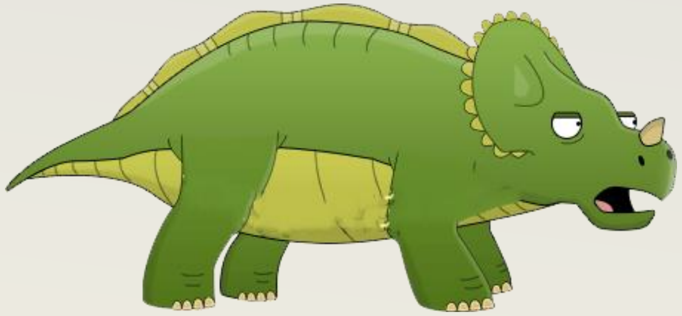
Ehsan Majidi

9917051059

Amir Shokri

Denoising and Segmentation of Epigraphical Estampages by Multi Scale Template Matching and Connected Component Analysis

Third International Conference on Computing and Network Communications (CoCoNet'19)



Abstract :

Epigraphy refers to the study of Epigraphs, or inscriptions. Estampage or stamping, is a term that is commonly used in epigraphy in order to get an exact copy of an inscription.

An estampage is typically derived by applying wet paper onto the rock face, over which any ink material is wiped. This process also results in the appearance of unwanted noise which arises due to the texture of the rock, human errors or a variety of different factors.

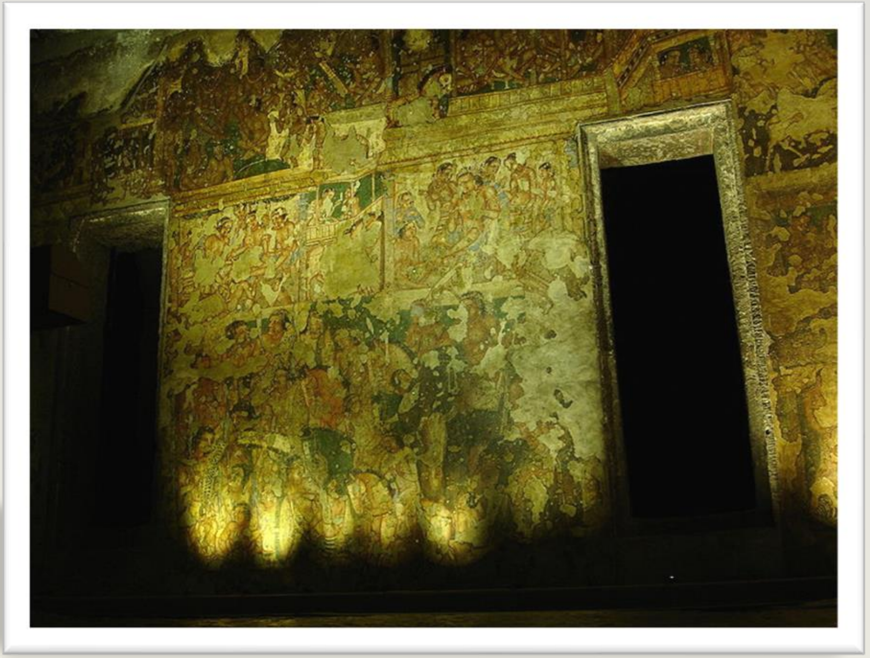
In this paper, we propose CCD: Connected Component Denoising , a novel approach for denoising and character segmentation via connected component analysis and multi scale template matching. A dynamic threshold is obtained via histogram density projections to effectively remove noisy areas. A comparative study is then performed between the different approaches (MSTM , CC , Hybrid).

Introduction :

Epigraphy is the detailed analysis of inscriptions on stones, pillars, coins and other surfaces.

The processing and preservation of these scriptures in India is done by the Archeological Survey of India. The engravings of interest can be spotted even today on a multitude of temple walls such as the famous Ajanta and Ellora caves.

