Segmentation of Epigraphical Documents using Filtering Approaches

Manjunath B
Dept. of MCA, Maharaja Institute of Technology Mysore
Mandya, India

Sharathkumar Y H
Dept. of ISE, Maharaja Institute of Technology Mysore
Mandya, India

Abstract

In this work, importance of preprocessing and segmentation of epigraphical documents have been discussed and techniques are proposed to address the same. Preprocessing includes the enhancement of epigraphical records, noise elimination and binarization. The experimental results of preprocessing which involves enhancement and noise removal stages are tested on 250 camera-grabbed and scanned epigraphical images. Ancient epigraphical documents are enhanced by using suitable Spatial filtering techniques. Mean, Median, Gaussian Blur, Bilateral, Laplace filter, Unsharp Masking (USM) filters are explored with different filter sizes and filter parameters. Thus, the enhancement of historical records - transforms the degraded input document into a better perceivable image. The system performs well for preprocessing of ancient documents and provides flexibility to the user in controlling the process of image enhancement to obtain desired output.

1. Introduction

Kannada is a historical speaking and writing official and administrative language of Karnataka state in India. It is a language of pride and royalty about 100 million people are using Kannada as their daily conversations language in India and various parts of the world viz USA, Canada, Gulf and Asian countries of Middle-East, UK, Australia and etc..Kannada was the court language of the most powerful empires of South and Central India, such as Ashoka, Shaatavaahana, Kadamba, BaadamiChaalukya, Raashtrakuta, KalyanaChaalukya, Hoysala, Vijayanagara and Mysore. The Government of India has declared that the Kannada language was to be identified as one of the classical languages of India based on the report of Director of the Central Institute of Indian Languages (CIIL). The character set of Kannada language is similar to other languages in India. The language uses phonemic letters, divided into three groups such as: Vowels, Consonants, and others (neither vowel nor consonant) which are called as Swaragalu, Vyanjanagalu and Yogavaahakagalu respectively in Kannada language.

Kannada has its own scripting style, which has strongest words and verbs, nouns, adjectives and pronouns .Kannada language script has divided as three eras such as Old Kannada also called Halegannada from 450 AD –1400 CE, Middle Kannada also known as Nadugannada from 1400–1800, and the 3rd era called Modern Kannada from 1800 to the present. The old Kannada and middle Kannada are together called as ancient Kannada (from 450 AD – 1800 CE). Karnataka state has its own and rich historical and cultural heritage. The history is well preserved with immune able inscriptions and other historical artifacts and literature. The inscriptions or epigraphs are the sources of writing ancient Kannada script. Understanding inscriptions or epigraphs is the challenging and tedious task for this generation. This is a major drawback of inscriptions or epigraphs. With the digital technology advancement, this major drawback or limitation can be automated by recognizing the inscriptions or epigraphs.

Epigraph and its Evolution

Epigraph is an engraved inscription, which is written on rock, metal like bronze, iron, gold, silver, copper, and brass, wooden planks, skins of animals, crystals, cones and the leaves like bursary, palm, and papers to retain for longer period as much as possible. These kinds of inscriptions are engraved by the kingdoms to convey their orders, to keep their history records, and to promote their achievements, cultural activities, social life, educational information's, languages used, scripts adopted, tax and economical details, administration protocols, religious behaviors, health related information during their period. In the universe, India is one of the most significant ancient nations with unique culture, civilization and its knowledge. However, within India, Karnataka is one of the leading ancient states, which has very wide epigraph compared to other states. For Indian Script, the evolution of Kannada script has massive scope. The oldest Kannada inscription was discovered at the small community of Halmidi and dates to about 450 CE. As per the various records and statistics, there are over 50000 epigraph records [3] came into existence in Karnataka state, which attracts most of the citizens who are very curious about our history. Figure 1 depicts some of the sample inscriptions images of Kannada language.

Evolution of Kannada Epigraph

The Kannada language scripting has evolved by the Brahmi scripting family in the 3rd century in the form of epigraph. Further, it has been undergone with the many changes with various distinct stages and grown to present form [1, 2] and it has been classified as follows:

- The Pre-Ancient Kannada (PoorvaHalegannada)
- Ancient Kannada (Halegannada)
- Middle Kannada (Nadugannada)
- Modern Kannada (Hosagannada)

Details of the above classification such as scripting existence century, their respective rulers names and their sample scripting styles are given in table 1. Table 1 shows, the oldest Kannada script, i.e., the pre-ancient Kannada was evolved in the 3rd century and existed through the 8th century. This script was ruled by the various kingdoms such as Ashoka, Shaatavaahana, Kadamba and BaadamiChaalukya. The 2nd version of Kannada script i.e., the ancient Kannada, which was existed between the 9th and 14thcentury. The various kingdoms such as Raashtrakuta, KalyanaChaalukya, and Hoysala were ruled this Kannada script. The 3rd version of Kannada script (just the previous to modern Kannada script) i.e., the Middle Kannada, was ruled by the Vijayanagara and Mysuru kingdoms between 15th to 18thcentury. The 18th century Kannada script is called as Mysuruwadeyarkannada, which is almost similar to the present Kannada script and can be read easily. The modern Kannada script is the present Kannada script, which came into existence from the 19th century and presently all the people who are reading, writing and speaking are following the modern Kannada. The evolution of Kannada alphabets of various stages discussed in this section can be seen in Figure 2.

Figure 1: Sample Inscriptions Images

Table 1. Evolution of Kannada Script

Classification	Period (Century)	Rulers Name
Pre-Ancient Kannada	3rd BC	Ashoka
	2nd AD	Shaatavaahana
	4th–5th AD	Kadamba
	6th AD	BaadamiChaalukya
Ancient Kannada	9th AD	Raashtrakuta
	10th AD	KalyanaChaalukya
	12th AD	Hoysala
Middle Kannada	15th AD	Vijayanagara
	18th AD	Mysore
Modern Kannada	19th AD and onwards	Present Kannada

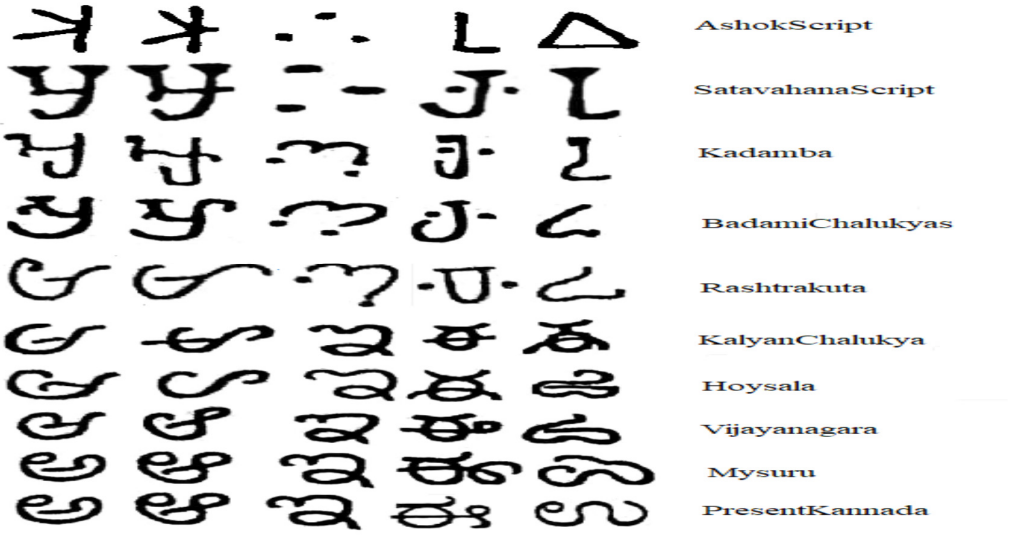


Figure 2. The evolution of sample Kannada alphabets of various stages

Epigraphy is a process of studying epigraphs or inscriptions, which plays a significant role to human society. Epigraphy is a primary tool of archaeology when dealing with literate cultures. Epigraphs are degraded or occluded due to depraved storage. Epigraphs are very hard and tedious task to read and understand by the normal citizens and even by the highly educated people. However, there are some people they can read and understand these epigraphs. These

people are called as epigraphers. However, nowadays there are no or very less epigraphers are available and may prone to error. These are the major drawbacks to the people they are very interested to know our history which is written through inscription. Hence, it is very necessary to automate the epigraphy by using the recent and advanced technology. In this work, an automated and efficient Optical Character Recognition (OCR) System has been proposed to recognize the ancient Kannada scripts.

