RECOGNITION OF HISTORICAL RECORDS USING GABOR AND ZONAL FEATURES

Soumya A[1] and G Hemantha Kumar[2]

[1]Dept. of Computer Science & Engg, R V College of Engineering, Bangalore, India
[2]Dept. of Studies in Computer Science, University of Mysore, Mysore, India

ABSTRACT

The paper addresses the automation of the task of an epigraphist in reading and deciphering inscriptions. The automation steps include Pre-processing, Segmentation, Feature Extraction and Recognition. Pre-processing involves, enhancement of degraded ancient document images which is achieved through Spatial filtering methods, followed by binarization of the enhanced image. Segmentation is carried out using Drop Fall and Water Reservoir approaches, to obtain sampled characters. Next Gabor and Zonal features are extracted for the sampled characters, and stored as feature vectors for training. Artificial Neural Network (ANN) is trained with these feature vectors and later used for classification of new test characters. Finally the classified characters are mapped to characters of modern form. The system showed good results when tested on the nearly 150 samples of ancient Kannada epigraphs from Ashoka and Hoysala periods. An average Recognition accuracy of 80.2% for Ashoka period and 75.6% for Hoysala period is achieved.

KEYWORDS

Optical Character Recognition (OCR), Gabor Features, Zonal Features, Artificial Neural Network (ANN), Epigraphy

1. INTRODUCTION

Character recognition is getting more and more attention since last decade due to its wide range of applications. Epigraphy is the study of inscriptions on rocks, pillars, temple walls, copper plates and other writing material. It is the science of identifying graphemes, clarifying their meanings, classifying their uses according to dates and cultural contexts, and drawing conclusions about the writing. Conversion of handwritten ancient characters is important for making several important documents related to history, such as manuscripts, into machine editable form so that it can be easily accessed and preserved. Independent work is going on in Optical Character Recognition that is processing of printed/computer generated document and handwritten and manually created document processing i.e. handwritten character recognition. There are many OCR systems available for handling printed and handwritten documents of modern form, with reasonable levels of accuracy. However there are not many reported efforts at developing OCR systems for ancient Indian scripts especially for a South Indian script like Kannada.

The current work focuses on developing an automated system to convert an epigraphic document to modern Kannada form, taken from two different periods Ashoka and Hoysala, so as to optimize the recognition rate without much human intervention.

DOI : 10.5121/sipij.2015.6405

Proposed work helps in realizing and understanding the culture and civilization of ancient period. Thus facilitates the archaeologists to explore their research further. Therefore the developed product can be used by archaeologists and historians to identify the characters of different eras and hence know the cultural heritage of the civilization in that era.

This paper is organized as follows: The related works in the area is discussed in Section 2. The system architecture is presented in Section 3. The techniques used in the system, with related theory and mathematical background are covered in Section 4. The methodology used in the current work is given in Section 5. Section 6 covers the experimental results and performance analysis and Section 7 provides concluding remarks.

2. RELATED WORK

Some of the works reported on recognition of characters of Indian and foreign languages are discussed in the section :

The paper [1], investigates the use of moments features on Kannada Kagunita. Kannada characters are curved in nature with some symmetry observed in the shape. Moments and statistical features are extracted from original images, directional images and cut images. These features are used for both vowel and consonant recognition on Multi-Layer Perceptron with Back Propagation Neural Network. The recognition result for vowels are average 86% and consonants are 65% when tested on separate test data. The confusion matrices for both vowels and consonants are analyzed.

A novel zone based method has been presented for recognition of handwritten characters written in Kannada language [2]. The normalized character image is divided into 64 zones each of size 8x8 pixels. For each zone, from left to right and from top to bottom, the crack code, representing the line between the object pixel and the background (the crack), is generated by traversing it in anticlockwise direction. A feature vector of size 512 is obtained for each character. A multi-class SVM is used for the classification purpose. Five-fold cross validation is used for result computation that yielded 87.24% recognition accuracy.

Combining statistical, structural Global transformation and moments features to form hybrid feature vector is reported [3]. The work combines SVM and KNN Classifiers for achieving high accuracy for Devanagari Script. To remove the hitch of misclassification and increase the classifier accuracy SVM and KNN are combined together.

