

Extracting Old Persian Cuneiform Font Out of Noisy Images (Handwritten or Inscription)

Seyed Muhammad Hossein Mousavi

Department of Computer Engineering

Bu Ali Sina University

Hamadan, Iran

h.mosavi93@basu.ac.ir

Vyacheslav Lyashenko

Department of Informatics (INF)

Kharkiv National University of Radio Electronics

Kharkiv, Ukraine

lyashenko.vyacheslav@gmail.com

Abstract— The process of converting the text in the digital image to font (encrypted text) is called OCR. This paper is involved with extracting inscribed texts on the Achaemenid inscriptions. This is the first proper quality example of using OCR to recognizing Achaemenid scripts. There are different approaches to recognizing characters, of which we have chosen open source Tesseract engine for segmentation, learning and classification in this research. Due to existence of noise (stone crack) in inscriptions, this paper uses some image processing techniques to eliminate noises. This system's final output includes: extraction of cuneiform font, Persian and English transcription of sentences, sentence pronunciation and translation of a substantial number of extracted Persian and English words, which makes us better understand the way they spoke in that era. Acquired results of validation and result section indicates that this system has been able to properly cope with the recognition of cuneiform characters and has classified all characters of test data properly with about 92% accuracy. The acquired results are promising that they are able to make and improve NLP in this area.

Keywords—OCR; Achaemenid inscriptions; Tesseract engine; Image processing techniques; Cuneiform font

I. INTRODUCTION

Discovering and reading ancient scripts has always been a wish to human being, and along with time, by discovering valuable historical mysteries, we have been able to realize some aspects of ancient civilizations and yet it is going on. Cuneiform is a kind of script which bearing symbols looking like wedge. There are different kinds of cuneiform script, which are used to write different languages. Old Persian cuneiform, Sumerian, Akkadian, Elamite, Babylonian and Avgryty are examples of Cuneiform script. All discovered ones are written from left to right. This script which some believe has ideogram basis, has been used in all ancient western Asia civilizations. This script dates back to sixth century BC and has an age of 2600 years. This script most probably had been innovated by Cyrus the great. Old Persian inscriptions are written by this script. Discovered inscriptions in recent times from: Dasht-e Morghab, Bistun, Alvand, Susa, Persepolis, Romania, Armenia, Hamedan and along the Suez Canal basin, were engraved by the command of Cyrus the great and Xerxes [1] [2] [3] [4]. The importance of work where heightens which this system causes the pace, ease and functionality of text translation enhance for archeologist's

work and other researchers'. In the section II some works done on ancient and new scripts in field of OCR, will be discussed. Section III will completely explain proposed method with image processing technique's details which has been used. Section IV includes experimental results, test data and achieved results of their validation. Section V concludes mentioned discussions and offers suggestions to improve proposed method.

A. Optical Character Recognition(OCR)

OCR is auto-recognition of digital picture's script and converting it to searchable and editable script for computer. Picture of document is usually prepared by the scanner or digital camera and bears some colorful pixels with a variety of brightness. From human viewpoint, a document may have a lot of information value, but for a computer an image does not differ with other ones, due to seeing it as a bunch of pixels. In order to utilize the image script of document, these pixel scripts should in a way be recognized by the computer. This is done by OCR software. Nowadays, OCR is mostly used to recognize printed documents such: books, magazines and printed letters. Image recognition and pattern recognition are two underlying factors of these systems. The complexity of these systems differs in respect of the language. For instance, Latin due to being written separately, is easier to recognize than Farsi and Arabic, which are written more connectedly. In Figure 1, OCR steps are displayed. For more information on this, refer to [5] [6].

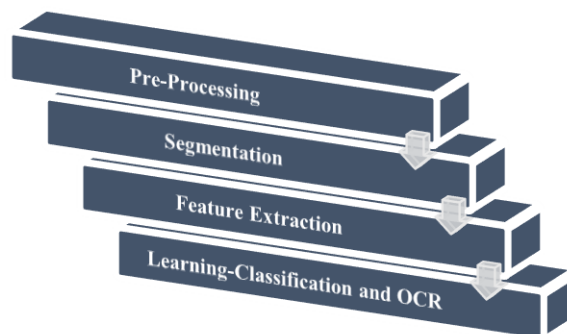


Fig. 1. OCR procedure

B. Pattern recognition

Pattern recognition is a branch of machine learning. Most of researches in the field of Pattern Recognition are in regard to supervised leaning and unsupervised learning. Pattern recognition methods separate intended patterns out of a bunch of data, using prior knowledge about patterns or statistical information of data. After acceptable results are achieved by a set of training patterns, system will be engaged to experiment real data. The function quality of recognition is widely specified by training pattern similarity and real data which are faced along operations [7].