2. Related Work

An OCR document binarization technique for searching of keyword from historical printed documents is presented [4]. This technique is considered as a preprocessing step in OCR. In this document binarization technique, they have developed the several capable techniques such as pre-filtering, error diffusion binarization, multi resolution version of Otsu's binarization, denoising by the method of despeckling, and post-banalization denoising. There are six newspapers are utilized for experimentation and obtained the best average results using classic Otsu method and Otsu based methods. A new framework to pre-process the historical documents of large database is presented, which includes the selection and evaluation phases. There are about 900 printed and handwritten documents from Google-Books databases were used for experimentation. In selection phase, the best binarization technique is selected and for validation of the selected binarization technique, they have proposed an evaluation phase. This frameworks yields a very promising results for all the datasets [5]. In [6], an innovative supervised binarization technique is presented. Hierarchical Deep Supervised Network (HDSN) architecture is used to predict the pixels of the text at various levels of features. The text pixels from the background noises are differentiated by the HDSN with the high level features. With this, the major degradations available in documents are managed. On the other hand, the prediction of text pixels from the foreground maps at the lower level features are present in advanced visual quality of the boundary. The proposed technique obtains the better results with largely used database compared to other existing techniques. In [7] authors proposed a technique for document image binarization, which majorly uses a non-local means methodology for the elimination of noise existing in the document image. This removal of noise from the document image is considered as a step of preprocessing. Here, thresholding techniques are used to binarize the document. To achieve the better binarization outcomes, it is necessary to take post processing at the end. This process involves the following measures such as despeckle, preserve stroke connectivity and text regions quality improve. The experimentation is carried out on broken and degraded books, documents files, and magazines.

At last, the proposed technique provides a noteworthy results. The phase based features dependent binarization framework is developed to preprocess and binarize the ancient epigraphical document images by using an Expectation Maximization (EM) algorithm and preprocessing. This framework includes the phase congruency and Gaussian Mixture Model (GMM) based background elimination after the image denoise through log Gabor wavelets and edge detection. Also to smothout the output images, the adaptive Gaussian filters were utilized. For complete background elimination, the EM algorithm is used. This framework is experimented with various inscriptions and epigraphs datasets. At last, the framework achieves an outstanding results compared to other existing binarization techniques[8]. In [9], an adaptive document image binarization approach is proposed by considering the hybrid approach and the class properties of the document region. Authors also addressed about the various problems arise due to noise, degradations, and illumination. To determine the local threshold, there are two new algorithms are employed for every pixel of the image. The significant part of this algorithm is the quality result validation using ground truth. For experimentations, they have used the test image datasets consisting pictorial, textual, and document images with degradations and ground truths. The outcome of this approach gives the qualitative and quantitative results compared to earlier approaches. Souwya, [10] proposes a method to preprocess the camera captured epigraphs and also to segment the handwritten Kannada document scripts. In preprocessing task, the two main tasks carried out are removing noise and enhancement of degraded images. The distinct filters such as Unsharp mask, Laplacian filter, and Gaussian blur are applied for smoothing and sharpening the images. Image enhacment is performed using the specific filter by providing varied mask size and parameter values. For image binarization, the Otsu technique is used. At last, Kannada hand written document is developed by the connected component technique. This segmented output will be used for future stages of OCR. This proposed method is suitable for preprocessing and flexibility in control the process of image enhancement for obtaining the expected output. The proposed technique provides a better results comparing with existing systems. An attempt has been made to design an approach for noise removal from an input image. In this technique, there are two kinds of images are used viz noisy and noseless. Initially, noisy printed document images and noisy epigraphical document images used for binerization. Later noiseless images are considered for binerization. The noise from the document images are identified and removed by counting the vertical and horizontal run length. The Speckle, Gaussian and Poisson noise models are used to derive the images consisting synthetic noise. It is found that the algorithm shows the efficient result for removal of noise[11]. Jundale,[12]developed a technique for

detection of skew and correction at word level in handwritten Devanagari manuscript by using Hough Transform (HT) algorithm. In this technique, an image is obtained by the camera of scanner. Once the image is obtained, pre-processing task is carried out for removing the noise. After removal of noise, every word from the document is identified individually. At the final stage, HT algorithm is employed on every identified word to detect the skew. The rotation transformation is utilized to correct the skew. As a result of this developed technique, an average accuracy of 97% and 92.88% is achieved for identification of words and skew correction respectively. In 2016, a novel and fast skew correction method and baseline detection techniques are presented for Arabic documents using randomized HT[Hough Transform] algorithm. This algorithm is developed based on detecting lower baselines of the text lines from the Arabic records. Skew angle is obtained by detecting lower baseline pixels from the lower edge of the word images using randomized HT algorithm. Text or word baselines are identified by the 'y' intercept histogram with or without correcting the skew. Also, this approach is used to extract the text lines from the skewed text documents of any language[13]. A robust skew detection approach termed "Principle-axis farthest pairs Quadrilateral (PFPQ)" has proposed to estimate the skew angle from the various types of Telugu documents and other languages of India of any type documents such as images, diagrams, mathematical or scientific formulas, trigonometric functions, textual, statistical tabular or pictorial regions [14]. Specifically, principle-axis-oriented quadrilateral painting and directional smearing approach is developed to detect the skew angle. The specialty of this approach is that it works without any angle restriction and independent on line spacing between text lines, font size etc. This developed approach is experimented on various five kinds of documents such as magazines, newspapers, handwritten documents, textbooks, and social media. In 2015an inventive approach for identification of text line and correcting skew using Euclidean Distance (ED) algorithm appropriate for handwritten text documents by calculating the distance between the words is presented [15].In this approach, the main aim is to handling single and multi-skew of distinct writers' hand written text documents. For removal of noise, binarization, and baseline correction are carried out using Discrete Cosine Transform (DCT), K-means Color Quantization, and gradient or slope of a line respectively. Based on the distance calculated using ED algorithm, lowest base lines from the document and skew correction is carried out. This approach was developed on more than 90 distinct script and achieved an average accuracy of 99.2%. Alaei [16] proposed a technique to estimate the skew from the scanned document image using Piece-wise Painting Algorithm (PPA). The considered documents for this algorithm are flow of writing and content independent. At first, PPA is

working in both horizontal and vertical directions of the document image to achieve the distinct horizontally and vertically painted images. This proposed algorithm was experimented broadly on 3 kinds of datasets consisting different types of document images and obtained an inspiring results with better skew estimation time. A new method called a segmentation-free word spotting method, which is efficient and applied in ancient document collections. In this method, a patch based outline is used. Here Local patches are represented through the bag-of-visual words model which is employed by the SIFT descriptors. The document information is efficiently indexed both in terms of time and space. There are four distinct group of epigraphical documents are used to evaluate the proposed algorithm. The explained method achieves a better result both on typewritten and handwritten scenarios[17]. An approach, which segments the text images of Urdu into lines. The split position of the text line, overlapping, merged ligatures, and cross line split position problems are successfully overcome by the curved line split algorithm and classical horizontal projection-based segmentation techniques. There are 10063 text lines of Urdu printed text images consisting 332000 connected components are used and obtains an average accuracy of 99.80%. The proposed segmentation algorithm is experimented with Arabic Urdu Nastaleeq text and extracted the lines very efficiently[18]. The three stage text line segmentation technique is proposed Vishwas [19], to segment the Kannada epigraphical document images. The datasets used here are contains more issues and complexities because they consists touching and overlapping of characters and extra modifiers. The binarization of the document image is achieved by the Sauvolas technique and morphological thinning process. For assigning the labels to individual components, the connected component labeling technique is utilized. The text line location localization and then neighborhood search in vertical and horizontal direction for assigning the characters to their representative lines of the text by obtaining the horizontal projection. The proposed algorithm shows an efficient result in resolving the extra modifiers and interline gaps of the variable. A new segmentation technique is presented by Kavitha, to segment the text and non-text documents by employing a novel combination of laplacian and algorithms to enhance the degraded low contrast picture elements. Then it generates the skeletons to reduce the computational burdens in enhanced images for text documents. This helps to study the component structures effectively. For removing the non-true text components, the study has been taken on the corrosiveness of components, which is depending on branch info. To group up the components available in the same line, which outcomes in clusters, the nearest neighbor algorithm is used. The proposed algorithm also classifies the clusters as text and non-text cluster based on features of text components. A huge dataset consisting a categories of images

for the experimentation and an effective and efficient outcomes comparing with the earlier techniques in terms of precision and recall[20]. In 2020,[21] a new line segmentation technique is presented for Devanagari epigraphy scripts by splitting the image as vertical strips and by piecewise horizontal projection profiles. The documents are taken from distinct museums and libraries. The proposed algorithm is experimented on text lines of 1500 line. Under or over segmentation of text lines are identified for epigraph manuscript recognition based on the calculated average line height. This approach can be also employed for other languages of India. The average segmentation accuracy of this algorithm is 97%.

