# Automatic Decipherment of Ancient Indian Epigraphical Scripts - A Brief Review

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

[1]Department of Computer Science & Engineering, R V College of Engineering, Karnataka, India,
[2] Department of Studies in Computer Science, University of Mysore, Karnataka, India,
[1]soumyaa@rvce.edu.in
[2]ghk2007@yahoo.com

***Abstract:*** The history of writing in India dates back to the 3rd millennium BC as is evident from the seals and clay pottery fragments bearing short inscriptions discovered in various parts of India. These seals and various artifacts are known to belong to the ancient civilization of Indus Valley; the mature phase of this civilization is recognized as Harappan Civilization, spanning period between c.3500 and 1700 BC. There are thousands of inscriptions found across various regions in India. Importance of inscriptions to mankind is remarkable. Although the claims of decipherment are made, no acceptable reading of the inscriptions is yet possible. The scripts of modern Indian languages have evolved over centuries. We can observe changes in characters during the phase of evolvement. Many difficulties are faced by modern readers in interpreting an ancient script. To decipher ancient script initially the era to which a given ancient script belong to has to be predicted, followed by automatic recognition of ancient script. This knowledge can be used by archaeologists and historians for further explorations.

***Keywords:*** Inscription, Epigraphy, Paleography, Document Image Analysis, Character Recognition.

## 1. Introduction

Among many ancient societies, writing held an extremely special and important role. A writing system as a set of visible or tactile signs used to represent units of language in a systematic way. It is true that many non-writing cultures often pass long poems and proses from generation to generation without any change, and writing cultures can't seem to do that. But writing was a very useful invention for complex and high-population cultures. Writing was used for record keeping to correctly count agricultural products, for keeping the calendar to plant crops at the correct time, for religious purpose (divination) and socio-political functions (reinforcing the power of the rulers). In past centuries, scientists had used writing as one of the "markers" of civilization [1].

Scripts denote the writing systems employed by the languages to represent the sounds which form the phonetic base of the language. In India, prior to invention of writing or printing papers, Palmyra leaves and birch leaves were used for writing purposes. As they could not be long lasting, engraving on rocks, pillars and plates made of copper/ gold/ silver came into practice. Epigraphy (derived from two Greek words viz., **epi** meaning **on or upon** and **graphie** meaning **to write)**, is the study of inscriptions engraved on stone or other durable materials, or cast in metal. It is the science of classifying inscriptions according to cultural context and date, elucidating them and assessing what conclusions can be deduced from them. The person studying this is called an **epigrapher or epigraphist**. Many of the inscriptions are couched in extravagant language, but when the information gained from inscriptions can be corroborated with information from other sources such as still existing monuments or ruins, inscriptions provide insight into India's dynastic history that otherwise lacks contemporary historical records [2]. The inscriptions provide valuable information about history, culture, astronomy, medicine, management, political, religious, social, economic, administrative and educational conditions that prevailed during ancient periods.

Many inscriptions do not contain enough historical details to fix their authorship conclusively. For example, from the inscriptions with the name Rajaraja, it is not very clear whether Rajaraja the First or the Second or the Third is intended. To assign dates to such inscriptions and to identify the rulers, palaeography is the main tool. **Paleography** is the study of ancient handwriting and the practice of deciphering and reading historical manuscripts. The paleographer must have the knowledge of: first, the language of the text and second, the historical usages of various styles of handwriting, common writing customs, and scribal/notarial abbreviations.

Language and Script are two different entities. The relation between a language and a script is neither 'original' nor 'fixed'.Any language can be written in any script. Having or not having 'own script' is neither a status nor any hurdle for a language. Three important varieties of scripts that were prevalent in ancient India were: Indus valley script, Brahmi Script and Kharosti script. The scripts of modern Indian languages have evolved from one of these ancient scripts over the centuries [3]. The evolution of the script is dependent on many factors: the writing material, (Stone, Copper, Palm leaf, Paper etc), writing tools, modes of writing and the background of the scribes. Important inventions with advanced technology such as those of paper, printing, typing and the fonts used in computers have had their own influences over a period of time. In India currently there are 13 Scripts and 23 official languages for communication at state level. Apart from these, there are many languages and dialects, used by a number of people.