The Neural Network based Bilingual OCR system which can read printed document images, written in two scripts of English and Kannada languages is reported [4]. Document image pre-processor, accepts the bilingual document image and performs grey to two tone conversion, segmentation into lines and words. Dynamic feature extractor extracts distinctive equal number of features from each separated word irrespective of size of the word. These features are accepted by probabilistic neural classifier and are sorted by script, Kannada and Roman. Developed Kannada character recognition system accepts these words and further segments each word into characters and maps the recognized characters into corresponding ASCII values of the chosen Kannada font. Similarly specifically developed English character recognition system, segments English words into characters and maps to corresponding ASCII

value of the specific English font. Thus recognized English and Kannada characters are written into separate ASCII files language wise.

Directional spatial features like stroke density, stroke length and the number of stokes are employed as potential features to characterize the handwritten Kannada numerals and English uppercase alphabets [5]. KNN classifier is used to classify the characters based on these features with four fold cross validation. The proposed system achieves the recognition accuracy as 96.2% and 91.04% for handwritten Kannada numerals and English uppercase alphabets respectively.

An ancient document recognition process consisting of two stages: training with collected character image examples and classification of new character images [6] is reported. The proposed OCR builds fuzzy membership functions from oriented features extracted using Gabor filter banks. Results on a significant test led to a character recognition success rate of 88%.

The problem of recognizing early Christian Greek manuscripts written in lower case letters [7] is reported. Based on the existence of closed cavity regions in the majority of characters and character ligatures in these scripts, a novel, segmentation-free, fast and efficient technique that assists the recognition procedure by tracing and recognizing the most frequently appearing characters or character ligatures. The proposed method gives highly accurate results and offers great assistance to old Greek handwritten manuscript OCR.

The script identification problem of handwritten document images, which facilitates many important applications such as sorting, transcription of multilingual documents and indexing of large collection of such images, or as a precursor to optical character recognition (OCR) [8] is reported. The script identification scheme proposed in this paper has two phases. First phase reports the script identification of text words using global and local features, extracted by morphological filters and regional descriptors of three major Indian languages/scripts: Kannada, Roman and Devnagari. In the second phase Kannada and Roman handwritten numerals script identification is carried out. For classification of text words and numerals, a K nearest neighbor algorithm is used. The proposed algorithm achieves an average maximum recognition accuracy of 96.05% and 99% respectively for text words and numerals with fivefold cross validation test.
The method of segmenting Nom historical documents and clustering character patterns to build a Nom character pattern database [9] is reported. A projection profile based method is used for segmenting hundreds of pages into individual characters. Then, a combination of Chinese OCR-based clustering and K-means clustering to group characters into categories are implemented. The experiment shows that the proposed system can help collecting the characters patterns effectively.

3. SYSTEM ARCHITECTURE

The architecture of OCR system for historical records is shown in Figure 1. The intermediate components of the system consist of: Preprocessing, Segmentation, Feature extraction, Recognition and Post-processing [11, 12]. The input to the system is ancient Kannada epigraphic documents.

- **Preprocessing:** This sub-system performs noise removal, deblurring, filtering and binarization on the input image. Next samples out characters from preprocessed ancient documents.

- **Feature Extraction:** This component extracts features from the input image and stores the extracted features in a feature vector.
.
- **Classification:** Artificial Neural Network (ANN) classifies the segmented characters of the test image.

- **Post-Processing:** This subsystem maps the results of classification to modern form.

