International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 11, No. 1 (2018), pp.37-44
http//dx.doi.org/10.14257/ijsip.2018.11.1.03

A Systematic Approach in Transforming Inscriptions into Modern Text – Review

Preethi Padmaprabha[1] and Mamatha Hosalli Ramappa[2]

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

Abstract

Inscriptions are the part of history all over the world. The information on the inscriptions are very important to the mankind, but understanding and transforming it is need of today as epigraphists are in extinct condition. Many researches are voted for the restoration, segmentation, and classification of such inscriptions and it is still in progress. In making an attempt, a transformation system is proposed which Normalizes, Segments and classifies the characters on the inscriptions to modern readable characters.

***Keywords**: Inscriptions; Estampages; Kurtosis; Mamdani Fuzzy Classifier; Artificial Neural network; Decision tree; Bayesian Network; Nearest Neighbor clustering; Transductive Support Vector Machine*

1. Introduction

Communication plays an important role in exchanging information. We use email, phone Calls, messages, and letters to pass on the information and make sure that it reaches in an effective way to the recipient. In case of exchanging information on a large scale to a wide range of people - mass communication via Television, Satellite and Telephone play foremost role. Centuries ago, Kings used to mass communicate their achievements, donations and important acumen via inscriptions on copper plates, palm leaves and also on the rock bed. The scripting language used mainly depends on the region and century they belong to. Now in current scenario, reading and understanding such inscriptions by common man is almost impossible. To understand the gist associated with the inscriptions, we need an expert epigraphist who is seldom found. Across the world, automation of inscriptions are in high demand to know the meaning and also to convert the inscriptions to modern readable languages for the future mankind.

The transformation of the inscription has to undergo measures of digital image processing. The eminence of the inscriptions is deteriorated due to ageing, constant effects of sunlight, rain, storage place and material used. To enhance its features and to remove outliers the filtering techniques are applied as a preprocessing technique. Binarization method is applied on the input image to highlight the foreground statistics. Segmentation is a process of dividing the input image into meaningful lines and then into characters for the automation of the optical character recognizer. Finding the features of the characters, extracting them to store as training and testing data set helps in finding and converting the input - inscriptional image into modern readable characters.

In spite of several works carried out on the field of document analysis and transformation, continuous research is in progress all over the world on inscriptions available locally as they can easily understand the language. The list goes beyond with the names like, Greece, Italy, Thailand, Srilanka, Indonesia, America and India. Majority of

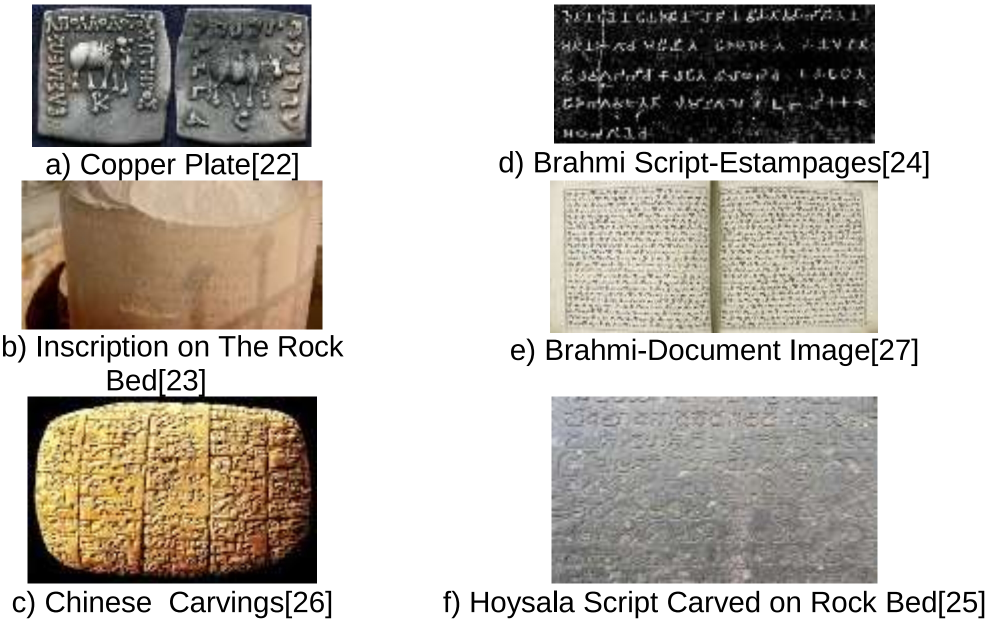
Received (October 8, 2017), Review Result (December 1, 2017), Accepted (December 8, 2017)

