International Journal of Database Theory and Application
Vol.9, No.4 (2016), pp. 143-150
http://dx.doi.org/10.14257/ijdta.2016.9.4.13

# A Review on Automation of Ancient Epigraphical Images

Preethi.P[1] and Mamatha.H.R[2]

[1]*Assistant Professor, Dept., of CSE, K.S School of Engineering*
[2]*Professor, Dept., of ISE, PES Institute of Technology*
[1]*ppreethijain@gmail.com*
[2]*mamathahr@pes.edu*

***Abstract***

*Inscriptions are the main source of historical study available throughout the world in constricted language. Epigraphy is the study of such inscriptions and the one who reads and understands (epigraphists or epigraphers) are in extinct condition, due to lack of knowledge transfer and interest. New inscriptions are found during excavation and finding expert epigraphers these days is a real challenge. Due to this, Researchers from the digital enhancement domain are actively involved in the decipherment of inscriptions all over the world. Automation Techniques uses Optical character recognition to convert the inscriptions into intelligible language. Currently works are witnessed in the literature survey on the global languages like Greek, Italian, Japanese, Russian, Latin, Iranian and Indian languages like Kannada, Tamil, Devanagari, Brahmi, Hoysala, and Pali. This paper reviews the techniques followed in automation of epigraphical scripts.*

***Keywords****: Epigraphy, Inscriptions, Optical Character Recognizer, FRBR (functional Requirements for bibliographic records)*

## 1. Introduction

Inscriptions are the foremost resource of information which narrates about the history and it is found all over the world. Inscriptions are written using regional languages of their derivation. In India most of the inscriptions are found at historical places and Archaeological department has given eminence in preserving and translating them. Inscriptions are written on stone, metal, rock, embossing on cast metal, copper plates and also on palm trees. But, the media used to write and forms of graphemes pose a tough challenge to epigraphists due to extensive diversity style and inscriptions usually bear scratches and cracks due to aging.

The Archaeological department has taken initiative on preservation and development of epigraphy. The movement of computerization or digitization of inscriptions, removal of noise and breakages by applying image processing techniques plays a vital role in the era of computers. Preserving the epigraphy as they are is not the solution for the problem, but also need to find epigraphers who can read them.

Document analysis and recognition is to extract and classify data meaningfully from paper based documents. This includes recognition of text, characters, lines, symbols, images, handwriting, signature and graphics. Algorithms and techniques are applied to images of documents to gain a computer understandable description from pixel data. The byproduct of document image analysis is Optical Character Recognition software which recognizes characters in the scanned document. This digitization helps in globalizing the data, ease of access and permanent storage.

ISSN: 2005-4270 IJDTA

## 2. Related Work in the Field of Deciphering Epigraphy

Inscriptions are the main source of information about personages and proceedings of history. Inscriptions are given importance as they are found in every part of the world as a communication media. These inscriptions are known to belong to Indus valley civilization which ranges from C.E 3500 to 1700 B.C. The writing style and scripts used to convey cultural significance, political documents on medium like stone, leaf and metal vary from region as well as period. This section reviews the work done on deciphering epigraphy all over the world and discusses its importance to the mankind.

Greek –A history of languages and speakers with the Mycenaean civilization in the early thirteenth century. Linear B is the script used to write the oldest surviving Greece alphabetic inscriptions as shown in Table 1.1(g). Segmentation free optical character recognizer for identification of Greek manuscripts is discussed. Considering the closed cavities in the characters written which helps in the segmentation of the characters, the features are derived from the protrusions in the outer contour. Artificial neural network classifications are applied to find the characters with cavity. The characters like O,D,B,Q,R,P and 6,8,9,0 are easily identified.[16] In analyzing the historical document, a novel feature extraction on recursive subdivision of the character image based on the center of masses at different levels of granularity is developed.

Digitization has become a progress in preserving inscriptions, but preservation is not the stopping rule. Processing and recognition has evolved to infer the inscriptions of engraving periods supporting linguists and historians [2]. Thai lamma or Fakkham script inscriptions are mostly found in the temples of Thailand shown in Table 1.1(f). The proposed method creates metadata schema, stores a set of character images for reusing. As output, the FRBR model recognizes the origin and evolution of the alphabet, and also infers the engraving period using stored metadata in the database.