3. Proposed Model

The objective of the work is two fold: i) to enhance ancient documents with varying level of degradations and bring them in to computer readable format. ii) Followed by this to perform segmentation of ancient documents in order to extract characters from it which can be further subjected to recognition. The proposed model for Preprocessing and Segmentation of epigraphic documents is presented in Figure 3. The model comprises of the components – Enhancement, Binarization and Segmentation. The input to the system is ancient epigraphic documents of varying amount of degradation. Preprocessing here mainly deals with the noise removal and enhancement of degraded ancient epigraphical images, for better human perception and also to transform the input into computer recognizable form. Enhancement is achieved through different Spatial filtering methods for smoothing or sharpening namely Mean, Median, Gaussian blur, Unsharp mask,

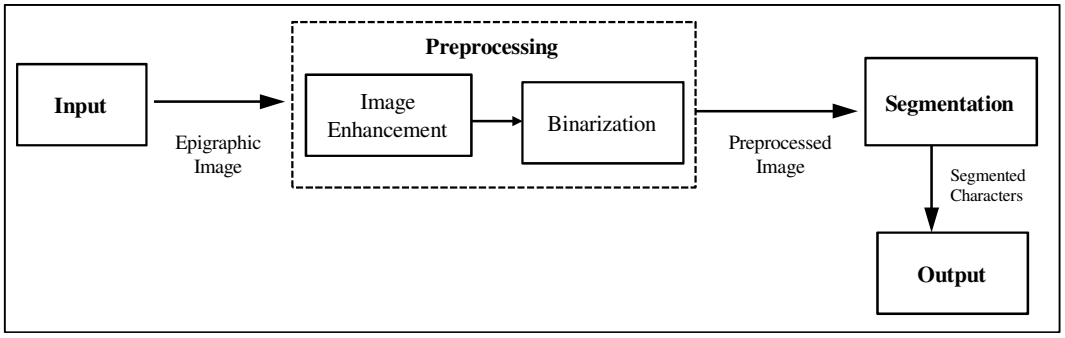


Figure 3: Proposed Model for Preprocessing and Segmentation of Epigraphs

Laplacian, and Bilateral filters. These filters are used with different mask sizes and parameter values which can be specified by the user, according to the nature of image quality and varying amount of degradation. Contrast enhancement is performed using Yu and Bajaj's algorithm and binarization of the enhanced image using Otsu thresholding, Brensen, Niblack, Sauvola

and Savakis algorithm. Characters are subjected to thinning using Guo Hall algorithm. Finally, segmentation of handwritten Kannada epigraphic documents is carried out using Connected component labeling, Contour based Convex Hull, Nearest Neighbor Clustering, Drop Fall and Water reservoir techniques to obtain sampled characters. The output from segmentation phase is fed to further stages of character recognition system.

3.1 Segmentation

Segmentation is the method of extracting objects of interest from the image. The general steps in the segmentation of document images are identifying lines, the words in each line and the individual letterings in each word. Segmentation of text into individual letters is a major problem in recognition of handwritten documents.

Connected-component Labeling is an algorithmic application of graph theory, which is used the segmentation of scanned handwritten documents into characters. This approach is used in document images to segment the characters using connectivity among the components of the image. The distribution of bounding boxes [26] describes the segmentation of an image consisting of characters. By calculating adjacency relationships merging can be performed, or their size and aspect ratios can be used for splitting mechanisms. Much of the segmentation task can be accurately performed at a low cost in computation. The approach used for segmenting the characters is a combination of Contour detection and Bounding Box Algorithm [25]. First Contours are retrieved from the binary image. Once all the contours are obtained, each contour is used by Convex Hull algorithm which generates the Convex Hull objects for each character. Then the outline around each character is drawn using boundary points in a Convex Hull object. The entire image is componentized with the help of Bounding Box Algorithm. Thus, the rectangles or the bounding box are drawn across the character. To form meaningful characters or syllables some of these boxes are merged to form a single box [27]. The majority of the epigraphical scripts contain touching lines and touching characters. Hence, a novel approach Nearest neighbor clustering method [22] for line and character segmentation is proposed, which also works even for the skewed document. A survey on methods and strategies on touched characters segmentation is covered in [28]. Drop fall algorithm [21] is based on the principle that an equally ideal cut between two touched characters can be created, if a hypothetical marble is rolled off the top of the first character and create the cut where the marble falls. The position where to drop the marble from is important because if the algorithm starts at the wrong place, the marble can simply roll down the left side of the first letter or the right side of the second letter and hence, it would be completely unsuccessful. The best

approach to start drop falling process is the point at which two characters are touched. In this process, the pixels are scanned row by row until a black boundary pixel with adjacent black boundary pixel to the right of it is identified, where as the two pixels are separated by white space. This pixel is used as a point to start the drop. The direction that the algorithm will move is according to the current pixel position and its surroundings. The larger space generated by touching numeral is analyzed with the help of water reservoir concept [29]. When water is poured from the top (bottom) of a component, the regions of the component where water will be stored are considered as a top (bottom) reservoir. The reservoirs obtained in this procedure are not considered for further processing. Those reservoirs whose heights are greater than a threshold value T1 are considered for further processing. When two characters touch each other, they create a space (reservoir) between the characters. This space is very important for segmentation because,

- As cutting points are concentrated around the base of the reservoir, and hence, decreases the search area.
- The cutting points lie on the base of the reservoir.
- Space attributes (center of gravity and height) aid to go near the best touching position.

The touching position between characters is to be determined. The best reservoir for touching is determined. The base-line (lowermost row of the reservoir) of the best reservoir is then identified. To find the touching position in the components, the morphological thinning operation is applied to touching components for further processing. For feature points extraction the touching position is renowned. The leftmost and rightmost points of the base-line of considered reservoirs are the feature points. These points are initial feature points. With this initial feature points, the best feature point is chosen for segmentation.

4. Algorithmic Models

The detailed design and the algorithmic models of the proposed work are presented in this section.

Preprocessing of Ancient Epigraphical Images

***Algorithm*: Enhancement (Epigraph_Image) Input:** Ancient epigraphical image

Functionality: Enhances epigraphic image of medium-level degradation using spatial filters namely Mean, Median, Gaussian blur, Unsharp mask, Laplacian , Bilateral filter, with varying mask size and filter parameters, depending whether smoothing or sharpening is the requirement, for better human perception. The enhanced image is subjected to binarization and thinning.

Output: The enhanced and binarized image with reduced noise.

Method: The steps towards Enhancement of Ancient Epigraphs are as follows:

Step 1: [*Read Image*]: Read epigraphical image of varying amount of degradation

Step 2: [*Enhancement*]: The input ancient epigraph is enhanced using any of the following filtering methods, designed and implemented with different mask sizes,

if (method = 1 or method = 2)

Read the size of the mask nxn , where n = 2, 3, 4, 5

Read value of the Standard Deviation, σ

Filter the input image using Gaussian blur method.

Filter the input image using USM method.

Else if method = 3, (For Laplace Filter)

Read and Display the Diffusion constant value D

Read and display the step size of the mask, No. of Steps

Filter the input image using Laplacian method.

Else if method =4 (For Mean Filter)

Filter the input image using Mean filtering method

Else if method = 5(For Median Filter)

Filter the input image using Median filtering method

Else if method = 6(For Bilateral Filter)

Filter the input image using Bilateral filtering method

Step 3: [*Binarization*]: The enhanced images are converted to binary image consisting of ones and zeroes.

Use Otsu thresholding, Brensen, Niblack, Sauvola and Savakis approaches

Step 3: [*Contrast Enhancement*]: Use Yu and Bajaj's algorithm

Step 4: [*Thinning*]: Employ Guo-Hall algorithm.

Segmentation of Epigraphical Images

The algorithms for segmentation are presented here in the section.

