ISSN 2278-3091
Volume 10, No.3, May - June 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse201032021.pdf
https://doi.org/10.30534/ijatcse/2021/201032021

Sindhi Handwritten Text Recognition Using SVM

Shafique Ahmed Awan[1], Fida Hussain Khoso[2], Aijaz Ahmed Arain[3], Abdullah Lakhan[4], Shah Zaman Nizamani[5], Kirshan Kumar Luhana[6]
[1,4]Benazir Bhutto Shaheed University Lyari Karachi Sindhi Pakistan; shafique.neduet@gmail.com abdullahrazalakhan@gmail.com
[2]Deptt.of Basic Science, Dawood University of Engineering & Technology Karachi, Pakistan; fidahussain.khoso@duet.edu.pk ,
[4]Department of Computer Science, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Pakistan; aijaz@quest.edu.pk
[5]Department of Information Technology, Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Pakistan; shahzaman@quest.edu.pk
[6]Department of Computer Science, University of Sindh Jamshoro,Pakistan: kirshan.Luhano@usindh.edu.pk

ABSTRACT

In Sindhi Language, handwritten text feature extraction is such a challenging task for all scholars, because different people write in different styles or manners, to analyze each text is such a complex problem. Feature extraction of text segmentation, classifying each character and labelling for training data to recognize text for different handwritings and testing for analyzing features of providing handwritten text data .In this research, SVM (support vector machine) is used for analyzing and tokenizing each character or word of Sindhi Language text and transform into suitable information with efficiency & accuracy. The research is not only useful for improving the knowledge of Sindhi Handwritten Text Recognition but it can be beneficial for other recognition systems

Key words: Handwritten text recognition, SVM, feature extraction, python, Sindhi language, native language.

1. INTRODUCTION

Sindhi Language (SL) is an Indo-Aryan Language spoken by Sindhi people. Many researchers assume that the history of Sindhi Language was about 15000 BC back [1]. The strong root of SL is with rich culture, tradition, and civilization. In Pakistan, Sindhi is an official language in the province of Sindh. In a Report of 2007, there are 25 million people speak or write Sindhi language all over the world [2].

Intelligent Characters Recognition (ICR) is an upgraded technology of Optical Characters Recognition (OCR). The main difference between OCR & ICR: OCR works on printed characters recognition while ICR is used to analyze handwritten text characters recognition. In this research, we are just focusing on ICR because handwritten text recognition of Sindhi Language. The major problem arises to write any text in SL is because that SL contains lots of dots or points (Nokta is a Turkish word which means dot or point) to write a perfect sentence. If we Neglect these points in our written text then the meaning of sentence could be changed and pronunciation also makes no sense. Analyzing these dots without ignoring it, we need a strong algorithm which have an ability to understand each character without discarding anything. Support Vector Machine (SVM) is a machine learning technique for supervised learning, which belongs to classification and regression technique. Most of the text analysis researchers use SVM for text recognitions. The question is why we use this algorithm in our research? SVM provides us a complete understanding of text recognition without discarding dots or points. By this act, we get accurate result with less complexity and minimum amount of time. In the next paragraph, we discuss some related literature and write some critical reviews of the research.

Claus Bahlmann et.al elaborate the importance of SVM in handwritten text recognition. In this research they create a model called kernel SVM (Gaussian dynamic time warping) which recognizes text and compares it with HMM for accuracy [3]. Djeddi, Chawki, et al intended a unique technique, which is based on independent text of multi scripting. Handwritten text recognition of Greek and English language use KNN and SVM algorithms. The idea is to analyze 126 writers of different hypothesis and compare writing technique of each sample with each other. By this we can easily identify that which writer writes this paper or script. Make four samples for each writer. The short writing implies to resemble in real life environment, forensic experts use such technique to identify short piece of text which can easily be recognized by a writer [4].

