

A Symbolic Representation In Historical Kannada Document

Ravi .P ^{1*}, C. Naveena ², Sharathkumar.Y .H ³

^{1,2} Department of Computer Science And Engineering, SJB Institute of Technology,
Bengaluru,Karnataka,India.

¹ Department of Computer Science And Engineering, Vidyavardhaka College of Engineering, Mysuru,
Karnataka, India.

³ Department of Information Science And Engineering Maharaja Institute of Technology, Mysuru,
Karnataka, India.

^{1,2,3} Affiliated to Visvesvaraya Technological University, Belagavi, Karnataka,India.

* ravipbympr@gmail.com

Abstract











In India, the Kannada is one of the historical and formal languages in Karnataka. The historical Kannada documents gives us information about education, legislation, culture and traditions that have been practiced. Getting such information from stone carvings and palm leaves and other sources are improves our knowledge of Kannada language. Extracting information from historical records is very challenging because of poor quality, variability, contrast and envelope of characters. In this work, the given document is pre-processed using connected component analysis. The features like LBP, Gabor filter and GLTP are extracted. The features are stored in knowledgebase by aggregating and representing as inter-value type data. Symbolic classifier is introduced for the classification purpose. Experiment was conducted on our database to verify the performance of the presented method.

Keywords: Historical Kannada, Symbolic classifier, Ancient scripts, Inscription.

1. Introduction

Kannada is one of the ancient languages of India. Kannada has its own style in scripting and originated before 230BC in the form of ancient scripts, such as epigraphs or inscriptions. Ancient scripts are the primary resources to enhance our knowledge about ancient civilization, which includes traditions, education, legislation, medicine and so on. Ancient scripts are totally different than the current scripts, it is shown in table 1 and To read and identify ancient scripts is not straightforward task because of scripting style but we can identify ancient scripts by epigraphers. This manual recognition method is a time-consuming and tedious task. It is a good idea to develop OCR to automatically identify Kannada ancient scripts to alleviate the difficulties of the manual recognition method. The OCR (Optical Character Recognition) is an essential part of a manuscript image processing method. The LBP, Gabor filter and GLTP features with symbolic classifier to recognize ancient Kannada characters. This automated epigraphs recognition system, it reads the epigraphs, extract the significant features, based on the extracted features performs the classification, and finally recognition the epigraphs. The epigraphs have set of ancient characters. In this presented work, our efforts has been made to recognize the Kannada characters as shown in table 1 through classification and recognition techniques. The paper is structured as follows: in section 2, the related work is presented, in section 3, gives the details proposed recognition system, in section 4, about Symbolic representation, in section 5, illustrates the results and discussion and finally section 6 conclusion.

Table1: Evolution of Character ‘Aha’ from 3rd B.C to Present Kannada.

Sample Scripting Style	Period (Century)	Rulers Name
	3 rd BC	Ashoka
	2 nd AD	Shaatavaahana
	4 th -5 th AD	Kadamba
	6 th AD	Baadami Chaalukya
	9 th AD	Raashtrakuta
	10 th AD	Kalyana Chaalukya
	12 th AD	Hoysala
	15 th AD	Vijayanagara
	18 th AD	Mysore
	After 18th AD	Present Kannada

2. Related Work

The authors proposed symbolic representation [1] technique to extract symbolic features from 2D shapes images and numerous experiments conducted on good number of dataset with good results. Ehtesham et al.[2] presented symbol classifier to recognize character of Gujarati language by using multiple kernel learning. Here 3 different feature depictions applied for symbol images of Gujarati script and got good results. D S Guru et al.[3] proposed a method symbolic classifier for a text document, experiment was performed on a vehicle Wikipedia database for capturing the features and revealing the results obtained with the existing results, it takes relatively less time for text classification. In [4] reported symbolic classifiers for classify text documents by using Symbolic clustering approaches for different measures. The authors [5] presented MKFC-Means method for symbolic features of text documents, proposed method frequency vector, mean and standard deviation were consider for clustering the standard dataset. Writer dependent features discussed [6] in online signature verification application by different filter based features collection technics. Experiment was conducted on standard database MCYT and discussed about importance of symbolic representation, feature vector and relevancy in signature verification problems. A symbolic classifier [7] has used to classification of unlabeled text documents, this work translate the imbalance classes into multiple smaller subclasses by class-wise clustering. After that each subclass denoted as feature vector form and stored in a knowledgebase, the results of proposed technique is better than the other existing methods. To solve the problem of image classification and authentication in document by using feature values from the image documents by class 1 classification built on the symbolic representation is proposed [8]. The authors [9] presented work on recognition method for dissimilarities in font style and size of Kannada character set by using symbolic representation. The Hybrid approach has been reported in [10] for recognition of Amazigh character, in this work the Hough transformation used to extract the directional primitives from pre-processing character image and then trained the Hidden Markov Models by using directional primitives for recognizing the Amazigh character. In [11], the authors presented a system to classify Greek Inscriptions with different classification techniques.