C. The structure of the Old Persian cuneiform

The Old Persian cuneiform with 36 alphabets has been the last kind of cuneiform (in 6 BC). This script has been used in Achaemenid inscriptions. The Achaemenid cuneiform is a semi-alphabetical semi-syllabic script, plus having 8 ideograms which are used to describe useful words such as: king, country, Ahura Mazda and more. Its symbols are wedge-like and written from left to right [4] [8] [9]. This is one of the simplest cuneiforms. Only Ugaritic cuneiform is simpler and being all alphabetical. This script is now accepted in Unicode standard (4/1) and a domain is allocated to it. Table I and II respectively indicate the alphabets and ideograms with numbers of this language in cuneiform, English and Persian.

II. SOME OF PRIOR ACTIVITIES

In recent decades, auto alphabet recognition using software has become very widespread; in particular, ancient ones. Also different algorithmic methods used for learning and classifying these languages, have been checked and their results have been compared. In 2006, Ntzios, K, et. al, succeeded developing a method to recognizing Greek language characters relating to the first years of Christianity, and their main focus was on the lower case alphabets, which was indicating good results [10]. In 2013, Naktal M. Edan succeeded making a system to recognize ancient Sumerian language. He also used K-means method to cluster similar symbols, and used multilayer perceptron neural network for classification of clusters [11]. In 2016, H Ashari, M., Asadi, et. al succeeded to increase the processing speed for Persian and Arabic alphabet recognition by 5.5 times, using parallel processing of Graphic card and CUDA platform. They also got help of Derivative Projection Profile feature extraction method and hamming neural network classification [12]. In

2012, Mishra, N, et. al succeeded designing a system to increase the recognition accuracy of Hindi alphabets [13]. They used Tesseract engine for character recognition. In 2015, Adnan, Khashman and Mostofi, Fahimeh succeeded developing a very simple system for recognizing Old Persian cuneiform characters [14]. But this system used only 46 training data, and owing to this, was not able to deal well with character recognition of stone inscriptions and handwritten scripts which possessed different shapes. This system was only able to eliminate the Gaussian noise which was pointless in regards to stone inscriptions. They used neural networks for learning and classification of characters. Also in 2015 Bogacz, Bartosz, et al, made an OCR system based on HOG features, ICP algorithm and HMM classifier for handwritten cuneiform scripts [20]. The proposed system, possessing training data with 1500 characters of different shapes, and eliminating of Gaussian and salt and pepper noises. Also with eliminating small objects which are bigger than noises, is able to cope with recognizing most of the inscription and handwriting of this old script. This system after recognition, is able to pronounce these characters in Persian and English, and even is able to pronounce acoustically sentences and translate them, which is unique in its kind.

III. PROPOSED METHOD

This study introduces a character recognition over Old Persian (Achaemenid) cuneiform. The work commences with pre-processing such as image normalization like brightness intensity and image resizing. In the case, the taken pictures of inscriptions or handwritings have noise, noise reduction takes place using special methods. Thus, first, images which are polluted by Gaussian and salt and pepper will be manipulate using median filter and then Sharpening , then small objects will remove with a morphology technique. Next, image will be restored by other morphology technique, called Dilation. Eventually, by applying fuzzy edge recognition method, intended character will be extracted [15] [16]. Pre-processing stages are shown completely in Figure 2. Segmentation, learning and classification stages take place using open source Tesseract engine v.3.04. The segmentation which is applied on input images by Best-Frist-Search graph or A*, has application just in learning section and is not used in test phase. For more on Tesseract you can refer to [17]. There are some important hints about image learning of this engine such:-

