International Journal of Trend in Scientific Research and Development (IJTSRD)

International Open Access Journal

ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 1 | Issue – 5

An Efficient OCR System based on the Regional Feature using the ASVM as Classifier

Maninder Kaur
Research Scholar, ECE Department,
Rayat and Bahra University, Mohali

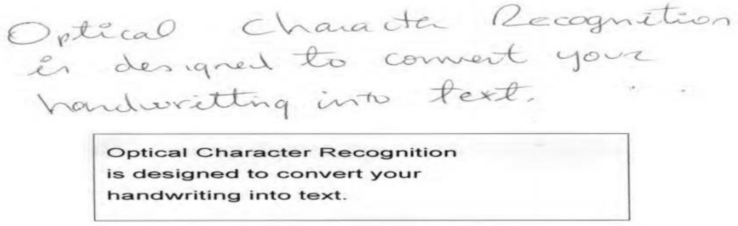
Ms. Manjeet Kaur
Assistant Professor, ECE Department,
Rayat and Bahra University, Mohali

ABSTRACT

In Image Processing, sometimes due to poor handwriting, the writer left some gap between diacritics and character or between diacritics and header line due to which small text blocks gets created which leads to improper text line segmentation and hence leads to wrong results and overlapping. As a result accuracy of the algorithm degrades. In proposed work Adaptive SVM will be used to improve accuracy of the system.

INTRODUCTION:

Optical character recognition which is also commonly known as optical character reader is the process of converting the mechanical and electronic images into the handwritten, printed text etc.OCR is a course by which focused software is used to change the skimmed pictures of manuscript to electronic text so that digitized data can be examined, indexed and recovered. The OCR are basically design to settled and improved the multiple real world applications such as mining data from business documents, checks, passports, invoices, bank statements, insurance documents, license plates etc. Each and every application contains the processing data sets that contains the hundreds and thousands of scanned documents of the images in order to train and enhance the systems and the processing of the drill data set is naturally done by humans in order to provide accurate data that can be used by the engine to learn and apply, makes it smarter. OCR is mainly used as a form of information entry from printed paper data records, whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation. OCR is mostly used in the area of the computer vision, artificial intelligence, and pattern recognition. Handwriting text recognition (HTR) can be defined as the ability of a computer to transform handwritten input represented in its spatial form of graphical marks into equivalent symbolic representation as ASCII text. Usually, this handwritten input comes from sources such as paper documents, photographs or electronic pens and touch-screens.



Optical Character Recognition
is designed to convert your
handwritting into text.

Optical Character Recognition
is designed to convert your
handwriting into text.

Fig.1 Handwritten document Convert into Text

Approaches for learning Optical character recognition:-The following are the approaches for learning Optical character recognition.

- **Histogram Approach:-** This method is based on the pixel histogram in which a Y-histogram forecast is achieved which results in text line position and to divide the line into different areas a threshold is applied.

- **Header line and base line detection based method: -**This method calculates the header line and base line of a text document for line segmentation.
- **Line Segmentation:** This algorithm is based on the projection profile method. This algorithm professionally pacts with skewed text as well as with the overlapped and touched text lines.
- **Character Segmentation:** To extract characters with overlapping, this method helps to removes the vowel converters and consonant transformers. By removing the consonant and vowel converters, the word image contain only the base characters with clear paths between them.
- **Overlapping and Touching of Characters:-** Due to overlapping and touching of characters, there may contains no important break between text lines and when the two or more text lines comes in a same text block then they leads to wrong results.

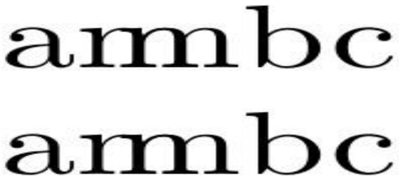


Fig.2 Overlapping of Character

Machine learning techniques used in hand written recognition:

Artificial Neural Network (ANN) An Artificial Neuron is basically an engineering approach of biological neuron. ANN consists of a number of nodes, called neurons. Neural networks are typically organized in layers. In neural network each neuron in hidden layer receives signals from all the neurons in the input layer. The strength of each signal and the biases are represented by weights and constants, which are calculated through the training phase. After the inputs are weighted and added, the result is then transformed by a transfer function into the output.