Authors [12] introduced OCR system, which has pre-processing, segmentation, feature extraction and classification for different font sizes of Devnagari characters using different methods for each stage and the recognition rate found to be quite high. Identifying the ancient inscription using methods of image processing, pattern recognition has been reported [13]. The authors proposed [14] a recurrent CNN for text classification and applied a recurring structure to capture related dataset. Using a max-

pooling layer that automatically determines which words play an important role in text classification to capture key points in the text. A text classification from essay dataset by using CNN and RNN approach has been reported[15],The authors conclusion was RNN done better than CNN for essay dataset. Garz et al. [16] presented, identifying text areas and decorative features in ancient scripts and robust method was encouraged by objects recognition method. SIFT descriptors were chosen to identify interest regions, which is used for localization. The authors [17] used local features for effective layout analysis of olden scripts. Here identifies and localizes the layout units in scripts. Hence, the textual units were disassembled into sections and part based finding has done, which employs local gradient features from object recognition fields, SIFT to define these structures. The authors proposed [18] a system to detect and recognize the text regions from scanned images by maximally stable external sections and a trained Convolutional Neural Networks, pre-processes the image, after that find out MSERs, and resulted values saved individually. Then classifier is trained by using resulted values for analyses the characters successfully with error rate less.

3. Proposed System

In this proposed method, shown in figure 1, which described following section. The features like LBP, Gabor filter and GLTP are extracted. Subsequently extracted features will feed into symbolic classifier to obtain class label

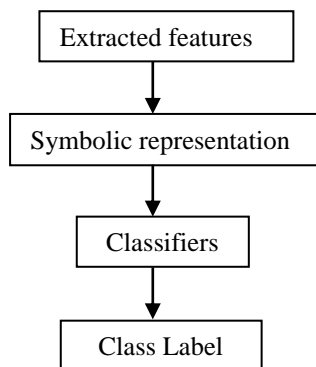


Fig.1. Block diagram of the proposed work

3.1 Pre-processing

In the pre-processing phase, we apply connected component mechanics to identify the elements in the image manuscript. After that, the bounding box is drawn and filled with color for each of the identified components. Then apply horizontal projection so that we can more accurately identify each text line using these methods.

The steps are in pre-processing phase

- Using Sauvolas approach to pre-process the historical image.
- To use connected component mechanics to identify the elements in the binary image.
- To draw bounding box for each recognized component and fill with color.
- A horizontal projection is applied on it and identify each text line, which is shown in figure 2

3.2 Feature extraction

We extracted the features such as LBP, GFR and GLTP from the segmented Kannada character, which are explained in the following section.