Ajanta



Ellora

Literature Survey :

Connected components labeling scans an image and groups its pixels into components based on pixel connectivity, i.e. all pixels within a connected component have pixel intensity values similar to each other and are in some way connected.

Once the scan is complete, each pixel is given a label or a value corresponding to the component it belongs to.

Extracting and labeling of various disjoint and connected components in an image is central to many automated image analysis applications.

Histogram equalization is a technique which is frequently used for improving image quality. Smoothening and Continuous intensity allocation provide stronger averaging effects and reduce the effect of outliers .

There are various metrics for evaluation of image quality. Image quality evaluation methods are divided into objective and subjective methods

Background Concepts :

Understanding the various individual pieces of the complete algorithm is important to its working. The pipeline can be broken down into 3 major pieces :

- **White Background Removal**
- **Average Character Area Finder**
- **Connected Component Analysis**

In the first step there is a need for some form of a bounding box solution. Initially few simple morphological operations are performed to remove certain noise.

After which contours have come of use in the process of finding the smallest box that encloses the image of interest

$$\Sigma_{in\ row}\, pixelval\ (\ ) = Horizantal\ projection$$

$$\Sigma_{in\ col}\, pixelval\ (\ ) = Horizantal\ projection$$

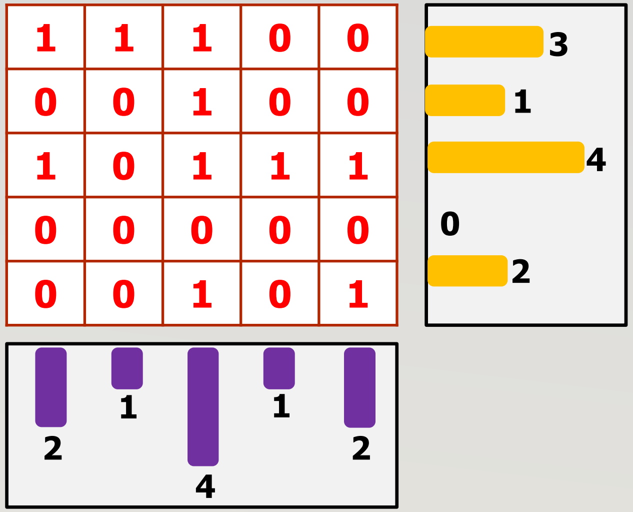


Fig 1 : Horizontal and Vertical projections of an image

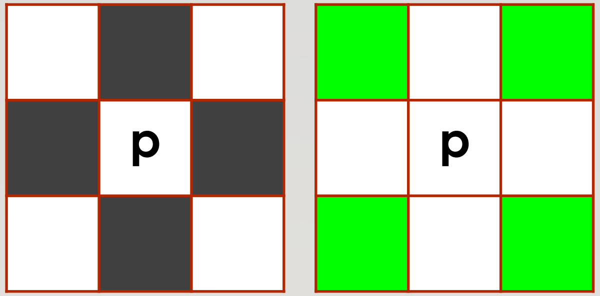


Fig 3 : 4 neighbourhood and d- neighbourhood of a pixel p

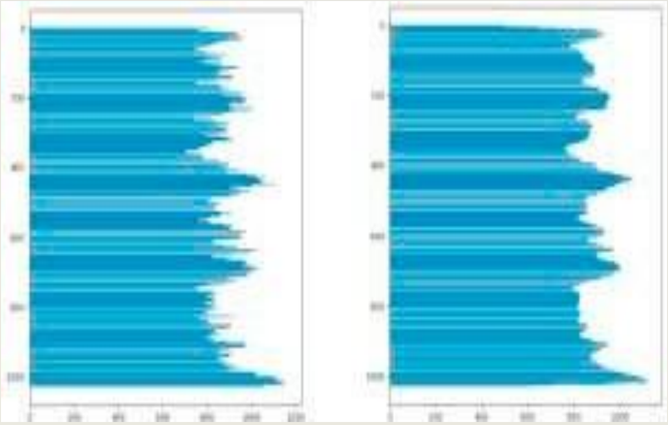


Fig 2 : Original Projection and Smoothened Projection

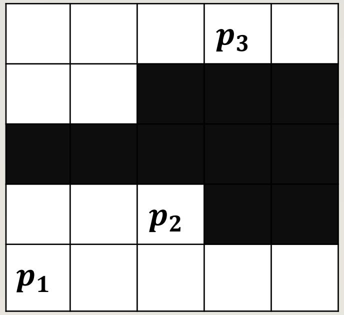


Fig 4 : Connectivity and connected components

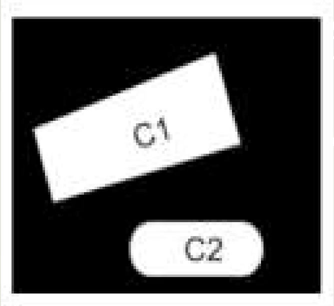


Fig 5 : 2 Connected Components in an image

Connected Components：

Union-Find Data Structure and Algorithm

A disjoint-set , also known as Union-Find data structure is a data structure that keeps track of a set of elements partitioned into a collection of disjoint (non overlapping) subsets. The algorithm performs 2 operations :

1. Find : This operation determines which subset a particular element belongs to. Can be used to find some form of similarity between elements belonging to the same subset. It thus answers the question " Are two grid points connected ? ".
