From Stone Inscriptions to Digital Text: Enhancing Archaeological Scripts and Automated Conversion

P. Vasuki, Ashwini T, Dharagesh T, Deiva Kauvya M, Divyadarshini G

Department of IT, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India, VasukiP@ssn.edu.in

Analysis of archeological inscription plays a crucial role in understanding the values of historical, cultural and heritage information. Many of the available inscriptions are damaged and deteriorated in color and clarity, thus the identifications become a challenge. In this work, Various filtering techniques used to enhance the Tamil stone inscription images are analyzed empirically. Performances of filtering techniques are compared and evaluated in terms of certain features of an image, specifically the Peak Signal-to-Noise Ratio (PSNR)value, Signal-to-Noise Ratio (SNR) and Mean Squared Error (MSE) value. The effectiveness and efficiency of these techniques are tested using different categories of input images; thereby the most efficient filtering technique can be determined based on the image. As the ancient inscriptions are of different formats, we have devised a conversion system which converts Brahmi scripts to modern Tamil script. OCR is used to recognize the Brahmi characters and converted into Tamil characters using code point toning technique. The proposed system uses Tesseract OCR, which has been trained manually for the recognition of Tamil Brahmi Letters. The core functionality of the OCR depends on LSTM. LSTM cells store information about the input for a specified time period which makes it suitable for language training. Further to transcript the recognized 1 Brahmi letters to present age Tamil letters, code points are toned with the present age Tamil letters using Unicode-Character lookup table. The overall character and word error rate of the system has been reduced to 0.0086 and 0.057%

Keywords: Signal-to-Noise Ratio (SNR), Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR).

1. Introduction

Writing in India has a rich historical backdrop, dating back to the illustrious era of the Indus civilization, spanning approximately from 2500 to 1500 B.C. The early period inscriptions show the existence of writings since the ancient periods. In the period of Ashoka from 269 to 232 B.C., is a pivotal era for inscriptions where the widespread use of scripts such as Brahmi, which found its way across the territories under his rule, including Sri Lanka and Tamil Nadu. These Inscriptions found in rock-shelters and potsherds are unearthed and have been evolved to many languages like Tamil & Prakrit words showing the diverse nature and dynamic evolution of letters of language.

One of the most significant developments during this era was the emergence of the Tamil-Brahmi script, which is believed to have originated in Tamil Nadu before the Asokan period. Also known as Dhamili or Tamil, this script laid the foundation for the development of subsequent scripts in the region. During the Mauryan era, the Tamil script underwent significant adaptations, especially when it had to be adapted to accommodate the Prakrit language. This period was also marked by the influence of Asokan Brahmi, leading to the coexistence of multiple script systems in Tamil Nadu, including the Tamil-Brahmi II system[29].There are many variations in the Brahmi scripts due to amalgamation of diverse systems. Tholkappiam has given standard frameworks like consonants and vowels.

There are 3 different systems of Tamil, Brahmi I & II and Pulli systems that were present in earlier days. Amongst them, Tamil Brahmi -II scripts were widely used[29].

2. LITERATURE REVIEW

Ancient scripts are valuable treasures which gives information about our language, culture, lifestyle and knowledge of that contemporary generation Siromoney et al., 2020). Primary challenges in this process are cleaning the images and efficiently converting Brahmi characters to Tamil characters.Museums and several historical centres implement many steps to preserve these ancient documents. The amount of degradation suffered by such documents needs to be denoised so as to retrieve the information from these documents. This section illustrates the various methodologies used for the enhancement of inscription images and an insight of the problem domain. It describes various image enhancement techniques in detail.

Histogram Equalization (HE) has become popular due to its simplicity and efficacy. Using contrast adjustment, the clarity and vibrancy of images are enhanced(Guo et. al,2013). Linear cumulative histogram yield an enhancement in images and is applied across the domains such ad speech recognition, medical image processing and text synthesis(Pizer et al., 1987). HE initializes the image by finding the pixel value of the gray level image (gray level is 0 - 255). From the histogram of the image (plotting the pixel value of the image), cumulative distribution function values are calculated. Finally the enhanced image is displayed by plotting the cumulative distribution function value.