TABLE I. ACHAMENID CUNEIFORM ALPHABETS WITH ITS ENGLISH AND PERSIAN PRONUNCIATION

Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Cuniform	𐎠	𐎡	𐎢	𐎣	𐎤	𐎥	𐎦	𐎧	𐎨	𐎩	𐎪	𐎫	𐎬	𐎭	𐎮	𐎯	𐎱	𐎲	𐎳	𐎴
English PRON	a	i	u	ba	pa	ta	tu	sa	ja	ji	xa	da	di	du	ra	ru	za	sa	sha	fa
Persian PRON	ا	ای	او	ب	پ	ت	تو	ث	ج	جی	خ	د	دی	دو	ر	رو	ز	س	ش	ف
Number	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	-	-	-
Cuniform	𐎵	𐎶	𐎷	𐎸	𐎹	𐎺	𐎻	𐎼	𐎽	𐎾	𐎿	𐏀	𐏁	𐏂	𐏃	𐏄	𐏅	-	-	-
English PRON	ka	ku	ga	gu	la	ma	mi	mu	na	nu	va	vi	ha	ya	C	cha	𐎶𐎠	-	-	-
Persian PRON	ک	کو	گ	گو	ل	م	می	مو	ن	نو	و	وی	ه	ی	ث	چ	𐎶𐎠	-	-	-

TABLE II. DISCOVERED IDEOGRAMS IN ACHAEMENID CUNEIFORM AND NUMBERS

Number	1	2	3	4	5										
Cuniform	𐎧𐎫	𐎧𐎫𐎧𐎫	𐎧𐎫𐎧𐎫	𐎧𐎫𐎧𐎫	𐎧𐎫𐎧𐎫𐎧𐎫										
English Translation	king	country	land	god	Ahura mazda										
Persian Translation	شاه	کشور	زمین	خدا	اهورامزدا										