Scripts have evolved over centuries to the present form In every century each letter was written in a particular style. During the regime of a single ruler, inscriptions may have different features for the same characters. Each inscriber had his individual style and there was a good deal of diversity in style in a given period. There was also a certain amount of regional variation [3], [4], [5]. Thus even for the experts, it is difficult to assign dates to many inscriptions, whether complete or fragmentary It is observed that many of the Indian languages evolved since 3rd century B.C. and the characters have assumed different shapes over the centuries. Modern readers find difficulties in interpreting an ancient script. The expert epigraphists decipher these scripts and translate them into the regional languages. These expert epigraphists are few and it is expected that they could become extinct in near future and also the significance of inscriptions to mankind is enormous. Hence there is a dire need for the automation of deciphering the inscriptions into an understandable form, which would help archaeologists and historians to know the cultural heritage of the civilization, so as to enable further explorations.

The image of inscriptions captured are subjected to various types of degradations like erased characters, broken characters, touching characters, non-uniform spacing of the text between lines and characters, unwanted marks engraved , add complexity in segmenting the text into lines, words or characters. Inturn poor results of segmentation, affect the classification and recognition accuracy. Nevertheless, the classification and recognition of epigraphical document image remains to be one of the most challenging problems in Pattern recognition and Image analysis.

## 2. Related work

Extensive research has been carried out on Optical Character Recognition (OCR) in the last few decades. Many commercial and accurate systems are now available for machine-printed character recognition. Unfortunately, the success obtained with the machine-printed OCR systems has not readily been transferred to the handwriting recognition arena. High accuracy OCR systems are reported for English with excellent performance in presence of printing variations and document degradation. Recognizing English characters is much simpler as there are only 26 letters and each letter is quite distinct from others compared to recognition of Indian language characters. For Indian and many other oriental languages- OCR systems are not yet able to successfully recognize printed / handwritten document images of varying scripts, quality, size, style and font. Many researchers have been working on script recognition for more than three decades but there are very few tools to identify these scripts. Compared to European languages, Indian languages pose many additional challenges. Indian languages are characterized with the properties: (i) large number of vowels, consonants, and conjuncts. (ii) Have a base character along with vowels attached, forming single character called compound character. (iii) Most scripts spread over several zones. (iv) Lack of standard test databases (ground truth data) of the Indian languages. In India also pioneering work has been done on several scripts like Bangla, Devanagari, Telugu, Tamil, Kannada, etc.

These conventional OCRs address the recognition of characters of various scripts of modern period. It is observed that many of the scripts have been evolved from the Brahmi script which is assumed to be present in 3rd century B.C. Since then, the evolution in scripts has been taken and there are many scripts today. The reported work in the field of developing a computer-based system for recognizing the text of epigraphical documents is very less. There are only few works carried out in this area on Indian script in general. Hence, it is the need of the time for the complete automation of deciphering epigraphical scripts written in olden days. The scripts of modern Indian languages have evolved to the present form over the centuries, leading to changes in characters over a period of time. Hence initially the dating of given input inscription is to be done, so as to have an idea of which character set to be applied for automatic reading of inscriptions Age identification and recognition of ancient epigraphical scripts is a problem under the genre of pattern recognition and image analysis.

Over the last few years, several of the major epigraphic corpora have begun digitization projects. The epigraphic community also hopes to create a unified database of information about all known Greek and Latin inscriptions [6]. A digitized corpus of inscriptions can include several different representations of the inscriptions: photographs of inscriptions; photographs of 'squeezes' of inscriptions, which are casts of the stone made in a flexible material like paper or latex; diplomatic transcriptions; edited texts; translations; commentaries. Many projects also find it convenient to store meta-data about the inscriptions in a database, to facilitate searching. The most useful meta-data fields include the date of the inscription, its language, the types of letter forms in use in it, where it was found, what material it is on, and its size.