Ostu’s extension of binarization of document images method’s adaptive and parameter-free extension addresses the parameterization issues in binarization methods (Sezgin & Sankur, 2004) by incorporating grid-based modeling and background map estimation. The parameter behavior has been enhanced by estimation of average stroke width and line height and generalized to multiscale binarization to handle diverse document images effectively.

Praveen et al. (2013) introduced a robust technique using adaptive image contrast for binarizing document images to address the difficulty of separating text from severely degraded document backgrounds. This method generates an adaptive contrast map, binarizing the map and overlap with edge maps to detect text stroke edges, facilitating precise text segmentation.

Llados et al. (2009) introduced binarization techniques to enhance the degraded images, where the phase information of images incorporates Gaussian filtering and median filter to improve output quality and eliminate noise.

Unlike traditional thresholding approaches, He and Wang (2015), proposed a spatially adaptive statistical method for binarizing degraded document images. utilizes maximum likelihood classification and spatial relationships on the image domain (He & Wang, 2015). This method extracts text from the images using soft-max techniques.

Some researchers use Deep learning based information retrieval from inscripts(Devi et al., 2018). The performance based on these systems are less, since training in normal images may not be sufficient to retrieve information from the deteriorated inscription. The inscription has different structured letters at different periods. Thus within a specific era available inscriptions are not sufficient to train the deep networks.The same problem arises for the researchers when working with template matching systems to decipher Brahmi to Tamil text (Gautam et al., 2019).

Researchers have explored various methods for Brahmi character recognition, including the coded run method and preprocessing techniques like thinning and thresholding (Siromoney et al., 2020). But it faces the challenges of recognizing non connected characters.

As the Deep learning aspects have these issues, we preferred to go with traditional filtering techniques and an adaptive method is used to choose an appropriate filter for the enhancing given image. Tesseract OCR is trained using LSTM to recognize the Brahmi script.

3. PROPOSED WORK

This work is aimed to decipher the ancient inscriptions to equivalent Tamil text. The system addresses the challenges used in deciphering the text from the noisy and blurred images. The work has been divided into 2 phases. In the first phase the image has been enhanced using various filtering techniques. The optimal filtering technique is chosen based on adaptive technique. In the second phase, the Brahmi to Tamil text conversion system has been implemented. The Tesseract OCR method has been trained with Brahmi text using LSTM.

A. Proposed Methodology

Image Enhancement:

The proposed system aims to enhance the inscriptions which are degraded due to noise using various filtering techniques.

DATASET

The number of inscriptions is large which encompass a wealth of knowledge-rich cas well the content are rich in information. The Tamil Kalvettukal book and web sources contain a sufficient number of these inscriptions.

IMAGE PROCESSING

The acquired images has been converted into JPEG format and then converted into gray scale. The gray scale image entails eliminating hue and saturation information but retaining luminance. The gray scale image is more suited for morphological operations and image segmentation.

The Histogram on images shows the distribution of intensities within an image. The 'imhist' function illustrates the intensity variation of the processed image at each stage.

Contrast enhancement techniques are adjusting the brightness values and optimizing color images. The contrast enhancement improves the image's analyzability and interpretation. Histogram equalization adjusts the intensity distribution, by spreading out most frequent intensity values. Thus increasing global contrast and lowering local contrast. Contrast stretching stretches the intensity of the images to the full range of pixel values, permitted by the type of image.

Contrast Stretching is obtained using the formula

$$s = (r - r_{min}) \quad \frac{(I_{max} - I_{min})}{(r_{max} - r_{min})} + I_{min}$$

For I_{min} = 0 and I_{max} = 255 (for standard 8-bit grayscale image)

Histogram equalization and contrast stretching are employed to enhance the input image before proceeding to noise removal.

CONTRAST STRETCHING

Contrast stretching, also known as normalization, enhances image contrast by stretching the range of intensity values to the full range of pixel values permitted by the image type. This simple technique aims to utilize the entire dynamic range of pixel values, thereby enhancing image contrast for better visualization and analysis.