Muhammad Tanvir Parvez et.al describe a comprehensive survey on handwritten text recognition of Offline Arabic

Language. There are lots of classifications and feature detection algorithms are available like: HMM, SVM, KNN, NN etc. But in recent database of Arabic Language like: Sadri et.al [6], Alaei et.al [7], Alamri et.al [8] these all scholars used SVM for handwritten text recognition of Arabic Language. By using SVM, the result describes overall efficiency as-well-as accuracy rate, which is too higher than other algorithms [5] [6] [7] [8]. Another approach applying on Offline Arabic Handwritten Text Recognition Database is the combination of CNN capacities and SVM classification. This system is unique and it is provides more accurate results than traditional classification approaches [9]. In Bangla handwritten text recognition using HMM but for classifying lower and upper zone components are recognized by SVM [10]. SVM also contributes in Online Gujrati Handwritten Text Recognition and finds out the features of text. The system aims to classify handwritten text, 3000 training sample datasets are given by the system and test by 100 different writers [11]. John et.al define the importance of SVM in Malayalam handwritten text [12]. So, it is clear that SVM contributes more in handwritten text recognition and it is widely used in different perspective of different languages. The accuracy & efficiency rate is very high that is why it is more suitable for applying on Sindhi Language.

2. METHEDOLOGY

The proposed solution describes each and everything in detail. It starts with the input given by the user manually handwritten text of SL. Gathering users input text with the help of scanner, preprocessing is the first step which removes noise, skew detection, boundaries correlation and normalized text characters for further processing. After preprocessing, segmentation is the second step, segmentation is used for finding coherent objects. In segmentation, there are number of segmentation methods like line segmentation of text recognition, word segmentation, ligature segmentation, character segmentation. In this scenario, each paragraph is tokenized into words and each word is tokenized into characters. These characters are classified easily by using different algorithms. RGB is converted into color scanned image in gray scale (0-255 number of color shades), thresholding finds the intensity level of each word or characters, which contains high pixels rate or less pixels rates. The third step is feature extraction, which is the crucial part where we find projection, angle, movement variation, skeletonizing, correlation. The objective of feature extraction, selecting important feature helps to recognize handwritten text of SL. Finally, classification or recognition can possibly be matched maximum values by the help of SL handwritten text database (a training dataset). In testing each word or character is recognized accurately with minimum amount of time and gives editable Sindhi text as an output (shows in figure1)

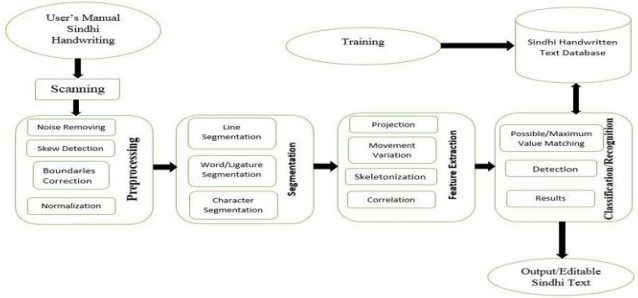

Figure 1: Methodology for Sindhi Handwritten Text Recognition

3. ALGORITHM

Following the feature extraction, the most important job is to classify and recognize the words. It could be possible by using SVM machine technique. The final step in An Intelligent Offline Sindhi Characters Recognition (ICR) System is the recognition or classification of the characters where characters are recognized by using various available classifiers. Classification is a common technique for categorization, where different characters or words can be recognized and differentiated and understood. Furthermore, categorization is a process through which objects are sorted and classified. There are diverse numbers of classifiers that can be applicable for word recognition. The final step in an Intelligent Offline Sindhi Characters Recognition (ICR) System which is the recognition or classification of the characters where characters are recognized by using various available classifiers. Here researcher used the Support Vector Machine (SVM) for supervised learning. SVM is the best approach for classification and recognition because SVM does not need many samples. Through SVM different numbers of Sindhi words and characters can be recognized.