1. ***Algorithm* : Drop_Fall**

(Epigraph_Image) Input: Noise-free

binarized epigraphical image

Functionality: The binarized image is segmented to characters using Drop Fall algorithm

Output: Segmented characters

[Step 1]: Find the size of the characters to find touched

characters Find the Height and Width of the touched

characters

[Step 2]: Apply Breadth First Search (BFS) algorithm to find the touched characters.

Search for the initial pixel; scan row by row until two black pixels are separated by white pixel.

[Step 4]: If found start the Drop fall.

drop falling algorithm will always move downwards, crossways down-wards, to the right, or to the left.

[Step 5]: Make the slice where marble parks.

Segmentation path for connected components is found

2. ***Algorithm* : Water_Reservoir**

(Epigraph_Image) Input: Noise-free binarized

epigraphical image

Functionality: Binarized image is segmented to characters using Water reservoir concept

Output: Segmented characters

[Step 1]: Find the size of the characters to find touched characters.

[Step 2]: The positions and sizes of the reservoirs are analyzed.

[Step 3]: A reservoir is detected where touching is made,

The initial feature points for segmentation are noted.

[Step 4]: The best feature points are noted from the initial feature points.

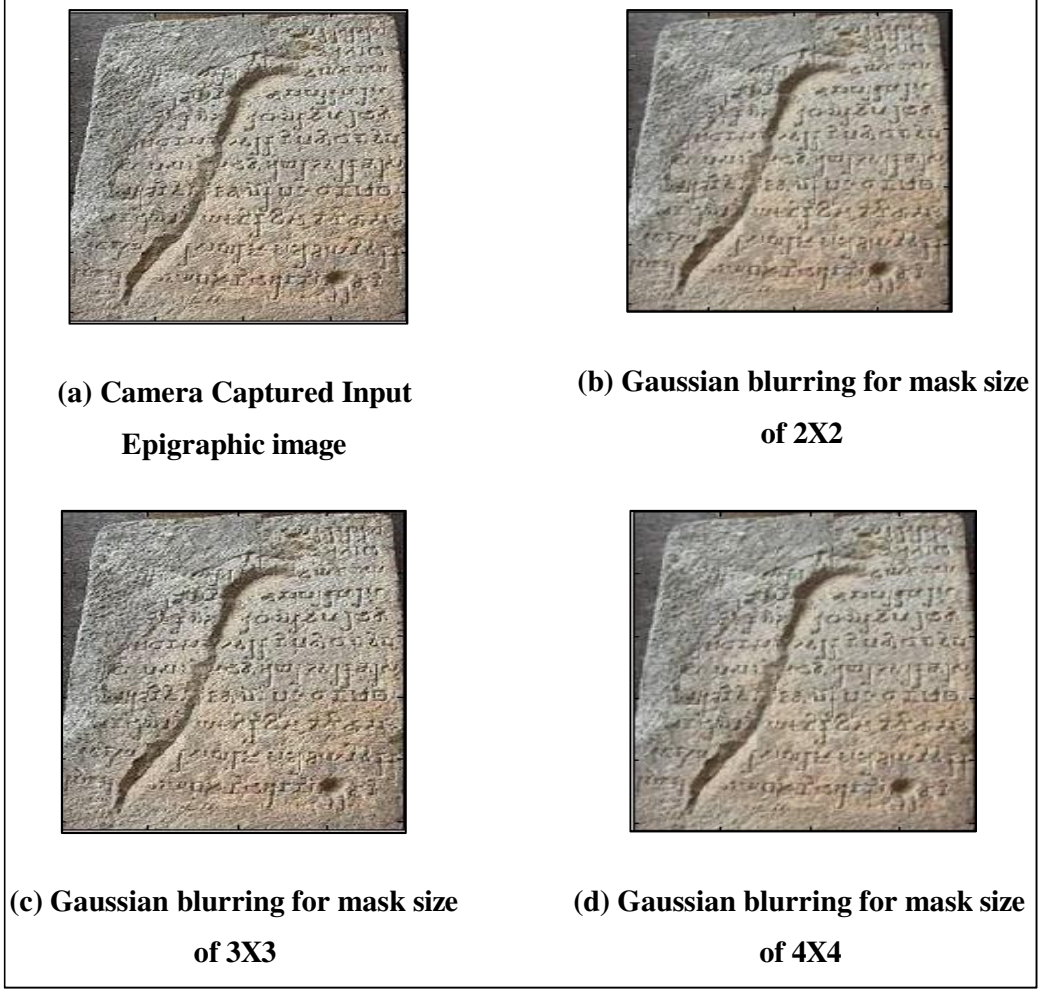
[Step 5]: Based on touching position, close loop positions and morphological

structure of touching region the cutting path is produced.

5. Experimentation

This section discusses the experimental results and comparative analysis of the approaches used in preprocessing and segmentation of epigraphic records. The experimental dataset consists of degraded ancient epigraphic document images for preprocessing from sources such as camera-captured and scanned records. The subsystem for noise removal and enhancement is tested on nearly 150 epigraphical images (100 camera-grabbed images of inscriptions and 150 scanned ancient epigraphic images) with a medium amount of degradation. Enhancement of Camera-captured Inscriptions and Analysis this section, the experimental results of image enhancement using the available spatial filtering techniques such as Gaussian blur, USM, and Laplacian are explored for camera captured inscriptions with varying quality. The results of these filtering operations for different mask sizes are discussed with illustrations. Figure 4 represents the results of Gaussian blur filtering, which successfully smoothens the input image and thus reduces the background noise. The Gaussian blur method is used to blur the sharpen image so that a less edge highlighted image is produced and this also reduces some amount of background noise for further stages of preprocessing and segmentation system. Figure 4(a) is the input epigraphic image and **t**he Figure 4 (b), (c), and (d) shows various results of Gaussian blur method for mask size of 2X2, 3X3 and 4X4 respectively. It is observed that when the mask size is moderate then the output image will appear to be clear, and use of a low mask or higher size results in the blurred image as in Figure 4(b) and (d). Unsharp Masking (USM) filter, on the other hand is a sharpening filter which enhances the blurred input image. Figure 5 shows the

results of USM filter for sharpening the input image. The results of applying USM filter for a sample image in Figure 5 (a) using mask size of 2X2, 3X3 and 5X5 are shown in Fig 5 (b), (c) and (d) respectively. If the mask size is 2X2 the image is more sharpened and as the mask size increases the image is less sharpened resulting in a good output image.



(a) Camera Captured Input Epigraphic image

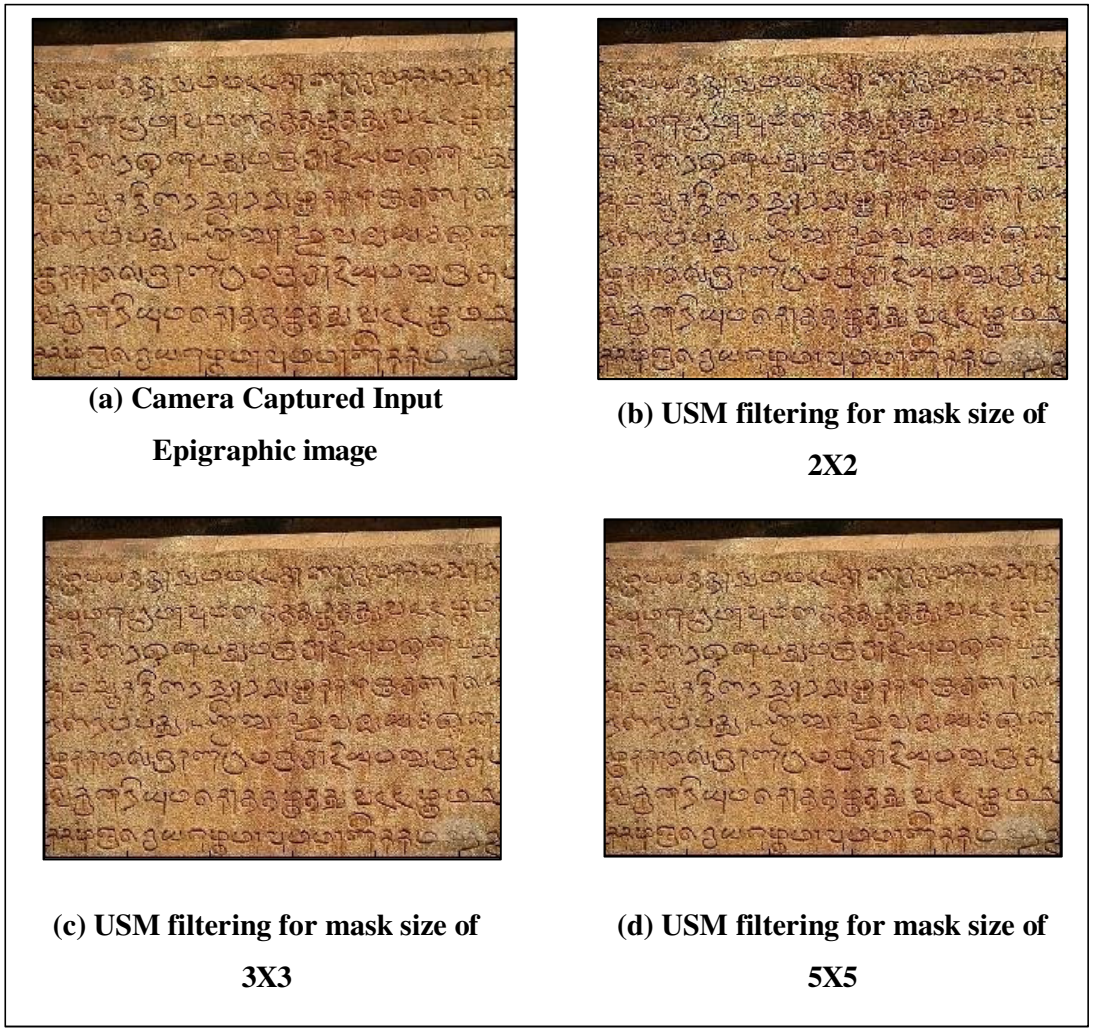
(b) Gaussian blurring for mask size of 2X2