Support Vector Machine: SVM is a non-linear classifier which is now mostly new in the machine learning which is used to solve the texture classification and pattern recognition problems.SVM is designed to work with only two classes by determining the hyper plane to divide two classes and the separation of these classes is performed by using different Kernels.

PROPOSED METHODOLOGY

OCR involved various steps to read the characters from a scanned Image. In proposed research, a model has been built for handwritten images. The system extracts the characters from handwritten images and writes into text file.

Flow Chart OF OCR Model

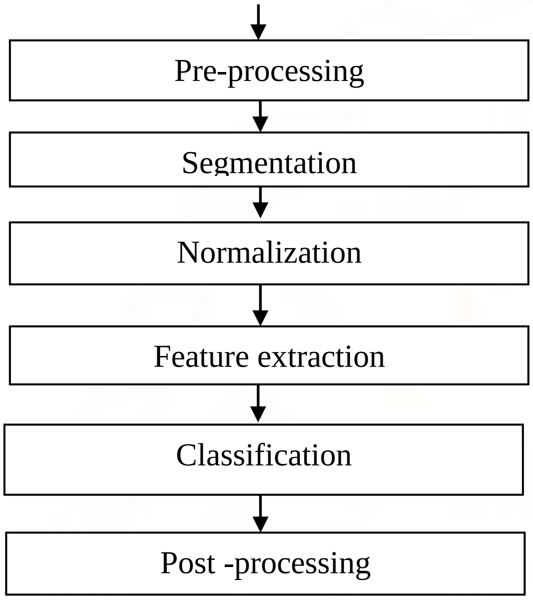


Fig.3 Steps of OCR System

Data Acquisition: Most Important initial phase in OCR is to gather the image from either device sensor like PDA or tablets in case on online recognition or getting the images containing characters directly for offline recognition. The image should have a specific format such as JPEG, BMP etc.

Pre Processing: The goal of pre-processing is to simplify the pattern recognition problem without missing any vital information. It reduces the noises and inconsistent data. It enhances and prepares it for the next steps.

Segmentation: Segmentation is an integral part of any text based recognition system. It assures efficiency of classification and recognition. Accuracy of character recognition heavily depends upon segmentation phase.

Normalization: The results of segmentation process provides isolated characters which are ready to pass through feature extraction stage, thus the isolated characters are reduced to a specific size depending on the methods used. The segmentation process essentially renders the image in the form of m*n matrix.

Feature Extraction: Feature extraction is the process of extracting the relevant features from objects/alphabets to form a feature vectors. These feature vectors is then used by classifiers to recognize the input unit with target output unit Feature extraction methods are based on 3types of features:

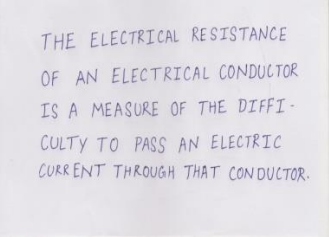
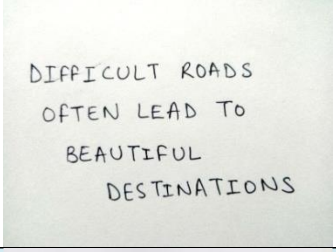
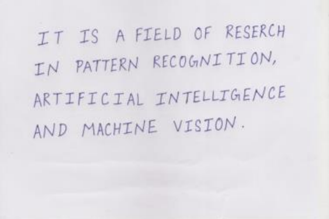
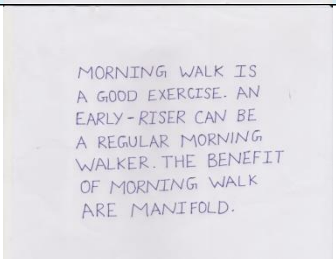
- Statistical
- Structural
- Global transformations and moments