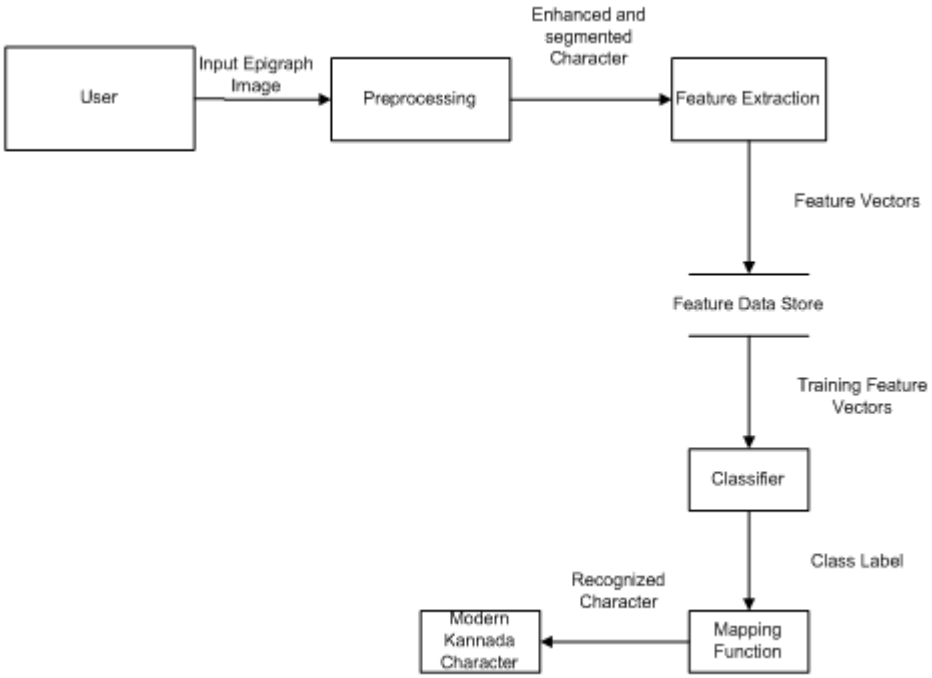

Figure 1. Architecture of OCR System

4. REVIEW ON THE APPROACHES USED IN PROPOSED SYSTEM

The methods used in current system are described in this section:

Gabor features and Zone based features are extracted from the characters and stored as feature vectors. Artificial Neural Network is used for classification of ancient characters.

4.1 Preprocessing and Segmentation

Preprocessing stage of ancient document images includes: enhancement and reduction in noise, which is achieved through different spatial filtering methods namely Bilateral, Mean, Median, and Gaussian Blur Filters. This is followed by binarization of the enhanced image to highlight the foreground information, using Otsu thresholding algorithm. Character segmentation is performed based on Drop Fall and Water Reservoir concept [14].

4.2 Feature Extraction

4.2.1 Gabor Feature Extraction

In image processing, a Gabor filter, named after Dennis Gabor, is a linear filter used for edge detection. Gabor filters act very similar to mammalian visual cortical cells so they extract features from different orientation and different scales. Frequency and orientation representations of Gabor filters are similar to those of the human visual system, and they have been found to be particularly appropriate for texture representation and discrimination.The filters have been shown to possess optimal localization properties in both spatial and frequency domain and thus are well suited for texture segmentation problems. Gabor filters have been used in many applications, such as texture segmentation, target detection, fractal dimension management, document analysis, edge detection, retina identification, image coding and image representation. A Gabor filter can be viewed as a sinusoidal plane of particular frequency and orientation, modulated by a Gaussian envelope. In the spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave.

The impulse response of gabor filter is defined by a sinusoidal wave (a plane wave for 2D Gabor filters) multiplied by a Gaussian function. Because of the multiplication-convolution property (Convolution theorem), the Fourier transform of a Gabor filter's impulse response is the convolution of the Fourier transform of the harmonic function and the Fourier transform of the Gaussian function. The filter has a real and an imaginary component representing orthogonal directions. The two components may be formed into a complex number or used individually.

Complex

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right) \quad (1)$$

Real

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

Imaginary

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \sin\left(2\pi\frac{x'}{\lambda} + \psi\right) \quad (3)$$

where

$$x' = x\cos\theta + y\sin\theta \quad \text{and} \quad y' = -x\sin\theta + y\cos\theta$$

In this equation, λ represents the wavelength of the sinusoidal factor, θ represents the orientation of the normal to the parallel stripes of a Gabor function, ψ is the phase offset, σ is the sigma/standard deviation of the Gaussian envelope and γ is the spatial aspect ratio, and specifies the ellipticity of the support of the Gabor function.

Figure 2 represents the Gabor Filters of a sampled character at two scales and four orientations.

