INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Multilingual OCR systems for the regional languages in Balochistan

Anwar Ali Sanjrani[1]*, Muhammad Shumail Naveed[1], Muhammad Sajid[1], Atiq Ahmed[1], Shafiq Awan[2], Awais Khan Jumani[3]

1 Department of Computer Science and Information Technology, University of Balochistan-Quetta, Pakistan
2 Department of Computer science, Benazir Bhutto Shaheed University Lyari, Karachi, Pakistan
3 ILMA, University Karachi, Sindh, Pakistan

OPEN ACCESS

Received: 27-03-2020
Accepted: 27-04-2020
Published: 24-06-2020

Editor: Dr. Natarajan Gajendran

Citation: Sanjrani AA, Naveed MS, Sajid M, Ahmed A, Awan S, Jumani AK (2020) Multilingual OCR systems for the regional languages in Balochistan. Indian Journal of Science and Technology 13(21): 2157-2167. https://doi.org/10.17485/IJST/v13i21.2

***Corresponding author**.
Anwar Ali Sanjrani

Department of Computer Science and Information Technology, University of Balochistan-Quetta, Pakistan
anwarali.cs@uob.edu.pk

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

Abstract

Background: There are various languages for which an optical character recognition technology has been developed but most of these address a particular language and thereby multilingual OCR remains a challenge. **Methods**: Development of multilingual OCR is one of a highly debated issue. Researcher are studying the feasibility and operational feasibility of multilingual OCR from technical as well as from viable aspects. Multilingual OCR includes printed or handwritten characters' form. In this paper, we study the significance, challenges and issues of developing multilingual OCR system for regional language based on Persio-Arabic script by conducting a comprehensive survery about the operational viability of mmultilingual OCR. **Findings**: A feedback of 339 participants is collected through an online surgery to find the scope and applicability of multilingual OCR. The respondents were from different linguistic background. The study identified that a large majority of participants are willing to use their native language for the accomplishment of their computational task and deemed that the support of multiple languages in a software would increase their productivity. **Novelty**: In current form, the study addresses the viability of multilingual OCR of regional language based on Persio-Arabic script. To the best of our knowledge, such kind of study has not been conducted for the domain of Pakistan.

Keywords: Multilingual; OCR; Multi-fonts; Omnifont; Regional languages

1 Introduction

The OCR technology is the emerging field and one of the most important computer software technique in the field of computer science. The aim of OCR is to process human readable image or text to translate into machine readable format for editing, and searching text. The commercial OCR packages are easily available for translation various natural languages such Arabic, Persian, and Chinese into machine readable codes efficiently and effectively. The most common available and well-known OCR engines are ABBY and Tesseract (Marcin Heliński). For example, ABBY [(1)] FineReader is an all-in-one

PDF and OCR systems overall supports more than 200 OCR languages such as the most common languages are Arabic Persian/Farsi, Chinese, Japanese, Korean, and Thai etc. and Tesseract [2] OCR engine is open source technology which is used to support wide variety of languages up to 250 languages to extract features typed text, printed text, and handwritten text. The Tesseract OCR engine open source system works on the variety of operating systems such as Windows, Linux, and Mac. Initially Tesseract was developed by HP and UNLV in 1990s and later Tesseract sponsored and developed as an application by Google in 2006 [3]**.** The applications of OCR include in banks for processing checks, post offices for processing handwritten addresses, business cards recognitions and office automations for processing all kinds of data entry forms and many more applications. The function of OCR [4] engines is to read handwritten or printed text image and convert it into machine readable and editable form. The images of multiple languages include letters, symbols, numerals, punctuations marks, diacritical marks, broken characters, and text line are easily and efficiently recognized through reading. The most common, powerful and commercial OCR engines are available in Desktop PC and mobile or handheld devices. The OCR engines are more useful for different languages and documents having multi directional text, and coloured text documents etc. The capability of OCR engine to ensure high speed image processing, automatic recognition, and searching, editing, and copying words in multiple documents which saves time.

The significance of regional languages in Balochistan (بلوچستان) has wide influence in culture, art, music, and literature etc. Balochi is the one of major language of the Balochistan province in Pakistan. Pashto is considered secondary language in the province of Balochistan. Brahvi and Sindhi languages are spoken widely districts of Balochistan. Balochi [5] is North Western Iranian Language. Balochi (بلوچی) is the main language of the Balochi people. It was not used for writing as i.e. unwritten language before 19th century. The writing system of Balochi is from right to left direction. The official language was Persian. Balochi has 30 alphabets/letters [6] for writing Balochi scripts shown in Figure 1.