General Formula for Contrast Stretching:

$$s = (r - r_{min}) \quad \frac{(I_{max} - I_{min})}{(r_{max} - r_{min})} + I_{min}$$

For I_{min} = 0 and I_{max} = 255 (for standard 8-bit grayscale image)

where,

r = current pixel intensity value

r_{min} = minimum intensity value present in the whole image

r_{max} = maximum intensity value present in the whole image

FILTERING TECHNIQUES

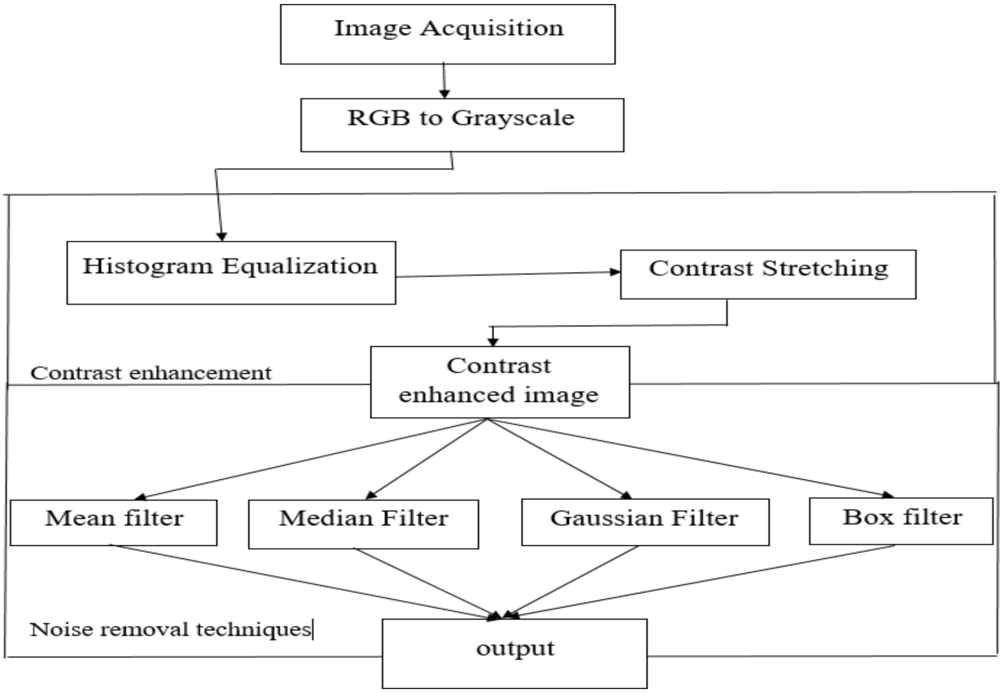

Figure 1: Selection of output image based on the quality of the output picture.

LINEAR FILTER

Various filtering techniques are used to process the images. High pass filters smoothen the image and low pass filter does edge detection whereas the various kinds of noises of images are removed using Mean, Median, Box and Gaussian filters. The filters have been designed with different mask sizes and parameter values.

In linear filtering the output is generated as a linear combination of input pixel's neighborhood values. These filters linearly add the time varying input signals with linearity constraints.. Mean filter, Gaussian filter, and Box filter are the kinds of linear filters.

Mean filtering is used to reduce the noise by removing the intensity variations between adjacent pixels. The value of the pixel is mapped to the average value of the neighboring pixels within the specified boundary.

The intensity of a pixel (i, j) of an image I is replaced by

$$I(i,j) = \frac{1}{M}\sum_{(x,y)\in N} I(x,y)$$

Convolution mask is used to calculate the new value where all coefficients are set to 1/M to represent the average value of the neighborhood pixels in N.

Gaussian filters are used to assign a mask and weights of the mask as Gaussian function's shape. The intensity of every pixel is changed to the weighted sum of the pixels covered in gaussian shape, where the weights are assigned based on spatial distance and pixel intensity. The calculated weights preserve the edges and cancel the noise of the type Gaussian.

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-(x-a)^2}{2a^2}}$$

The parameters of gaussians are tuned based on the training taken in the noisy and clear images synthesized using imagenet data set. The selection of appropriate value is crucial as the filter assigns weights to the mask fraction based on sigmas. As most of the signals are gaussian in nature this filter is a more suitable filter as well it is an computationally effective filter as the computation is space and time efficient. Gaussian smoothing facilitates the acquisition of a multi-scale space representation of an image and is effective for tasks such as edge detection or unsharp masking.