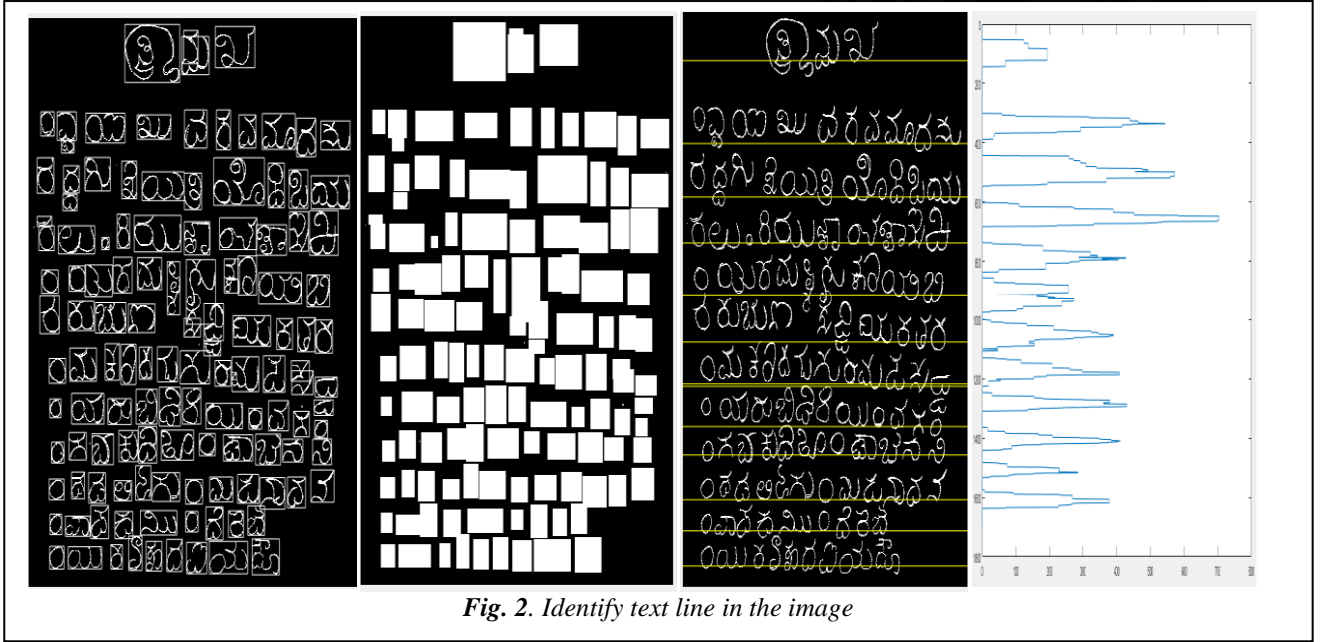


Fig. 2. Identify text line in the image

3.2.1 Local Binary Pattern (LBP)

It is a operator for texture, which symbolizes the spatial arrangement through local image textures. LBP allocates unique label pattern to each pixel dependent on neighbors. Here, g_c is the central pixel gray value of circularly symmetric neighbors g_p ($p = 0,1,\dots,P-1$), g_p is the gray value of its neighbors, P is the number of neighborhood and R is the radius of the neighbors.

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p$$

where $s(g_p - g_c) = \begin{cases} 1, & (g_p - g_c) \geq 0 \\ 0, & (g_p - g_c) < 0 \end{cases}$

3.2.2 Gabor Filter Response (GFR)

The human visual system resembles the representation of the occurrence and direction of a Gabor filter. The Gabor filter is represented by the harmonic function multiplied by the Gaussian function. Because of the multiplier convolution property, the Fourier transform of the Gaussian filter's impulse response is the Fourier transform of the Harmonic and Gaussian function, which is given by

$$g(x, y, \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x}{\lambda} + \psi\right)$$

3.2.3 Gray Level Local Texture Pattern (GLTP)

GLTP was developed through a combination of texture spectrum and LBP. GLTP is a strong against changes in appearance. The number of transitions or breaks is determined by the GLTP in a given pattern. Here identical patterns are allocated through unique labels. All the other irregular patterns are together under a single group. The labeled images are represented by a 1D histogram with the abscissa shows the GLTP label and its occurrence. Here Δg is a +ve value which denotes a desired gray value and forming identical patterns. Uniform measure (U) matching to number of regional transitions in a circular way to form a pattern string and it is well-defined as

$$GLTP_{P,R}^{riu3} = \begin{cases} \sum_{p=1}^P s(g_p, g_c) & \text{if } U \leq 3 \\ P \times 9 + 1 & \text{otherwise} \end{cases}$$

Where

$$s(g_p, g_c) = \begin{cases} 0 & \text{if } g_p < (g_c - \Delta g) \\ 1 & \text{if } (g_c - \Delta g) \leq g_p \leq (g_c + \Delta g) \\ 9 & \text{if } g_p > (g_c + \Delta g) \end{cases}$$