Lagrange M., and Renaud, H have simulated reasoning by means of an expert system in archaeology from France using computer [7].

A few projects linking Indian epigraphy with the computer technology have been proposed and implemented with a fairly high degree of success. Siromoney, G [1975] has demonstrated the use of computer techniques for enhancement of an image and information retrieval for reading ancient Tamil inscriptions [8]. A statistical analysis of personal names in ancient Indian inscriptions has been reported by Karashima and Subbarayalu [1976]. Siromoney G., Chandrasekaran M. and Chandrasekaran R [1981] have shown that, to assign approximate date to ancient and medieval Tamil inscriptions of unknown authorship found in the southern part of India and recognition of an ancient Tamil script of the Chola period Indian Script, computer techniques may be used [8]. Chandrasekaran, R. [1982] has worked on recognition of certain ancient and modern Indian scripts using computer techniques [8]. A work on automated recognition of ancient Indian Scripts in general and ancient Brahmi script in particular by Anasuyadevi [2000] has been reported. She has proposed a fuzzy neural network for the recognition of Brahmi characters [17].

Concept of Component analysis is applied to the study of South Indian sculpture. An expert system was developed for Indian epigraphy used to assign probable dates to medieval Tamil inscriptions [18]. The program was developed in BASIC on a Genie-1 Microcomputer [1985].

K Harish Kashyap, Bansilal, P Arun Koushik [2000] have proposed a hybrid neural network architecture for age identification of ancient Kannada scripts, which focuses on classification and age identification of different characters by a hybrid model. After pre-processing the characters, the work is implemented in two phases. The first phase- identifies the base character, incorporates an Artificial Neural Network (ANN). ANN is trained by Back propagation algorithm to identify the present day base character corresponding to input character. In the second phase - for identification of age pertaining to the base character, a Probabilistic Neural Network (PNN) - a Bayesian classifier is used, taking the advantage that no training is involved prior to classification [19].

The research work carried out by Srikanta Murthy K [2005] provides novel methods for preprocessing - for removal of noises, segmentation of lines and characters, thinning and finally classification of the epigraphical documents belonging to different periods. The work aims at transformation of the epigraphical object into an image of readable form. Two preprocessing techniques for removal of noise have been proposed – the first algorithm is based on a rectangle fitting wherein the height of the character to be retained is assumed to be greater than the noisy pixel. The second algorithm employs a template to obtain the minimum majority of noisy pixels. Segmentation of lines and characters are carried out using- a Partial Eight Direction Based Line Segmentation (PEBLS) algorithm wherein horizontal Projection profile is applied to identify the base and supplementary reference lines. The second approach is based on Nearest Neighbor Clustering (NNC), which could be used even when the document is skewed. Three thinning algorithms - two-step algorithm, fully parallel thinning algorithm and rotation invariant four- step algorithm have been designed. Classification of the epigraphical document belonging to different period is carried out using - a method based on texture features, a method based on invariant moments which are invariant to rotation, translation and scaling, and for accurate estimation of the period, a neural network based approach is adopted [20].

## 3. Overview of the system for deciphering ancient scripts

The complete automation of classification and recognition of ancient epigraphical scripts involves the following steps and the workflow is as shown in Figure 1.

- The input image of the inscription may be degraded due to the presence of the broken characters, erased characters, touching characters, distortion due to fossils settled, irrelevant symbols engraved by the scribes and so on. Also the non uniform spacing between the lines and characters of epigraphical document and the skew could complicate the process of deciphering the script. Input epigraphical image has to be subjected to pre-processing stage initially. Hence suitable preprocessing techniques for removal of noise and segmentation of lines and characters are to be devised.