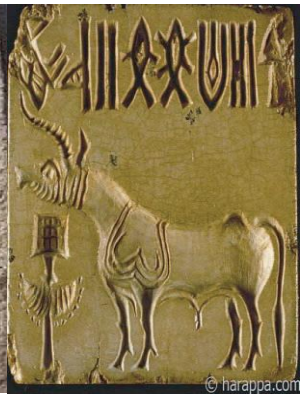
In Chinese [4-3], Ancient Chinese Tablets are the emblem of Chinese history depicted in Table 1.1(a). It has an extreme aesthetic value in building and maintaining the Chinese history. Automation of character reading is must in terms of character identification and preservation due to its high noise on the image and degradation. A method to enhance the image and segmentations are applied as a basic step during processing of inscriptions. Stroke based filters are used for the enhancement process and energy projection and propagation method for the segmentation. Dataset of 2,451 characters are formed by 60 Chinese tablets and the average precision ratio is 80%. The large cracks in the tablet are identified and fixed in the step of preprocessing. The incomplete characters can be extracted from damaged tablets.

Srilanka being the spring of diversified Historical sources, all these sources are available in the form of inscriptions. The continuous ravage by humans and nature destroys the literary sources and fails to exhibit its inestimable value. Computerizing the inscriptions has moralistic and pedagogic objective in briefing the evolution of Sinhala in south Asia. Brahmi being the oldest script became the matrix for Sinhala, Devanagari, Burmese, Khmer, Thai and Laotian scripts. A sample images is shown in Table 1.1(e). Perceptive approach in the automation of Sinhala [4], by segmenting the character using Adobe Photo shop tool and comparing the characters segmented with training data using correlation function.

**Table 1.1 Inscriptions Available in Different Parts of World**

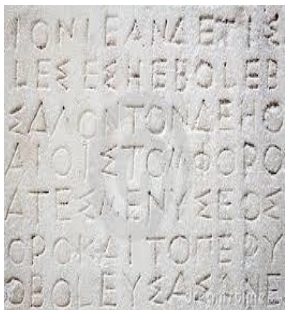
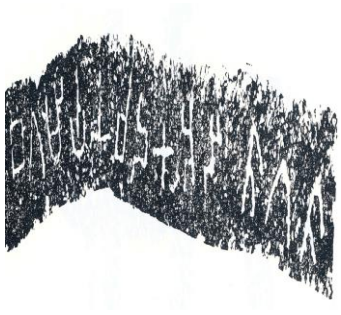





(a) Chinese Tablet [4] (b) Brahmi Script [11] (c) Harappan Inscription [5]



(d) Tamil Inscription [18] (e) Srilankan Inscription [4] (f) Thai Lamma Inscription [2]


(g) Greek Inscription [16] (h) Srilankan Script Image [17] (i) Hoysala Inscription [12]

Harappan civilization on the Indus river banks used a script which has survived on Seals, Pottery, Temple wall inscriptions, and copper. The script occurs mainly in the sign form having very minimal number of numeral and sign sets ranging up to 400. With no proper knowledge of the scripts or sign being used, the decipherment has become a tiresome process. Using bigram, trigrams, the correlation between the characters are estimated which in turn helps in analysis [5] depicted in Table 1.1(c). Bigram model consider two adjacent tokens to compute correlation and entropy, restores the signs if not understandable to corpus due to damage. N-grams application helps in identifying the token and creating the corpus which helps in decipherment of the Indus script.

Coming to India, being south Asian vast country with diverse landscape ranging from Himalaya to Indian Ocean follows multilingual scripts in different regions. The old scripts

have taken modern forms, which are now easily readable but the rulers and historians have recorded the meaningful information on the inscriptions. In the following survey, Digitization of various epigraphical scripts of different regions discussed.

Tamil nadu is famed for Dravidian style Hindu temples rich with inscriptions. Many research works are conducted to read such inscriptions and a sample is shown in Table 1.1(d). Considering the scribed character on the inscription, prediction of the era and relating the history was of major concern. The characters of inscriptions are examined for meaningful features using machine intelligence [5-7, 18-19]. The system follows image acquisition, binarization, preprocessing, feature extraction and classification. The nearest neighbor algorithm is used to segment the inscriptions into lines and characters. The meaningful data is then extracted for features like structural, syntactical, statistical, and loop features using Fourier wavelet transform. The classification uses support vector machine, genetic algorithm and transductive support vector machine. The research also includes reading on palm leaves, where the identification is based on structural, syntactical, statistical and loop features. Genetic algorithm is used to train the samples cropped but the outcome of the system is concluded to be very poor as the Tamil characters are critical and joined. Application of Boolean matrix for recognition on the palm leaves resulted in good results compared to previous one.