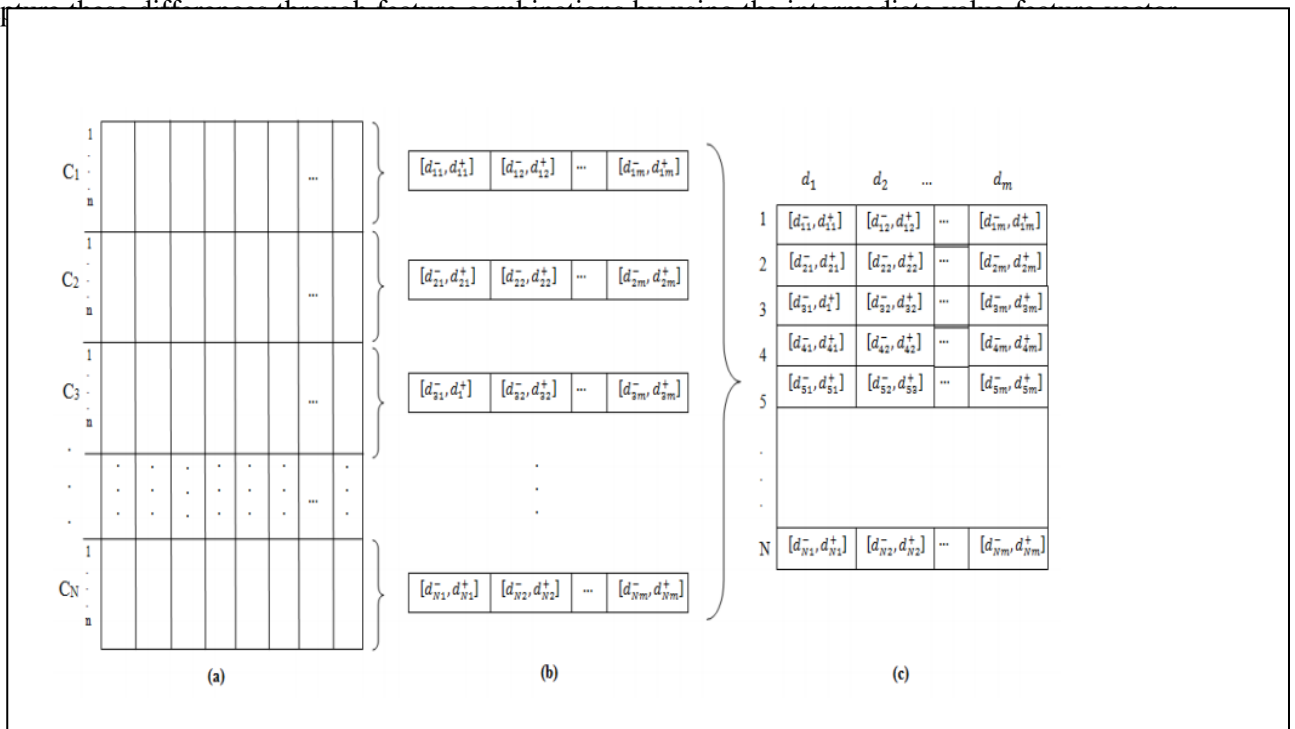
$$p=1,2,3,\dots,P$$

$$U = f(s(g_p, g_c), s(g_1, g_c)) + \sum_{p=2}^p f(s(g_1, g_c), s(g_{p-1}, g_c))$$

$$\text{where, } f(X, Y) = \begin{cases} 1 & \text{if } |X - Y| > 0 \\ 0 & \text{Otherwise} \end{cases}$$

4. Symbolic representation

In this section, we used various features from manuscript samples for symbolic representation. As Kannada records have significant inter-class differences in every subgroup through traditional data characterization, these differences are hard to preserve. Hence, the presented method aims to use unconventional data processing known as symbolic data study, which has the potential to more efficiently reserve the variations in data. In proposed method, symbolic representation is adopted to capture the differences between features by using the interval of the features as



4.1 Classification

In this part, we use symbolic classification to classify the Kannada historical documents. In classification method, an unknown test samples is defined by a set of m distances of type crisp and matches it with the equivalent interval type features of the corresponding symbolic reference samples RF_j , which is stored into knowledgebase unit to ascertain the efficiency.