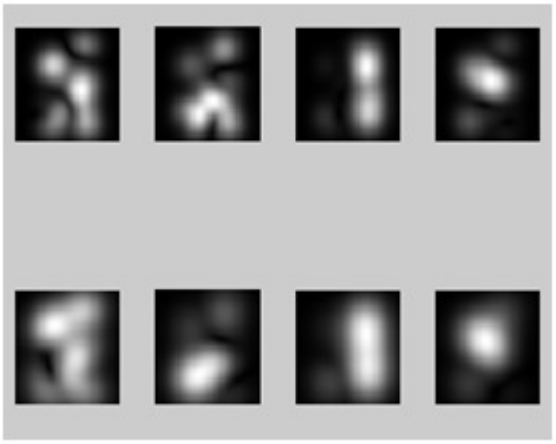

Figure 2(a) Magnitudes of Gabor Filter

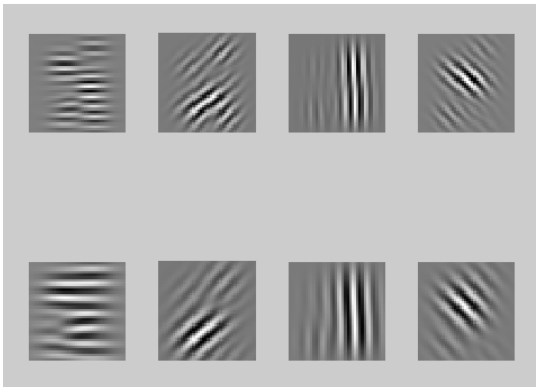

Figure 2(b) Real parts of Gabor Filter

Figure 2. Gabor Filtered images of a sampled character

4.2.2 Zone-based Feature Extraction

In the field of OCR zoning is used to extract topological information from patterns. The segmented image is divided into some 'n' zones and from each zone statistical features like number of horizontal/vertical/diagonal lines, length of horizontal/vertical/diagonal lines, total number of intersection points are extracted.

4.3 Classification

Artificial neural networks (ANNs) are computational models inspired by human central nervous system (in particular the brain) which is capable of machine learning as well as pattern recognition. Artificial neural networks are generally presented as systems of interconnected neurons which can compute values from inputs.
ANNs possess the following characteristics:

- consist of sets of adaptive weights, i.e. numerical parameters that are tuned by a learning algorithm, and
- are capable of approximating non-linear functions of their inputs.

The adaptive weights are conceptually connection strengths between neurons, which are activated during training and prediction.

For example, a neural network for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function (determined by the network's designer), the activations of these neurons are then passed on to other neurons. This process is repeated until finally, an output neuron is activated. This determines which character was read.

5. PROPOSED SYSTEM AND METHODOLOGY

5.1 Structure Chart OCR System

Figure 3 represents the Structure chart of the OCR system developed. It consists of five modules:

- Image Acquisition
- Pre Processing
- Segmentation
- Feature Extraction
- Classification
- Post processing

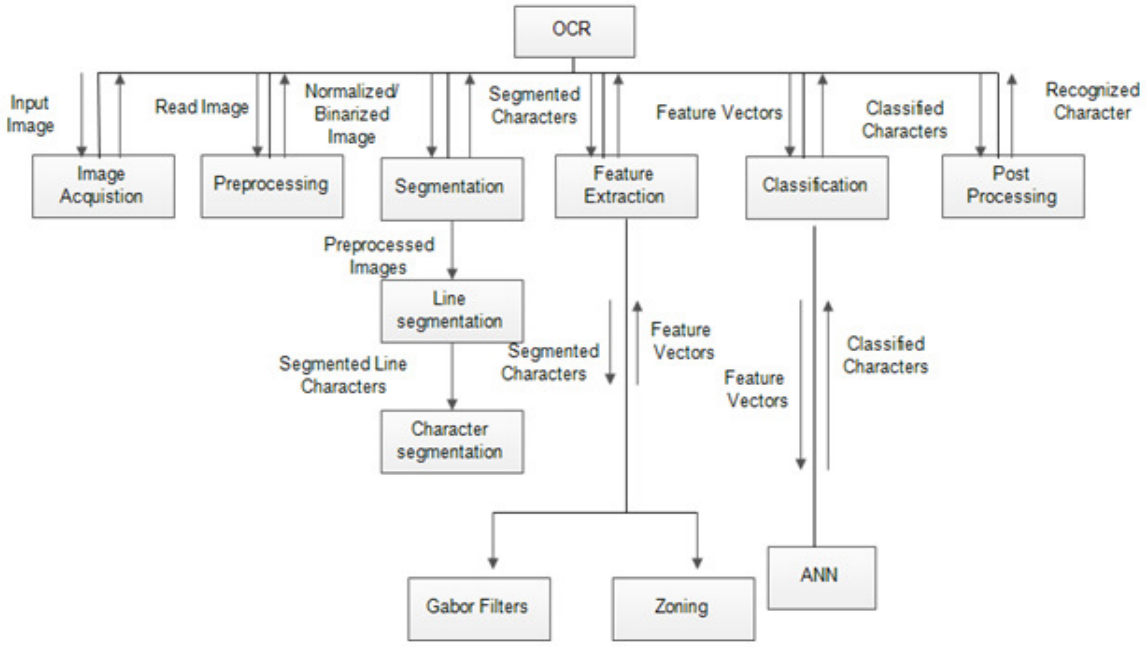

Figure 3. Structure Chart of OCR System

The image is acquired from the file store, which contains character images of two different dynasties namely Ashoka and Hoysala. The input image is fed to the Preprocessing module, that converts input epigraphic image to Gray scale, enhances the image by reducing noise, and performs binarization. Pre-processed epigraph is sent to Segmentation module which samples out the characters from the document. Next the Feature Extraction module extracts relevant features for the sampled characters. Finally Classification module labels the test character using ANN classifier. The classified characters are sent to Post processing module which maps the classified character to the present Kannada character.