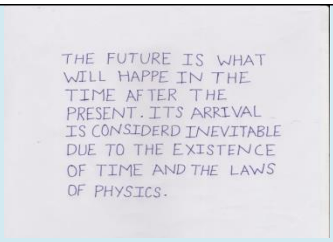
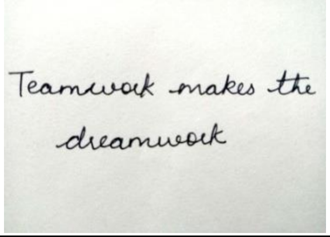
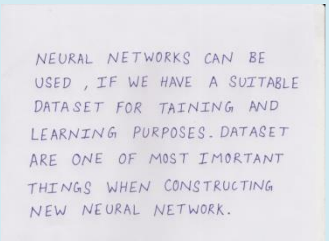
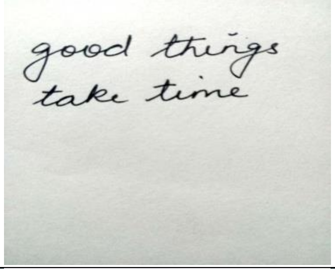
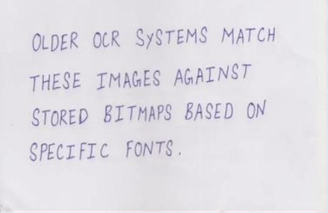
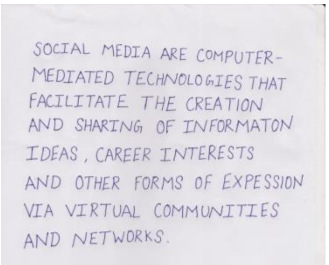
Classification: The results Classification is the last stage where we train the neural net using the feature vectors obtained during feature extraction method against the required targets

Post Processing: The goal of post processing is the incorporation of context and shape information in all the stages of OCR systems is necessary for meaningful improvements in recognition rates.

Feature extraction: The following is the feature matching and classification algorithm for matching the extracted plant disease image with the different images of same plant, which are taken at different times, from different viewpoints, or by different sensors.

RESULT AND DISCUSSION:

No.	Images	Real Text	Recognized Text	Accuracy (%)
1	 THE ELECTRICAL RESISTANCE OF AN ELECTRICAL CONDUCTOR IS A MEASURE OF THE DIFFI- CULTY TO PASS AN ELECTRIC CURRENT THROUGH THAT CONDUCTOR.	The electrical resistance of an electrical conductor is a measure of the diffi-culty to pass an electric current through that conoctor.	The electrical resistance of an electrical conductor is a measure of the oiffi . Culty to pass an electric current through that conoctor	**96.52**
2	 DIFFICULT ROADS OFTEN LEAD TO BEAUTIFUL DESTINATIONS	Difficult roads often lead to beautiful destinations	Difficult roads often lead to beautiful destinations	**100**
3	 IT IS A FIELD OF RESERCH IN PATTERN RECOGNITION, ARTIFICIAL INTELLIGENCE AND MACHINE VISION.	It is a field of reserch in pattern recognition, artificial intelligence and machine vision.	It is a field of reserch in pattern recognition, artificial intelligence and machine vistion	**97.50**
4	 MORNING WALK IS A GOOD EXERCISE. AN EARLY-RISER CAN BE A REGULAR MORNING WALKER. THE BENEFIT OF MORNING WALK ARE MANIFOLD.	Morning walk is a good exercise. An early-riser can be a regular morning walker.The benefit of morning walk are manifold.	Morning walk is a gooo exercise. An early-rtser can be a regular morning walker the benefii of morning walk are manifold	**95.09**