the transformation is into English (Global Language) but India being the multilingual country with 200+ regional languages has manuscripts available in local scripting styles and need conversion also to local languages.

This paper works ahead for the strategies and techniques which are implemented in the area of automation of epigraphical images, the challenges associated in conversion into modern language and a design based approach to face the challenges in automation. Section 2 and 3 discuss about the challenges and the work carried out in transforming old script in brief. Section 4 elaborates the new approach of designing an effective model in converting old script into modern script. Section 5 concludes with a concise outline.

2. Challenges in the Field of Inscription Analysis

The automation of inscriptions entails the input data collected from different sources, devices and Estampages. Diverse data obtained by these procedures are directly subjected to climatic condition, photo effect and noise due the coloring agent used in Estampages which introduces challenges for the automation procedure. Quality of the input image depends on the device in which, the image has taken and if the same object has been captured in diverse devices then a standard normalization is must before completion. Due to ageing and constant degradation owing to natural calamities, the actual impressions of the script on the inscriptions are degraded and faded. The excavation, mining and renovation activity near the temples and historical places can also form cuts and bruises on the inscriptions. The climatic condition on the day of photo shoot, material used for Estampages, extra or flash light used to brighten the image may also harm and introduce unwanted noise in the input picture.



a) Copper Plate[22]

d) Brahmi Script-Estampages[24]

b) Inscription on The Rock Bed[23]

e) Brahmi-Document Image[27]

c) Chinese Carvings[26]

f) Hoysala Script Carved on Rock Bed[25]

Figure 2.1. [a, b, c, d, e, f] Depicts the Different Types of Available Input and Languages for the Research

Challenges posed for Automation are:

- Removal of background noise and outliers using filters and to construct strong foreground details using masking and binarization techniques.
- Filling of cuts and bruises to maintain connectivity in the image, facilitate segmentation.
- Segmenting the scripts into meaningful objects like lines, words and then characters. Preparation of training and testing samples from the input script image.

- The detailed features describing the specific structures are Extracted, Normalized and useful patterns are identified for the subsequent process.
- Classification plays a major role in identifying and matching the segmented characters into modern readable script format.

Keeping confront in mind, and to discuss the techniques used to attain these challenges, the literature survey is discussed.

3. Literature Survey

Automation of degraded historical documents is of great interest around the world. Many works are carried out in enhancing and deciphering such inscriptions. Below are the epigrammatic studies of the techniques used by the researchers in meeting the endeavor. The survey has been made on the three major areas in processing epigraphical scripts which include preprocessing, feature extraction and classification.

Undesirable statistics on the epigraphical images decreases the recognition accuracy. To highlight the foreground and to filter out the noise associated, preprocessing techniques of image processing are applied. The very popular OTSU threshold binarization method has become a global algorithm in refining the background noises by setting threshold value nominated automatically by discriminate standard. To remove the uneven illumination effects caused by flash light, climatic condition, binarization techniques using local minimum and maximum process are adopted. Iterative global threshold is another method of binarization to remove the cuts, bruises and other degradations on the image iteratively considering foreground and back ground information on each step. Majorly concentrates on uneven patterns of the image. The edges of the characters in the historical degraded images are blunt and disturbed due to its storage. Application of masking and filtering techniques can be helpful in restoring and preparing strong edge. Adaptive filters and linear unsharp masking can highlight the edges and increases the accuracy rate. In [1-5] authors discuss about the various binarization techniques and masking methods as initial step of preprocessing.

