasties based on their age-type is: Kadamba 97.83%, Badami chalukya 97.78%, Kalyana chalukya 97.92%, Hoysala 97.87%, Vijayanagara 100%, Mysore wodeyars 97.87% and Aadhunika Kannada 97.96%, which is shown in the Table 4. The average classification accuracy of F1 (5 features) and F2 (3 features) sets on dataset 1 and dataset 2 using K-NN classifier is graphically represented in the Fig. 3.

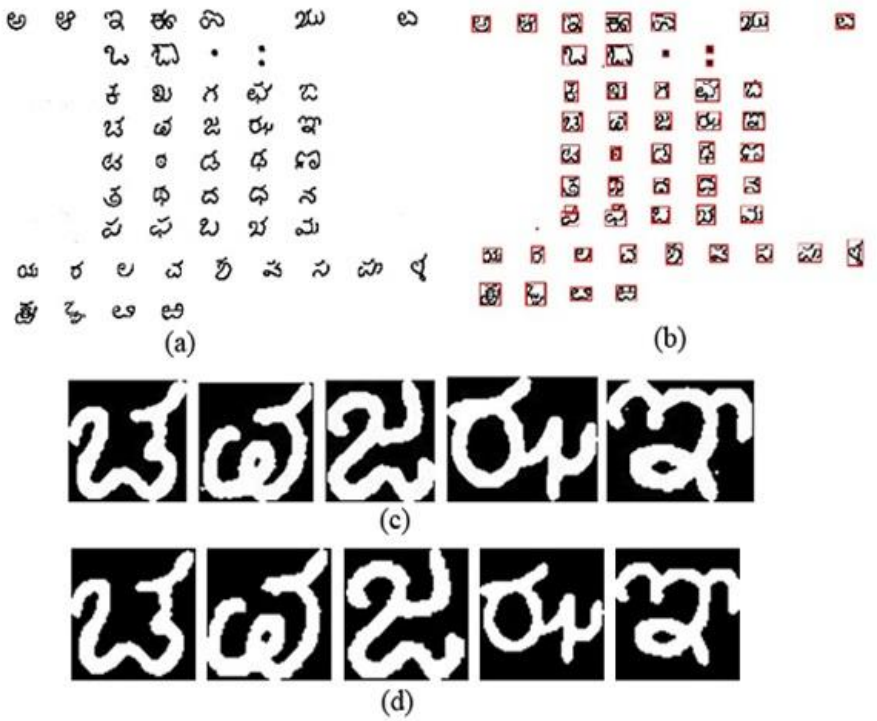

Fig. 2: *Historical Kannada Handwritten Character Recognition:* ***(a)*** *Original Image,* ***(b)*** *Applied Bonding Box Method to Original Image,* ***(c)*** *Sample Segmented Characters Using Bonding Box Method,* ***(d)*** *Sample Normalized Image of 64×64.*

Table 1: *The Datasets Containing Different Historical Kannada Handwritten Characters Used from Various Dynasties for Experimentation.*

Name of Dynasty	Dataset 1	Dataset 2
Kadamba	460	180
Badami Chalukya	450	170
Kalyana Chalukya	480	200
Hoysala	470	190
Vijayanagara	480	220
Mysore Wodeyars	470	190
Aadhunika Kannada	490	230

Table 2: *The Classification Accuracy Computed by Extracting F1 (5 Features) and F2 (3 Features) Sets on Dataset 1 using K-NN Classifier.*

Dynasties	Recognition Rate of Accuracy (%)									
	F1 (5 Features)					F2 (3 Features)				
	K=1	K=2	K=3	K=4	K=5	K=1	K=2	K=3	K=4	K=5
Kadamba	97.83	97.83	97.83	97.83	97.83	95.65	95.65	95.65	95.65	95.65
Badami Chalukya	97.78	97.78	97.78	97.78	97.78	97.78	97.78	97.78	97.78	97.78
Kalyana Chalukya	97.92	97.92	97.92	97.92	97.92	97.92	97.92	97.92	97.92	97.92
Hoysala	97.87	97.87	97.87	97.87	97.87	97.87	97.87	97.87	97.87	97.87
Vijayanagara	100.00	100.00	100.00	100.00	97.83	97.92	97.92	97.92	97.92	97.92
Mysore Wodeyars	97.87	97.87	97.87	97.87	97.87	97.87	97.87	97.87	97.87	97.87
Aadhunika Kannada	97.96	97.96	97.96	97.96	97.96	97.96	97.96	97.96	97.96	97.96

Table 3: *The Classification Accuracy Computed by Extracting F1 (5 Features) and F2 (3 Features) Sets on Dataset 2 using K-NN Classifier.*

