Translation of Brahmi Inscriptions to Sinhala using Natural Language Processing

Nethmi Hettiarachchi
Department of Information Technology
Sri Lanka International Buddhist Academy
Pallekele, Sri Lanka

Indrachapa Pathiraja
Department of Information Technology
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka

Dilum Madawala
Department of Information Technology
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka

ABSTRACT
The recognition and translation of handwritten characters that were written without constraints is challenging. In this particular domain of interest "ancient inscriptions", character recognition and translation is more curtail due to the wide variety of endemic writing styles. Research project named "Brahmi to Sinhala Translator", is a result of an idea of coming up with a software solution to translate the ancient inscriptions written in "Mula Brahmi" language to Sinhala. According to the facts that found through the researchers, it is clear that, there is a huge gap between the Archeology and modern technology. Therefor archeologists still follow manual procedures to get their work done. "Brahmi to Sinhala Translator" is based on image processing, character recognition, text mining and natural language processing [1]. This project has been divided into four different functionalities. Removing the noise of the scanned image of the stain paper, recognition of letter patterns of Brahmi language, identification the corresponding Sinhala letter to the Brahmi letter and performing word and sentence break down and represent Brahmi Script in Sinhala. Since the proposed solution is a step by step approach, it will be able to provide a user friendly environment yet robust and accurate. Therefor this will be a great innovation for not only the field of archeology but also for the information technology.

Keywords
Image Processing; Wavelet Decomposition; Character Recognition; Text Mining; Natural Language Processing

1. INTRODUCTION
Automatic translation is a main way which will provides one of the most natural ways for people to interact with computers, without any extra skills such as typing. People have been doing research in this area for more than three decades. Different approaches, such as statistical, syntactic and structural, and neural network approaches, have been proposed. Language wise, the shapes of the characters are warring. Normally characters are consisting of line segments and curves. In order to recognize a character, it should be 1st find out the structural relationships between the elements which make up the character of the feature value of the most important features of the letter. And also language wise grammar patterns are also varying. Different languages have different grammar rules. In order to do successful translations, 1st have to identify and define correct grammar patterns of the language. To make handwritten character recognition and translation feasible, speed, accuracy, and flexibility should be highly considered. High speed and accuracy are always the key characteristics of any system, while flexibility is also an important concept due to variations of the handwriting. In other words, an ideal handwriting recognizer should be able to quickly and accurately recognize a reasonably wide range of hand- writing input.

In this paper author will propose a simple, yet robust structural approach for translating "Brahmi Language" inscription in to Sinhala. Authors approach is to achieve reasonable speed, fairy high accuracy and sufficient tolerance to variations. At the same time, it maintains a high degree of reusability and hence facilitates extensibility. Since the proposed solution is a step by step approach, it will be able to provide a user friendly environment to user. Another useful feature of the propose solution is, no one has successfully used above mentioned technologies to implement a translation software yet. Therefor this will be a great innovation for not only the field of archeology but also for the information technology environment.

Currently, archaeologists do their processes manually to read and find out the real meaning of stone inscriptions. It is not an effective and accurate way. The manual process is very time consuming. Archaeologists in Sri Lankan archaeology department are fed up with these manual tasks to fulfill their business process. Sometimes they had failed with these manual processes and at finally they had not earned any business value to them. And also no any computerized, effective way to complete all manual things efficiently and find out the accurate output. Some researchers had tried out to define computerized system. When they continue their process, they had failed at some steps. In recent past, substantial amount of research effort has been applied for character recognition in various languages such as English, Tamil, Bengali, and Sinhala, unfortunately there are no published similar systems developed to recognize the Brahmi characters in ancient Sri Lankan inscriptions. When consider about the manual process, first of all archaeologists should copy the inscription which is on the stone, to the stain paper. They had used black ink to do this task. Then they should scan the stain paper and read the scanned paper spending huge effort. Therefore, the main purpose of the project carried out by our group is to solve the above problem in a user-friendly and efficient way.

The proposed system has special condition where there is considerable amount of noise in the given input. Therefor some of the noisy patches on the stones are identified as characters and produced incorrect result. Image processing, character recognition, text mining together with natural language processing technologies are used to develop the proposed system.