Segmentation is a process of deriving interesting and meaningful objects in the image. It plays a major role in the automation of the degraded scripts. Due to connectivity in the writing style of the characters, cuts and bruises, writing material used, skew in the writing style and depth of the carving, the accuracy of the segmentation varies for each type of input images. The discussion below summarizes the segmentation algorithms applied on these input images.

The idea of using histogram projections on binarized image to segment lines, words and characters on handwritten text document could be really great for the epigraphical images. The binary image undergoes filtering, erosion and dilation to have clear text without any cuts and bruises. The pixel which describes the characters on the image is certainly counted throughout the projections along horizontal and vertical axis of the image. A projection profile along the horizontal axis is a histogram which gives the count of ONE pixel along the parallel lines in a one dimensional array. Similar projection profiles are applied on vertical axis to get the column sum. The minima in the horizontal projection profile separates lines in the given input and the vertical projection profile applied on input lines delivers words and characters vertically. Valleys of projection profile methods to compute row wise sum of black pixels, to denote boundary line, to figure column wise sum of black pixels are applied in [6] to segment the given document into lines, words and characters.

The historical documents have connected components and tracing and segmenting is a real challenge. The drop fall algorithm can cut the connected component by tracing image pixel by pixel, just to find boundary pixel (white pixel) which acts as a separator to start segmentation. The procedure is to simulate falling drop on the binarized image and the cut tracing is defined based on the information of the neighboring pixels and next to neighbor pixels considering five adjacent pixels and the current pixel and the six probable cases the drop moves to the next possible location. The traditional drop fall has certain limitation with respect to the segmentation process which may cause wrong split due to case one where it can even go to right or left but it moves to down and if all the adjacent pixels are black then selection won’t happen and drop will fall on the ground. To expand the concept of drop fall, inertia and weight are introduced. When the neighboring pixels are black the drop fall gets inertia due to its weight, moves along the smooth surface and cuts the connected component, makes the segmentation possible and is described in [7].

Finding the base line and touching edges for the segmentation is a tedious process, but the water reservoir method discover them by finding the cutting points on the image. The cutting points are concentrated on the base of the characters as water at the bottom of the reservoir, and the space attributes like center of gravity and height helps in finding touching edges which made segmentation easy [8]. The water stores on certain regions of a component forming a reservoir, when it is poured on top of it. Assigning some threshold for those reservoirs whose heights are greater forms a mark stone. The center of gravity attribute place a major role in estimating the base line and touching position. For the segmentation process morphological thinning operation is applied to get the best line features and confidence value.

The nearest neighbor algorithm traces line and character segmentation for the given input. The procedure starts from the first black pixel and iteratively traces the entire black pixel from left to right of the image. The centroid of the characters is logged to find the distance between the characters, words and line. [7] Nearest neighbor algorithm works well on the disconnected characters on the image and gives best results and it fails to work on connected component. The algorithm scans the given input image from the left corner. When it encounters the first black pixel, it identifies the complete character through connected component. This character is segmented and placed at different location. The centroid of the character is computed. Similarly the second character is identified and the centroid is computed. The Euclidean distance between the cancroids is computed to know whether the character belongs to the same line or next line. This is determined based on the threshold which is based on the assumption that the space between the text lines is greater than between the characters. In this way, the text lines and characters are segmented which could be used for the classification process.

The segmentation uses a spectral partitioning approach [9] that tries to maximize the proximities across them. This class of algorithms computes a pair wise similarity matrix built over every pair of components (pixels) from the image. The idea is to find an indicator vector from the spectrum of this matrix which can be the threshold to partition the set. Given a document image with number of connected components, algorithm finds coherent partitions using Eigen values. The results of nearest neighbor algorithm were found appreciable on Epigraphical images.