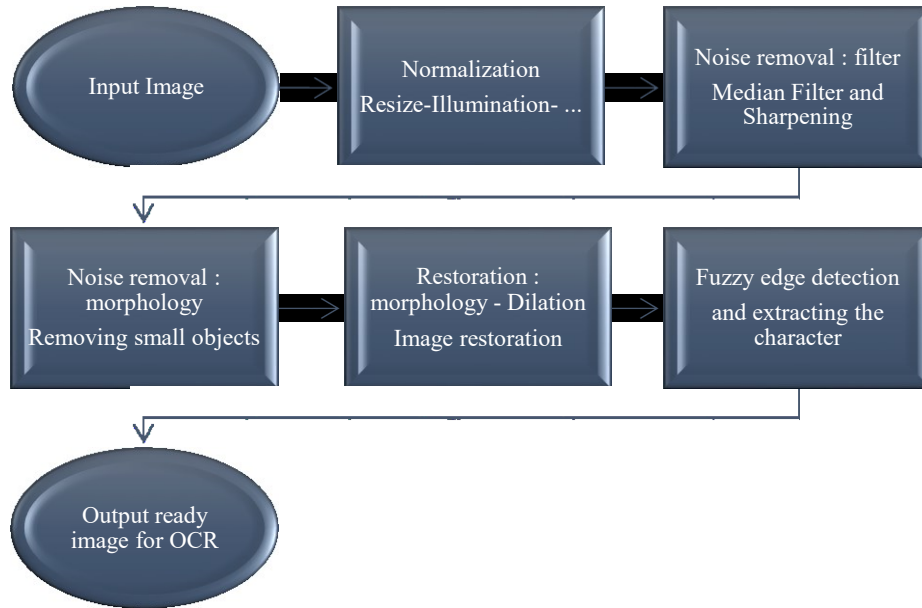


Fig. 2. Pre-processing stages to image preparation for learning and test



Fig. 3. X-Height size to learning by Tesseract engine

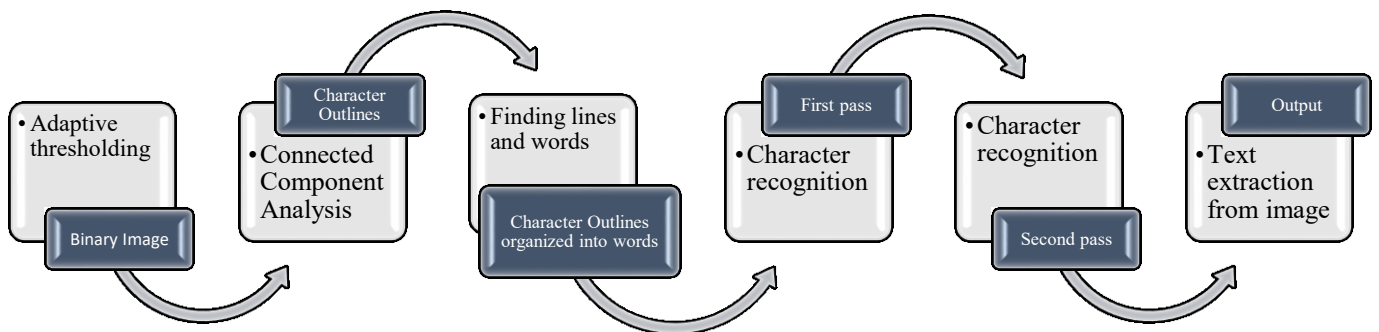


Fig. 4. The structure of Tesseract open source character recognition engine

Every rotation in the image must be corrected. Otherwise the character would not be recognizable. Low frequency changes in the smooth images have to be corrected by high frequency

filters or sharpening, and blacked edges in the image must be eliminated, since these edges would be wrongly recognized as characters. Despite of the troubles in learning stage, this

engine has a great capability in classification on different characters. In learning stage, if a character has separate sections, and these sections have significant distances of each other; the engine will assume each as a distinguished character or as so-called, chops them [17], unless in a way, they approach each other or make connections, or even manually classify and label them. The function structure of the Tesseract character recognition engine is observable in Figure 4.

After reading input image, it will converted to binary image using placement method of Adaptive thresholding [18]. The next stage deals with extracting outlier of character. In this stage, all the walls are converted to blob by nesting action. It means that all the lines will convert to completely closed objects. These blobs will be organized in a line. In the next stage, the lines and closed areas convert to a proper text and eventually, the acquired text will divide into different words using fuzzy area. The recognition stage is including two sections. In the first pass, an effort takes place to organize every word in turns. In the first pass of recognition, every word is sent to adaptive classifier [17] as training date. In the second pass of recognition, the compatible classifier repeats the recognition to find the words it couldn't find in the first pass, and detects them. The final phase includes resolving the problems of fuzzy spaces and checking the issue of X-height for the texts shorter than 20 pixels and between mean-line and baseline [17] [19]. After classification stage, in order to level off the acquired strings and its readiness for pronunciation (TTS), post-processing will take place. The pronunciation of acquired sentences is achieved by the Microsoft SAPI service and immediately shows the translations of some important sentences. Figure 5 indicates the procedure of main image conversion to outlier lines and blobs, and finally the feature extraction of the final model shape, and matching operation for comparison. Figure 6 also presents an example of all the pre-processing and processing stage which the procedure levels are represented by numbers. In level 2, some salt and pepper noise has been added and everything else is according to Figure 2. The sampled section out of inscription, has been extracted at random and does not have any special meaning, and here just presents how system works on the inscription. Following examples present meaningful texts.

Table III, includes post-processing operations of Figure 6 and respectively from right to left presents: Farsi pronunciation, English pronunciation and recognized cuneiform image, which is automatically generate by the system. An edge is a boundary between two uniform regions. You can detect an edge by comparing the intensity of neighboring pixels. However, because uniform regions are not crisply defined, small intensity differences between two neighboring pixels do not always represent an edge. Instead, the intensity difference might represent a shading effect. The fuzzy logic approach for image processing allows you to use membership functions to define the degree to which a pixel belongs to an edge or a uniform region. For more information on this, refer to [15] [16] [21] [22].