Dynasties	Recognition Rate of Accuracy (%)									
	F1 (5 Features)					F2 (3 Features)				
	K=1	K=2	K=3	K=4	K=5	K=1	K=2	K=3	K=4	K=5
Kadamba	94.44	94.44	94.44	94.44	94.44	94.44	94.44	94.44	94.44	94.44
Badami Chalukya	94.12	94.12	94.12	94.12	94.12	94.12	94.12	94.12	94.12	94.12
Kalyana Chalukya	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00	95.00
Hoysala	94.74	94.74	94.74	94.74	94.74	94.74	94.74	94.74	94.74	94.74
Vijayanagara	94.74	94.74	94.74	94.74	94.74	100.00	100.00	100.00	100.00	100.00
Mysore Wodeyars	95.45	95.45	95.45	95.45	95.45	95.45	95.45	95.45	95.45	95.45
Aadhunika Kannada	95.65	95.65	95.65	95.65	95.65	95.45	95.45	95.45	95.45	95.45

Table 4: *The Average Classification Accuracy of F1 (5 Features) and F2 (3 Features) Sets on Dataset 1 and Dataset 2 using K-NN Classifier.*

Dynasties	Recognition Rate of Accuracy (%)			
	F1 (5 Features)		F2 (3 Features)	
	Data Set 1	Data Set 2	Data Set 1	Data Set 2
Kadamba	97.83	94.44	95.65	94.44
Badami Chalukya	97.78	94.12	97.78	94.12
Kalyana Chalukya	97.92	95.00	97.92	95.00
Hoysala	97.87	94.74	97.87	94.74
Vijayanagara	100.00	94.74	97.92	100.00
Mysore Wodeyars	97.87	95.45	97.87	95.45
Aadhunika Kannada	97.96	95.65	97.96	95.45

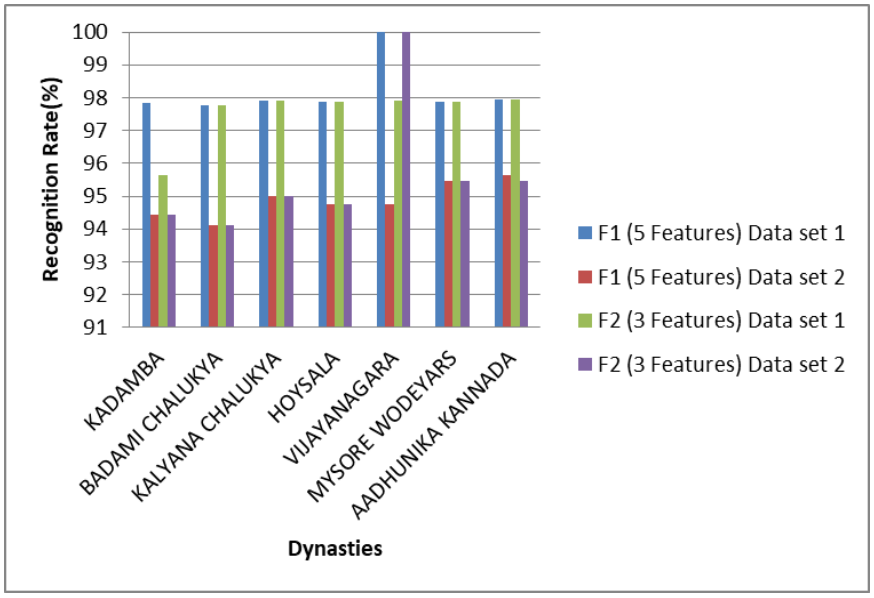

Fig. 3: *The Average Classification Accuracy of F1 (5 Features) and F2 (3 Features) Sets on Dataset 1 and Dataset 2 using K-NN Classifier.*

CONCLUSION

In this paper, we proposed an algorithm to recognize handwritten characters from the historical Kannada handwritten script by applying bonding box segmentation method and extracting geometric features in a two different sets, F1 and F2 and accuracy is calculated by using K-NN classifier. The feature sets are determined based on the experimentation of historical Kannada handwritten characters. The feature set F2 gives better results compared to F1 with reduced features from 5 to 3. The average rate of accuracy of the different dynasties based on their age-type is: Kadamba dynasty 97.83%, Badami chalukya 97.78%, Kalyana chalukya 97.92%, Hoysala 97.87%, Vijayanagara 100%, Mysore wodeyars 97.87% and Aadhunika Kannada is 97.96%. Based on the experimental results, it is observed that, the historical Kannada handwritten characters of Vijayanagar dynasty is achieved 100% recognition results in both the datasets, i.e., F1 with 5 features in dataset 1 and F2 with 3 features in dataset 2, which demonstrate the efficacy of the proposed method. In the present study, only two different handwritten scripts are considered and collected from the historical book. However, the same algorithms need to be implemented in degraded historical Kannada handwritten documents of other writers will be considered in the future work.