Devanagari is the decedent of Brahmi script has traces in northern India. The temples of Gupta dynasty bear the inscriptions written using this script as shown in Table 1.1(b). Devanagari script has taken the form of Hindi, Marathi, Pali, Nepali, Konkani, Sindi and many more in the new era [9]. The research on Devanagari script digitization include segmentation using line projection techniques, features are extracted using vertical, horizontal, right and left diagonal stroke density and K-nearest neighbor algorithm for classification [18]. Hybrid technique of combining the Features extracted for identification using intelligent classifier resulted in good optical character recognition.

The current Kannada script is evolved over centuries, and Modern readers find difficulty in interpreting old inscriptions written in different ancient scripts. To read ancient scripts the period has to be determined so has to have knowledge of which character set of ancient days to be taken for automatic reading. Prediction of the era of a given ancient script is a major component of the optical character recognition. SVM classifier is used to classify and period identification of various ancient Kannada scripts. The scripts automatically matched with the characters available for different periods using machine intelligence. This classifier is tried and tested on many ancient scripts of different languages like Tamil, Srilankan scripts.

Brahmi is the most ancient script in south Asia, it became the base script for more than 15 scripts including Kannada, Tamil, Sanskrit, Hindi, and also produced Burmese, Khemer, Thai alphabet, *etc*., as shown in Table 1.1. Evolution of the Sinhala script is studied on the basis of inscriptions found in Sri Lanka with the modern techniques of computer image processing. In [11] the study is concerned on the shape of ancient letters and fonts of early Brahmi script have been produced.

The automatic processing of the Kannada inscriptions plays a vital role in digitizing the content written on inscriptions. The writing styles followed in inscriptions are shown in Table 1.1(i) of hoysala script. Preprocessing of camera captured inscriptions is really an important and complex task due to its dreadful conditions and storage. An effort has been implied to preprocess the image by removing background noise and improving the quality of the image for better individual acuity and also to computer recognizable form. The smoothing or sharpening of the input image is done through Gaussian blur, Unsharp mask and Laplacian filter. In [11] Filters are applied based on the varying amount and nature of image quality. Augmentation to the image is achieved through suitable filter, provided with different mask sizes and parameter values which can be precise by the user and then

followed by binarization of the enhanced image using Otsu thresholding algorithm to highlight the foreground information.

Prediction of period is really an important task in knowing to which period the ancient script belongs to. The characters from the inscriptions are examined using machine intelligence and coordinate with the characters belonging to different periods in [20]. The system follows image acquisition, binarization, preprocessing, feature extraction and classification using Transductive Support Vector Machine. The experimental result shows tremendous growth in period identification of ancient Kannada script when compared with Support vector machine.

A phase –based binarization model for ancient document to improve the clarity of the document image. The work is to uplift the heritage and to know ground truth behind the ancient period. Preprocessing binarization and post processing are the major steps involved using adaptive Gaussian and median filters are considered in [8]. This process mainly works on the degraded document images.

Segmentation being the important part of optical character recognition, analysis of the features and to classify them accordingly yields a good result. Mamdani based fuzzy classifier is adopted on the statistical features extracted from the segmented character set. The features are mean, variance, standard deviation, skewness, kurtosis, entropy and GLCM features like energy, homogeneity, correlation, contrast are extracted from the Brahmi and Hoysala script images. In [12] the data extracted is matched with training data set using machine intelligence to present modern script.

## 3. Decipherment Process Model

The hierarchy of the decipherment directly depends on the input image. The epigraphical inscriptions are degraded due to natural erosion of layers, uneven color distribution, cuts and bruises. The input image undergoes three stages *i.e*., Preprocessing, Feature Extractor and Classification. Expected output is to print the recognized characters of the inscription and the complete processing is as shown in Figure 1.1.

Historical documents which are degraded and of poor quality, undergoes some process to improve the readability of the input image. For better human perception the epigraphical script images are subjected to noise removal using smoothening and sharpening filters. To highlight the foreground information which helps in detailing, binarization techniques, skew removal techniques are equipped. Segmentation algorithms are used to segregate the lines and characters and finally feed to feature extractor. Preprocessing endow with suppleness to the user in controlling the image enhancement process for the desired output.