After segmentation the output derived is fed into extraction process to obtain the silent features of the characters and words.

Feature extraction may be the intermediate step in transforming historical documents into readable form. However the parameters are derived at the beginning of the process to create training data set and test data set. Some of the features applied on historical documents and their results are discussed. The fuzzy features such as Mean, Kurtosis, Variance, Standard deviation, GLCM features are extracted to predict the class labels for the segmented characters using Mamdani Fuzzy Classifier in [11]. To increase the accuracy in classification Fourier wavelet features to find the local variation of the characters are used. The features like centroid, wavelet radius, polar coordinates and spectrum are computed by 1-d Fourier transform techniques [12]. The zonal statistical features [13] like number of horizontal/vertical/diagonal lines, total number of intersection points and length of horizontal/vertical/diagonal lines are extracted from the input image. Single class of features is unfeasible to predict the classification labels in case of historical degraded images. A combination of two or more features from different class is selected to form a hybrid features which leaves strong impact on the classification and transformation. Statistical, structural and moment features are combined together to form a hybrid feature vector in [14] to derive a better class labels.

Classification being the final stage of transforming historical documents uses the data from Feature extraction. The methodologies include Bayesian networks, Decision trees, Artificial Neural networks, Support vector machine and Transductive support vector machine.

The support vector machine is the data points that lie closest to the decision surface and most difficult to classify. Support Vector Machines maximize the margin around the separating hyper plane which classifies the objects clearly, still the decision from the SVM are having erroneous prediction. Hence a study on semi supervised method like Transductive support vector machine to maximize the separating margin which helps in better classification. The iterative TSVM uses unlabeled patters to draw a separation line which is considerably easier than labeled patterns. In [16] prediction of Tamil scripts using TSVM has proven accuracy than the other classification Support Vector Machine which uses labeled patterns. A brief description of the transductive learning setting can be formalized as follows. The input documents are represented by a scaled and normalized histogram of words in vector space. The learning algorithm observes one input element as training set and remaining set as test set. TSVM access both labeled training vectors and also unlabeled test set to predict the result. The erroneous predictions are reduced in terms of TSVM than SVM [17].

The two main functions of the genetic algorithm a) chromosome generation function for each character and b) chromosome evaluation functions are used to classify the characters for recognition. Combining the feature extracted from the previous step of image processing chromosomes for each character is generated. Based on the zeros and ones in the bit wise representation of characters, fitness value is calculated to match and evaluate in the next section of genetic algorithm. This method is applied and tested on Tamil epigraphical scripts and the results were well appreciated in [18].

In identifying characters on palm leaves we must exercise extreme care in extracting the features as each character inscribed on the same source may have varying qualities. The 3-D features like depth of the pixel *i.e.* is the amount of pressure applied by the scriber while writing and other statistical features are merged to derive feature extraction. The decision tree which generates a set of rules at each hierarchy of the tree is considered for classification in [17]. This classification method builds a concise model with two attributes numerical and categorical attributes to derive the class labels. Without much computation a clear prediction comes out as an effect of derived rules. Decision trees are assumed to be expensive and work well on smaller set of data attributes. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called “root” that has no incoming edges. All other nodes have exactly one incoming edge. Internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. Each leaf is assigned to one class representing the most appropriate target value. Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along with the path, its branches are labeled with its corresponding values.

Classifying the historical documents using hierarchical method uses granularity features with three steps to create training and one step for testing the inputs. Initially confusion matrix is created using K-fold cross validation process for each level using features extracted. Highest recognition rate is considered as threshold, this threshold is compared with confusion matrix and the lower range is selected for next iteration. Classifier is trained at next level and the patterns are tested for the new class labels and the classifier decides the recognition result in [19].

Literature survey elaborated the current work in progress and the work carried out in the domain of automation of ancient script. Some of achievements are segmenting the characters, finding the era of the inscriptions, applying image processing techniques to improve the appearance for the readability of the inscriptions. Identifying the characters, filling the cuts and bruises, transforming the characters into new writing style is in process.