ACKNOWLEDGEMENT

The authors are grateful to The Chairman, Department of P. G. Studies and Research in Kannada, Gulbarga University, Kalaburgi and Dept. of Hastaprati, Kannada University, Hampi for providing the historical Kannada handwritten documents and visualization of manual results.

REFERENCES

1. Victor Lavrenko, Rath Toni M, Manmath R. Holistic Word Recognition for Handwritten Documents. *DIAL'04 Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)*. 2004; 278p.
2. Joan Andreu Sanchez, Vicent Bosch, Verónica Romero, *et al.* A Handwritten Text Recognition for Historical Documents in the TranScriptorium Project. *DATeCH'14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*. 2014; 111–117p.
3. Dinesh Dileep Gaurav, Renu Ramesh. A Feature Extraction Technique Based on Character Geometry for Character Recognition. *CoRR*. 2012.
4. Veronica Romero, Nicolas Serrano, Toselli Alejandro H, *et al.* Handwritten Text Recognition for Historical Documents. *Proceedings of Language Technologies for Digital heritage and Cultural Heritage Workshop*. 2011; 90–96p.

5. Guanglai Gao, Xiangdong Su, Hongxi Wei, *et al.* Classical Mongolian Words Recognition in Historical Document. *IEEE, ICDAR*. 2011; 692–697p.
6. Soumya A, Hemantha Kumar G. Performance Analysis of Random Forests with SVM and KNN in Classification of Ancient Kannada Scripts. *International Journal of Computers and Technology*. 2014; 13(9): 4907–4921p.
7. Gangamma B, Srikanta Murthy K, Punitha P. Curvelet Transform Based Approach for Prediction of Epigraphical Scripts Era. *IEEE International Conference on Computational Intelligence and Computing Research*. 2012.
8. Soumya A, Hemantha Kumar G. SVM Classifier for the Predication of Era of an Epigraphical Script. *International Journal of Peer to Peer Networks (IJP2P)*. 2011; 2(2): 12–22p.
9. Soumya A, Hemantha Kumar G. Recognition of Historical Records using Gabor and Zonal Features. *Signal and Image Processing: An International Journal (SIPIJ)*. 2015; 6(4): 57–69p.
10. Mohana HS, Navya K, Srikanth PC, *et al.* Stone Inscripted Kannada Character Matching Using SIFT. *Proceeding of IRF International Conference*. 2014; 126–131p.
11. Verma Brijesh, Blumenstein Michael, Kulkarni S. Recent Achievements in Offline Handwriting Recognition System. *International Conference on Computational Intelligence and Multimedia Applications*. 1998.
12. Parashuram Bannigidad, Chandrashekar Gudada. Age-type Identification and Recognition of Historical Kannada Handwritten Document Images using HOG Feature Descriptors. *Proceedings of The International Conference on Computing, Communication and Signal Processing (ICCASP-2018)* Advances in Intelligent Systems and Computing, vol. 810. Springer, Singapore. 2018; 1001–1010p.
13. Parashuram Bannigidad, Chandrashekar Gudada. Identification and Classification of Historical Kannada Handwritten Document Images Using GLCM Features. *International Journal of Advanced Research in Computer Science*. 2018; 9(1): 686–690p.
14. Parashuram Bannigidad, Chandrashekar Gudada. Restoration of Degraded Historical Kannada Handwritten Document Images using Image Enhancement Techniques. *Proceedings of the Eighth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2016)*, Advances in Intelligent Systems and Computing, vol. 614. Springer, Cham. 2016; 498–508p.
15. Parashuram Bannigidad, Chandrashekar Gudada. Restoration of Degraded Kannada Handwritten Paper Inscriptions (Hastaprati) using Image Enhancement Techniques. *Proceedings of the IEEE International Conference on Computer Communication and Informatics (ICCCI-2017)*. 2017; 1–6p.
16. Parashuram Bannigidad, Chandrashekar Gudada. Restoration of Degraded Non-Uniformally Illuminated Historical Kannada Handwritten Document Images. *International Journal of Computer Engineering and Applications*. 2018; XII(I): 1–13p.
17. Manjunath MG, Devarajaswamy GK. *Kannada Lipi Vikasa*. Jagadhguru Sri Madhvacharya Trust, Sri Raghavendra Swami Matta, Mantralaya; 2004.
18. Narasimha Murthy AV. *Kannada Lipiya Ugama Mattu Vikasa*. Mysore: Kannada Adhyayana Samsthe, Mysore University; 1968.
19. Devarakonda Reddy. *Lipiya Huttu Mattu Belavanige-Origin and Evolution of Script*. Bangalore: Kannada Pustaka Pradhikara (Kannada Book Authority).

Cite this Article
Parashuram Bannigidad, Chandrashekar Gudada. Historical Kannada Handwritten Character Recognition using K-Nearest Neighbour Technique. *Journal of Artificial Intelligence Research & Advances*. 2019; 6(1): 23–29p.