Early researches on languages translator did not linked with image processing and preprocessing techniques. And most

probably, research efforts had been applied for character recognition of various languages such as Tamil, Arabic,

Bengali and so on others without Brahmi. Brahmi Language is very primitive language which earns huge ancient value towards Sri Lanka. When concern on history of Sri Lankans, more than 90% stone inscriptions are written using Brahmi language. No any existing system to preprocess scanned images of Brahmi stone inscriptions using image processing techniques. Researchers had not tried out to process on images of stone inscriptions and separate the existing Brahmi letters in the inscription.

RESEARCH METHODOLOGY

2. SYSTEM OVERVIEW

In order to make the proposed system, a reality, 1st will have to take the scan images of the stain paper into the system. Then system should clear all the noises of the image. After clearing them system should identify the patterns correctly with the corresponding Sinhala letter. Identifying it will not be enough. Then system will have to break the content into separate words and sentences. After all system will translate whole content into meaningful Sinhala sentences.

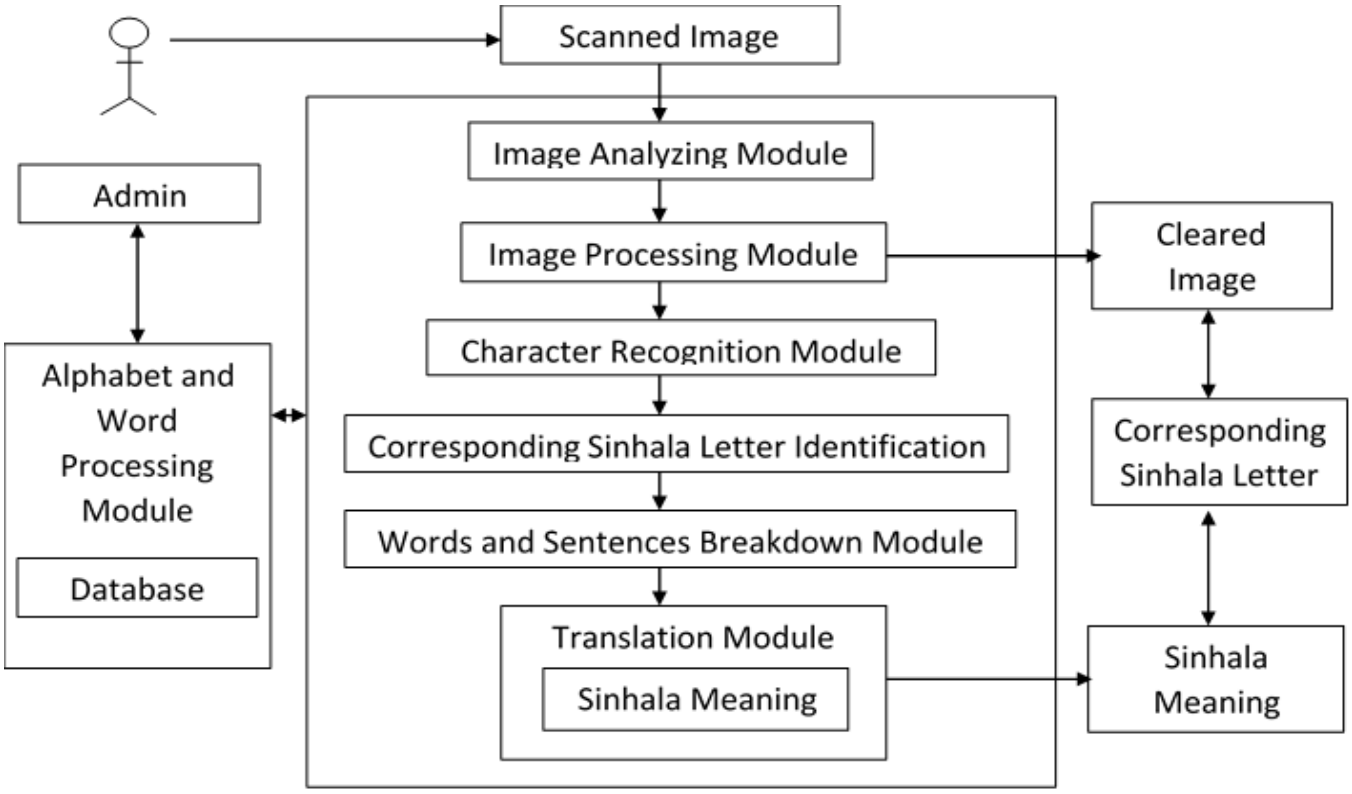

Figure 1: High Level Diagram

3. SYSTEM IMPLEMENTATION

Below given are the main modules of the system.