4. SUPPORT VECTOR MACHINE (SVM)

The most important and an efficient classifier is Support Vector Machine (SVM) for supervised learning. SVM represents a class of very dominant, broadly used tool, that have been effectively applied for prediction, classification and clustering problems. It can be used to recognize Sindhi characters. It is exceptionally useful to numerous example order issues, for example, image recognition, speech recognition, text recognition, face detection, and faulty card detection. SVM fits a function (hyperplane) that attempts to separate two classes of data that could be of multiple dimensions. Prior training, collected data is randomly divided into the training set and test sets and the ratio of training testing data is 70% and 30%. When training phase is completed the SVM has power. Once the training phase is

complete, the SVM has ability to test the rest of data. Testing of data is done against the training data. Here we used the supervised learning. Each data has been labelled by using the supervised learning. Using SVM approach each word and character are easily mapped as point in space, and finally a clear gap available which is widely acceptable. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

5. TRAINING AND TESTING

In machine learning, the training and testing of the machine is very important to recognize and predict the available data. The final model has been developed from the different multiple datasets. The training and testing of supervised data is easy as compared to unsupervised data. Training is done by labelling the data. While 30% of text images were used for testing and 70% of text images were used for the training.

6. RESULT

The efficient accuracy rate is achieved by several factors starting from the capable preprocessing, effective's segmentation, competent feature extraction, and finally SVM classifier for recognition. Sindhi Intelligent Characters Recognition (ICR) System has been compared with the different national and international languages.