ا ب پ ت ٹ ج چ ح خ د ذ ڈ ر ڑ ز ژ س
ش ف غ ک گ ل م ن و ھ ء ی ے

Fig 1. Balochi alphabet characters

Taimur Mengal introduced Balochi letters in his pamphlet ***"Balochi Nama Qasim"*** published in 1987. He also published same alphabets in his article***"Balochi Mund Likh"***. The Balochi scholars follows Urdu Arabic script. Mir Gul Khan Nasir published his first Balochi poetry collection in Urdu Arabic script in 1951. The "Father of Balochi" Sayyad Zahurshah Hashemi wrote the comprehensive guidance in Urdu Arabic script and standardized Balochi in Pakistan and Iran. According to Sayyad Zahurshah Hashemi, there are 26 characters [7] in Balochi alphabet as shown in Figure 2.

ا ب پ ت ٹ ج چ د ڈ ر
ز ژ س ش ک گ ل م ن
و ہ (ھ) ء ی ے

Fig 2. Balochi alphabet characters

Brahvi is Dravidian language which is spoken by Baloch people. Brahvi people Pakistan ethnic group of about 2.2 million people [8] found in the Balochistan, Paksitan. The mainly areas of Brahvi people in Balochistan are Bolan Pass and Ras Muari. The mostly Brahvi people areas in Balochistan region are Kalat, Mastung, Khuzdar, Bolan Pass, and some parts District Quetta speaking Brahvi language predominantly. The Brahvi mostly found in Balochistan region and some areas of Sindh in Paksitan. The Brahvi [6] speakers are also found in Afghanistan, Iran, Iraq, Qatar, and UAE. Brahvi has 39 alphabetic characters/letters. The writing script is Arabic and Latin. It has no official status and neither used in education nor Government.

In Persian literature Pashto is known as Afghani and in Urdu and Hindi is known as Pathani language. Pashtoon speakers are also known as Pashtuns and Puktuns. Pashtoon is known sometime Afghans or Pathans. Pashto belongs to Indo-European family and it is Eastern Iranian language. Pashto is one of the official language among two largest languages in Afghanistan. Pashto is the second regional largest language among the regional languages of Pakistan. The total number of population of

Pashto speaker is about 45-60 million people in all around the world. The Pashto is the national language of Afghanistan. Pashto has 44 letters in its alphabetic haracterset. It has 4 diacritical marks. The writing [9] system of Pasto is right to left. The characteristic property of Pashto language is bidirectional as Urdu and Sindhi Languages in Pakistan. There are various writing styles of cursive script language such as Nashk, Kofi, Naataliq Thuluth, Diwani, and Riqa [10]. The primarily two writing styles for the cursive scripts such as Nash script and Nasataliq script are most commonly used. Urdu, Brahvi and Balochi are written in Nastaleeq script. These both writing styles are commonly used for the cursive languages such as Arabic, Persian, Urdu, Sindhi, Balochi, Brahvi and Pashto.

The word Sindhi (سنڌي) is an adjective which means "belonging to Sindh". The word Sindhi has been derived from the word Sindhu (و سنڌ) which is name of Indus River in Pakistan Sindh (سنڌ). The meaning of word Sindhu is an ocean or river. The fundamental ancient languages of Sindhi derived from Sanskrit, Prakrit, Arabic, Farsi, Dravidian (such Brahvi in Balochistan). The Sindhi is the one of the famous language in Sindh province of Pakistan. It is the official and state language of the Sindh province in Pakistan. The history of Sindhi language is as old as the civilization of Moen-jo-Daro. The Sindhi is used as a medium of instruction in Schools, offices, and in colleges. It has gained popularity in print media such as Sindhi news Kawish (كاوش), and electronic media such as KTN news etc. There are 52 characters in Sindhi language. The Sindhi language is cursive language and writing of direction is from right to left order. The Sindhi language has also bidirectional characteristic such as Arabic, Urdu, and Farsi etc. In [11] proposed Sindhi OCR for the recognition of Sindhi handwritten numerals and arithmetic string numerals without using input devices and memory. The regional languages for OCR are Sindhi and Pashto. But Balochi and Brahvi languages have not gained attention in the field of OCR. Urdu is the national language of Pakistan. Both Balochi and Brahvi languages use common font such as Noori Nastaleeq.

2 Background

The regional languages such as Sindhi, Pashto, Brahvi, and Balochi are the cursive script languages. There is no database for the regional languages such as Brahvi, and Balochi and no commercial OCR based engine in the field of computer vision and image processing. But the little attention has gained Pashto and Sindhi and no commercially software is available to best of our knowledge. The major languages spoken in Pakistan are Sindhi, Punjabi, Balochi, Pashto, and Urdu. The minor languages spoken in Paksitan are Brahvi, Kashmiri, Hindco, Siraiki, Gujrati, and Farsi etc. The mostly population of Pakistan is Muslims. Arabic is spoken to some extent. Because our Holy Quran is in Arabic language. Our regional languages based on Persio-Arabic script such Naskh and Noori Nastaleeq. Urdu, Brahvi, and Balochi use the Noori Nastaleeq Arabic font. Sindhi, Pashto, and Arabic use Naskh. The character set of regional languages such as Balochi, Brahvi, Pashto, and Sindhi as shown in above figures. Balochi consists of 26 alphabet characters, Brahvi has 39 characters, Pashto has 43 characters, and Sindhi has 52 characters set. These languages are written from right to left. Sindhi is superset of regional languages such as Balochi, Brahvi, and Pashto. The class family is the similar of the most of the characters of these cursive regional languages. The cursiveness refers to the joining and connection of characters of a ligature. The ligature or sub-word is connected part of a word without spacing in between characters. The word consists of one or more ligatures.