5	 THE FUTURE IS WHAT WILL HAPPE IN THE TIME AFTER THE PRESENT. ITS ARRIVAL IS CONSIDERD INEVITABLE DUE TO THE EXISTENCE OF TIME AND THE LAWS OF PHYSICS.	The future is what will happe in the time after the present.Its arrival is considerd inevitable due to the existence of time and the laws of physics.	The future is what will happe in the time after the present its arrival is considerd inevitable oue to the extstence of time and the laws of physics.	**97.56**
6	 Teamwork makes the dreamwork	Teamwork makes the dreamwork	Teamwoik mapeo the dremwoik	**80.00**
7	 NEURAL NETWORKS CAN BE USED , IF WE HAVE A SUITABLE DATASET FOR TAINING AND LEARNING PURPOSES. DATASET ARE ONE OF MOST IMORTANT THINGS WHEN CONSTRUCTING NEW NEURAL NETWORK.	Neural networks can be used , if we have a suitable dataset for taining and learning purposes. dataset are one of most important things when constructing new neural network.	Neural networks can be used , if we have a sutable dataset for taining and learning purposes dataset are one of most important things when constructing new neural network	**98.62**
8	 good things take time	Good things take time	Goool things tare time	**83.34**
9	 OLDER OCR SYSTEMS MATCH THESE IMAGES AGAINST STORED BITMAPS BASED ON SPECIFIC FONTS.	Older ocr systems match these images against stored bitmaps based on specific fonts.	Older ocr systems match these images against stored 8itmaps based on specific fonts.	**98.61**
10	 SOCIAL MEDIA ARE COMPUTER-MEDIATED TECHNOLOGIES THAT FACILITATE THE CREATION AND SHARING OF INFORMATON IDEAS , CAREER INTERESTS AND OTHER FORMS OF EXPESSION VIA VIRTUAL COMMUNITIES AND NETWORKS.	Social media are computer-mediated technologies that facilitate the creation and sharing of informaton ideas , career interest and other forms of expesion via virtual communities and networks .	Social media are computer-medtated technoloc-ies that facilitate the creation and sharting of informaton ideas , career interest and other forms of expesion via virtual communities and networrs .	**97.60**
Overall Accuracy of Proposed work				**94.48**