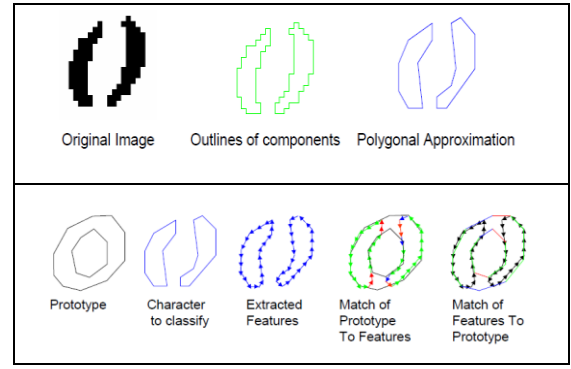


Fig. 5. The conversion procedure of the main image to outlier lines and blobs (for feature extraction)

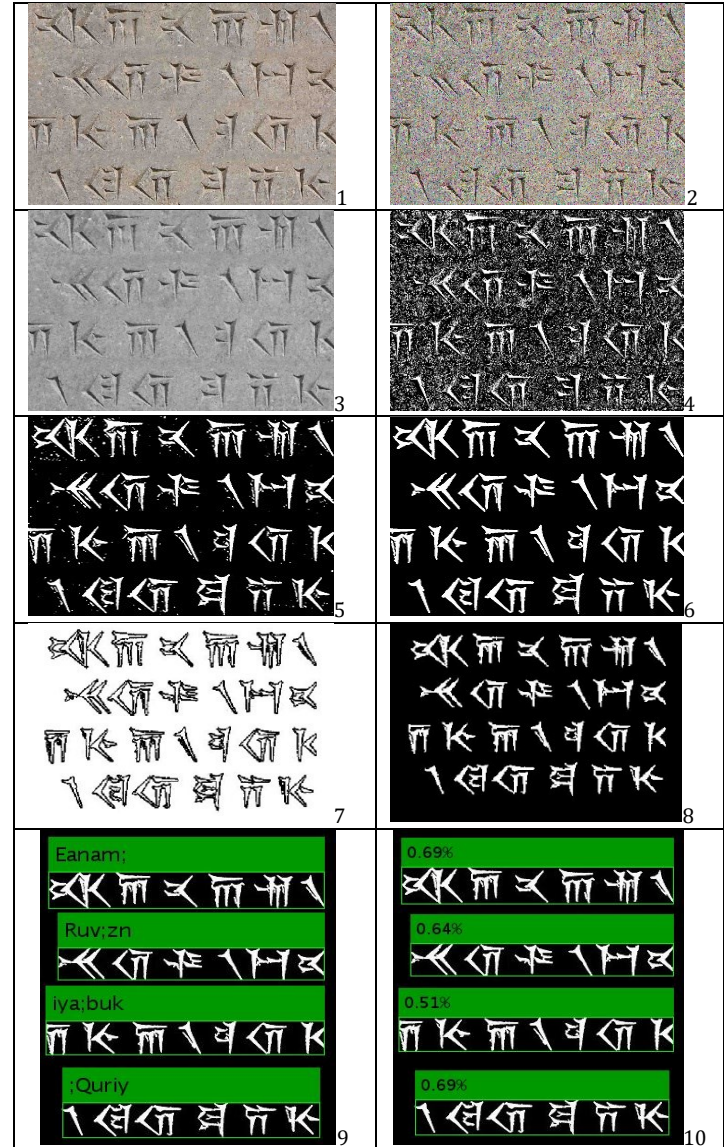


Fig. 6. An example of the entire pre-processing and main process on a portion of an inscription (1.RGB image, 2.Adding noise (salt and pepper), 3. Gray image, 4. Noisy gray, 5. Median filter and sharpening, 6. Morphology (dilation and removing small objects), 7. Fuzzy edge detection, 8. Image invert and connected component filling, 9. Extracted and recognized characters, 10. Recognition accuracy for every line)

ᐱᐱ ᐱᐱ	King a na a ma ," ru u va , za na ia ya a , ba u ka du u ra da ya , "	"پادشاه آنا ا ما ، رو یو وا ، زانا آی یا ا ، با یو کا دو یو را دا یا ،
--	---	--

The success percentage and recognition rate of the character recognition systems, in a relatively great extent, depends on their database. This project uses 23 different kinds of writing, which totally reaches 1500 training data. Due to the deficiency of training data for this ancient language, inevitably some of the training data were created by image processing techniques, and also changing the present samples. Using such a volume of training data for this old language is unique. Table IV indicates the results on the character recognition of a handwritten sentence which has been polluted by salt and pepper noise with intensity of 0.025. In this table from top to down (1. Noisy image, 2. Denoised image with median filter, 3. Extracted and recognized characters, 4. Recognition accuracy for each line)