Ligature is connected and is a joining part of the word. There is no space between characters of ligature. The word is a set of ligatures. Ligature is the subset of word. Ligature has two components such as primary ligature, and secondary ligature. The primary ligature is set of connected characters in a word. The secondary ligature consists of dots or diacritical marks in a word. The word Balochistan and Pakistan is shown in Figure 3. Balochistan word has three ligatures such as "Balo" (بلو), "chista", (چستا), and " n" (ن) is shown in Figure 3. Pakistan word has also three ligatures is shown in Figure 2(b). The ligature consists of primary and secondary parts. The primary ligature refers to the continuous subpart of the word without spacing between characters. The secondary ligature may have dots or diacritical marks. The position and placement of dots or diacritical marks is above and below the ligature. Some characters are distinguished with "Tauy", "Hamza", and "Mada" for the regional languages.

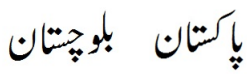

Fig 3. Ligature example of Balochistan and Pakistan

The complex ligature has two or more connected and joined sub words in the word. Context sensitivity refers to the multiple glyph of the characters, ligatures and words. The different glyphs of a character are formed while joining the standalone or final character, or initial character, and isolated character. The different glyphs/shapes of character are shown in Table 1. There are four different basic shapes of a character "Alif" (ا) "Bay", (ب), and "Jeem" (ج) with initial, medial, final, and isolated character and a few more characters etc. The shape and position changes when ligature is formed by joining the characters.

Table 1. Basic four shapes of characters in ligature

Standalone	Final	Medial	Initial
ا	ا	ا	ا
ب	ب	بـ	بـ
ج	ج	جـ	جـ
س	س	سـ	سـ
ک	ک	کـ	کـ
ل	ل	لـ	لـ
ي	ي	يـ	يـ

The regional languages such as Sindhi, Pashto, Brahvi, and Balochi are the cursive in nature. Cursiveness refers to the joining characters in the writing ligatures and words. These regional languages are written from write to left. The behavioral characteristic writing system of these regional languages based on Urdu, Arabic, and Farsi script languages. The ligatures or sub-words, words, and sentences of these regional languages are shown in Figure 4.

منهجو نالو انور علي سنجراني آهي.
زمانُوم انورِ علی سنجرانی دے
کناپن انورعلی سنجرانی اے
منی نام انورعلی سنجرانی انت

Fig 4. Cursiveness examples of regional languages translated as "My name isAnwar Ali Sanjrani"

The property of the cursive regional languages is bi-directionality. It is inherited from the Persio-Arabic script languages. These cursive regional languages mainly read and written from right to left. The numerals read and written from left to right direction. Figure 5 illustrates bi-directionality example of cursive regional languages. The numerals in box read and written from left to right and the rest of text is read and written from right to left direction.

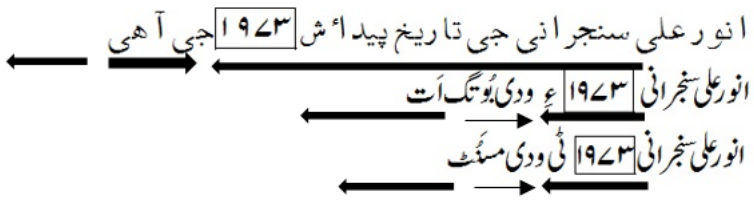

Fig 5. shows Bi-directionality translated as: Anwar Ali Sanjrani born in 1973.

Diagonality is the attributed characteristic of the Noori Nastaleeq font which is written from top right base line to the bottom left at certain variable tilted angle with well-defined rules. Figure 6 shows the demonstration of diagonality characters and diagonality ligatures

Stretching refer to elongating of the characters. There are two types of stretching: horizontal, and vertical stretching. The length and size of a character is increasing across the line horizontally is known as horizontal stretching. The length of a character is extending vertically is known as vertical stretching.