A box linear filter, calculates the average value of every pixel. Equal weight is given for all surrounding points, which is the simplest filter and also used to calculate Guassian blur. Box filter is a low-pass filter, effectively blurring the image.

NONLINEAR FILTER

In nonlinear filters, the pixels are ranked based on intensity. The target pixel has been replaced by either minimum, maximum or the middle value of the pixel among the neighborhood pixels in a specific locality based on the chosen rank.

The median filter is a nonlinear filter that removes the impulse noise, often referred to as "salt and pepper" noise, from images. It operates as a nonlinear signal processing technique for noise suppression. The process involves:

1. Reading the pixel values.

2. Sorting the list of pixel values from the surrounding neighborhood.

3. Selecting the median, which represents the central value in the sorted list.

4. Replacing the central pixel of the mask with this median value.

5. Moving the window by one pixel and repeating the process until the entire image is processed.

The median filter is particularly adept at preserving image details while effectively reducing noise. The new value of pixel located at x,y is given by

$$f(x,y) = median_{(s,t)\in S_{xy}}\{g(s,t)\}$$

f(x,y) is the median value, which has been calculated from the image g(s,t) where the area is defined by Sx,y.

The figure 2 shows the inclusion of various filters in image enhancement techniques.The system is used to choose the best method out of all methods said above. The input image is given to all the filters. The SNR of the resultant image, which has come as output of all filters, is calculated and the final output is chosen to be the one which produces less SNR.