<p>1</p>	<p>1</p>	<p>1</p>
<p>2</p>	<p>2</p>	<p>2</p>
<p>3</p>	<p>3</p>	<p>3</p>
<p>4</p>	<p>4</p>	<p>4</p>
<p>"Was ; Cyrus ;ra ; the Most Great ;Army"</p>	<p>"aha , kuuruusha , aita , masaishata , kaara"</p>	<p>"أها ، کو یورو یوشا ، آای تا ، مانا ای شاتا ، کا آرا بود ؛ کوروش ؛ را ؛ بزرگ ترین ؛ سپاه"</p>

TABLE V. THE RESULTS OF THE PROPOSED SYSTEM ON TYPED DATA
(LAST 12 LINES OF CYRUS THE GREAT'S INSCRIPTION)

Zatiy

daryvuS

xSayZiy;

mna;aurmzda;upstam;

blauv;

hda

VizibiS

bgibiS

uta;

imam;dhyaum;aulmzda;

paTuv;hca;hinaya;

hca;

QuSiyala

hca;druga;abiy;

imam

dhyaum;ma;

ajMiya;

ait;

aim

yanm;jDiyaMiy;

aitmiy

ddaTuv

0.82%

0.9c0.86%

0.9c0.84%

0.77%

0.84%

0.66%

0.9c0.78%

0.9c0.80%

0.9c0.85%

0.69%

0.69%

0.69%

0.79%

0.9c0.65%

0.78%

0.61%

0.70%

0.76%

0.79%

0.77%

0.74%

0.95%

0.79%

"Kl;h

-Table V. Continue



<p>"Zatiy ; daryvuS ; xSayZiy; mna;aurmzda;upstam; blauv; hda ; ViZibiS ; bgibiS ; uta; imam;dhyaum;aulmzda; paTuv;hca;hinaya; hca; QuSiyala ; hca;druga;abiy; imam ;dhyaum;ma; ajMiya; ait; aim ;yanm;jDiyaMiy; aitmiy ; ddaTuv "</p>	<p>"نآ تا ای یا ، دا آ را یا وا یو شا ، خا شا آ یا تا ای یا ، ما نا ، آ یو را ما زا دا آ ، یو پا سا تا آ ما ، با لا آ یو وا ، ها دا آ ، وی آ ای تا ای با ای شا ، با گا ای با ای شا ، یو تا آ ، آی ما آ ما ، دا ها یا آ یو ما ، آ یو لا ما زا دا آ ، پا آ تو یو وا ، ها چا آ ، ها ای نا آ یا آ ، ها چا آ ، دو یو شا ای یا آ لا آ ، ها چا آ ، دا را یو گا آ ، آ با ای یا ، آی ما آ ما ، دا ها یا آ یو ما ، ما آ ، آ جا مو ای یا آ ، آ ای تا ، آ ای ما ، یا آ نا ما ، جادی آ ای یا آ مو ای یا ، آ ای تا ما ای یا ، دا دا آ تو یو وا "</p>
<p> s1.mp3</p> <p> s2.mp3</p>	
<p>"Says ; Darius ; King; to Me;Ahura Mazda;Helper; Does; with ; All ; Gods ; and; this;Country;Ahura Mazda; Keep and Save;from;Enemy; from; Drought ; from;Lie;to; this ;Country;not; Become; This ; Mine ;Forgiveness;Ask; This to ; Give "</p>	<p>"گوید ؛ داریوش ؛ شاه ؛ مرا ؛ اهورامزدا ؛ یآوری ؛ برد ؛ با ؛ همه ؛ خدایان ؛ و ؛ این ؛ سرزمین ؛ اهورا مزدا ؛ باید ؛ از ؛ دشمن ؛ از ؛ خشکسالی ؛ از ؛ دروغ ؛ بهسوی ؛ این ؛ سرزمین ؛ نه ؛ آید ؛ را ؛ من ؛ بخششی ؛ درخواست می کنم ؛ را این مرا ؛ بدهد "</p>

Table V includes: recognized characters, confidence value, extracted Cuneiform characters, English and Persian pronunciation, acoustic English pronunciation and translation, English and Persian translation of sentences.