The multifont, multisize, and multiscript problems have not been addressed for the regional languages in Balochistan. The

Fig 6. Diagonality of Balochi script

mulifont, omnifont, multisize character, and compound character with varying shape and style is still a challenge for the both handwritten and printed characters [1]. These are the open challenges for the natural languages OCR systems for Urdu, Balochi, Pashto, and Sindhi in Pakistan. Omnifont refers to the recognition [12] of any font size that includes writing style, shape, size, weight, width, cursive glyph or character, and word for handwritten and printed scripts. The single font and a sing size characters have achieved a promising accuracy in the field of character recognition. The chaining code and zoning features for 30 Tamil characters using Support Vector Machine (SVM) [13] classification learning algorithm have implemented with reported accuracy of 88% by [2]. Natural language processing, image processing, computer vision, and pattern recognition have been successfully applied in the field of character recognition aka is optical character recognition [14,15] . The OCR research is still active [16] and challenging field in the natural language processing. The various historical documents, books, manuscripts, and other handwritten scripts are available for digitization for the information retrieval. The OCR technology is used to search, edit, and recognize keywords from the digitized and scanned documents. The single value decomposition (SVD) factorization matrix, Discrete Wavelet Transform (DWT) is similar to Fourier Transform (FT), and Projection Profile (PP) that includes both Horizontal Projection Profile (HPP), and Vertical Projection Profile VPP) features have been used. HPP and VPP experimented and appended as feature descriptor for the character image of size 32 x 32 using classifiers such as KNN and SVM. The accuracy for Telugu characters using SVD is 92.62, with PP 90.13, and with DWT 95.47 in k-NN classifier. The accuracy for Telugu characters at SVD 96.71, with PP 92.48, and with DWT 97.77 in SVM classifier. Overall SVM outperformed as compared to k-NN classifier using SVD, PP, and DWT for Telugu characters [4].

3 Multilingual Regional Languages Characters Family

The character family of regional languages in Baluchistan, are shown in Table 2. Sindhi and Pashto uses the Naskh font family, and Brahvi and Balochi uses the Nastaleeq font family. The writing system of the cursive regional languages is from right to left.

The number of dots with characters are zero characters, one dot characters, two dot characters, three dot characters, and four dot characters are shown as in Table 3. The dot or period (.) sign is known as **"nukta"** in regional languages.

4 Structural Framework

Formal description of multilingual OCR is important for developing its application environment. The generic framework is shown as in Figure 7 and each step of OCR design is explained briefly. OCR is a great successful application osf computer vision and pattern recognition.

The process of acquiring input images from various sources such as offline and online through some electronic devices is known as image acquisition. The most famous electronic devices digital camera and scanners are commonly used in obtaining the images in electronic format. There are two methods of image acquisition in digitizing the input image in stored template format.

1. Online image acquisition
2. Offline image acquisition

The online image can be obtained by connecting digital camera to the computer system and then transform image into electronic format for further processing. The offline image can be scanned to transform input image into digital format using scanner or digital camera. These images are offline because images could not be obtained directly some specialized electronic device. The digital image is obtained and stored into the computer databases. The database of input images is prepared for preprocessing operation. The input training image is supplied to the system for the recognition and classification. The input images match with stored template database to determine the class of that input image.

Preprocessing is the important operation after the image acquisition process. The process of removing noises, imperfections, and improving properties from the input image is achieved by the preprocessing technique. There are many preprocessing algo-

Table 2. Multilingual Characters Classes

S.No	Class	Sindhi	Balochi	Brahvi	Pashto
1	ا	ا	ا	ا	ا
2	ب	ٮ ٹ پ ٻ ٻ ٽ ٺ ٿ ٽ	ب پ ت ٹ	ب پ ت ٹ ٹ	ب پ ت ٹ ټ
3	ج	ج ڄ ڃ ڇ ڇ ڃ ڄ ڇ ڃ	ح ڇ ج	ج ڇ ح خ	ج څ خ ڇ څ ڇ
4	د	ڈ ڊ ڈ ڈ د	د ڈ	د ڈ ڈ	د ډ ڈ
5	ر	ز ڑ ڑ	ر ز ڑ	ر ڑ ز ڑ	ر ړ ز ڑ ږ
6	س	ش س	س ش	س ش	س ښ
7	ص	ض ص	-	ص ض	ص ض
9	ط	ظ ط	-	ط ظ	ط ظ
10	ع	غ ع	-	ع غ	ع غ
11	ف	ڦ ف	-	ف	ف
12	ق	ق	-	ق	ق
13	ک	ک	-	-	-
14	ک	ک	ک	ک	ک
15	گ	ڳ گ ڱ گ	گ	گ	ک
16	ل	ل	ل	ل ڵ	ل
17	م	م	م	م	م
18	ن	ڻ ن	ن	ن ں	ن ڼ
19	و	و	و	و	و
20	ھ	ھ	ھ	-	-
	-	-	ہ	ہ	ہ