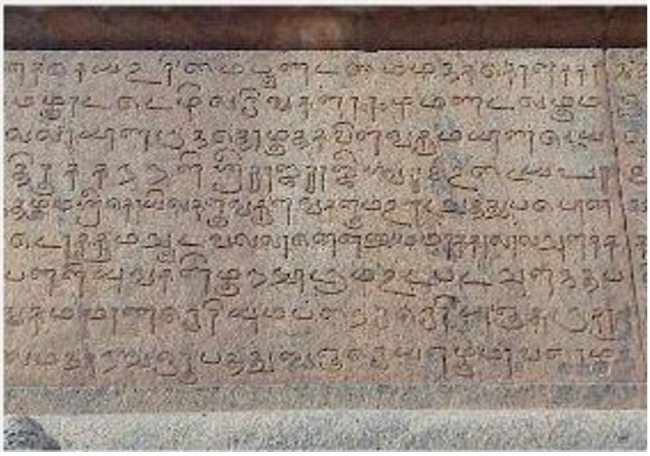

Fig 2: Input image

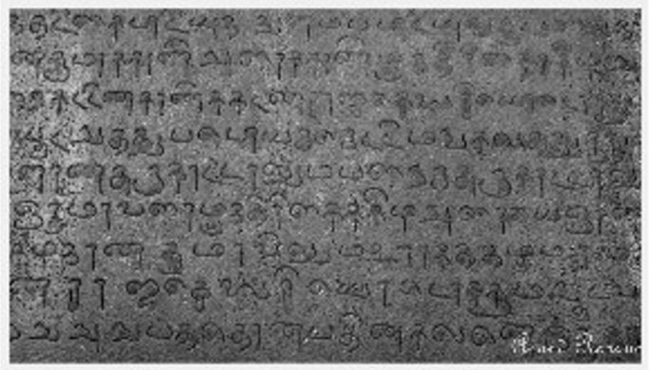

Fig 3: Gray Scale Image

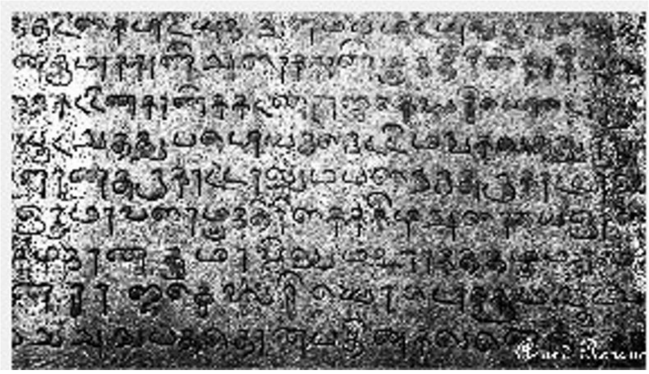

Fig 4: Histogram Equalisation

Fig 5 Contrast Stretching

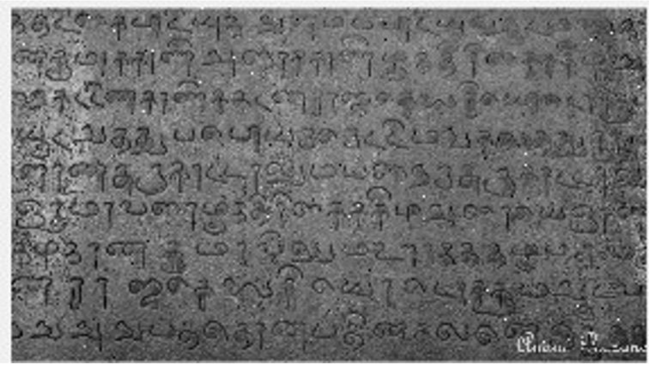

Fig 6: : Noise removal using Mean filter (Image 1)

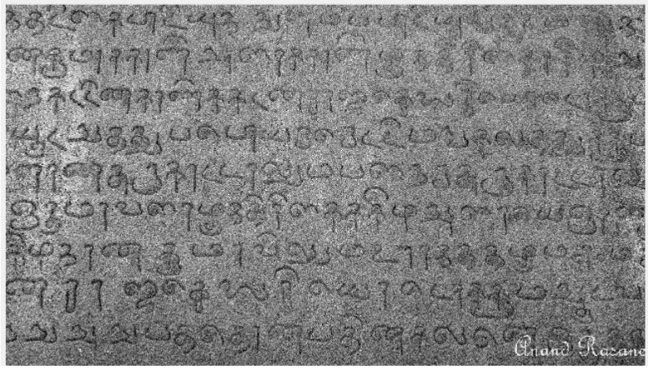

Fig 7: Noise removal using Median filter (Image 1)

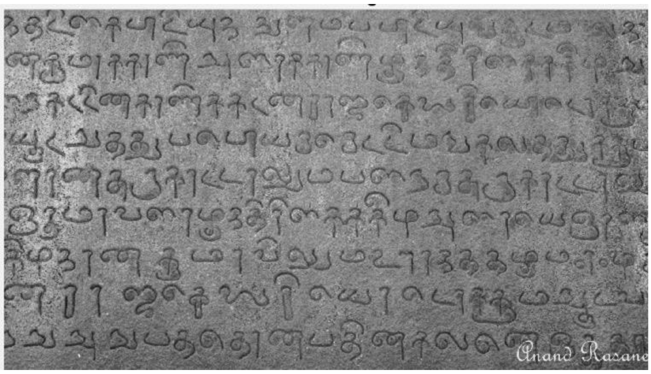

Fig 8: Noise removal using Gaussian filter

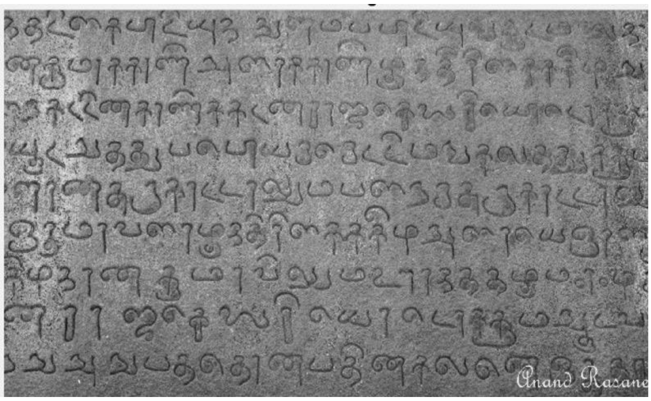

Fig 9: Noise removal using Box filter

The appropriate filter has been chosen based on the SNR ratio of the input and output image. Different images are degraded with different noise values, thus the appropriate filter has been chosen based on the performance of the various filter on the given image.

PERFORMANCE MEASURES

The signal-to-noise ratio (SNR), the peak signal-to-noise ratio (PSNR) and mean-squared error(MSE) are used to compare image compression quality.

SIGNAL TO NOISE RATIO

Signal-to-noise ratio is a measure used to compare the level of a desired signal to the level of background noise. SNR is defined as the ratio of signal power to the noise power, often expressed in decibels. Signal-to-noise ratio (SNR) is used in imaging to characterize image quality. The sensitivity of the digital file is typically described in the terms of the signal level that yields a threshold level of SNR.