(c) Gaussian blurring for mask size of 3X3

(d) Gaussian blurring for mask size of 4X4

Figure 4: Results of Gaussian blur Filtering on Camera Captured Epigraph

The Laplacian filter is another smoothing filter. Figure 6 shows the results of Laplacian filter for smoothing the input image. Figure 6 (b), and (c) shows results of applying a Laplacian filter on the input image shown in Figure 6 (a) using the diffusion value of -0.01 and +0.01 respectively. For D-value of -0.01, the input image is slightly sharpened and as the D-value increases the image smoothened. As a result for D-value of +0.01, the input image smoothened. As compared to other filters, Laplacian filter shows less difference in output compared to input.



(a) Camera Captured Input Epigraphic image

(b) USM filtering for mask size of 2X2

(c) USM filtering for mask size of 3X3

(d) USM filtering for mask size of 5X5

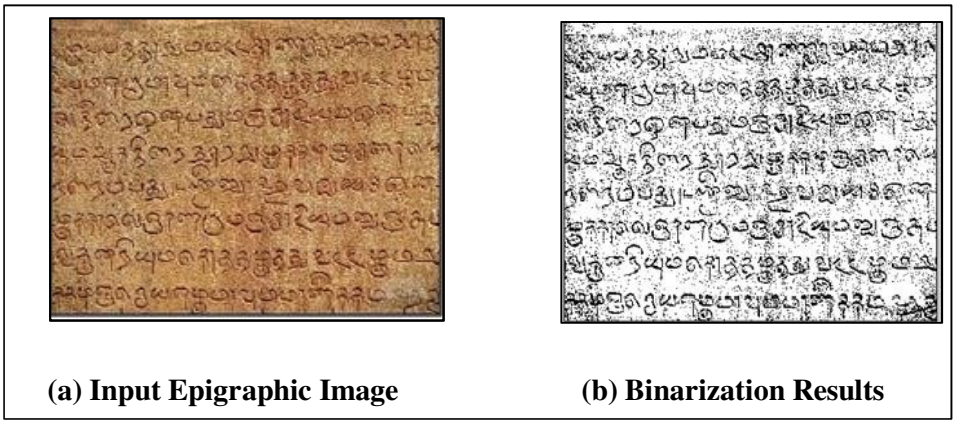
Figure 5: Results of USM Filtering on Camera Captured Epigraph

(a) **(b)** **(c)**

Figure 6: Results of Laplacian Filtering on Camera Captured Epigraph
(a) Camera Captured Input Epigraphic image; (b)Laplacian filtering for D-value of -0.01, (c) Laplacian filtering for D-value of +0.01

The smoothing filters used in this system will result in good accuracy if the edges of the input

image are very thick, where as in the case of sharpening filter namely, USM filter better output is achieved for the degraded input image. The two smoothing filters used in the proposed system are a Gaussian blur and Laplacian filter, depending on the input image the required filter can be used with different mask sizes to get the filtered image. The enhancement is found to be appreciable for mask size of 3X3 for Gaussian blur, 5X5 for USM filter and in the case of Laplacian filter input image is sharpened for negative values of the diffusion constant, smoothened for positive values. The enhanced image is binarized using Otsu's method and the result is shown in Figure 7.

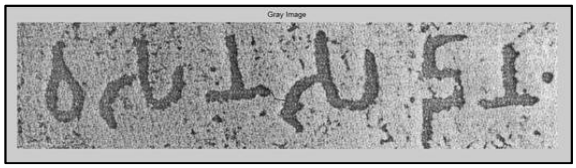


(a) Input Epigraphic Image **(b) Binarization Results**

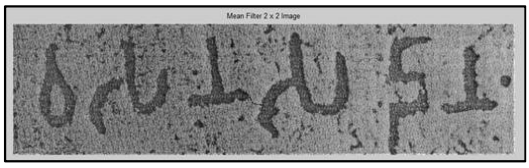
Figure 7: Result of Binarization using Otsu's method

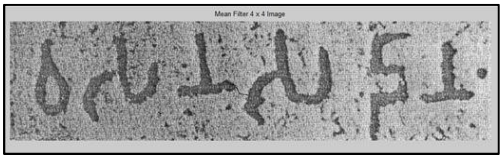
5.1 Results

The spatial filtering techniques are explored on nearly 150 scanned ancient epigraphic images of varying image quality and degradations. Figure 8 and 9 shows the input ancient historical record with varying- writing material and degradation, which is analyzed for different spatial filtering techniques.



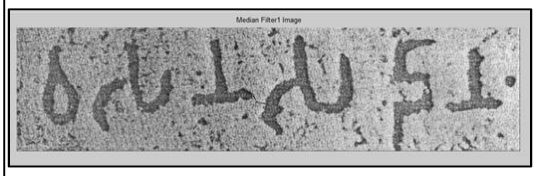
(a) Scanned Input Epigraphic Image - Sample1 (S1)

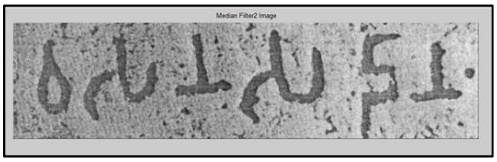




(b-i) Mask size 2x2 **(b-ii) Mask size 4x4**

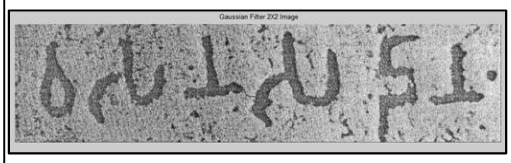
(b) The results of Mean Filtering S1 for Mask size 2x2 and 4x4

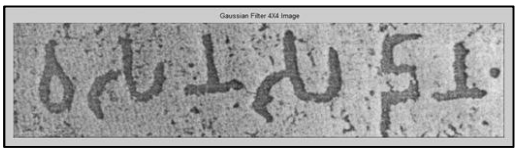




(c-i) Mask size 2x2 **(c-ii) Mask size 4x4**

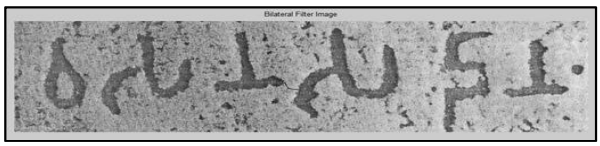
(c) The results of Median Filtering for Mask size 2x2 and 4x4