A feature vector which is an identity of the segmented characters extracted based on the representation of the pixels, helps in the recognition rate using training data set. The features subjected to shape, size, histogram, statistical values, structural, global transformations and moment features are extracted.

Machine intelligence adopts major classifiers such as, Artificial neural network, Clustering algorithm, K-nearest neighbor, Bayes classifiers and Support vector machine are used with the training data set. Comparison between the training data set and features of the segmented character, results in recognition.

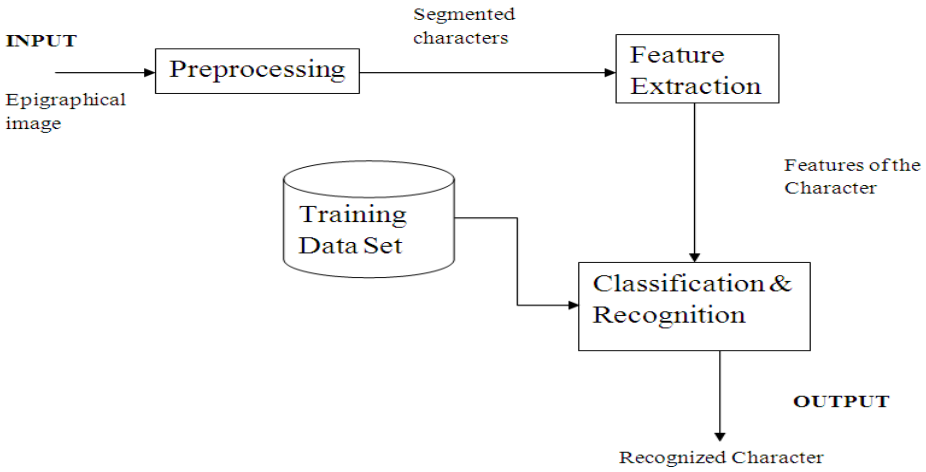



**Figure 1.1 Model of Automated Decipherment Process for Epigraphical Scripts**

## 4. Overall Review

This discussion is about, the implementation of optical character recognizer for the epigraphical script images and concise about various concepts involved and heighten further advances in the area. The recognition rate is directly proportional to the input type, texture, and its quality. The process "Decipherment of Epigraphical Images" is being implemented all over the world and literature reviews its application on Thai, Chinese, Greek, Italian, Burmese and Sri Lankan Scripts. In India the application is executed on Devanagari, Tamil, Telugu and Kannada. The review establishes a complete system that recognizes the characters of the given historical scripts. The studies imply that selection of relevant feature extraction and classification techniques plays an efficient role in the performance of the decipherment process. The research in this area is upholding new challenges day by day due to new excavations by the Department of Archeology and unavailability of the epigraphers. The existing systems are good enough to identify to which era the script belongs to and major work is required in identifying the written characters on the script in future. This material can serve as guide and update manual for the readers working in the area of optical character recognition of epigraphical scripts.

## References

[1] B. Gatos, K. Ntzios, I. Pratikakis, S. Petridis and S. J. Perantonis, "An Efficient Segmentation-Free approach to assist old greek handwritten manuscript OCR", Pattern Analysis and Applications (PAA), vol. 8, no. 4, **(2006)**, pp. 305-320.
[2] C. Techawut, P. Inkeaw, J. Chaijaruwanich and T. Hutangkura, "The Metadata Schema Design and Utility Architecture for Thai Lanna Inscription Collection", Springer International Publishing, LNCS, vol. 8279, **(2013)**, pp. 157–160.
[3] W. J. Teahan, Y. Wen, R. Mcnab and I. H. Witten, "Compression based algorithm for Chinese word Segmentation", Computational Linguistics, vol. 26, no 3, **(2006)**, pp. 375-393.
[4] D. Bandara, N. Warnajith, A. Minato and S. Ozawal, "Creation of precise alphabet fonts of early Brahmi script from photographic data of ancient Sri Lankan inscriptions", Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition, vol. 3, no. 3, **(2012)**, pp. 33-39.
[5] N. Yadav, H. Joglekar, R. P. N. Rao, M. N. Vahia, R. Adhikari and I. Mahadevan, "Statistical Analysis of the Indus Script Using n-Grams", PLoS ONE, vol. 5, no 3, **(2010)**, pp.1-15.
[6] S. V. K. Kumar and T. V. Poornima, "An Efficient Period Prediction System for Tamil Epigraphical Scripts Using Transductive Support Vector Machine", International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 9, **(2014)**, pp. 7999-8002.
[7] S. R. Kumar and V. S. Bharathi, "An Off Line Ancient Tamil Script Recognition from Temple Wall