Table 3. Regional languages characters with dots, without dots, and positions

Characters	Sindhi	Balochi	Brahvi	Pashto	Dots position
Zero dot characters	ح د ر س ص ا ھ ڳ ک ط ع ک گ ل م و ء	ا ح د ر س ک گ ل م و ھ ء ی ے	ا ح د ر س ص ط م ع ک گ ل ں و ہ ء ی ے	ح ا د ر س ص ط ع ک ل ہ ۂ ے	Without dots and
One dot characters	ض ج ه ب ج خ ذ ز ظ غ ف ں	ب ز ں	ب ج ذ ز ض ظ غ ف ں	ب ج خ ذ ز ض ظ غ ں ف	Above Below inside
Two dot characters	ڊ ڏ ڃ ٺ ت پ گ ي ڳ ڃ ق	ت	ت ق	ت ٻ ښ ي ى	Above Below Inside
Three dot characters	پ ٽ ڃ ڈ ش ٿ	پ ڃ ش	ٹ پ ڃ ڑ	پ ٹ څ ڃ ڑ	Above Below Inside
Four dot characters	پ ٽ ڃ ٿ ڙ ڤ				Above Below
Diacritics characters	ٺ	ٹ ڈ	ٹ ڈ ڵ ڑ	خ ډ ړ ک ی	Above Below
Diacritics and dot				ټ ڼ	Above Below

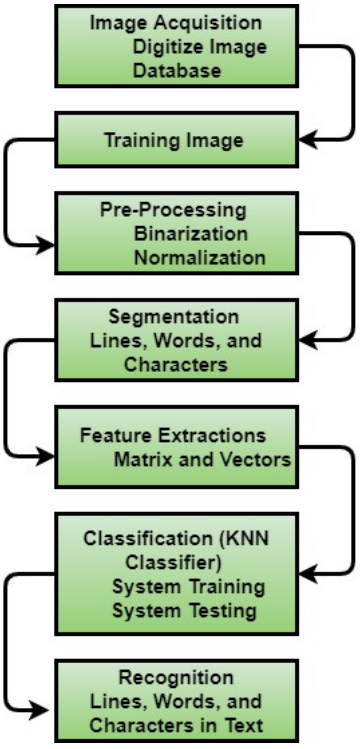

Fig 7. General OCR framework for the regional languages

rithms such as image binarization, thresholding [17] and image normalization etc. The process of transformation input image into gray levels and binary image is the first preprocessing operation. There are two gray levels values (0,255) for monochrome images and then two binary levels (0,1) for black and white image. The cut off value can be obtained using thresholding preprocessing technique for the binary image. In simplified terms, the thresholding [17] is cut off value of the binary image. Preprocessing operations involve noise removal, smoothing, edge detection [18] [9], slant correction and skew detection [19], image normalization [20] [5], thinning [21] [22] or keletonization, and baseline detection [23] [24]. The process of separating [25–27] and partitioning the images is known as segmentation [28] [29]. Segmentation is the still challenging and complex problem in the field of image processing. The segmentation algorithms divide the whole image into sub images with help of different segmentation algorithm into lines, words, and then characters. These segmentation algorithms provide the regions of interest [8] (ROI) for the recognition of isolated images.

The properties and characteristics of the segmented objects are the important features and information is in the form numerical values is the feature extraction technique. The shape, pattern, intensity distributions, and texture are features of an identified and isolated image after segmentation of an image. The features refer to the numerical values in the form of vectors. The extracted features of an isolated image are given to the classifies for the recognition and classifications. The classifier such as Neural Networks [30], and Convolution Neural Networks [31] (CNN) will identify and recognize the class of objects and image. The machine-readable image will be classified and recognized using state of art classifier such as Support Vector Machine [32] (SVM) into human readable format. The classification is sometimes being referred to as recognition. The classification and recognition terms can be used as interchangeably. After classification, the image is recognized from machine readable format to human readable format. The process of identifying object using classification technique is known as recognition. Use of deep neural network is more preferable to use in multilingual OCR as it is already very successful in different areas including healthcare [33], face verification and person identification [34].

5 Design Materials and Methods

Balochistan is multilingual province in the region of Pakistan. There are numerous languages spoken in Pakistan but few regional languages such as Balochi, Brahvi, Pashto, and Sindhi has not gained a due attention in the field of character recognition. There is a need to survey and identify problem related to the cursive OCR languages. Balochi, Brahvi, Pashto and Sindhi languages are

cursive languages. These cursive languages have not gained popularity in the field of computer vision and pattern recognition. Our focus is to determine the problems and issues in these regional languages in the context of OCR. The goal of OCR is to develop a system that reads these cursive regional languages and transform into machine readable format. The most commonly used commercial languages such as Chinese, Japanese, and English has gained popularity in computing field. In these regional languages, The Pashto and Sindhi has gained little attention but Balochi and Brahvi still has not gained attention in computer vision and image processing. In order to identify the viability of multilingual OCR system, a tiny study is conducted. The study population comprised of 339 participants with different linguistic background. The number of participants and their mother tongues is shown in Figure 8.