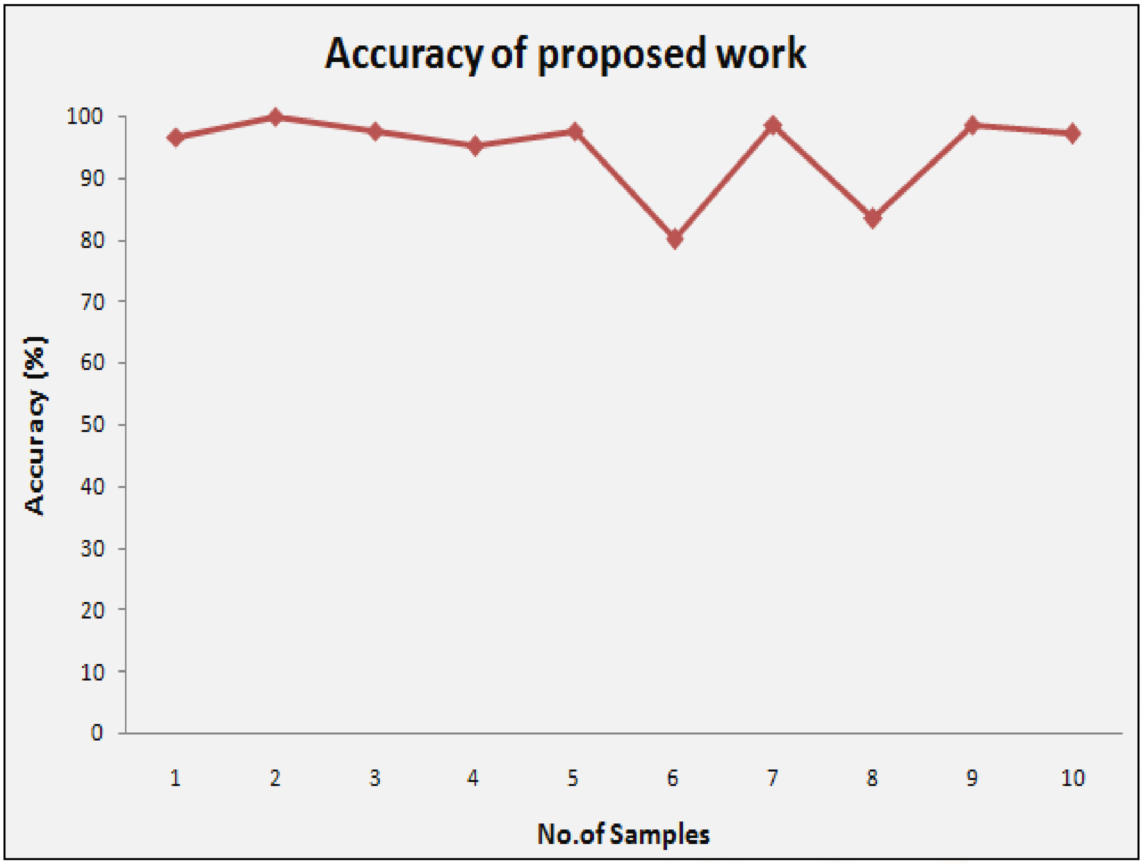


Fig.4 Accuracy of proposed work

In above figure, the accuracy of proposed work is represented in the form of graph. In graph, X-axis denotes the number of samples which are included in the proposed work for the testing and Y-axis denotes the accuracy of proposed work in percentage. Form the above graph; it has been observed that the average percentage of accuracy is more than 94% with handwritten images.

Comparison between Classifier			
Classifier	**ANN**	**SVM**	**ASVM**
Accuracy (%)	**86**	**93**	**94**

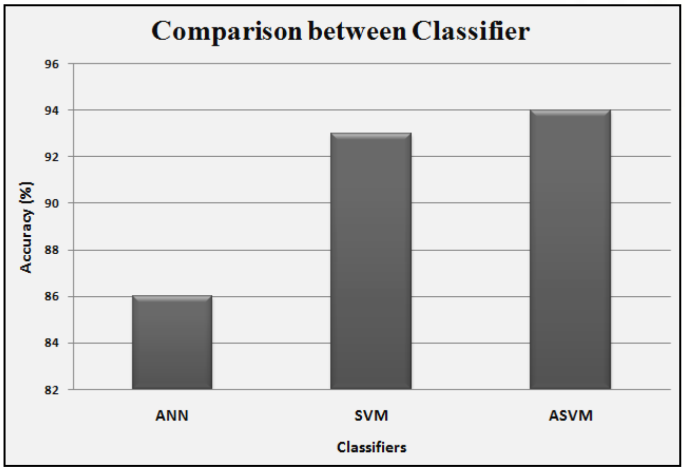


Fig.5 Accuracy comparison between Classifier

In above figure, the accuracy comparison between artificial neural network (ANN), support vector machine (SVM) and adaptive support vector machine (ASVM) is represented in the form of bar graph. From the figure we observe the accuracy of proposed character recognition system with ASVM is better than ANN and SVM classifier due to the best training.

Conclusion: Due to overlapping and touching of characters, there remains no significant gap between the text lines and hence two or more text lines comes in a same text block which leads to wrong results. The main focus in this research project is to experiment deeply with, and find alternative solutions to the image segmentation and character recognition problems within the Overlapped Character Recognition. In the existing work, SVM classifies is applied but it has less accuracy. So in future, Adaptive SVM will be applied to improve better accuracy of the system.

Future work: In future, we can use the artificial neural network along with the optimization algorithm to achieve the better results by minimizing the more noisy data from the images for the character recognition system. The combination of the artificial neural network as classifier instead of SVM with optimization technique the precision of character recognition will have to increase and the rate of noise will decreases.

