Evaluation of the Descriptive Type Answers using Neural Network Based OCR and Self-Organizing Map

Suma D K
Department of Telecommunication Engineering
Siddaganga Institute of Technology
Tumkur, India
sumadk232@gmail.com

Naveenkumar M
Department of Telecommunication Engineering
Siddaganga Institute of Technology
Tumkur, India
navinvcm@gmail.com

***Abstract*— Most of the colleges question papers contain a substantial portion of subjective questions. The pattern of these types of question papers is usually fixed to certain extent. Typically, the distinct kinds of answers are assessed physically by various teachers with nearly a similar measure of involvement and scholarly capabilities with the potential outcomes of deviation in the assessment of a similar answer-content. This project concentrates on the automatic assessment of answers which avoids the error in manual assessment. Assessment of answer sheets automatically can be performed utilizing Optical Character Reader technique and SOM (Self-Organizing Map) mechanism. Neural Network (NN) inputs the answer sheets in this process. NN helps in processing of the content from the handwritten answer sheet. NN is trained on the handwritten text and that helps in proper classification of test answer sheet to digital content. The proposed technique is tried with short answers composed by students of the class. This technique has various advantages like improved quality of results, decreased time and exertion, less stress on the staff and productive utilization of assets.**

Keywords—Optical Character Recognition (OCR), Neural Network (NN), Self-Organizing Map (SOM)

I. INTRODUCTION

The invention of paper held one of the greatest revolutions of humanity. Currently, it is still a form of media that accumulates more information, although not the most efficient. The paper brings several disadvantages, such as large physical space required to store it, which increases exponentially with the amount of information. Also, to make analysis on it, manual intervention is must. Classification of these systems can be made into 2 types that are online and offline systems. Online approach is executed in real time during the time when writers compose the text. It is less critical as it can catch the momentary or time- driven data for example speed, count of strokes, heading of composing of strokes and so forth. Also, there exist no requirements for thinning methods as the edge of the pen is n o t many pixels wide. Offline detection system works on static information for example if the input information is a bitmap. Consequently, it is exceptionally hard to perform identification. OCR empowers countless applications. Amid the good old days, optical character reader has been utilized for email arranging, bank check perusing and check signature matching [1].

In addition, OCR can be utilized by associations for automatic processing of paper forms where large volume of printed text is available. Different applications of OCR can be handling service charges, visa approval, pen figuring and automatic number plate identification and so forth [2]. Another valuable utilization of optical character reader is helping blind and eyes disorder patients to read the text. Post 90, visual computing methods and automatic identification of patterns were collaborated through AI. Due to conventional approaches in 90s, the robustness of OCR was limited and was applicable for few applications only such as, typewritten printed documents [3].

This work proposes a scheme in which Neural network is trained with some handwritten script data and would convert images of answers to digital text. Predefined keywords and their weightage (marks) would be helpful in computing the score of a student for particular answer. SOM technique is helpful in clustering of text and mapping the answer text with the predefined keywords and thus, marks would be calculated upon number of keywords match.

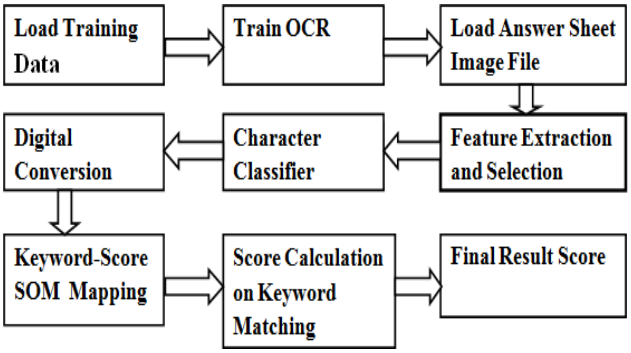


Fig. 1. Block Diagram of evaluation process of the descriptive Type Answer Sheet

In the above block diagram of evaluation process is depicted diagrammatically. The process steps are load training dataset, train the OCR Model, load answer sheet, feature extraction aselection, character classification, convert answer sheet to digital file, SOM keyword and answer mapping, score calculation on keyword matching and final score calculation.