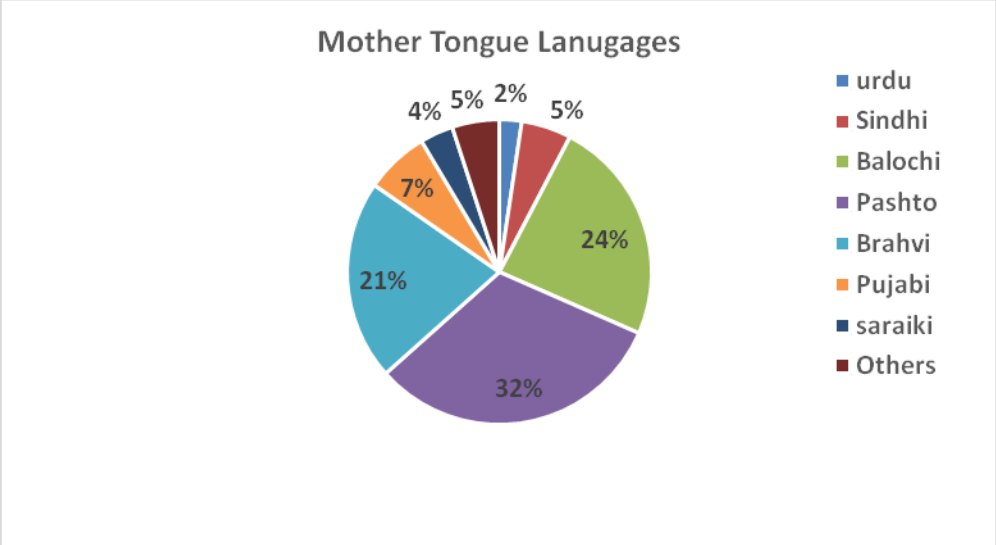

Fig 8. Mother tongues of participants

During survey most of the natural language speakers participated in the study and this indicates the wideness of study. During study, the participants were asked that which natural language they want to use in the accomplishment of their computational task. The feedback received from the participants is shown in Figure 9.

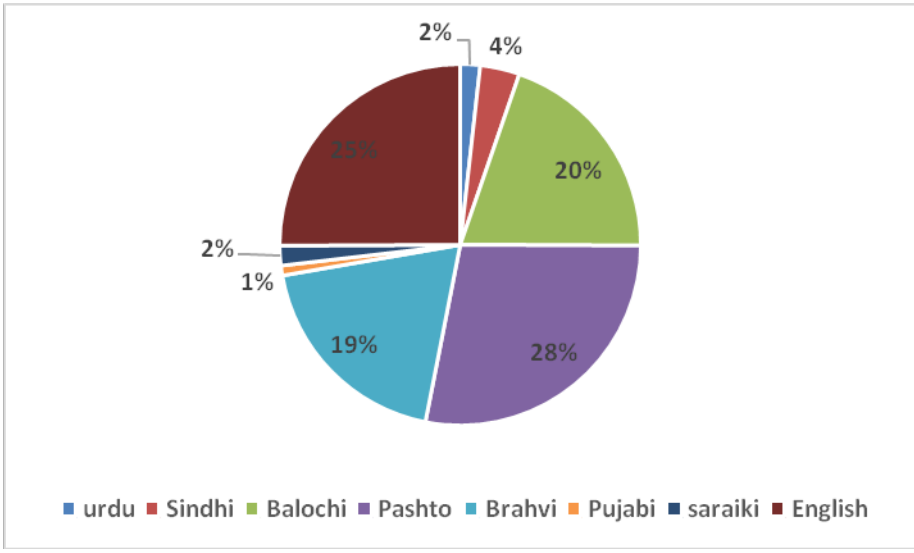

Fig 9. Interest of using natural language in computational domain

It can be seen that a large of participants are keenly willing to use their natural language in computational domain which

natural indicates the scope and viability of multilingual OCR system. During survey the participants were asked whether the use of natural language in computer would simplify their work. The feedback received from the participants is shown in Figure 10.

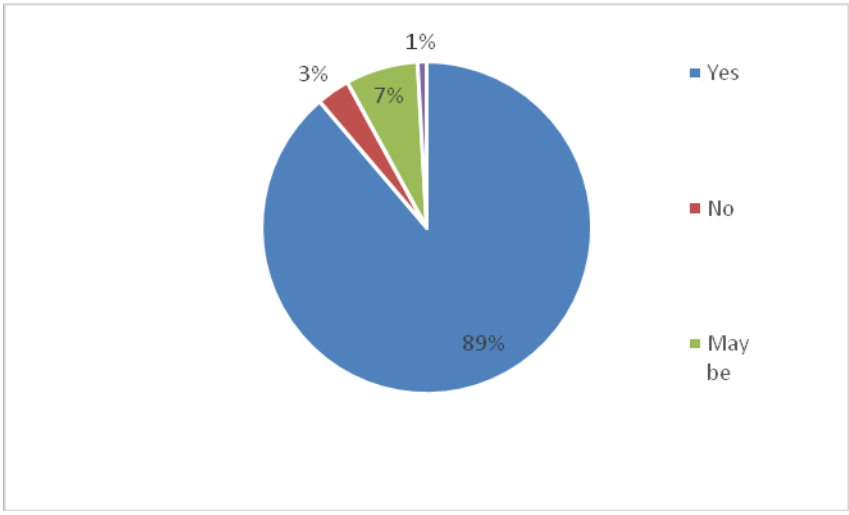

Fig 10. Participant response on the impact of naturallanguage

During study the participants are also asked whether the use of software that support multiple natural languages would simplify the work and increase their performance/productivity. The feedback received from the participants is shown in Figure 11.