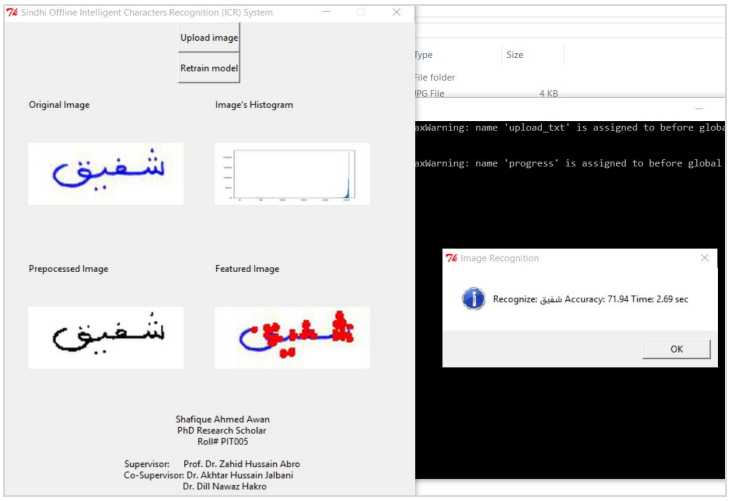

Figure: 2 Recognition of Sindhi Text

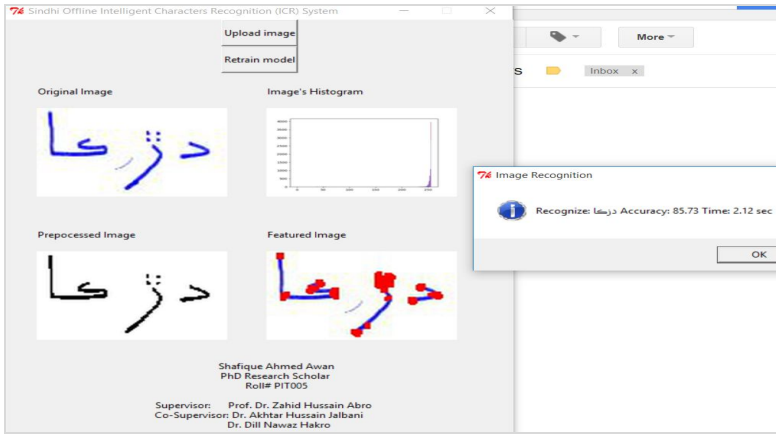

Figure: 3 Recognition of Sindhi Text

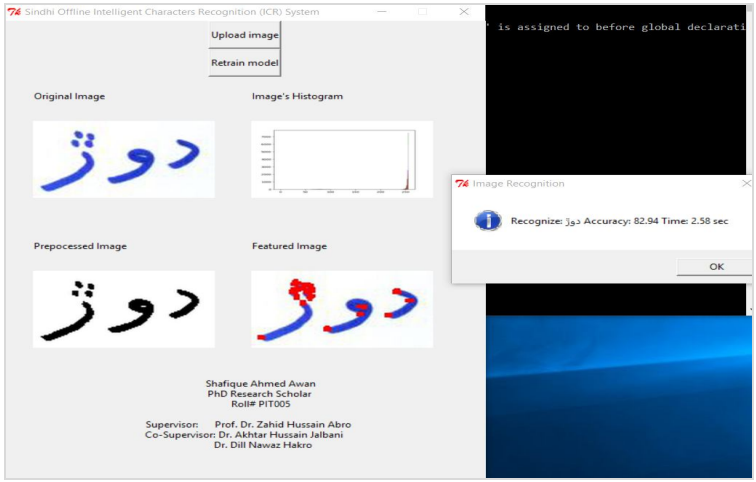

Figure: 4 Recognition of Sindhi Text

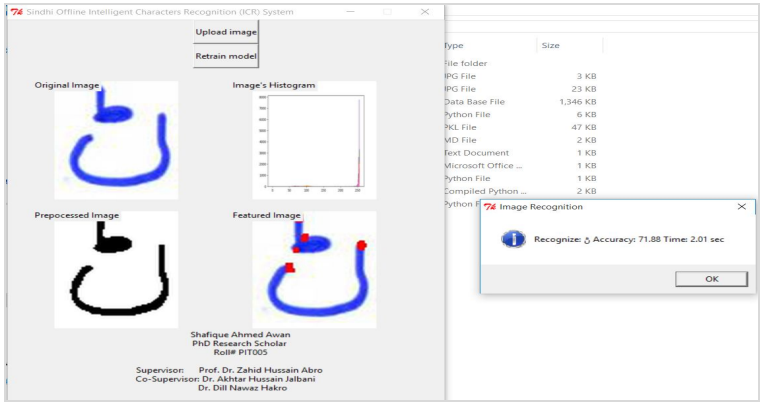

Figure: 5 Recognition of Sindhi Characters

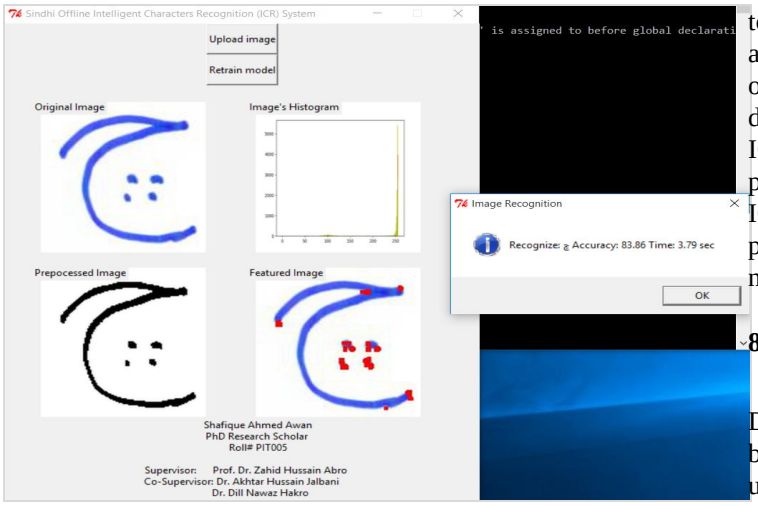

Figure 6: Recognition of Sindhi Characters

When Sindhi ICR System has been compared with Arabic language, rate of recognition is different forms shows in Table1.

Table 1. Difference between the Sindhi and Arabic Recognition System

Recognized	Arabic Text Recognition	Proposed Sindhi Handwritten Text Recognition	Difference
Word	81%	73%	-8%
Ligatures	91%	78%	-13%
Ligature and Isolated characters	93%	82%	-11%
Isolated Characters make words	96%	81%	-15%
Characters	93%	88%	-5%

7. SCOPE & LIMITATION

Intelligent Character Recognition (ICR) is a functioning Area of research. This Sindhi ICR system is the bottom for many diverse types of applications in various fields, which can be used for day to day activities. Sindhi ICR System can be used to recognize the cheques signatures and users writing. This application can be used in post office to recognize the address of sender and receiver. Sindhi ICR systems can reduce the data entry time, storage space required by documents. Sindhi ICR System can facilitate to recognize the Sindhi number plates. Handwritten forms can be processed by using Sindhi ICR system. Sindhi handwritten legal documents can be processed. Sindhi ICR System can recognize only limited number of writer and data because it is supervised learning.