1.1 Module 01: Image Processing and Image Enhancing Module

Image processing and image enhancing module is where the noise remove of the scanned image and highlight the existing characters of the scanned image. User can simply enter the scanned image as the input. There are several image processing and enhancing technologies behind the module.

1. Invert the colors
2. Background removing
3. Skeletonisation
4. Noise removing

As the first input, scanned images are entered to the system. There are several types of scanned images according to noise level of the images. All of the scanned images are black and white images.

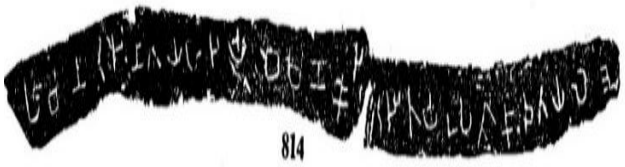

Figure 2: Scanned Image

In this phase, pixel density of letters and noise are mainly concerned. Threshold value is defined by considering the noise level of each and every scanned image. Final output of this module is inscription image where the noise is minimized and background is removed.

1.1.1 Invert the colors

Inversion is done as a support to character recognition

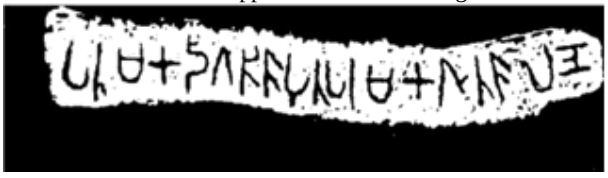

Figure 3: Negative Image

1.1.2 Background removing process

To remove the unnecessary black background of the image, defined algorithm is used. The system will consider each black pixel in each horizontal raw around the white patch in the image as well as each black pixel in each vertical column. And also it should clearly mention, the explained process is stop when it moves to any white pixel. System is scanned for white pixels left to right, right to left, top to bottom and also bottom to top. All found black pixels are converting to white. Traversing technique is applied to move each and every black pixel.

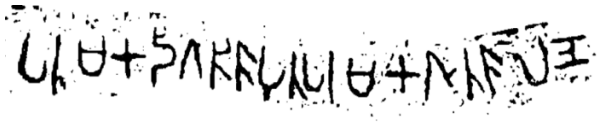

Figure 4: Background Removed Image

1.1.3 Skeletonization

The background removed image is then sent to the skeletonization process. The basically skeletonization is used to reduce the thickness (pixel density) of the image. Skeletonization is applied for both letters and noise in the image, so it will effect to the pixel density of the noise also. In this module, all defined algorithms are traversing each and every black pixel one by one, so it is better to reduce the pixel density to proper applications of defined algorithms.

Figure 5: Skeletonized Image

Skeletonization is applied to the background removed image, without damaging to the existing letters.

1.1.4 Noise removing

Noise removing process is applied for skeletonized image. There was a problem with applying predefined filters such as Min filter, Max filter and also Median filter. Though input image is not colored image, applying min & max filters are not 100% effective. Mainly, when considering features of the scanned images of stone inscriptions, there are several types of images when categorizing on noise level such as law, medium and high. So the predefined noise filters are suitable for the some of the input images and they were not suit for several input images. Categorizing images according to noise level and find out the suitable noise filter for relevant image category was more time consuming process.