2. Union : Combines two individual subsets into a single subset. It thus acts as an operation that adds an edge between two points.

Proposed Method ：

The Archeological Survey of India (ASI) has provided a set of 144 epigraphs for testing. Each of these images are in binary black and white format or grayscale format. Most of these images contain significant amounts of noise and preprocessing must be done in order to serve as an input image to the next stage (Segmentation).

- **The source image used is the output images of the MSTM algorithm. These images serve as ideal inputs since our previous approach effectively removed all small noise and also helped splitting large chunks of character components into smaller more useful connected components. The results section further explains as to why this ensemble technique is more efficacious than just connected component denoising.**
- **The source image is then projected on to the X-axis and Y-axis. A histogram relating the axis to the pixel density is plotted and an estimation of the average character size is obtained.**
- **The projections observed tends to be erratic in nature and thus to find clear line separation or character spacing , the projections are smoothened.**
- **A percentage of this obtained character area is then used as a threshold value for the connected component analysis. The threshold is an indicator for meaningful component preservation.**
- **8 neighbor connected component labelling is performed and each pixel is assigned a respective value based on the component as it is located in. Connected component analysis is then performed to remove all components that fall below the threshold value.**
- **The output image is then obtained and image quality is assessed using PSNR and SSIM values.**
- **The output image was segmented using the connected components method by using gradients to colorize the different components found.**
- **Then the gradient range was divided into k bins , where k represents the no of lines of text in the image.**
- **Each gradient range was then looped through and the lines of text were separated from the rest.**

Connected Components

Now a connected component as defined earlier is set of pixels that are belonging to each others neighbourhood forming a chain of connectivities. For the sake of the algorithm we define a connected component simply as a set of mutually connected vertices.