Inscription using Fourier and Wavelet Features", European Journal of Scientific Research, ISSN 1450-216X, vol. 80, no.4, **(2012)**, pp. 457-464.
[8] A. K. N. Halambe and R. C. Thool, "Combining Multiple Feature Extraction Techniques and Classifiers for Increasing Accuracy for Devanagari OCR", International Journal of Soft Computing and Engineering, vol. 3, no. 4, **(2013)**, pp. 38-41.
[9] M. Hangarge, B. V. Dhandra, "Offline Handwritten Script Identification in Document Images", International Journal of Computer Applications, vol. 4, no. 6, **(2010)**, pp. 6-10.
[10] A. Sowmya and G. H. Kumar, "Preprocessing of camera captured inscriptions and Segmentations of handwritten kannada text", International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 5, **(2014)**, pp. 6794-6803.
[11] A. Sowmya and G. H. Kumar, "Automatic Decipherment of Ancient Indian Epigraphical Scripts - A Brief Review", International Journal of Computer Science & Emerging Technologies, vol. 2, no. 1, **(2011)**, pp. 139-144.
[12] B. Gangamma, K. S. Murthy and A. V. Singh, "Restoration of Degraded Historical Document Image", Journal of Emerging Trends in Computing and Information Sciences, vol. 3, no. 5, **(2012)**, pp. 792-798.
[13] A. Soumya and G. H. Kumar, "SVM Classifier for the prediction of era of an epigraphical script", International journal of peer to peer networks (ijp2p), vol. 2, no. 2, **( 2011)**, pp.12-22.
[14] K. S. Murthy, G. H. Kumar, P. S. Kumar and P. R. Ranganath, "Nearest neighbor clustering based approach for line and character segmentation in epigraphical script", European Journal of Scientific Research, **(2009)**.
[15] G. Vamvakas, B. Gatos and S. J. Perantonis, "A Novel Feature extraction and classification methodology for the recognition of historical documents", IEEE computer society, **(2009)**.
[16] X. Lu, Z. Tang, Y. Liu, L. Gao, T. Wang and Z. Wang, "Stroke-based Character Segmentation of Low-quality Images on Ancient Chinese Tablet", IEEE 12th International Conference on Document Analysis and Recognition, **(2013)**.
[17] A. Shaus, E. Turkel and E. Piasetzky, "Binarization of First Temple Period Inscriptions –Performance of Existing Algorithms an d a New Registration Based Scheme", International Conference on Frontiers in Handwriting Recognition, **(2012)**.
[18] P. Subashini, M. Krishnaveni and N. Sridevi, "Period Prediction System for Tamil Epigraphical Scripts Based on Support Vector Machine", Information Systems for Indian Languages - Communications in Computer and Information Science, **(2011)**.
[19] A. Sowmya and G. H. Kumar, "Recognition of ancient Kannada Epigraphs using fuzzy-based approach", International conference on contemporary computing and informatics, **(2014)**.

## Authors

**Dr. Mamatha H. R.**, received her B E degree in Computer Science and Engineering from the Kuvempu University in 1998 and M.Tech degree in Computer Networks and Engineering from the Visvesvaraya Technological University in 2006. She obtained her Doctoral Degree from Visvesvaraya Technological University. She has total 18+ years of teaching experience. Her current research interests include Pattern Recognition and Image Processing. She has published 35+ international papers. She is a life member of Indian Society for Technical Education, MIR Labs and IACSIT. She is a reviewer and session chair for various international conferences and journals. She has mentored students for various competitions at international level including the Windows Embedded Students Challenge Competition-2006 held at Microsoft Campus, Redmond, Seattle, USA. Currently she is working as Professor in the Department of Information Science and Engineering, P E S Institute of Technology.

**Mrs. Preethi P.**, received her B E degree in Computer Science and Engineering from the Visveswaraya Technological University in 2004 and M.Tech degree in Computer Science and Engineering from Visveswaraya Technological University in 2013. She has total 10+ years of experience including teaching and industry. She is a part time research scholar in the field of Pattern Recognition and Image Processing under Visveswaraya Technological University. Currently she is working as Assistant Professor in the Department of Computer Science and Engineering, K S School of Engineering and Management.