REFERENCES

1) Chame, Shivadatt D., and Anil Kumar. "Overlapped Character Recognition: An Innovative Approach." Advanced Computing (IACC), 2016 IEEE 6th International Conference on. IEEE, pp.464-469, 2016.
2) Garg N.K., Kaur L., Jindal M.K. “The segmentation of half characters in Handwritten Hindi Text”, Springer Verlag Berlin Heidelberg, pp.48-53, 2011.
3) Bansal G., Sharma D., “Isolated Handwritten Words Segmentation Techniques in Gurmukhi Script”, International Journal of Computer Applications, vol.1, issue.24, pp. 104-111, 2010.
4) Kumar M., Jindal M.K., Sharma R.K., “Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition”, International Journal Information Technology and Computer Science, pp.58- 63, Feb, 2014.
5) Kumar R., Singh A., “Algorithm to Detect and Segment Gurmukhi Handwritten Text into Lines, Words and Characters”, IACSIT International Journal of Engineering and Technology, vol.3, issue.4, 2011.
6) Kumar R., Singh A., “Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text” Institute of Electrical and Electronics Engineers, pp.353-356, 2010.
7) Mangla P., Kaur H., “An End Detection Algorithm for segmentation of Broken and Touching characters in Gurumukhi Word”, Handwritten Institute of Electrical and Electronics Engineers, pp.1-4, 2014.
8) Mehta B., Rani S., “Segmentation of Broken Characters of handwritten Gurmukhi Script”, International Journal of Engineering Sciences, vol.3 pp.95-105, 2014.
9) Kumar R., Singh A., “Challenges in Segmentation of Text in Handwritten Gurmukhi Script” Proceedings in BAIP 2010, CCIS 70, Springer-Verlag Berlin Heidelberg, pp. 388-392, 2010
10) BinnyThakral, Manoj Kumar, “Devanagari Handwritten Text Segmentation for Overlapping and Conjunct Characters- A Proficient Technique”, pp.1-4, IEEE 2014.
11) M. A. Massoud, M. Sabee, M. Gergais, R. Bakhit, “Automated new license plate recognition in Egypt”, Alexandria Engineering Journal, vol.5, issue.2, pp.319-326, Science Direct, 2013.
12) Ching-Liang Su, “Car Plate recognition by whole 2-D image”, Expert Systems with Applications, vol.38, pp.7195-7200, Science Direct, 2011.
13) Dening Jiang, Tulu Muluneh Mekonnen, Tiruneh Embiale Merkebu, Ashenafi Gebrehiwot, "Car Plate Recognition System", ICINIS, vol.1, pp.9-12, IEEE, 2012.
14) Setumin, U.U. Sheikh, S.A.R Abu-Bakar, "Car Plate Character Extraction and Recognition using Stroke Analysis", SITIBS, vol.1, pp.30-34, IEEE, 2010.
15) Pei-Chen Tseng, Jiun-Kuei Shiung, Chun-Ting Huang, Shih-Mine Guo, Wen-Shyang Hwang, "Adaptive Car Plate Recognition in QoS-Aware Security Network", SSIRI, vol.1, pp.120-127, 2008.
16) Ping Wang, Wei Zhang, "Research and Realization of Improved Pattern Matching in

License Plate Recognition", ISIITAW, vol.1, pp.1089-1092, 2008.

17) Ratree Juntanasub, Nidapan Sureerattanan, "Car License Plate Recognition through Hausdorff Distance Technique", IICTAI, vol.1, pp.645-651, IEEE, 2005.

18) Benjapa Ratchata sriprasert, Kittawee Kongpan, Paruhat Punyarprateep, "License Plate detection Based on Template Matching Algorithm", ICCCT, vol.1, pp.139-143, 2012.

19) Clemens Arth, Florian Limberger and Horst Bischof, "Real-Time License Plate Recognition on an Embedded DSP-Platform", Proceedings of IEEE conference on Computer Vision and Pattern Recognition, pp.1-8, June 2007.

20) Halina Kwasnicka and Bartosz Wawrzyniak, "License plate localization and recognition in camera pictures", AI-METH 2002, November pp.13-15, 2002.