(d-i) Mask size 2x2 **(d-ii) Mask size 4x4**

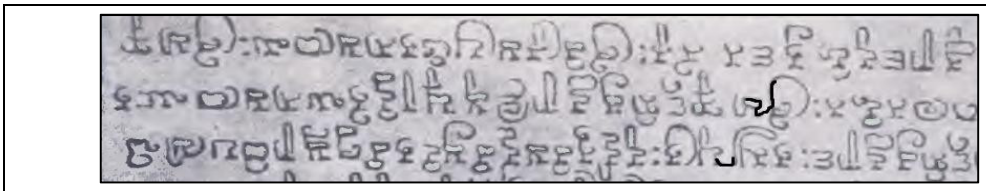
(d) The results of Gaussian Blur Filtering for Mask size 2x2 and 4x4



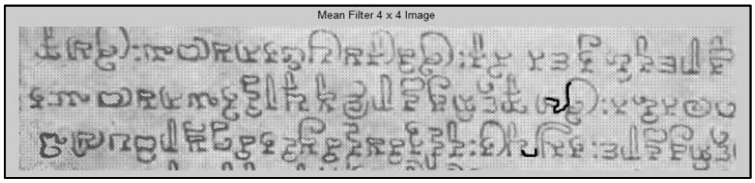
(e) The result of Bilateral Filtering

Figure 8: Results of Spatial Filtering Scanned Epigraphic Image (S1)

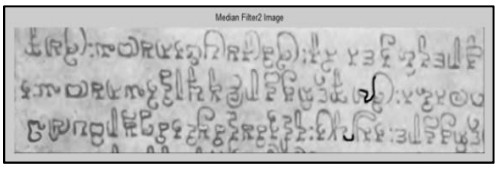
Figure 8(a) shows the input epigraphic image Sample S2 subjected to various spatial filtering techniques in proposed model and results are depicted in Figures 8(b) to 8(e).



(a) Input Epigraphic Image – Sample 2 (S2)

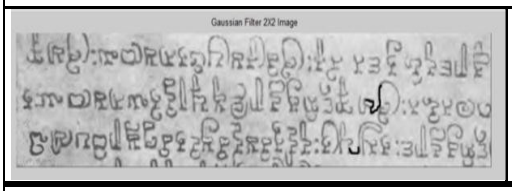


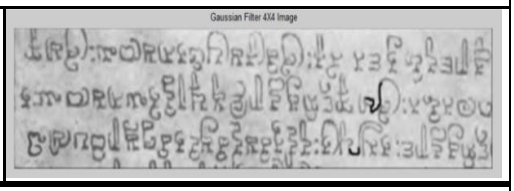
(b) Mean Filtering for Mask size 4x4



(c-i) Mask size 2x2 **(c-ii) Mask size 4x4**

(c) The results of Median Filtering for Mask size 2x2 and 4x4





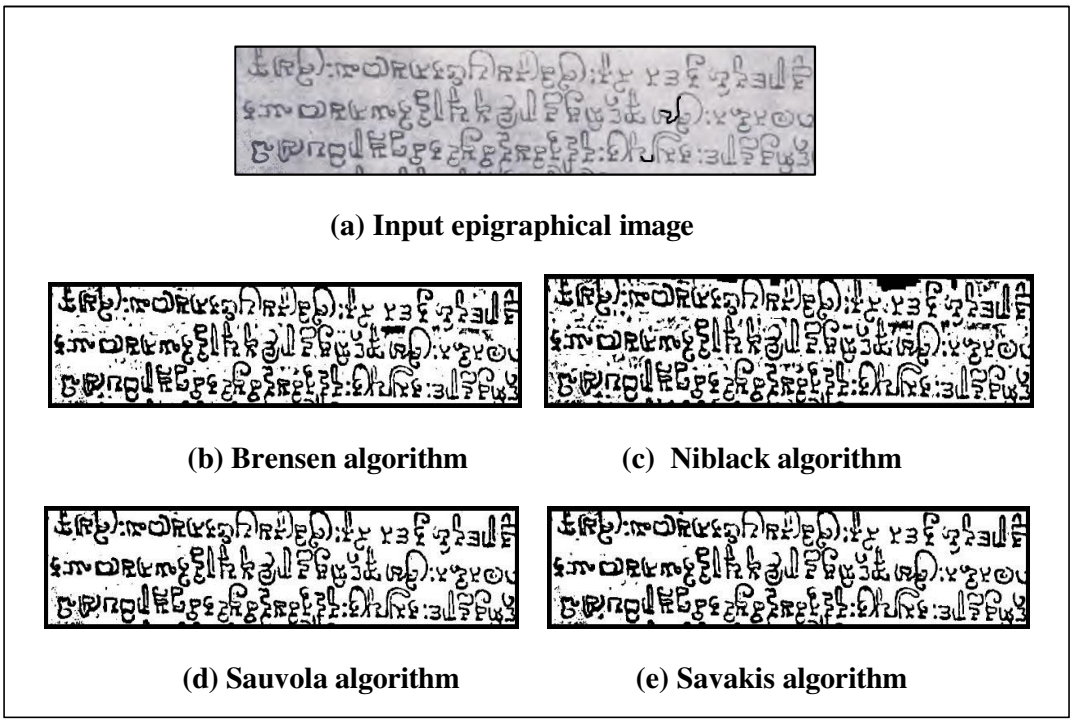
(d-i) Mask size 2x2 **(d-ii) Mask size 4x4**

(d) The results of Gaussian Blur Filtering for Mask size 2x2 and 4x4

(e) The result of Bilateral Filtering

Figure 9: Results of Spatial Filtering Scanned Epigraphic Image (S2)

Various binarization techniques Brensen, Niblack, Sauvola and Savakis are tested on scanned epigraphic images and the results are compared. This is illustrated for the sample input epigraphic document shown in Figure 10(a). Figure 10(b)-10(e) shows the result of the binarization using Brensen, Niblack, Sauvola and Savakis algorithm respectively. All the above algorithms for binarization are tested on 150 epigraphic images of a medium amount of degradation. The Sauvola algorithm performs comparatively better than Brensen, Niblack and Savakis for most of the images.



(a) Input epigraphical image

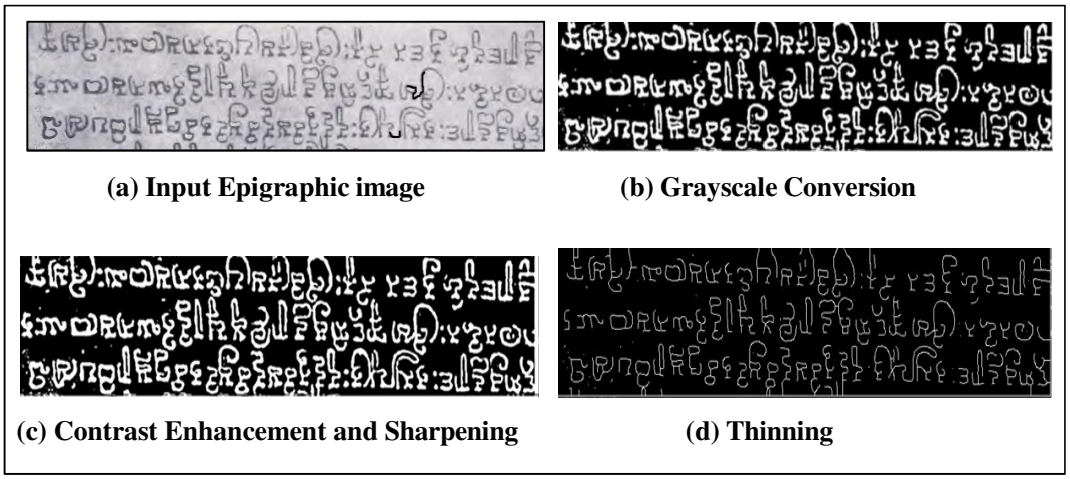
(b) Brensen algorithm **(c) Niblack algorithm**

(d) Sauvola algorithm **(e) Savakis algorithm**

Figure 10: Results of Binarization using different approaches

Here, illustrates the results of preprocessing a sample epigraphic document image depicted in Figure 11(a). Preprocessing involves the steps: gray- scale conversion, Contrast enhancement using Yu and Bajaj's technique, Sharpening using Unmask filtering technique and Thinning using Guo-Hall method. The results of these steps are shown in Figure 11(b)-11(d) respectively. Figure 11 (b) shows the result of the gray-scale conversion. The image is blurred by a small amount and hence is fed to the contrast enhancement and sharpening procedure. Figure 11(c) shows the result of the contrast enhancement and sharpening of the grayscale image. The output image highlights the edges of each character in the text. Figure 11(d) shows the thinned image in which the characters are thinned using the Guo Hall algorithm. Thinning helps to improve the efficiency of processing and assist in next step of obtaining sampled

characters to a certain extent by addressing the touching characters.