5.2 Methodology

The functionality of the phases – Preprocessing, Segmentation, Feature extraction and Recognition is covered in this section. The ancient input image is first converted to grayscale, enhanced and noise is reduced, binarized and the characters are sampled out from the preprocessed document [14]. Gabor and Zonal features are extracted for the characters and stored as feature vectors. ANN is trained with these feature vectors in training phase and later used for classification of test characters. Finally the results of classification are displayed in modern form.

The proposed system includes following processing steps: Image acquisition, Pre-Processing, Segmentation, Feature Extraction, Classification and Post-Processing.

- **Image acquisition:** Scanned handwritten documents of ancient Kannada epigraphs are read as input to the OCR system.
- **Preprocessing:** Preprocessing involves operations like RGB to gray scale conversion, noise removal, filtering, deblurring, binarization of the input image to make it suitable for the next phase, segmentation.
- **Segmentation:** Segmentation is the process of extracting the characters in pre-processed image.
- **Feature Extraction:** Zone-based and Gabor features are extracted for segmented characters and stored in feature vector.
- **Classification:** Artificial Neural Network is trained with feature vectors of the sampled handwritten characters. The trained ANN is then used to classify the segmented characters from test images.
- **Post-Processing**: The classified ancient characters are mapped to modern form.

5.2.1 Pre-Processing

The design and implementation of the phases – Preprocessing and Segmentation is carried in the work [14]. The ancient input image is first converted to grayscale image. The image is enhanced and noise is reduced by applying four different filters with mask size of 2x2 and 4x4. Next, the enhanced image is converted to binary image using Otsu's approach. Lastly the characters in the document are sampled out using Drop Fall and Water Reservoir approaches.

- **Image Enhancement**

The input ancient epigraph is enhanced and noise is reduced [14]

- **Input:** Degraded Gray Scale image

- **Functionality:** Enhances epigraphic image of medium-level degradation using spatial filters namely Median, Gaussian blur, Mean, Bilateral filter on the input image
- **Output:** The enhanced image with reduced noise.

➢ **Binarization**

The enhanced image is binarized using Otsu's Thresholding Algorithm [14].

- **Input:** Enhanced Image
- **Functionality:** The enhanced image are converted to binary image using Otsu's approach
- **Output:** Binarized image

5.2.2 Segmentation

The segmentation of characters in the document image is carried out using Drop Fall and Water Reservoir Approaches [14].

- **Input:** Binary epigraph image
- **Functionality:** The binarized image is segmented to characters
- **Output:** Segmented characters of the input epigraph.

5.2.3 Feature Extraction

- **Input:** Segmented character
- **Functionality:** Gabor filter and Zone based approach used for extracting features from the segmented character
- **Output:** The creation of feature vector

- **Algorithm for extraction of Gabor Features**

  [Step 1]: Compute the orientation

  [**Step 2]:** Compute the gabor filter bank

  [Step 3]: Convolve it using conv2 function.

  [Step 3]: Downsample the image by factors of the size of image

  [Step 4]: Store the resultant value in a feature vector.

- **Algorithm for extraction of Zonal features**

  [Step 1]: Input image is divided into 9 zones of equal size

  [Step 2]: From each zone following features are extracted.

  The number of horizontal lines
  The total length of horizontal lines,
  The number of right diagonal lines,
  The total length of right diagonal lines,
  The number of vertical lines,

The total length of vertical lines,
The number of left diagonal lines,
The total length of left diagonal lines and
The number of intersection points

[Step 3]: Extracted features are stored in new feature vector.

The features from both techniques are concatenated and used in the next phase, classification.

5.2.4 Classification

Artificial Neural Network (ANN) produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

- **Input:** The features obtained from the feature extraction stage.
- **Functionality:** ANN is trained with the feature vectors created in the previous phase and later used in classification of test data.
- **Output:** The output is the class label of each character present in input image
- **Algorithm for Classification**

[Step 1]: ANN is trained with the feature vectors of characters during training phase.