An algorithm had been defined to detect and identify noise and letters separately by comparing pixel density. Particular value had defined as threshold value as the main decision point in noise removing phase. Several sample input images are tested under all requirements and threshold value is decided. Several high noisy level images, law noisy level images and also medium noisy level images were in the sample set. So the defined threshold value is suitable, if input image is any type of image. System scanned each and every black pixel in the skeletonized image and compare with the defined threshold value. Traversing technique is used to each and every movements of scanning of black pixels. When scanning black pixels, if system meets one to one connected black pixels it checks out all connected black pixels and count total number of black pixels and compare with defined threshold value. If system meets isolated black pixel, it is also checked out by the system and black pixel count is compared with the threshold value. Clock wise scanning process is done around one black pixel and it is not a problem if there is isolated black pixel or connected black pixels. When comparing total pixel count with defined threshold value, if total pixel count is less than or equal to threshold value (pixel count <= threshold value), system detects it as noise. If total pixel count is greater than defined threshold value (pixel count > threshold value), system detects it as letter. Then identified noise is converted to white color and identified letters are kept same as input image.

Figure 6: Noise Removed Image

1.2 Module 02: Character Recognition Module

Inside the character recognition module, main consideration is variation of the existing letters in the noise removed image. So Wavelet method will be the most appropriate method. Since Wavelet method has a mathematical backbone, researchers can generate a series of values for each and every variant of letter and do the recognition based on that. Brahmi language has an alphabet with 38 basic characters.

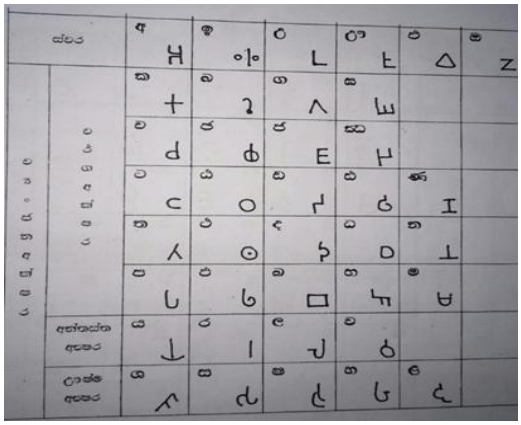

Figure 7: Brahmi Alphabet

So the possible set of variations for all the 38 characters have been identified in the module.

Example of possible variations of a letter.

Figure 8: Variation of Letters-1

Sub modules inside the character recognition are below.

1.2.1 Segmentation of letter string (input) in to single letters

In order to apply Wavelet method, first of all input letter string should be segmented in to single letters. Then can

easily apply Wavelet for each and every character to generate values and can extract features easily. [1]

1.2.2 Feature extraction of letters

Under this part extracting features from the character image will be done. Basically feature extraction will be done in three ways such as Horizontally, Vertically, and Diagonally. From this part it will generate three values. By referring to those values system can do the recognition.

Figure 9: Character Segmentation

Ex: Horizontal Value

Vertical value

Diagonal value

Every variant of same letter will provide different values in feature extracting. So value range is considered when doing the character identification.