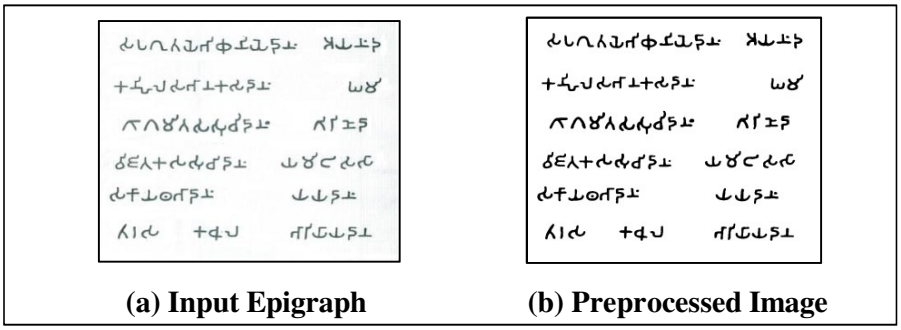


(a) Input Epigraphic image **(b) Grayscale Conversion**

(c) Contrast Enhancement and Sharpening **(d) Thinning**

Figure 11: Preprocessing of Scanned Epigraphic Image (S2)

The preprocessing sequence that will address the scanned epigraphic images of good quality is presented here. The scanned input epigraphic image is shown in Figure 12(a) and the Figure 12(b) illustrates the final Pre-processed image which is the outcome of Spatial filtering namely Gaussian, Median and Bilateral followed by morphological operations erode operation and dilation operation.



(a) Input Epigraph **(b) Preprocessed Image**

Figure 12: Result of Preprocessing Scanned Epigraphic Image (S3)

The Preprocessing steps are tested on 50 scanned ancient epigraphic images with minimal noise and it is found that the combination of the three filters used namely Gaussian, Median and Bilateral, remove the noise from the image while preserving the edges. The morphological operations namely erode and dilate improves the clarity of edges in turn textual writing as a whole, thereby further improving the accuracy of segmentation and recognition.

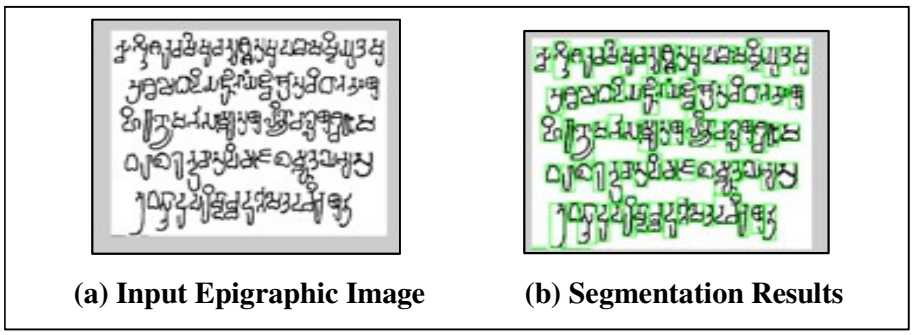
5.2 Segmentation Approaches – Results and Performance Analysis

The segmentation techniques in the proposed work are tested on nearly 150 noise-free epigraphic document images. This section illustrates the experimental results and highlights the inferences drawn in testing these approaches for segmentation of epigraphic records.

The metric used to evaluate the accuracy of Segmentation phase is the Segmentation Rate given by the Equation 1

$$\text{Segmentation Rate} = \frac{\text{Number of Correctly Segmented Characters}}{\text{Total Number of Characters in input document}}$$

The segmentation algorithm Connected component and bounding box, is tested on the text of ancient periods. The results of character segmentation of sample ancient epigraphic image are demonstrated in Figure 13



(a) Input Epigraphic Image **(b) Segmentation Results**

Figure 13: Segmentation Results using Connected Component and Bounding Box Method - Sample (S4)

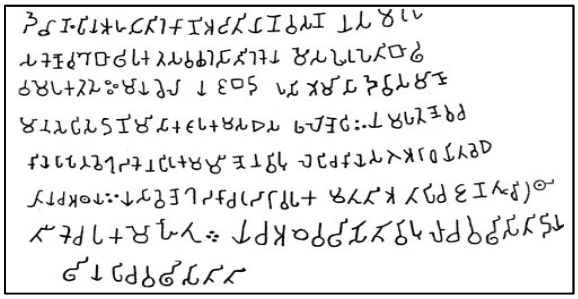
Performance Analysis

Segmentation using Connected Component and Bounding Box Method is tested on nearly 150 samples of reconstructed epigraphic images from different periods. The accuracy of the result is mainly dependent on appropriate connectivity of the pixels which forms the complete character in the document. It segments the base characters and compound characters correctly when the character sub-components have complete pixel connectivity. Few cases where in connectivity is not present, the compound character is segmented separately and output as disjoint characters. If the subcomponents that form a compound character are connected, then the character is segmented correctly. If there is disconnectivity of the character components, the output of segmentation results into two disjoint character components. If two characters are touching or overlapping then the

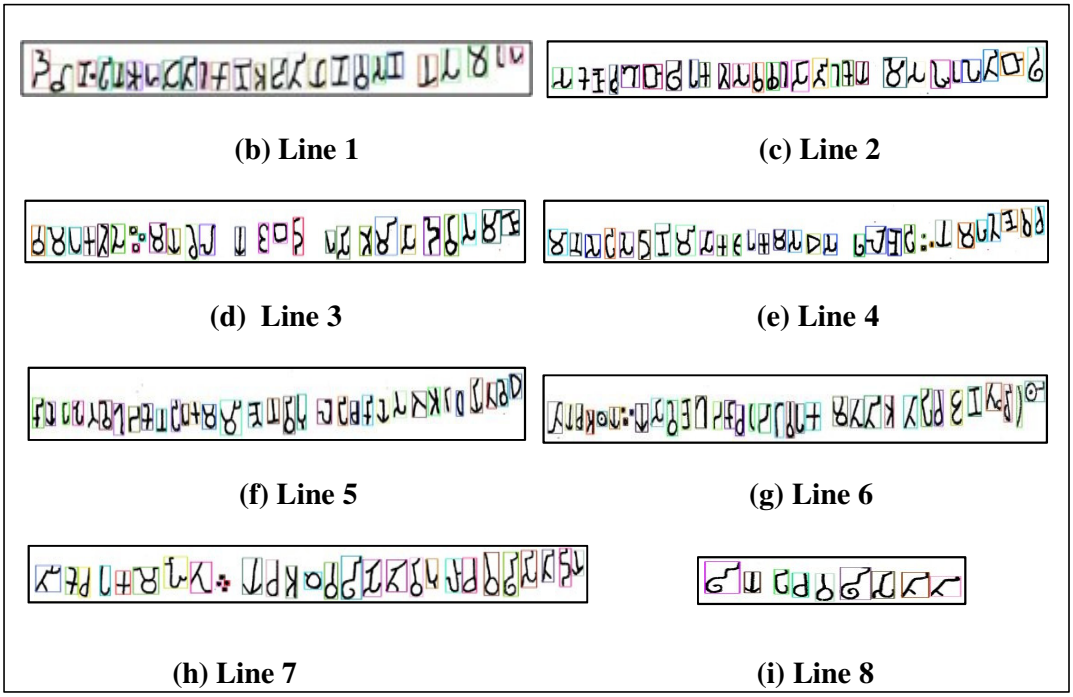
characters are considered as a single letter and segmented incorrectly as a single character. The accuracy rate of the segmentation around 83.5% is achieved for Brahmi script and 67.5% for other ancient Kannada scripts. The result for segmentation of Brahmi script is appreciable when compared to other ancient Kannada script from different periods because of the nature of the script. The complexity of writing style in Brahmi script is less compared to the other scripts from different times.

Contour based Convex Hull and Bounding Box technique

Figure 13(a) is the reconstructed input epigraphic image. The line segmentation is performed to obtain individual lines of text and next is subjected to character segmentation to sample out the characters. The results of character segmentation after line segmentation are shown in Figure 13(b) – 13(i) using Contour Detection and Minimum Bounding Box approach. The characters are segmented irrespective of the skew present in the document.