4. Importance of Automation and System Design

The archeological department has taken initiative in the conservation and development of epigraphy. The crusade of computerization or digitization of inscriptions for the removal of noise and breakages, by applying image processing practices plays a vital role in the era of computers. New excavations are found by the archeological department and hard-hitting to read by common people and preserving them as they are is not the solution for the problem.

Over the decades many research have been conducted to restore and transform the historical documents and a list of them are specified above in the literature survey, which actually conclude the importance of continuing the research for the next level of competence. The semantic gap in the identification of ancient characters is growing day by day, as finding epigraphists are difficult.

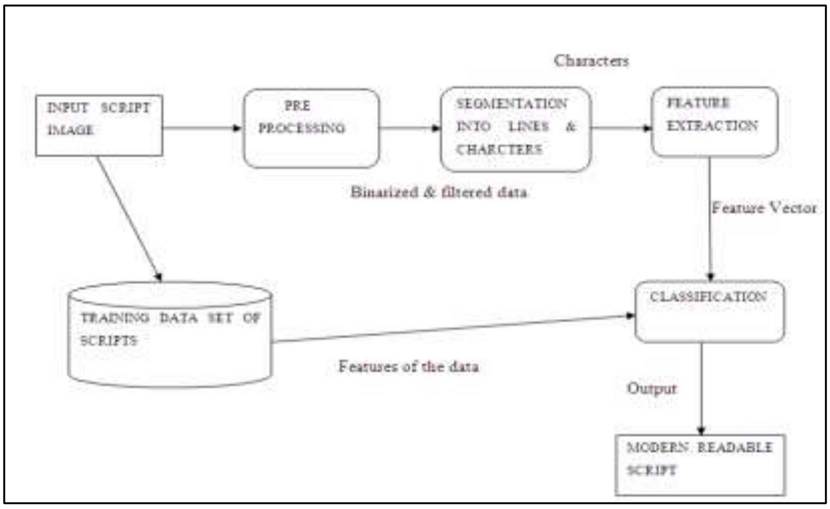


Figure 4.1. Depicts the Stages of Transformations

A system which automates the identification of the characters written, scribed on the inscriptional images to modern readable characters is the motto behind the work. The methodology of the proposed work will include preprocessing techniques such as binarization, noise removal, filtering and other preprocessing techniques to improve the quality of the images, Figure 4.1 explain the stages of transformation. The methodology will primarily target at the development of various analytical model for mitigating segmentation, feature extraction and classification. Segmentation being the important step in classifying the characters, a combination of two or more segmentation techniques is combined to create training and test set. Local features are combined with global features to form a strong feature extractor. Classification and identification of characters on the inscriptional image is the last step in interpreting the result.

5. Overall Review

The above Literature survey has detailed explanation of the importance of inscriptions to the mankind and the design espoused to automate the inscriptions. To have a meaningful impact of inscriptions on history, deliberation of the proper input is necessary. Research issues addressed here are to develop a systematic approach to decipher and to recognize the characters would assist historians and archeologists in prospect explorations. The paper depicts the importance of automation in reading and transformation of inscriptions.

References

[1] A. Sowmya and G. Hemanth Kumar, "Enhacement and Segmentation of Historical Records", Airccj, wireilla Publications, vol. 5, **(2015)**, pp. 95-113.

[2] M. R. Gupta, N. P. Jacobson and E. K. Garcia, "Binarization and Image Preprocessing for searching Historical Documents", University of Washington, Seattle, Washington 98195, **(2006)**.

[3] B. Su, S. Lu and C. Lim Tan, "Binarization of Historical Document images Using the local Maximum and Minimum", **(2010)**, pp.9-11.

[4] N. Venkata Rao, A.V. Srinivasa Rao, S. Balaji and L. Pratap Reddy, "Cleaning of Ancient Document Images Using Modified Iterative Global Threshold", IJCSI international Journal of Computer ScienceIssues, ISSN (online): vol.8, issue 6, no. 2, **(2011)**, pp.1694-0814.