II. LITERATURE REVIEW

According to [4] at the present time there is a large volume of digitally available information and most of this information is made up of scanned document images. Because of this huge amount, it is necessary that there are tools for that besides the documents are available in digital format, are also searchable and editable. This became possible with the development of Optical Character Recognition systems. OCR is a technology that allows you to convert different types of documents, such as scanned paper, PDF files and images captured with a digital camera into searchable data. Manually written text identification systems accuracy depends on the feature strained and selected and type of classifier used.

To perceive cursive letter sets, researchers [5] introduced holistic approach. The proposed procedure used in representing phrase through various transforming stages like features, forms, letters, expression and focuses. Feature vector is only made from picture to use arithmetical trusts among features and character, halfway determined characters are recognized by assessing through dictionary. Dictionary includes just 130 words, in this way limited number of words are distinguished. Classifiers aren't just utilized for identification of characters, a rank is given to each area which is isolated with beginning segmentation step using character hypothesis and they are distinguished based on most upper raking value [6].

The presented approach utilizes comprehensive method for identification of cursive letter set. They reduce a few features from the frame of char set. The feature vector is generated from the boundary information of chars that involves boundary position associated with 4 situation lines, its curve degree to the boundary and so on. A 10 dimensional feature vectors are created in their work. To perceive both cursive and separated manually written letter sets, the author of [7] had performed detection procedure by applying HMM. Hybrid system is used to get the match less quality of HMM. OCR picture is investigated in 4 different ways to main highlights from it. To use graph search mechanism, precise segmentation points are made.

Research works [8] are offered to recognize manually written English cursive letters using segmentation strategy. The essential plan uses blend of NN (Neural Network) with HMM for identification and in subsequent plan discrete HMM is used for identification. In first approach, pre segmentation of letterset is achieved through segmentation graph. NN ascertains the likelihood of each letters verification in graph and in this manner Hidden Markov Model decides probability for each letters in dictionary by including the likelihood along each predicted path in graph. In second approach, one hundred and forty geometric qualities are retrieved from every segment which are isolated by pre-segmentation. Using VQ i.e. vector quantization, this quality is altered to individual symbol and eventually by ascertaining the likelihood for each letters in a vocabulary character is identified.

Authors in [9] expressed segmentation technique to distinguish cursive letter-sets. According to their approach, cursive letter sets are first partitioned into specific characters, which are then recognized and converged to make significant expression by dictionary comparison. The thesaurus used in this work exists of 26 words. Henceforth, scope of this study is restricted to just 26 phrases.

III. PROPOSED METHEDOLOGY

A. OCR (Optical Character Recognition)

OCR stands for Optical Character Recognition. Optical Character Recognition is a process of classification of optical patterns that are included in a digital image corresponding to alphanumeric characters or other characters.

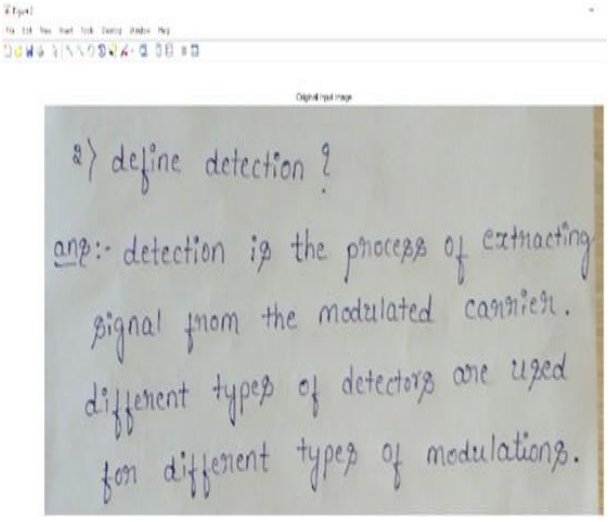


Fig. 2. Input Handwritten Image for OCR steps

Fig2, shows the handwritten text image, that is taken by digital camera and it will be in jpeg format. Currently, there is a lot of information available digitally, and most of this information is made up of scanned document images. Because of this large amount, it is necessary that there are methods to enable research in these documents, these functions are provided by Optical Character Recognition tools (OCR). There are several OCR offers available on the market today, however, none of these tools dominates as the best market option.