(a) Reconstructed Input Epigraphical Document Image (S5)



(b) Line 1 (c) Line 2

(d) Line 3 (e) Line 4

(f) Line 5 (g) Line 6

(h) Line 7 (i) Line 8

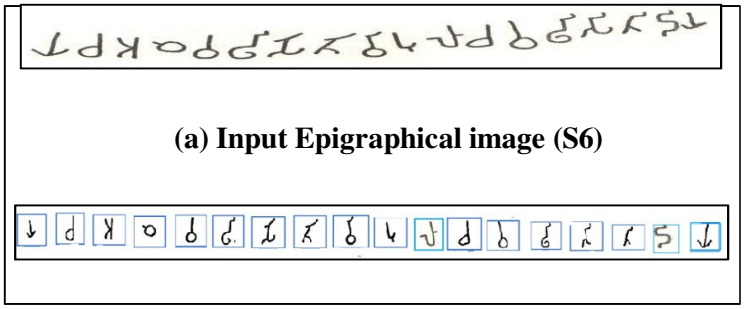
Figure 14: Results of Line and Character Segmentation using Contour Detection and Minimum Bounding Box Method

Performance Analysis

The Contour based Convex Hull and Bounding Box segmentation algorithm is tested on 150 epigraphical images of different periods each with varying number of characters. It works well for a noise free image both for base and compound characters irrespective of the size of the image and skew present. The algorithm is tested on pre-processed images of Brahmi script from Ashoka period. The spacing of text lines and the nature of characters reduces the complexity in the segmentation process of Ashokan Brahmi script yielding encouraging results of an average segmentation rate 92%. However, for the segmentation of other ancient Kannada scripts, achieved a segmentation rate of 83%.

Nearest Neighbor Clustering method

Figure 3.13 depicts the result of segmentation using Nearest Neighbor Clustering. The method segments the image into lines and characters.



(a) Input Epigraphical image (S6)

(b) Segmentation of Characters

Figure 15: Segmentation Results of Nearest Neighbor Clustering Method

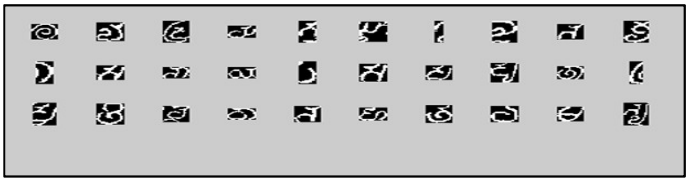
Performance Analysis

The nearest neighbor clustering based method segments the line and character and works even on the skewed document. The method segments the epigraph into lines and characters. The technique works fine for the skewed document. The result of the segmentation is dependent on the cuts and bruises in the input script image. A Segmentation rate around 87% is achieved for Brahmi script and 83.5% for other ancient Kannada scripts. The segmentation results for Brahmi script are appreciable when compared to other scripts because of the less complexity nature of characters.

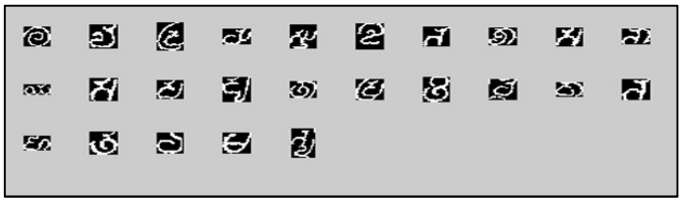
Drop Fall and Water Reservoir Segmentation Techniques

Segmentation is also carried out using Drop Fall and Water Reservoir algorithms especially to address touching characters. A preprocessed and binarized epigraphical image shown in Figure 3.14(a) is considered as input to segmentation phase. Figure 3.14(b) and Figure 3.14(c) represents the result of Segmentation of Characters using Drop Fall algorithm and Water Reservoir algorithm respectively. Segmentation is carried out using Drop Fall and Water Reservoir algorithms to address touching characters. The techniques segment the base and compound characters correctly when the connectivity is present. Few cases, where in the subcomponents of a compound character are disjoint, the compound character is segmented separately into two components. These techniques work well preferably for segmentation of touching characters. A segmentation rate of 85.6% for Drop Fall algorithm and 86.7% for Water Reservoir algorithm is achieved when tested on 150 preprocessed epigraphic documents.

(a) Input Binary Epigraphical image (S7)



(b) The Result of Segmented Characters using Drop Fall Algorithm



(c) The Result of Segmented Characters using Water Reservoir Algorithm

Figure 3.14 : Result of Segmented Characters using Drop Fall and Water Reservoir Algorithms

6. Summary

In this work, importance of preprocessing and segmentation of epigraphical documents have been discussed and techniques are proposed to address the same. Preprocessing includes the enhancement of epigraphical records, noise elimination and binarization. The experimental results of preprocessing which involves enhancement and noise removal stages are tested on 250 camera-grabbed and scanned epigraphical images. Ancient epigraphical documents are enhanced by using suitable Spatial filtering techniques. Mean, Median, Gaussian Blur, Bilateral, Laplace filter, Unsharp Masking (USM) filters are explored with different filter sizes and filter parameters. Thus, the enhancement of historical records - transforms the degraded input document into a better perceivable image. The system performs well for preprocessing of ancient documents and provides flexibility to the user in controlling the process of image enhancement to obtain desired output.

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 MKLas 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 withApplications,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, Povoa 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'. IJCSC, 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 Pureand 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 inElectronics, Computers and Communications (ICAECC), Bangalore, 2018.

[19] Van Phan, Truyen, Bilan Zhu, and Masaki Nakagawa. "Collecting handwritten Nom character patterns from historical document pages." In Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on, pp. 344- 348. IEEE, 2012.

[20] Chamchong, Rapeeporn, and Chun Che Fung. "Text line extraction using adaptive partial projection for palm leaf manuscripts from Thailand." In Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on, pp. 588-593. IEEE, 2012.

[21] Rao, Adabala Venkata Srinivasa. "Segmentation of Ancient Telugu text documents." International

Journal of Image, Graphics and Signal Processing 4, no. 6, pp 8-14, 2012.

[22] K. Srikanta Murthy, G.Hemantha Kumar, P.Shivakumar, P.R.Ranganath, “Nearest Neighbor clustering based approach for line and character segmentation in epigraphical scripts”, Proceedings of International conference on cognitive systems (ICCS-2004), New Delhi, pp. 1-5, 2004.

[23] Doreswamy, K.Srikanta Murthy, G.Hemantha Kumar, P.Nagabhushan “Partial Eight Direction Based Algorithm for Line Segmentation of Epigraphical Script Images” Proceedings of National Conference ETA-2003, Saurashtra University, Rajkot, July 11-13, 2003.

[24] Ramappa, Mamatha Hosalli, and Srikantamurthy Krishnamurthy. "Skew Detection, Correction and Segmentation of Handwritten Kannada Document."International Journal of Advanced Science and Technology 48 (2012).

[25] Kang, Le, David Doermann, Huaigu Cao, Rohit Prasad, and Prem Natarajan. "Local segmentation of touching characters using contour based shape decomposition." In Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on, pp. 460-464. IEEE, 2012.

[26] Kumar, S. Raja, and V. Subbiah Bharathi. "An off line ancient tamil script recognition from temple wall inscription using Fourier and Wavelet features."Eur. J. Sci. Res 80, no. 4 (2012): 457-464.

[27] Alirezaee, Shahpour, Hassan Aghaeinia, Majid Ahmadi, and Karim Faez. "Recognition of middle age Persian characters using a set of invariant moments." In Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on, pp. 196-201. IEEE, 2004

[28] Sridevi, N., and P. Subashini. "Combining Zernike moments with regional features for classification of handwritten ancient Tamil scripts using Extreme learning machine." In Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), 2013 International Conference on, pp. 158-162. IEEE, 2013.

[29] Bandara, Dammi, Nalin Warnajith, Atsushi Minato, and Satoru Ozawal. "Creation of precise alphabet fonts of early Brahmi script from photographic data of ancient Sri Lankan inscriptions." Can. J. Artif. Intell. Mach. Learn. Pattern Recognit 3, no. 3 (2012): 33-39.

[30] Meza-Lovon, Graciela Lecireth. "A graph-based approach for transcribing ancient documents." In Ibero-American Conference on Artificial Intelligence, pp. 210-220. Springer Berlin Heidelberg, 2012.