[Step 2]: Test character input is fed to the already trained ANN.

[Step 3]: Network predicts and outputs the class labels of the test characters.

6. EXPERIMENTAL RESULTS, ANALYSIS AND DISCUSSION

6.1 Experimental Results

The system developed is tested on nearly 150 ancient epigraphic images and the results are found to be satisfactory. Figure 4 shows a sample input ancient Kannada script of Ashoka period.

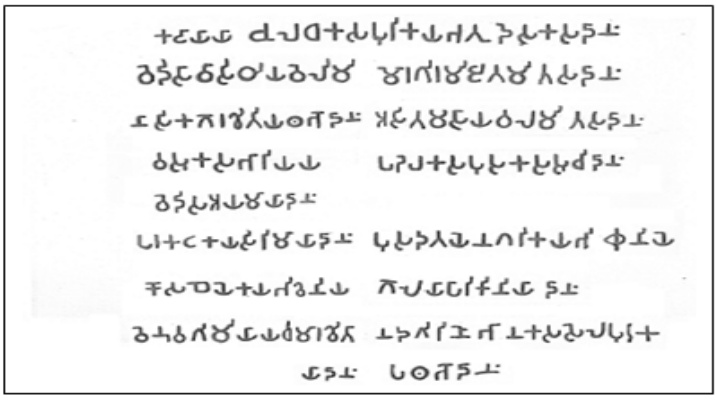

Figure 4. Sample Input of Historical Record from Ashoka period

The results of Pre-processing, Segmentation and Recognition of input image (in Figure 4) are depicted in Figure 5, 6 and 7 respectively.