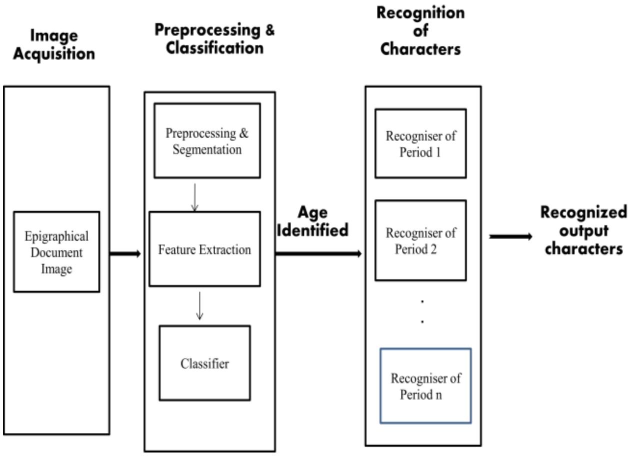
Image Acquisition
Preprocessing & Classification
Recognition of Characters
Epigraphical Document Image
Preprocessing & Segmentation
Feature Extraction
Classifier
Age Identified
Recogniser of Period 1
Recogniser of Period 2
Recogniser of Period n
Recognized output characters

**Figure1.** System architecture for Classification and Recognition of ancient epigraphical scripts

- Features have to be extracted for the segmented characters, so that the task of classifying the pattern is made easy by a formal procedure. Appropriate feature extraction methods have be devised for measuring the relevant shape information contained in a pattern.
- Characters have evolved over centuries to the current form undergoing several twins and turns. Different periods have different character set. Hence the period of epigraphical script has to be predicted so as to know which character set has to be used for supervisory reading of ancient epigraphy documents.
- Finally for automatic decipherment of ancient epigraphical scripts, recognizers are to be devised which takes the epigraphical document image of ancient script, whose period has been predicted as the input and outputs the text in a readable form.

Hence we need to seek new approaches for transforming ancient epigraphical script into recognizable form.

## 4. Conclusion

The epigraphical survey is of importance as inscriptions provide insight into history of the region during various dynasties which otherwise lacks historical records. It helps many scholars who are working in the field of history, archaeology and linguistics. Since the contribution of the inscriptions to the society is remarkable and the expert epigraphists could become extinct in future, a complete automated system with sufficient intelligence has to be developed to decipher the epigraphical documents. To sum up, the research issue addressed here is to produce a computer perceivable image from a raw epigraphical script which are the inscriptions on rocks or pillars or plate, then classification of the ancient script into respective periods and recognition of the characters which would assist historians and archeologists to know the cultural heritage of the civilization so as to enable further exploration.

## References:

[1] Writing Systems: tttp://www.ancientscripts.com/ws.html

[2] Possehl Gregory L. Indus Age: The writing System, University Pennsylvania Press, Philadelphia, (1996).

[3] A.V.Narasimha Murthy, “Kannada Lipiya Ugama Mattu Vikasa”, Kannada Adhyayana Samsthe, Mysore University, Mysore, (1968).

[4] Dr M G Manjunath, G K Devarajaswamy, “Kannada Lipi Vikasa, Jagadhguru”, Sri Madhvacharya Trust, Sri RagavendraSwamy Matt, Mantralaya.

[5] Dr. Devarakonda Reddy, “Lipiya Huttu Mattu Belavanige — Origin and Evolution of Script”, Published by Kannada Pustaka Pradhikara (Kannada Book authority), Bangalore.

[6] Electronic Textual Editing: Epigraphy [Anne Mahoney, Perseus Project & Stoa Consortium Tufts University] http://www.tei-c.org/About/Archive_new/ETE/Preview/mahoney.xml#body.1_div.3

[7] Lagrange, M., and Renaud, H, “Intelligent knowledge-based systems in archaeology: a computerized simulation of reasoning by means of an expert system”, Computers and the Humanities, Vol.19, pp. 37-49, (1985).

[8] Works on Epigraphy [online]: http://Dr Gift Siromoney/epigraphy.

[9] Siromoney, G., “Computer techniques of image enhancement in the study of Pallava Grantha inscription”, Studies in Indian Epigraphy 2, pp. 55-58, (1975).

[10] Siromoney, G., Chandrasekaran, M. and Chandrasekaran, R., “Computer methods of dating medieval Tamil inscriptions”, STAT-26/76, the Third Annual Congress of the Epigraphical Society of India at Udupi (March 1978).