Let $F_i = [d_{i1}, d_{i2}, d_{i3}, d_{i4}, \dots, d_{im}]$ be an m dimensional vector (of distances between points) labeling a test sample. Let $RF_c: c = 1, 2, 3, \dots, N$ is the representative vectors called symbolic feature stowed in knowledgebase. In classification procedure, each k^{th} feature value of the trial sample is matched with the particular intervals of all the representatives to test if the feature value of the trial sample lies within them. The trial sample is said to belong to the class with a maximum acceptance count A_c . Acceptance count A_c is given by,

$$A_c = \sum_{k=1}^m C(d_{tk}, [d_{tk}^-, d_{jk}^+])$$

$$C(d_{tk}, [d_{tk}^-, d_{jk}^+]) = \begin{cases} 1 & \text{if}(d_{tk} \geq d_{jk}^- \text{ and } d_{tk} \leq d_{jk}^+) \\ 0 & \text{otherwise} \end{cases}$$

Where

When the dataset happens to be vast, there is a probability for a trial to have a similar most exciting acknowledgment check with at least 2 classes. Below such circumstances we recommend to find the contention by use of the associated comparability size which records the similitude esteem between a trial and each one of the clashing classes state j^{th} class.

$$\text{Total_Sim}(F_i, RF_j) = \sum_{k=1}^m C(d_{tk}, [d_{tk}^-, d_{jk}^+])$$

Here $[d_{jk}^-, d_{jk}^+]$ represents the k^{th} feature interval of the j^{th} conflicting class, and

$$C(d_{tk}, [d_{jk}^-, d_{jk}^+]) = \begin{cases} 1 & \text{if}(d_{tk} \geq d_{jk}^- \text{ and } d_{tk} \leq d_{jk}^+) \\ \max\left(\frac{1}{1+|d_{tk}-d_{jk}^-|*\delta}, \frac{1}{1+|d_{tk}-d_{jk}^+|*\delta}\right) + 1 & \text{otherwise} \end{cases}$$

where δ is a normalizing factor.

5. Experimentation

In order for experimentation, The dataset of Kannada's historical letters are shown in figure 4. So as to substantiate the proficiency of the proposed methodology, we completed broad trials on various Character dataset viz. Ashok, Kadamba, Hoysala and Mysuru Scripts. Each character dataset contains 25 pictures shown in figure 5. In this section. We aimed to learning the classification accurateness under different features of LBP, Gabor and GLTP. We picked images arbitrarily from the dataset and experiment is carry-out in a dataset of 4 classes under 70, 50 and 30 different training models from each class. In addition, to exhibit the performances of classifiers and the testing is showed on Min-Max and Mean-Std Deviation representation by changing the training samples.



Fig. 4 . The dataset of Kannada’s historical letters.



Fig. 5. Character dataset of Kannada.

Figure 6 shows accurateness for dataset under changing training data for Min-Max representation. Figure 8 shows accurateness for dataset under changing training samples for Mean-Std Deviation representation. Figure 7 and 9 shows the accuracy of fusion of all the three features. The Graphical representation shows the accuracy for individual features. Further its can understand that Gabor filer achieves the maximum accuracy when compares to individual feature. In addition, we examined the fusion feature by combining LBP with Gabor and GLTP. By analyzing the graphical representation, it shows the fusion feature achieves maximum accuracy when compares to individual feature. The proposed work has shown that the mean-Std deviation representation achieves excellent accuracy compared to the Min-Max representation.

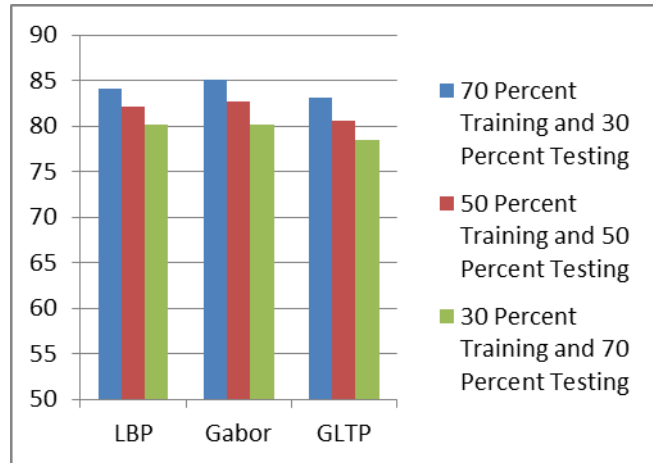


Fig. 6. The accuracy of changing training samples on individual feature for Min-Max representation.