PEAK SIGNAL TO NOISE RATIO

PSNR stands for Peak Signal to Noise Ratio. This ratio is often used as a quality measurement between the original and a compressed image. PSNR is another character of image which can describe the improvement of the contrast. Contrast enhancement adds noise to the image. For a good contrast enhanced image, the noise should be as minimal as possible.

PSNR value is inversely proportional to noise, this implies that PSNR value should be as high as possible.

The higher the PSNR, the better the quality of the compressed, or reconstructed image.

- PSNR is calculated as

$$PSNR = 10 \log_{10}\left(\frac{R^2}{MSE}\right)$$

where, R is the maximum fluctuation in the input image data type.

MEAN SQUARED ERROR

MSE stands for Mean Squared Error. It represents the cumulative squared error between the compressed and the original image.

The lower the value of MSE, the lower the error.

MSE is calculated as given in the following equation

$$MSE = \frac{\sum_{M,N} [I_1(m,n) - I_2(m,n)]^2}{M * N}$$

where, M=number of rows; N=number of column

CONVERSION OF BRAHMI INSCRIPTION TO TAMIL TEXT

Tamil is an ancient language which comes across various variations. There are many deep learning methodologies that convert input images into textual form as shown in the figure 11. The Tesseract engine preprocesses the image and does segmentation and processes the segmented letters to identify the digit. The system has been pre trained with different languages. We have LSTM to work with the pretrained modal.

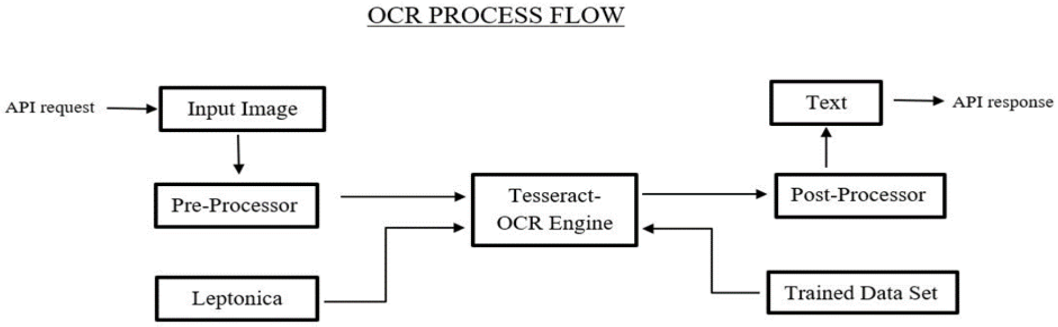

Figure 10: OCR Process flow

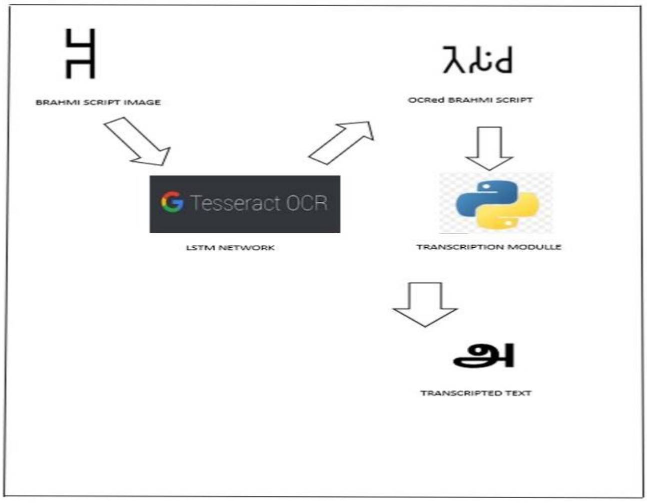

Figure 11. Conversion of Brahmi letter to Tamil letter

The System architecture depicted in Fig 10 explains the overall working of the proposed system. A Brahmi Script image is scanned first and is sent as an input to the pre-trained OCR engine. The OCR engine after recognition gives a Brahmi script text file containing the OCRed Brahmi letters from the image. This text file in turn is given to the Transcription Module which maps the Brahmi characters with the present age Tamil characters.