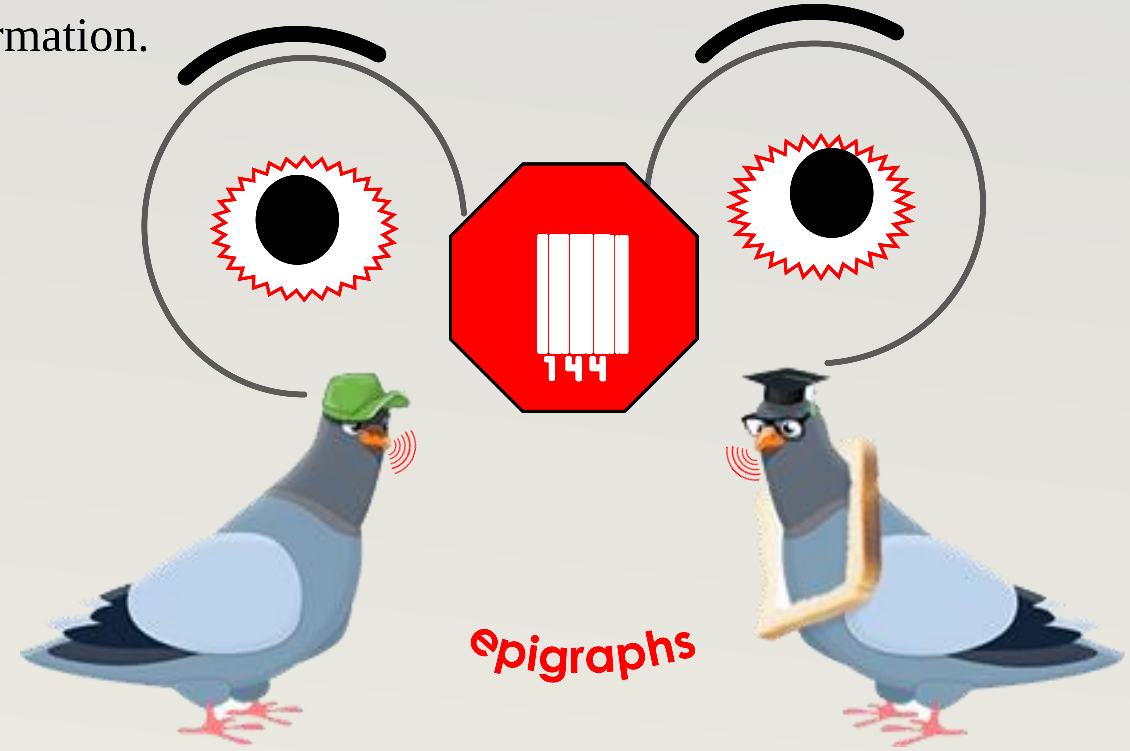
We start off with each pixel or vertex being a unique component. An observation can now be made that applying one union operation reduces the no of components by 1.

The algorithm in its simplest form, runs 2 passes on the image. The first pass in the classical algorithm makes use of dynamic programming for labeling. It goes through each pixel and looks at the labels that have already been assigned to 2 neighbouring pixels , one top and one left of the pixel of interest. This creates 2 cases :

1. In the case that the pixels are background pixels , create a new label for both of them.
2. In the case that either one or both are not background pixels , leave the labels as it is.

Results :

The proposed method was tested on a set of 144 epigraphs provided by the ASI. Since there does not exist an ideal or "noiseless" image for comparison, the values provided by the estimators do not provide a clear indicator to the image quality. However, a sufficiently high value of both these indicators indicate that the character integrity has been retained and that there is no loss in valuable information.



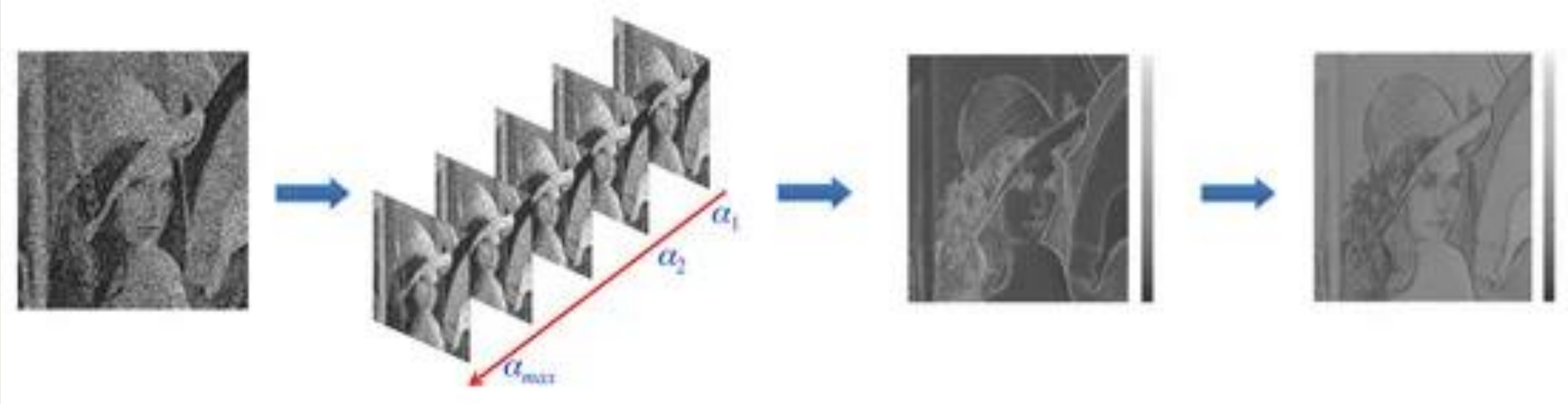
Method	Average SSIM	Average PSNR
Wiener Filter	0.9547	28.336
Median Filter	0.9410	26.033
CCD	0.8802	14.121