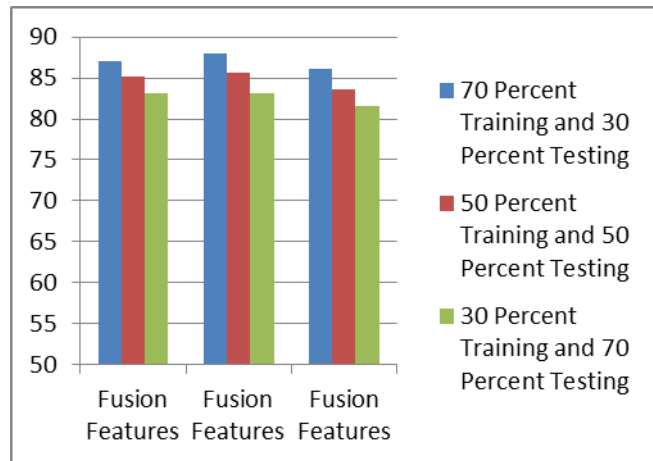


Fig. 7. The accuracy of changing training samples on fusion feature for Min-Max representation.

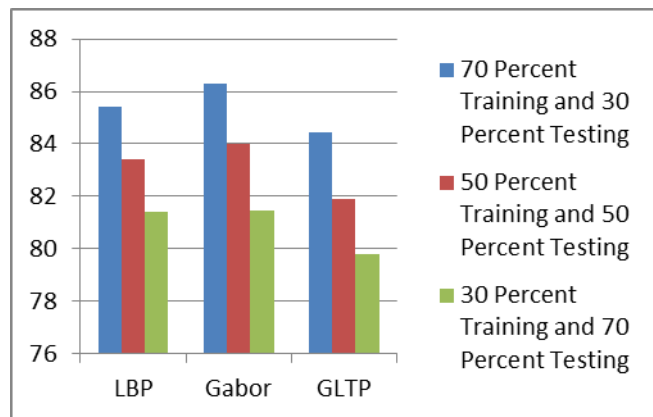


Fig. 8. The accuracy of changing training samples on individual feature for Mean-Std Representation.

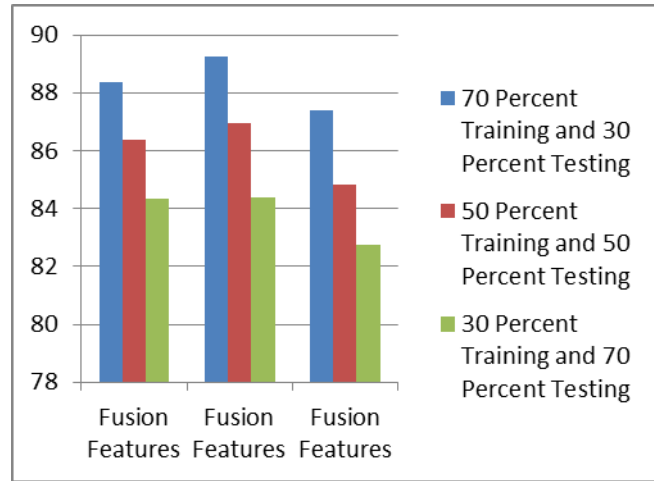


Fig. 9. The accuracy of changing training samples on Fusion feature for Mean-Std Representation.

6. Conclusion

Investigating Historical document is not straight advance procedure because of low quality, differentiation, contrast and covering of characters. In this analysis, the authors propose a LBP, Gabor filter and GLTP features with symbolic classifier to recognize Historical kannada characters. To begin with, the character is divided utilizing Connected Component Analysis and later the Different Features are detached. At long last, form a powerful Symbolic classifier with min-max and Mean-Std deviation representation to recognize the historical Kannada archives. Proposed tale schemes during the preprocessing stage to guarantee strong, precise and constant grouping. They assess their strategy all alone datasets their characterization results surpass 88% on all datasets, which are superior to the cutting edge in this space.

Acknowledgement

The authors acknowledge, Prof. Manjunath M.G. Head of the department, Prasaraaranga, University of Mysore for helping us in making the dataset and studying them.