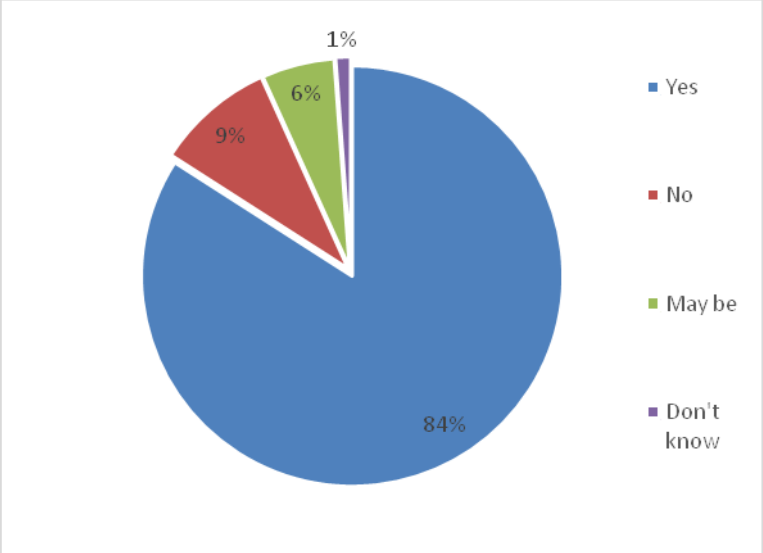

Fig 11. Participant response on the significance of multilingual applications

The overall feedback received from the participants indicates that the use of multilingual application is need of current era and thereby the multilingual OCR is extremely important and could improve the performance of users.

6 Conclusion and Future Work

In this study of survey paper, we studied regional languages such Balochi, Brahvi, Pashto, and Sindhi. Sindhi and Pashto have been considered for OCR but still there is no commercial system for recognizing these languages. However, there is no research in Balochi and Brahvi in the context of optical character recognition field of computing. Future work is continued in three ways: i) development of unified computable OCR framework of all the regional languages of Pakistan ii) Use of deep learning for the

recognition of multi-lingual OCR system iii) Bachmann–Landau analysis of multi-lingual OCR systems. In this study, we also combine OCR technology and Text to Speech Synthesizer (TSS) for multilingual characters and regional languages for visually impaired people using computer through voice interaction in the future work.

Acknowledgement

We would like to thanks faculty members and friends of Department of Balochi, Department of Brahvi, and Department of Pashto for support and suggestions