[11] Siromoney, G., Chandrasekaran, M. and Chandrasekaran R, “Computer recognition of an ancient Tamil script of the Chola period”, Journal of the Epigraphical Society of India, Vol. VI, pp 18-19, (1978).

[12] Siromoney, G., Chandrasekaran, M. and Chandrasekaran R, “Computer recognition of an ancient common Indian Script”, STAT-36/78, the Symposium on the Use of Indian Languages in Computer based Information Systems, (March 1978).

[13] Siromoney, G., M. Bagavandas and S. Govindaraju, “An application of component analysis to the study of South Indian sculpture”, Computers and the Humanities 14, pp. 29-37, (1980).

[14] Siromoney, G., M. Chandrasekaran and R. Chandrasekaran, , “Computer methods of dating Tamil inscriptions” , Proceedings of the Fifth International Conference-Seminar of Tamil Studies, Madurai, India, pp. 2.7-2.13, (1981).

[15] Chandrasekaran, R., “Computer recognition of certain ancient and modern Indian script”', Ph.D. Thesis, University of Madras, (1982).

[16] Siromoney, G., M. Chandrasekaran and R. Chandrasekaran, “Computer dating of medieval inscriptions: South Indian Tamil”, Computer and the Humanities, Vol. 17, pp. 199-208, (1983).

[17] Anasuya Devi H.K, , “Automated Recognition of Ancient Indian Scripts”, Proceedings of National workshop on Computer Vision, Graphics and Image processing, WVGIP, pp 216-219, (2002).

[18] Gift Siromoney, Chandrasekaran R. and Suresh D., “Developing an expert system for Indian epigraphy” at the Kibble Center for Statistical Computing at the Department of Statistics, Madras Christian College, (1985).

[19] K Harish Kashyap, Bansilal, P Arun Koushik, “Hybrid Neural Network Architecture for Age Identification of Ancient Kannnada Scripts”, Proceedings of 2003 International Symposium on Circuits and Technology (ISCAS 2003), Vol 5, Pg V661- V664, (2003).

[20] Srikanta Murthy.K, “Transformation Of Epigraphical Objects Into Machine Recognizable Image Patterns”, Ph.D Thesis, University Of Mysore, Mysore, (December 2005).

[21] Works on OCR for Indian Languages [online] http:// Indira Gandhi National Centre for the Arts (IGNCA's) Southern Regional Centre, Bangalore.

[22] S Pletschacher, J Hu and A Antonacopoulos, “A New framework for Recognition of Heavily Degraded Characters in Historical Typewritten Documents Based on Semi-Supervised Clustering”,10th

International Conference on Document Analysis and Recognition, (2009).

[23] Sheikh Faisal Rashid, Faisal Shafait and Thomas M. Breuel, "Connected Component level Multiscript Identification from Ancient Document Images", (2010).

[24] D Dayalan, "Computer Application in Indian Epigraphy", Bharatiya Kala Prakashan publication, (2005).

## Author Biographies

**Soumya A,** Assistant Professor, Department of Computer Science and Engineering, R.V College of Engineering, Bangalore. She has obtained M.S degree in Computer Cognition Technology, University of Mysore, Mysore and B.E in Computer Science and Engineering, Bangalore University, Bangalore. Her areas of research are Artificial Intelligence, Soft Computing, Pattern Recognition and Image Processing. She has guided several under graduate projects and 7 post graduateprojects.

**Dr G Hemantha Kumar,** Professor & Chairman, Department of studies in Computer Science, University of Mysore, Mysore. Serving as Course -Coordinator of Chinese B.Tech Programme, University of Mysore, Mysore. He was awarded Ph.D. for his thesis titled: "On Automation of Text Production from Pitman Shorthand Notes" from University of Mysore, Mysore. His areas of research are: Image Processing, Pattern Recognition, Numerical Techniques, and Bio-Metric. He has to his credits 31 publications in International / National Journals and 46 publications in International / National Conferences/Workshops. He has guided several PhD candidates and is presently guiding 5 PhD candidates.