Table 1 : Comparison with known algorithm

Image	SSIM	PSNR
Sample 1	0.87	14.27
Sample 2	0.85	13.25
Sample 3	0.92	16.37
Sample 4	0.87	13.97
Sample 5	0.88	14.21
Sample 6	0.95	17.54
Sample 7	0.94	16.32
Sample 8	0.88	13.55
Sample 9	0.81	12.04
Sample 10	0.92	15.32

Table 2 : SSIM and PSNR for a few images

wiener filter



Median Filter

Results

A comparative study was performed to find the best possible solution with the set of novel algorithms developed for the same. Multi Scale Template Matching was first used as a baseline to benchmark the clear images. In MSTM , a template is created which consists of the noise in the epigraph.

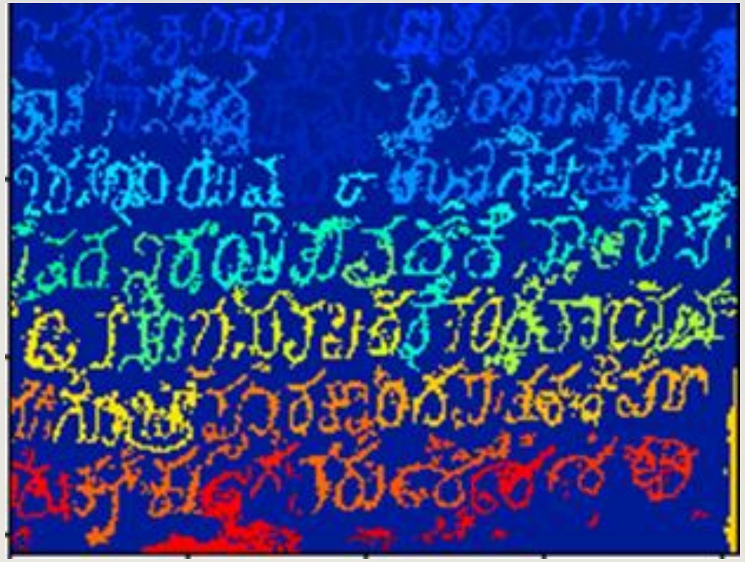


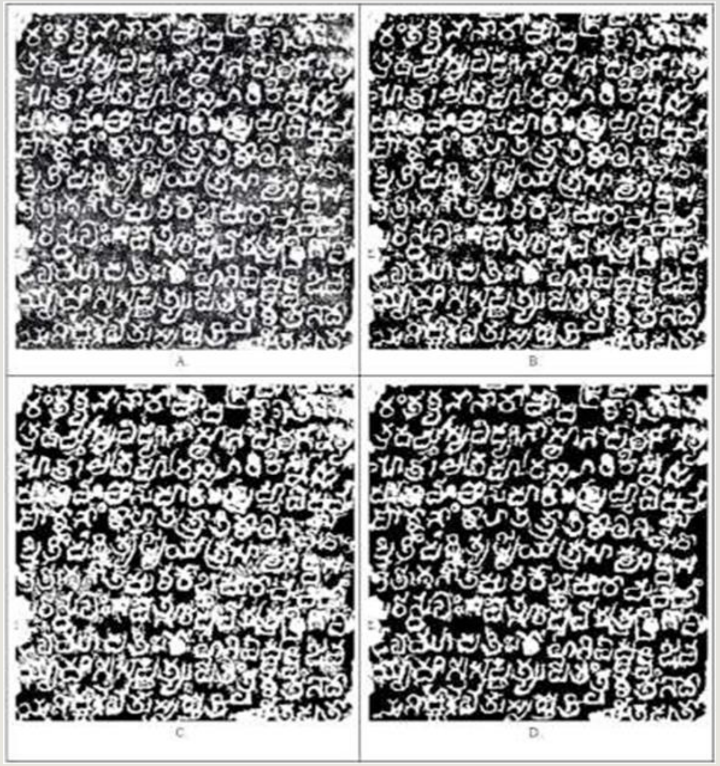
Fig 7: Gradient filled connected component

Results

The segmentation is not free of errors due to extremely small connections caused by rocky surfaces in the images. These cause false positives and misclassify certain components. But the complete task of segmentation itself can be considered as future enhancements that can complete the algorithm.

Fig 8 : Detected line 1 and line 2 of the image