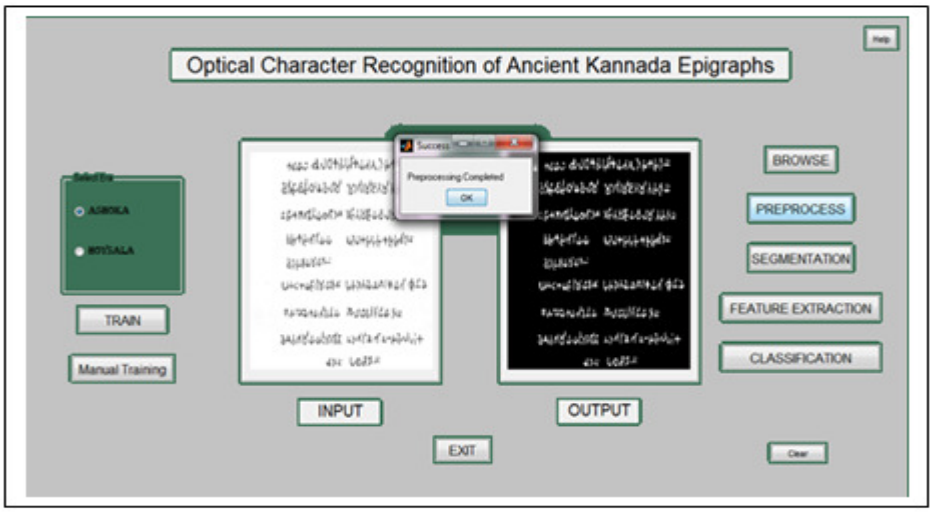

Figure 5. The results of Preprocessing the Historical record

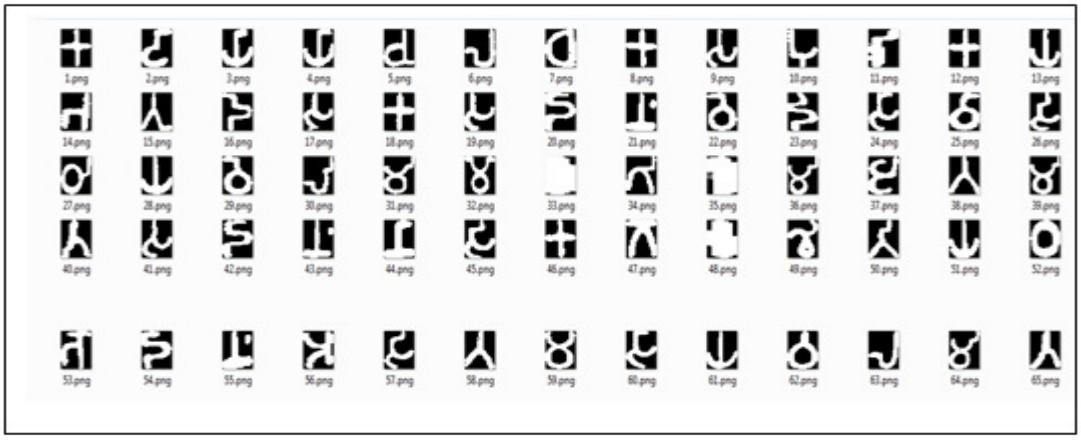

Figure 6. The result of Segmentation

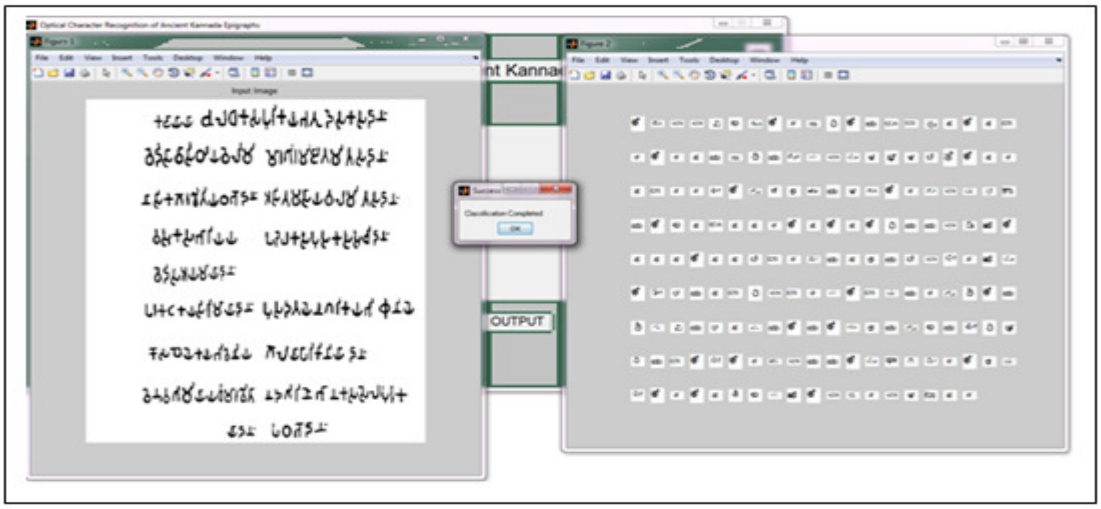

Figure 7. The result of Recognition

6.2 Experimental Data Set

- **Training Data**

The training database of Ashokan Brahmi script contains 8 vowels and 33 consonants. There are 258 different characters hence 258 class labels. Four instances of each character were used, so this gave raise to (250+8=258)*4=1032 characters which forms input training set.

The training database of Hoysala script contains 11 vowels and 34 consonants. There are totally 362 different characters hence 362 class labels. Four instances of each character were used, so this gave raise to (351+11=362)*4=1448 characters which forms input training set.

- **Test Data**

The system was tested with 150 ancient Kannada epigraphs from Ashoka and Hoysala period.

6.3 Performance Analysis

Recognition rate is the metric used to evaluate the performance of the developed system. The system was tested with 150 epigraphic images of Ashoka and Hoysala periods. The average Recognition rate of ancient Kannada Script is observed to be 80.2% for Ashoka period and 75.6% for Hoysala period. This is represented in the plot shown in Figure 8.