V. CONCLUSION AND SUGGESTION

In general terms, it can be said that the proposed system is not only a proper OCR system to recognizing Old Persian Cuneiform characters, but is also able to denoise salt and pepper and Gaussian noises. The use of a database, bearing 1500 training data of this script, which are created manually and transformed by graphical software, not only cause to recognize the major portion of writings, also, in its kind is the first proper database for this script. This system, plus extracting the font and pronouncing it in two languages, translates the sentences into two languages acoustically and all of these are automatically presented to the user. Removing speckle and poisson noises will include the Future activities. In general, this system is the first powerful system with large abilities for this old script, and the acquired results are very satisfying.

References

- [1] Kuhrt, A. (2013). The Persian Empire: A Corpus of Sources from the Achaemenid Period. Routledge. ISBN 978-1136016943.
- [2] Frye, Richard Nelson (1984). Handbuch der Altertumswissenschaft: Alter Orient-Griechische Geschichte-Römische Geschichte. Band III,7: The History of Ancient Iran. C.H.Beck. ISBN 978-3406093975.
- [3] Schmitt, Rüdiger (2000). The Old Persian Inscriptions of Naqsh-e Rostam and Persepolis. Corpus Inscriptionum Iranicarum by School of Oriental and African Studies. ISBN 978-0728603141.
- [4] Kent, Roland Grubb. Old Persian: grammar, texts, lexicon. Vol. 33. Eisenbrauns, 1953.
- [5] Schantz, Herbert F. (1982). The history of OCR, optical character recognition. [Manchester Center, Vt.]: Recognition Technologies Users Association. ISBN 9780943072012.
- [6] Cheriet, Mohamed, et al. Character recognition systems: a guide for students and practitioners. John Wiley & Sons, 2007.
- [7] Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern classification. John Wiley & Sons, 2012.
- [8] Windfuhr, Gernot L. "Notes on the old Persian signs." Indo-Iranian Journal 12.2 (1970): 121-125.
- [9] Daniels, Peter T., and William Bright. The world's writing systems. Oxford University Press on Demand, 1996.
- [10] Gatos, Basilios, et al. "An efficient segmentation-free approach to assist old Greek handwritten manuscript OCR." Pattern analysis and applications 8.4 (2006): 305-320.
- [11] Naktal, M. Edan, "Cuneiform Symbols Recognition Based on K-Means and Neural Network." AL-Rafidain Journal of Computer Sciences and Mathematics(2013).
- [12] Askari, M., et al. "Isolated Persian/Arabic handwriting characters: Derivative projection profile features, implemented on GPUs." Journal of AI and Data Mining 4.1 (2016): 9-17.
- [13] Mishra, Nitin, et al. "Shirokekha Chopping Integrated Tesseract OCR Engine for Enhanced Hindi Language Recognition." International Journal of Computer Applications 39.6 (2012): 19-23.
- [14] Mostofi, Fahimeh, and Adnan Khashman. "Intelligent Recognition of Ancient Persian Cuneiform Characters"(2015).
- [15] Rafael C.. Gonzalez, Richard E.. Woods, and Steven L.. Eddins. Digital Image Processing Using MATLAB®. McGraw Hill Education, 2010.
- [16] Marques, Oge. Practical image and video processing using MATLAB. John Wiley & Sons, 2011.
- [17] Smith, Ray. "An overview of the Tesseract OCR engine." (2007).
- [18] Chan, Francis HY, Francis K. Lam, and Hui Zhu. "Adaptive thresholding by variational method." (1998).
- [19] EL GAJOU, K. H. A. D. I. J. A., FADOUA ATAA ALLAH, and MOHAMMED OUMSIS. "Training TESSERACT Tool for Amazigh OCR." Recent Researches in Applied Computer Science: Proceedings of the 15 th International Conference on Applied Computer Science. 2015.
- [20] Bogacz, Bartosz, Michael Gertz, and Hubert Mara. "Character retrieval of vectorized cuneiform script." Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. IEEE, 2015.
- [21] Alshennawy, Abdallah A., and Ayman A. Aly. "Edge detection in digital images using fuzzy logic technique." World Academy of science, engineering and technology 51 (2009): 178-186.
- [22] Becerikli, Yasar, and Tayfun Karan. "A new fuzzy approach for edge detection." Computational Intelligence and Bioinspired Systems (2005): 675-709.