1.2.3 Plotting

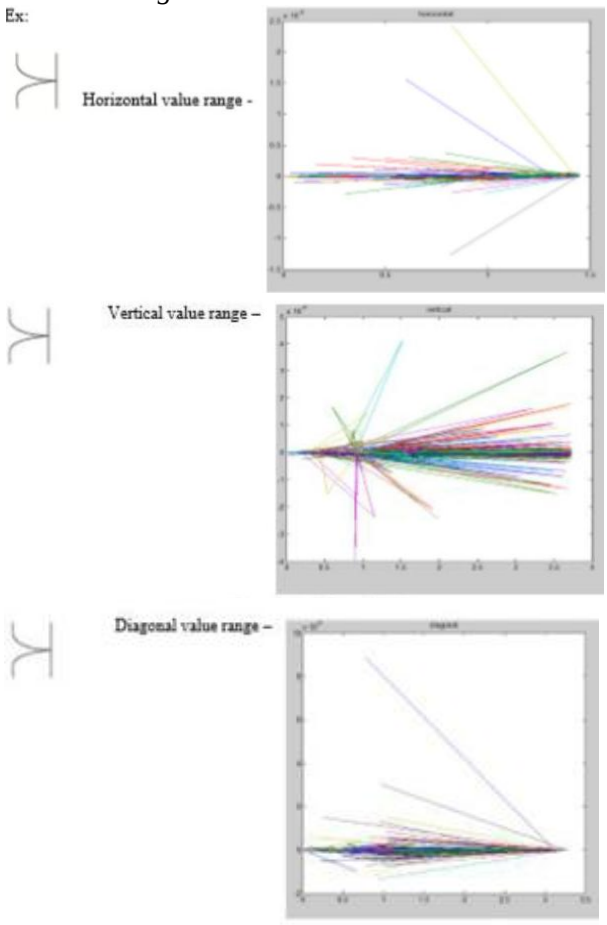

Figure 10: Plotting Charts

Since system should consider about range of values, when doing the identification, need to identify what is the exact value range for each and every character. To identify that plots did a huge contribution [2].

1.2.4 Applying Mohonolobius function

Purpose of this is to identify, the exact character that has been given as the input. From Mohonolobius function it will calculate the distance value between two characters. By maintaining those characters in a Metrix, system can exactly figure out, the given characters ID. Some sample Mohonolobius distance values as follows [3].

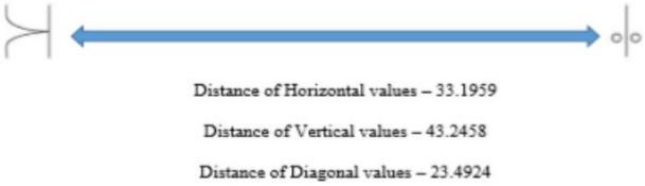

Figure 11: Mohonolobius function

1.3 Module 03: Word and Sentence Breakdown Module

Word and sentence breakdown module is designed to identifying corresponding Sinhala letter to Brahmi letter and division of that string of Sinhala letters in to Brahmi words and then prompting all possible combination of sentences which could be formed by using the identified words. Finally, it provides privilege to users to select the best option sentence out of the proposed suggestion sentences. The output of this function is a sentence with Sinhala letters with Brahmi word combination.

Some of the Characteristics of "Mula Brahmi" language are:

- Contains only few words in the vocabulary
- Not separated in to words and there were no punctuation marks used
- There are 38 basic row letters and most widely used 20 letters with syntaxes.
- Basic row letters and letters with syntaxes are seems to be very similar and those differ from each other by a slight property.

By using image concatenation technique, combined those slices of Brahmi letter images in to single image. That image is tally with the original stone inscription content. The usage of text mapping function to transliterate Brahmi letters in to its corresponding Sinhala letter along with the image concatenating. After performing one to one mapping, system generates the transliterated Sinhala string of the particular Brahmi string. The developed functionality is very less time consuming, high accurate and efficient [4].

The Brahmi letters in the scanned image of the stain paper are replaced with the Letter templates stored in the system as below.

Figure 12: Brahmi Letter Template

The transliteration function takes place and every Brahmi letter is mapped to its corresponding Sinhala letter as below.

Figure 13: Brahmi Letters Mapped to Sinhala Letters

Next functionality implemented in the module is word break down and generating suggestion sentences. After getting the input as a string of Sinhala letters, used regular expression matching to extract the words embedded inside the Sinhala string. There, first detected the basic Brahmi words with syntaxes using REGEX. Then apply regular expression matching again to select words without syntaxes. DB mapping with regular expression matching technique is applied as the sub module. And then undergo the process of spelling correction for those words without syntaxes and generated suggestion words for them.

Here the mapped Sinhala letter string is divided into Brahmi words by doing spelling correction and suggesting words as follows. Then form of all possible sentences by combining the identified words as below.

Figure 14: Word breakdown and generation of all possible sentences

Suggestions for some words have been given to improve the accuracy. That is because, during the letter recognition process, some letters with syntaxes may be misidentified as row letters. Letters may not be very clear to identify due to high noise available in the scanned image of the stain paper. That scenario is by expressing suggestion words by undergoing spelling correction process. When detected the words it can be displayed as Word 1 Word 2 Word 3 Word 4 Word5. Then just assume the Word 2 and Word 4 are incorrectly identified. Or those are suggestion words which can be generated from the same position of the string. So that system is capable of generating suggestion sentences by combining the identified expressions.