8. CONCLUSION

Different other hand written text recognition systems have been studied. Firstly we have collected the data from single user and with same handwriting. About 500 samples were collected from one writer and finally text images were created using the form filling. The given images are segmented by different techniques and then feature extractions have been done on these images. SVM classifier has been used using the supervised learning. Labelling of datasets are acheived by SVM. SVM has the best result on supervised learning. Sindhi Text images have been categorized into 70 and 30 ratio while 70% of data have been used training and 30% of testing. 83% accuracy rate has been achieved on Sindhi characters and 77% on Sindhi text. Accuracy rate can be improved by improving the segmentation.

9. FUTURE WORK

Sindhi ICR System can help to recognize the handwritten Shah-jo-Resalo. ICR System can be further modified for betterment using the deep learning. It can be modified to recognize the answer sheet of university students by using the Convolutional Neural Network (CNN) or caps net. It can be modified to recognize the grammar checking of Sindhi Language. ICR System recognizes the paragraph and sentence of each script. Multiple subject's handwriting can be developed by considering the proposed research.

REFERENCES

1. History of Sindhi Language, **Accredited,** link:https://www.alsintl.com/resources/languages/Sindhi/
2. National encyclopedia in Varldens 100 storsta sprak 2007 **The World's 100 Largest Languages in 2007.**
3. Bahlmann, Claus, Bernard Haasdonk, and Hans Burkhardt. **Online handwriting recognition with support vector machines-a kernel approach**. Frontiers in handwriting recognition, 2002. Proceedings. Eighth international workshop on. IEEE, 2002.
4. Djeddi, Chawki, et al. "**Text-independent writer recognition using multi-script handwritten**

texts." Pattern Recognition Letters 34.10 (2013): 1196-1202.

5. Parvez, Mohammad Tanvir, and Sabri A. Mahmoud. **"Offline Arabic handwritten text recognition: a survey."** ACM Computing Surveys (CSUR) 45.2 (2013): 23.

6. Alaei, A., Pal, U., and Nagabhushan, P. 2009a. **Using modified contour features and SVM based classifier for the recognition of Persian/Arabic handwritten numerals.** In proceedings of the 7th International Conference on Advances in Pattern Recognition (ICAPR). 391–394.

7. Alaei, A., Nagabhushan, P., and Pal,U. 2009b. **Fine Classification of Unconstrained Handwritten Persian/Arabic Numerals by Removing Confusion Amongst Similar Classes.** In proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR). 601–605.

8. Alamri, H., Sadri, J., Suen, C., and Nobile, N. 2008. **A Novel Comprehensive Database for Arabic Off-Line Handwriting Recognition.** In proceedings of the 11th International Conference on Frontiers in Handwriting Recognition (ICFHR). 664–669.

9. Elleuch, Mohamed, Rania Maalej, and Monji Kherallah. **"A new design based-SVM of the CNN classifier architecture with dropout for offline Arabic handwritten recognition."** Procedia Computer Science 80 (2016): 1712-1723.

10. Bhunia, Ayan Kumar, et al. **"A comparative study of features for handwritten Bangla text recognition."** Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. IEEE, 2015.

11. Naik, Vishal A., and Apurva A. Desai. **"Online handwritten Gujarati character recognition using SVM, MLP, and K-NN."** Computing, Communication and Networking Technologies (ICCCNT), 2017 8th International Conference on. IEEE, 2017.

12. John, Jomy, K. V. Pramod, and Kannan Balakrishnan. **"Unconstrained handwritten Malayalam character recognition using wavelet transform and support vector machine classifier."** Procedia Engineering 30 (2012): 598-605.