7. References

- [1] Guru, Devanur & Nagendraswamy, H.' Symbolic representation of two-dimensional shapes'. Pattern Recognition Letters. 2007,pp .144-155.
- [2] Hassan, Ehtesham & Chaudhury, Santanu & Gopal, Madan & Dholakia, Jignesh. 'Use of MKL as symbol classifier for Gujarati character recognition'.2010,pp.255-262.
- [3] Guru Devanur & Harish, B S & Shantharamu, Manjunath.'Symbolic representation of text documents'.2010,pp.1-8.
- [4] Harish, B S & M B, Revanasiddappa & Shantharamu, Manjunath . 'Document Classification using Symbolic Classifiers'. Proceedings of 2014 International Conference on Contemporary Computing and Informatics, IC3I 2014.
- [5] Harish, B S & M B, Revanasiddappa & Aruna kumar, S V.' Symbolic Representation of Text Documents Using Multiple Kernel FCM'. 2015,pp. 93–102.

- [6] D.S. Guru, K.S. Manjunatha, S. Manjunath, M.T. Somashekara, 'Interval valued symbolic representation of writer dependent features for online signature verification', Expert Systems with Applications, Volume 80, 2017, pp. 232-243.
- [7] Swarnalatha, K. and Guru, D. S. and Anami, B. S. and Suhil, M. 'Classwise Clustering for Classification of Imbalanced Text Data' In: Proceedings of International Conference Emerging Research in Electronics, Computer Science and Technology ICERECT 2018. 2019.
- [8] F. Alaei, N. Girard, S. Barrat and J. Ramel, 'A New One-Class Classification Method Based on Symbolic Representation: Application to Document Classification,' 2014 11th IAPR International Workshop on Document Analysis Systems, Tours, 2014, pp. 272-276.
- [9] T. N. Vikram, K. C. Gowda and S. R. Urs, 'Symbolic representation of Kannada characters for recognition' 2008 IEEE International Conference on Networking, Sensing and Control, Sanya, 2008, pp. 823-826.
- [10] M. Amrouch, Y. Es saady, A. Rachidi, M. El Yassa and D. Mammass, 'Printed amazigh character recognition by a hybrid approach based on Hidden Markov Models and the Hough transform' 2009 International Conference on Multimedia Computing and Systems, Ouarzazate, 2009, pp. 356-360.
- [11] C. Papaodysseus, P. Rousopoulos, D. Arabadjis, F. Panopoulou and M. Panagopoulos, 'Handwriting automatic classification: Application to ancient Greek inscriptions' 2010 International Conference on Autonomous and Intelligent Systems, AIS 2010, Povoia de Varzim, 2010, pp. 1-6,
- [12] Singh, Rahul R., Chandra Shekhar Yadav, Prabhat Verma and Vibhash Yadav. 'Optical Character Recognition (OCR) for Printed Devnagari Script Using Artificial Neural Network'. IJCS, 2010, pp. 91-95.
- [13] P. Rousopoulos et al., 'A new approach for ancient inscriptions' writer identification,' 2011 17th International Conference on Digital Signal Processing (DSP), Corfu, 2011, pp. 1-6.
- [14] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 'Recurrent convolutional neural networks for text classification'. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15). AAAI Press, 2015, pp. 2267–2273.
- [15] Kuttala, Radhika & K R, Bindu & Parameswaran, Latha. (2018). 'A text classification model using convolution neural network and recurrent neural network'. International Journal of Pure and Applied Mathematics. 2018. pp. 1549-1554.
- [16] A. Garz, M. Diem and R. Sablatnig, 'Detecting Text Areas and Decorative Elements in Ancient Manuscripts' 2010 12th International Conference on Frontiers in Handwriting Recognition, Kolkata, 2010, pp. 176-181.
- [17] Garz, Angelika & Sablatnig, Robert & Diem, Markus. (2011). 'Using Local Features for Efficient Layout Analysis of Ancient Manuscripts'. European Signal Processing Conference. 2011, pp. 1259-63.
- [18] S. Choudhary, N. K. Singh and S. Chichadwani, 'Text Detection and Recognition from Scene Images using MSER and CNN' 2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAEECC), Bangalore, 2018, pp. 1-4.