Word 1 [Sugestion1] [Sugestion2] Word3 [Sugestion3][Sugestion4] Word 5

So this can be displayed as,

System is generating a graph by considering the start position, end position, and length of each word like above and then apply Breadth First Search algorithm to generate all paths from start to end of the graph. This is the process of generating all possible combination of sentences in Sinhala letters with Brahmi words from a given string of Brahmi letters. Then the user has the privilege to select the best option sentence out of the suggestions.

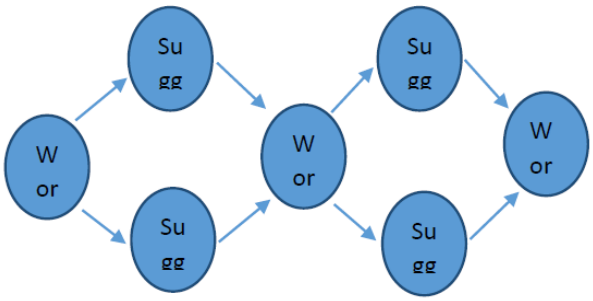

Figure 15: Logic of formation of graph

1.4 Module 04: Sinhala Meaning Generating Module

The main objective of this module is, gather words in to sentences in order to generate a meaning in Sinhala. To perform that, system should identify the meanings of Brahmi words in Sinhala. Then Sinhala words should be delivered to represent the correct meaning. To make this translation success, system will be feed by the grammar rules and meanings of Sinhal**a** language [5].

There are four sub modules in the main module.

1.4.1 Find all the relevant brahmi words and Sinhala meanings [6].

1.4.2 Categorize all the words as follows

Is_Verb

Is_Object

Is_Relation

Is_Livenoun

Is_Addingletter

Is_Name

Boolean fields are used for represent these categorized words.

1.4.3 Load the input to the database

Following are the example input word sets that are coming from word breakdown function. These are Brahmi words but they are in Sinhala letters.

ගාමිනි, පුත, රාජ, උටි, රාජ, උටි, පුත, අය අභයඝ, ජීත, අභි, අනුරාදි ය

Following are some of example words.

උටි - උට්ටිය ලෙලෙ - ලෙෙ දනව - දන්න රජ

Then map words from the database. But resulted words not in correct order as shown below

Input

ගාමිනි පුත රාජ උටි රාජ උටි පුත අය අභය ජීත අභි අනුරදි ය

Database mapping Output

ගාමිනි පුතා රජු උට්ටිය, රජු උට්ටිය පුතා කුමරු අභය කුමරිය අනුරාදි විය

By reading this output we can't get the exact meaning of the inscription.So for that it should be corrected to a particular pattern.

1.4.4 Defining rules

Following are the example grammar ruled that have defined. Rule based machine translation is using with Natural language processing techniques **[7]**. First studied all the inscriptions and identified some patterns inside each and every inscription **[8]**.

- *Rule No-01*

System is checking words two by two. If the nearby two words are,

Name (Ex- නිස , ධර්මරාජ)+Relation(Ex- පුත, ජය, ජිත)

Follow this order,

Name(Sinhala_meaning..)+ ලේ+ Relation(Sinhala_meaning.)

Ex-

ගාමිනි පුත

ගාමිනි + ලේ+ පුතා = ගාමිනිගේ පුතා

- *Rule No-02*

If the nearby three words are,

Live_noun (Ex- අය,අභි,රාජ) + Name (නිස,ධර්මරජ) + Relation(Ex- පුත, ජය)

Follow this order,

Name (Sinhala_meaning) +

Live_noun (Sinhala_meaning) + ලේ + Relation(Sinhala_meaning)

Ex-

රාජ උවි පුත

උවිටිය +රජු + ගේ +පුතා=උවිටිය රජුගේ පුතා

There are some example grammar rules. After applying all defined grammar rules user can get the exact meaningful Sinhala sentence.