[5] M. Trentacoste, "Unsharp Masking, Counter shading and Halos: Enhancements or Artifacts", 2012, The Euro graphics Association and Blackwell publishing Ltd., **(2012)**.

[6] L. Likforman-Sulem and A. Zahour,"Text line segmentation of Historical Documents: Survey", International Jounal on Document analysis and Recognition, Springer, **(2006)**.

[7] X. Wang, K. Zheng and J. Guo, "Inertial and Big Drop Fall Algorithm", International Journal of Information Technology, vol. 12 no.4, **(2006)**.

[8] A. Soumya and G. Hemantha Kumar, "Enhancement and Segmentation of Historical Records", Inetrnational Journal for Computer Science & Information Technology, **(2015)**, pp. 95–113.

[9] S. Khan. "Character Segmentation Heurisrtics for Check Amount Verification", Master Thesis,Massachusetts Institute of Technology, **(1998)**.

[10] N. Anupama, Ch. Rupa and E. Sreenivasa Reddy, "Character Segmentation for Telugu Image Document using Multiple Histogram Projections", Global Journal of Computer Science and Technology Graphics & Vision, vol. 13, **(2013)**.

[11] A. V. Srinivasa Rao, D. R. Sandeep, V. B. Sandeep and S. D. Jaya, "Segmentation of Touching Hand written Telugu Characters by using Drop Fall Algorithm", International Journal of Computers & Technology, vol. 3, no. 2, **(2012)**.

[12] C. Praveen kumar and Y. C Kiran, "Kannada Handwritten Character Segmentation using Water Reservoir Method", International Journal of Systems, Algorithms & Application, **(2012)**.

[13] A. Sowmya and G. Hemanth Kumar, "Recognition of ancient Kannada Epigraphs using fuzzy-based approach", International conference on contemporary computing and informatics, **(2014)**.

[14] S. Raja Kumar and V. Subbiah 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.

[15] A. Sowmya and G. Hemanth Kumar, "Recognition of historical record using gabor and zonal features", Signal & Image Processing: An International Journal (SIPIJ), vol.6, no.4, **(2015)**, pp.57-70.

[16] A. N Holambe and R. C Thool, "Combining Multiple Feature Extraction Technique and Classifiers for Increasing Accuracy for Devanagari scripts", IJSCE, vol. 3, issue 4, **(2013)**.

[17] S. Venkata Krishna 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, issue 9, **(2014)**.

[18] B. Gatos, K. Ntzios, I. Pratikakis, S. Petridis and S.J. Perantonis, "An Efficient Segmentation-Free approach to assist old greek handwritten manuscript OCR", Springer Verlag limited, **(2005)**, pp. 304-320.

[19] E. K. Vellingiriraj and P. Balasubramanie, "Recognition of Ancient Tamil Handwritten Characters in Palm Manuscripts Using Genetic Algorithm", International Journal of Scientific Engineering and Technology, vol. 2, issue 5, **(2013)**, pp. 342-346.

[20] https://www.ancient-asia-journal.com/articles/10.5334/aa.12317

[21] https://en.wikipedia.org/wiki/Brahmi_script#/media/File:Brahmi_script_on_Ashoka_Pillar,_Sarnath.jpg.

[22] https://bharatabharati.wordpress.com/2013/12/27/will-shrine-discovery-end-buddhas-birth-date-dispute-rediff/.

[23] http://karnatakaitihasaacademy.org.in/?p=2061.

[24] http://eduscapes.com/history/beginnings/3000bce.html.

[25] https://www.pinterest.com/pin/44262008810059167/.

Authors

Mamatha H R, she 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 36+ 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, and Seattle, USA. Currently she is working as Professor in the Department of Computer Science and Engineering, P E S Institute of Technology.

Preethi. P, she 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 11+ 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.