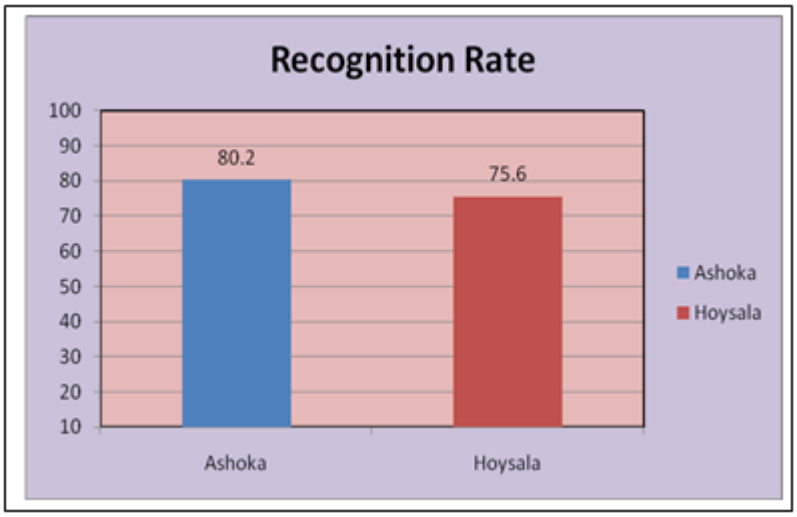

Figure 8. The Recognition Rate of the OCR System for Historical records

7. CONCLUSION

The developed system recognizes Kannada characters of two ancient eras Ashoka and Hoysala. The OCR system for Ancient Kannada Script takes epigraphic image from two different dynasties as the input, preprocesses the same for converting the image into individual characters, extracts features from segmented characters and classifies the characters of ancient period and finally maps into the modern Kannada characters. However, it is still challenging to design a system with high recognition accuracy, regardless of the quality of the input document and character font style variation. The system is designed to be independent of the font and size of

characters in the scanned hand written document and hence can be used with any kind of hand written document in Kannada. The system showed good results when tested on the 150 samples of ancient Kannada epigraphs. From the analysis a recognition rate of 80.2% for Ashoka period and 75.6% for Hoysala period is observed.

REFERENCES

[1] Leena R Ragha and M Sasikumar, "Feature Analysis for Handwritten Kannada Kagunita Recognition",International Journal of Computer Theory and Engineering, Vol.3, No.1, February, 2011

[2] Ganpat Singh G Rajput and Rajeshwari Horakeri "Zone based Handwritten Kannada Character Recognition Using Crack code and SVM",International Conference on Advances in Computing, Communications and Informatics,IEEE,2013.

[3] Anilkumar N Holambe, Ravindra C Thool,"Combining Multiple Feature Extraction Technique and Classifiers for Increasing Accuracy for Devanagari OCR",International Journal of Soft Computing and Engineering (IJSCE) , Sep 2013.

[4] Dr.S.Basavaraj Patil ,"Neural Network based Bilingual OCR System: Experiment with English and Kannada Bilingual Documents",International Journal of Computer Applications (0975 – 8887)Volume 13– No.8, January 2011.

[5] B.V. Dhandra ,Mallikarjun Hangarge and Gururaj Mukarambi," Spatial Features for Handwritten Kannada and English Character Recognition" IJCSE 2010.

[6] J. M. C. Sousa, J. R. Caldas Pinto, C. S. Ribeiro and J. M. Gil, "Ancient document recognition using fuzzy methods", IEEE 2005.

[7] B. Gatos, K. Ntzios, I. Pratikakis, S. Petridis, T. Konidaris and S. J. Perantonis, "An efficient segmentation-free approach to assist old Greek handwritten manuscript OCR" Springer-Verlag London Limited 24 August 2005.

[8] B.V.Dhandra and Mallikarjun Hangarge, "Global and Local Features Based Handwritten Text Words and Numerals Script Identification", IEEE 2007.

[9] Truyen Van Phan, Bilan Zhu and Masaki Nakagawa, "Collecting Handwritten Nom Character Patterns from Historical Document Pages", IEEE 2012.

[10] Panyam Narahari Sastry, Ramakrishnan Krishnan and Bhagavatula Venkata Sanker Ram, "Classification and Identification of Telugu handwritten characters extracted from palm leaves using Decision Tree approach", ARPN Journal of Engineering and Applied Sciences VOL. 5, NO. 3, MARCH 2010 ISSN 1819-6608.

[11] Soumya A and G Hemantha Kumar, "Zernike Moment features for the Recognition of Ancient Kannada Base Characters", International Journal of Graphics & Image Processing (IJGIP),Volume 4, Issue 2, pg 99-104, May 2014

[12] Soumya A and G Hemantha Kumar, "Recognition of Ancient Kannada Epigraphs using Fuzzy-Based Approach", pg 657-662, International Conference on Contemporary Computing and Informatics (IC3I-2014), IEEE publications ,SJCE, Mysore, 27th -29th Nov 2014

[13] Toufik Sari and Abderrahmane Kefali, "An MLP for binarizing images of old manuscripts" IEEE 2012.

[14] Soumya A and G Hemantha Kumar, "Enhancement and Segmentation of Historical Records", Vol 5, No.13, pg 95-113, Fifth International Conference on Digital Image Processing and Pattern Recognition (DPPR 2015), Computer Science Conference Proceedings – AIRCC publications, Chennai, 25th & 26th July 2015