4. RESULT AND DISCUSSION

The main objective of this research was to use the modern technology which is related to image processing, character recognition, text mining and natural language processing, to come up with a standalone application to support the archeologists. The standalone application is quite effective and efficiently supports the archeologists who are bothering with the Sinhala meaning of the stone inscriptions written in Brahmi language. The implementation of the application was done according to the "waterfall" software development methodology. The requirement gathering and analyzing was conducted beside of the image processing, as the research touches advance image processing techniques as well as other technologies of character recognition and natural language processing. So that, an in depth study about those areas was highly essential. The identified key concepts in these knowledge areas were the pace for the discovery of the new research outcomes.

As the research is mainly based on image processing, natural language processing and character recognition, huge number of scanned images were processed for the more reliable results. Scanned images were categorized by considering the noise level at the initial stage. A survey had been conducted to identify the accuracy level of the system. The results of the test are displayed below.

Table 1: The result of survey

Noise level of images	No of images taken	Identification of correct Sinhala meaning
0 – 30 %	10	90 %
30 – 60 %	25	75 %
60 – 100 %	30	60 %

Bases on the test results, it can mention that system is efficient and accurate under any of situation when considering the noise level.

5. CONCLUTION

With the rapidly developing technology, everything in the planet earth should go in a parallel line. If something is not aligning with the technology, there would be a huge difficulty for that field to gain its maximum achievements.

As per the description provided in above chapters this project is to implement translation software, which is a completely new concept to the domain of archeology. When this project completes a standalone application will be released, which will provide a huge contribution in translating ancient inscriptions. So having this kind of facility will be very important and easy. By having this kind of software solution archeology department can get lot of benefits as describe in above chapters in this document. Finally, the research group's main ambition is to give a better service archeologist through the proposed system and filling the gap between the field of Archeology and IT.

6. ACKNOWLEDGEMENT

Co-Authors: Nimesha Amarasinghe, Nipuni Kumarapeli (Sri Lanka Institute of Information Technology)

This research would not have been possible without the guidance and the help of several individuals who in one way or another helped and extended their valuable assistance doing the research. First and foremost sincere gratitude to goes to the supervisor Mr.Tharindu Perera who gave a constant encouragement and great support throughout the process and would not be able to success without that outstanding encouragement. Also sincere thanks goes to the lecturer in charge - Mr.Jayantha Amararachchi, for the continuous guidance given. The results embodied in this report have not been submitted to any other university or institution for the award of any degree or diploma. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

7. REFERENCES

[1] A. Manpreet Kaur, "Combination Method for Powerline Interference Reduction in ECG," International Journal of Computer Applications, vol. Volume 1– No.14, 2010.

[2] R. P. P. B. Anoop Kunchukuttan, "Brahmi-Net: A transliteration and script conversion system for languages of the Indian subcontinent," in Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, Denver, Colorado, 2015.

[3] A. l. o. o. p. MaesschalckD.Jouan-RimbaudD.L.Massart, "The Mahalanobis distance," Chemometrics and

Intelligent Laboratory Systems, vol. 50, no. Issue 1, 4 January 2000, p. 18, 4 01 2000.

[4] C. M. Sankari, "Object Matching using Skeletonization based," International Journal of Computer Applications , vol. Volume 28, no. No.7, 2011 August.

[5] P. K. B. Gaurav Kumar, "A Detailed Review of Feature Extraction in Image Processing Systems," in 2014 Fourth International Conference on Advanced Computing & Communication Technologies, Rohtak, India, 2014.

[6] D. H. C. L. N. U. Viraj Welgama, "Towards a Sinhala Wordnet," in Conference on Human Language Technology for Development, 2011.

[7] G. G. Chowdhury, Natural Language Processing, 2005.

[8] Parascript, "PARASCRIPT," 2014. [Online]. Available: https://www.parascript.com/signature-verification/. [Accessed 20 08 2015].

[9] T. S. J. P. Yannis Assael, "Restoring ancient text using deep learning: a case study on Greek epigraphy," Empirical Methods in Natural Language Processing (EMNLP) 2019, 2019.

[10] N. d. Silva, "Survey on Publicly Available Sinhala Natural Language Processing Tools and Research," arXiv:1906.02358v4 [cs.CL] 22 Jul 2019, p. 16, 2019.

[11] I. S. Nurmamatovna, "To recognize the manuscript texts of Arabic letters in ancient Uzbek script," World Scientific News, vol. 115, pp. 160-173, 2019.