Tesseract is an open source text recognition (OCR) Engine, available under the Apache 2.0 license. It can be used directly, or (for programmers) using an API to extract printed text from images.

4. CONCLUSION

The stone inscription images collected from various temples taken for enhancement of data were subjected to various contrast enhancement techniques like histogram equalization and contrast stretching and to remove the noise it has undergone various filtering techniques including mean filter, median filter, gaussian filter and box filter. The system helps to denoise the images and helps to retrieve the information from the same. The historical heritage of the inscriptions are thus preserved.

References

1. V. Singrodia, A. Mitra and S. Paul, "A Review on Web Scrapping and its Applications," 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2019, pp. 1-6, doi: 10.1109/ICCCI.2019.8821809.
2. G. V. Mantena, S. Rajendran, B. Rambabu, S. V. Gangashetty, B. Yegnanarayana and K. Prahallad, "A speech-based conversation system for accessing agriculture commodity prices in Indian languages," 2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays, Edinburgh, UK, 2011, pp. 153-154, doi: 10.1109/HSCMA.2011.5942384.

3. S. Thivaharan., G. Srivatsun. and S. Sarathambekai., "A Survey on Python Libraries Used for Social Media Content Scraping," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 361-366, doi: 10.1109/ICOSEC49089.2020.9215357.
4. Afzal, M. & Zubairi, Junaid & Akram, Attiya & Akram, Hazzi. (2015). AGRIKIOSK – A digital tool to reach farmers and rural communities. 10. 35650-35656.
5. M. Ramalingam, D. Saranya, R. ShankarRam, P. Chinnasamy, K. Ramprathap and A. Kalaiarasi, "An Automated Framework For Dynamic Web Information Retrieval Using Deep Learning," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-6, doi: 10.1109/ICCCI54379.2022.9741044
6. Xiaowei Sheng and Minghu Jiang, "An information retrieval system based on automatic query expansion and Hopfield network," International Conference on Neural Networks and Signal Processing, 2003. Proceedings of the 2003, Nanjing, China, 2003, pp. 1624-1627 Vol.2, doi: 10.1109/ICNNSP.2003.1281192
7. H. B. Sailor, H. A. Patil and A. Rajpal, "Unsupervised Filterbank Learning for Speech-based Access System for Agricultural Commodity," 2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR), Bangalore, India, 2017, pp. 1-6, doi: 10.1109/ICAPR.2017.8593040
8. S. Lunn, J. Zhu and M. Ross, "Utilizing Web Scraping and Natural Language Processing to Better Inform Pedagogical Practice," 2020 IEEE Frontiers in Education Conference (FIE), Uppsala, Sweden, 2020, pp. 1-9doi: 10.1109/FIE44824.2020.9274270.
9. A. S. Bale, N. Ghorpade, R. S, S. Kamalesh, R. R and R. B. S, "Web Scraping Approaches and their Performance on Modern Websites," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2022, pp. 956-959, doi: 10.1109/ICESC54411.2022.9885689
10. R. R. N. R, N. R. S and V. M., "Web Scrapping Tools and Techniques: A Brief Survey," 2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT), Kottayam, India, 2023, pp. 1-4, doi: 10.1109/ICITIIT57246.2023.10068666.
11. J. Meng, J. Zhang and H. Zhao, "Overview of the Speech Recognition Technology," 2012 Fourth International Conference on Computational and Information Sciences, Chongqing, China, 2012, pp. 199-202, doi: 10.1109/ICCIS.2012.202.
12. .Wenhao Ou, Wanlin Gao, Zhen Li, Shuliang Zhang and Qing Wang, "Application of keywords speech recognition in agricultural voice information system," 2010 Second International Conference on Computational Intelligence and Natural Computing, Wuhan, 2010, pp. 197-200, doi: 10.1109/CINC.2010.5643755.
13. C. Yu, G. Yang, X. Chen, K. Liu and Y. Zhou, "BashExplainer: Retrieval-Augmented Bash Code Comment Generation based on Fine-tuned CodeBERT," in 2022 IEEE International Conference on Software Maintenance and Evolution (ICSME), Limassol, Cyprus, 2022 pp. 82-93. doi: 10.1109/ICSME55016.2022.00016
14. Q. Zhu, X. Wang, C. Chen and J. Liu, "Data Augmentation for Retrieval- and Generation-Based Dialog Systems," 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 2020, pp. 1716-1720, doi: 10.1109/ICCC51575.2020.9344922
15. D. Kurniawati and D. Triawan, "Increased information retrieval capabilities on e-commerce websites using scraping techniques," 2017 International Conference on Sustainable Information Engineering and Technology (SIET), Malang, Indonesia, 2017, pp. 226-229, doi: 10.1109/SIET.2017.8304139.
16. S. Sharma and A. Bhagat, "Information Extraction from Web Pages Using Hyperlinks," 2022 10th International Conference on Reliability, Infocom Technologies and Optimization