References

1) Chandio AA, Leghari M, Leghari M, Jalbani AH. Multi-Font and Multi-Size Printed Sindhi Character Recognition using Convolutional Neural Networks. *Pakistan Journal of Engineering and Applied Sciences*. 2019;24(1).
2) Shyni SM, Raj MAR, Abirami S. Offline Tamil Handwritten Character Recognition Using Sub Line Direction and Bounding Box Techniques. *Indian Journal of Science and Technology*. 2015;8(S7):110–110. doi:10.17485/ijst/2015/v8is7/67780.
3) Mabee C. 2012. Available from: https://dev.panlex.org/wp-content/uploads/2014/03/ocr-survey.pdf.
4) Jyothi J, Manjusha K, Kumar MA, Soman KP. Innovative Feature Sets for Machine Learning based Telugu Character Recognition. *Indian Journal of Science and Technology*. 2015;8(24). doi:10.17485/ijst/2015/v8i24/79996.
5) Abu-Mostafa YS, Psaltis D. Image Normalization by Complex Moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1985;PAMI-7(1):46–55. doi:10.1109/tpami.1985.4767617.
6) Baloch HA. 1997.
7) Hashmi SZ. 2010.
8) Achanta R, Estrada F, Wils P, Süsstrunk S. Salient region detection and segmentation. In: International conference on computer vision systems. Springer. 2008;p. 66–75.
9) Canny J. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1986;PAMI-8(6):679–698. Available from: https://dx.doi.org/10.1109/tpami.1986.4767851.
10) Sanjrani AA, Baber J, Bakhtyar M, Noor W, Khalid M. Handwritten optical character recognition system for Sindhi numerals. *In2016 International Conference on Computing, Electronic and Electrical Engineering*. 2016;p. 262–267.
11) Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2000;22(8):888–905. Available from: https://dx.doi.org/10.1109/34.868688.
12) Jumani AK, Memon MA, Khoso FH, Sanjrani AA, Soomro S. Named entity recognition system for Sindhi language. In: International conference for emerging technologies in computing. Springer. 2018;p. 237–246. doi:doi.org/10.1007/978-3-319-95450-9_20.
13) Shujra AA, Rajper S, Jumani AK. Measurement of E-learners' level of interest in online course using Support Vector Machine. *Indian Journal of Science and Technology*. 2019;12(40):1–9. Available from: https://dx.doi.org/10.17485/ijst/2019/v12i40/147265.
14) Laghari AA, He H, Shafiq M, Khan A. Assessment of quality of experience (QoE) of image compression in social cloud computing. *Multiagent and Grid Systems*. 2018;14:125–143. Available from: https://dx.doi.org/10.3233/mgs-180284.
15) Karim S, Zhang Y, Laghari AA, Asif MR, IEEE. Image processing based proposed drone for detecting and controlling street crimes. In: and others, editor. IEEE 17th International Conference on Communication Technology (ICCT). 2017;p. 1725–1730. doi:https://doi.org/10.1109/ICCT.2017.8359925.
16) Siddiqui MF, Siddique WA, Jumani AK, Ahmed M. Face Detection and Recognition System for Enhancing Security Measures Using Artificial Intelligence System. *Indian Journal of Science and Technology*. 2020;13(09):1057–1064. Available from: https://dx.doi.org/10.17485/ijst/2020/v013i09/149298.
17) Otsu N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*. 1979;9(1):62–66. Available from: https://dx.doi.org/10.1109/tsmc.1979.4310076.
18) Perona P, Malik J. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1990;12(7):629–639. Available from: https://dx.doi.org/10.1109/34.56205.
19) Tsao YF, Fu KS. A parallel thinning algorithm for 3-D pictures. Elsevier BV. 1981. Available from: https://dx.doi.org/10.1016/0146-664x(81)90011-3. doi:10.1016/0146-664x(81)90011-3.
20) Finlayson GD, Schiele B, Crowley JL. Comprehensive colour image normalization. In: European conference on computer vision. Springer. 1998;p. 475–490.
21) Zhang TY, Suen CY. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*. 1984;27(3):236–239. Available from: https://dx.doi.org/10.1145/357994.358023.
22) JUMANI AK, MAHAR MH, KHOSO FH, MEMON MA. Online Text Categorization System Using Support Vector Machine. *SINDH UNIVERSITY RESEARCH JOURNAL -SCIENCE SERIES*. 2018;50(001):85–90. Available from: https://dx.doi.org/10.26692/surj/2018.01.0014.
23) Matas J, Chum O, Urban M, Pajdla T. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*. 2004;22(10):761–767. Available from: https://dx.doi.org/10.1016/j.imavis.2004.02.006.
24) Pan J, Tompkins WJ. A Real-Time QRS Detection Algorithm. *IEEE Transactions on Biomedical Engineering*. 1985;BME-32(3):230–236. Available from: https://dx.doi.org/10.1109/tbme.1985.325532.
25) Ibrar M, Mi J, Karim S, Laghari AA, Shaikh SM, Kumar V. Improvement of Large-Vehicle Detection and Monitoring on CPEC Route. Springer Science and Business Media LLC. 2018. Available from: https://dx.doi.org/10.1007/s13319-018-0196-5. doi:10.1007/s13319-018-0196-5.
26) Karim S, Halepoto IA, Manzoor A, Phulpoto NH, Laghari AA. Vehicle detection in Satellite Imagery using Maximally Stable Extremal Regions. *International Journal of Computer Science and Network Security*. 2018;18(4):75–78. Available from: http://paper.ijcsns.org/07_book/201804/20180413.pdf.
27) Karim S, Zhang Y, Yin S, Laghari AA, Brohi AA. Impact of compressed and down-scaled training images on vehicle detection in remote sensing imagery. *Multimedia Tools and Applications*. 2019;78:32565–32583. Available from: https://dx.doi.org/10.1007/s11042-019-08033-x.
28) Morar A, Moldoveanu F, Gröller E. Image segmentation based on active contours without edges. *In2012 IEEE 8th international conference on intelligent computer communication and processing*. 2012;p. 213–220. doi:10.1109/ICCP.2012.6356188.
29) Sun C, Si D. Skew and slant correction for document images using gradient direction. *Proceedings of the Fourth International Conference on Document Analysis and Recognition*. 1997;1:142–146.

30) Haykin S. A comprehensive foundation. *Neural networks*. 2004;2:41–41.
31) Krizhevsky A, Sutskever I, Hinton GE. In Advances in neural information processing systems. *Imagenet classification with deep convolutional neural networks* . 2012;p. 1097–1105. Available from: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.
32) Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning. Springer. 1998;p. 137–142.
33) Norgeot B, Glicksberg BS, Butte AJ. A call for deep-learning healthcare. *Nature Medicine*. 2019;25:14–15. Available from: https://dx.doi.org/10.1038/s41591-018-0320-3.
34) Naz S, Hayat K, Razzak MI, Anwar MW, Madani SA, Khan SU. The optical character recognition of Urdu-like cursive scripts. *Pattern Recognition*. 2014;47:1229–1248. Available from: https://dx.doi.org/10.1016/j.patcog.2013.09.037.