Results



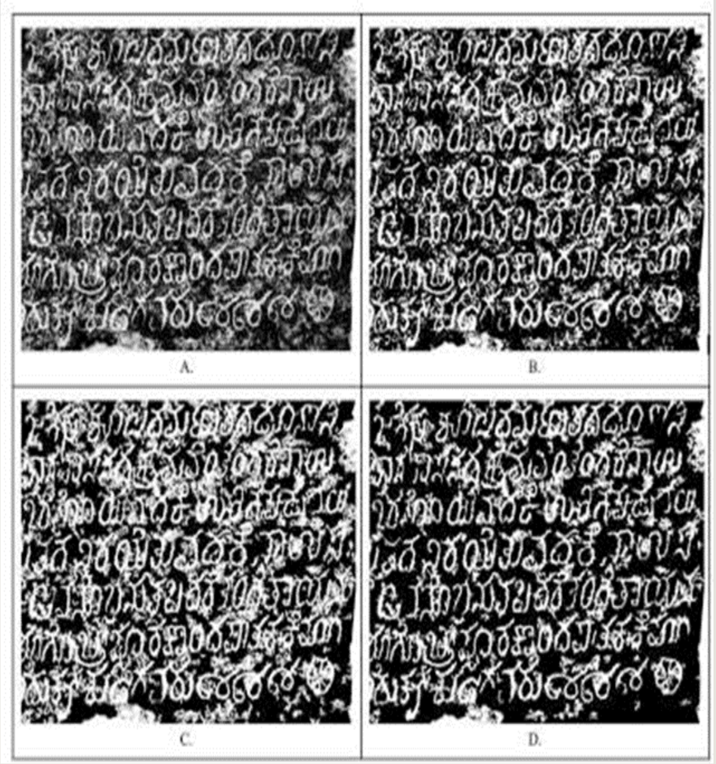


Fig : A: Input , B: MSTM output , C: CCD output , D: Hybrid approach output

Conclusion :

The objective of this research paper is to complete the initial phase of a process that starts with simple epigraphical estampages and finishes at character recognition. The precision of the OCR can be improved with higher quality “less noisy” images while also making sure of preserving character integrity and structure.

This paper thus focuses on creating a fully automated denoising algorithm which is time and space efficient in order to aid in further steps of processing such as segmentation and recognition.

The proposed method makes use of projections along with connected component labeling for denoising. The results have been evaluated using PSNR and SSIM metrics and are compared with standard denoising filters such as the median and wiener filters. It can be observed by the HVS that there is a noticeable difference in quality between the results from CCD and those of the standard filters, Thank you for your attention