(Trends and Future Directions) (ICRITO), Noida, India, 2022, pp. 1-3, doi: 10.1109/ICRITO56286.2022.9964792.

17. S. K. Patnaik and C. N. Babu, "Information Retrieval from web with Faster R-CNN Deep Learning Networks: A New Perspective," 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 2021, pp. 61-66, doi: 10.1109/UBMK52708.2021.9558956.
18. C. Saini and V. Arora, "Information retrieval in web crawling: A survey," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 2016, pp. 2635-2643, doi: 10.1109/ICACCI.2016.7732456.
19. Thenmozhi, D., & Aravindan, C. (2009). Tamil-English Cross Lingual Information Retrieval System for Agriculture Society.
20. Yadlapalli, Kasiviswanadham. (2012). K-RIAD Kiosk for Rural India Agricultural Development -Farmer to E-Farmer
21. Y. Ahn, S. -G. Lee, J. Shim and J. Park, "Retrieval-Augmented Response Generation for Knowledge-Grounded Conversation in the Wild," in IEEE Access, vol. 10, pp. 131374-131385, 2022, doi: 10.1109/ACCESS.2022.3228964.
22. S. Das, S. B. Partha and K. N. Imtiaz Hasan, "Sentence Generation using LSTM Based Deep Learning," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 1070-1073, doi: 10.1109/TENSYMP50017.2020.9230979
23. Siromoney, G., & Rajaram, S. (2020). A Study on the Character Recognition of Early Brahmi Script for Automatic Script Conversion. Journal of Digital Information Management, 18(5), 396-403.
24. Devi, K., & Kumar, S. (2018). Tamil-Brahmi script character recognition system using Deep learning technique. International Journal of Pure and Applied Mathematics, 118(20), 5119-5130.
25. Gautam, A., Sharma, S., & Hazrati, M. K. (2019). Conversion of Early Tamizh Brahmi Characters into Modern Tamil Characters Using Template Matching Algorithm . International Journal of Computer Applications, 182(11), 6-11
26. Jain, A. K. (1989). Fundamentals of Digital Image Processing. Prentice Hall.
27. Pizer, S. M., et al. (1987). Adaptive Histogram Equalization and Its Variations. Computer Vision, Graphics, and Image Processing, 39(3), 355-368.
28. Sezgin, M., & Sankur, B. (2004). Survey over Image Thresholding Techniques and Quantitative Performance Evaluation. Journal of Electronic Imaging, 13(1), 146-165.
29. Iravatham Mahadevan, "Occurrence of the pulli in the Tamil-Brahmi script", Indological Essays Commemorative Volume II For Gift Siromoney, edited by Michael Lockwood, Madras Christian College, 1992, p.141.
30. Praveen, K. P., et al. (2013). Robust Document Image Binarization Technique for Degraded Document Images. International Journal of Computer Applications, 65(18), 36-42.
31. Llados, J., et al. (2009). Historical Document Binarization Based on Phase Information of Images. Pattern Recognition, 42(7), 1417-1426.
32. He, K., & Wang, J. (2015). A Spatially Adaptive Statistical Method for the Binarization of Historical Manuscripts and Degraded Document Images. Pattern Recognition Letters, 52, 108-116.
33. Guo, Wen Ming, and Yan Qin Chen. “An Effective Method for Defects Detection in Radiographic Images of Welds Based on Edge Detection and Morphology.” Applied Mechanics and Materials, vol. 290, Trans Tech Publications, Ltd., Feb. 2013, pp.71–77.