The purpose of this section is to explain the working of OCR. OCR process comprises of the following basic phases: Optical scanning, Image Acquisition, Pre-processing, Extraction Characteristics, Classification, Post-processing.

a. Optical Scanning

The digital image of the original document is captured by the scanning process. Currently, it is also possible that the digital image is captured by a photo of the original document.

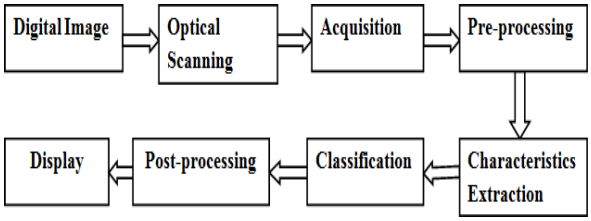


Fig 3. OCR Step

The scanners Optical generally consist of a transport mechanism which converts the light intensity in gray levels, being this process known as thresholding. The thresholding process is very important, because the results of the character recognition depend entirely on the image quality. The thresholding done in scanner is generally basic, being defined one fixed limit, where gray levels below the threshold are considered black and white levels above are considered to obtain good results.

b. Image Acquisition

In this character recognition stage, the image is segmented into characters before sent to the classification stage. The image splitting up can be performed explicitly or implicitly as a byproduct of classification stage [4]. In addition, the other stages of optical character recognition can help in providing circumstantial information useful for partition of image.

c. pre-processing

The pre-processing is to used improve the image quality by smoothing the characters scanned to try to eliminate noise resulting from the scanning process, such as stained or incomplete characters which hinder the performance of OCR. This step includes standardizing to obtain character size, uniform tilt and rotation. Skew correction of file is also a very vital part of pre-processing. There exist various kinds of skew degree assessment mechanisms such as, nearest neighborhood mechanism or projection profile. Image thinning also can be applied at initial phase for proper estimation. At last, lines of the text too can be considered as a component of pre- processing step. It can be performed using pixel clustering and projection methods.

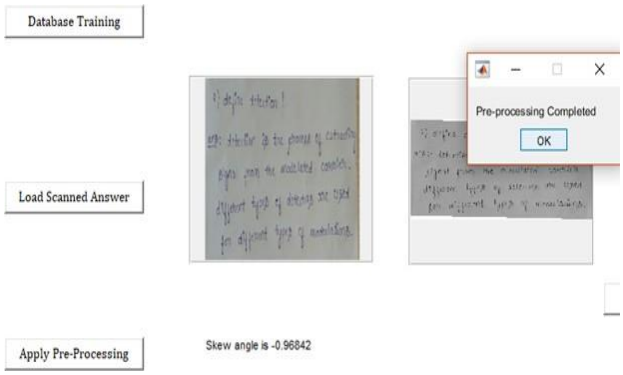
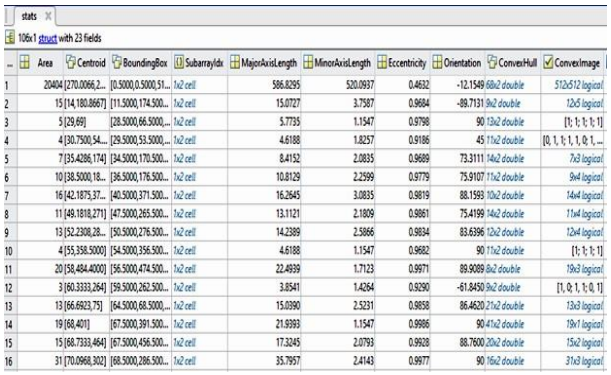


Fig. 4. Pre-Processing of Descriptive Type Answer Sheet

Fig4, shows the output window of pre-processing of descriptive type answer sheet automatic evaluation process with skew angle of handwritten text image, that is loaded by the folder.

d. Character Segmentation & Feature Extraction

In this step, segmentation is performed in order to separate the characters present in it. Various techniques can be used to perform segmentation including an algorithm of connected components, vertical and horizontal projections, an algorithm using contours and background information about the characters. The purpose of feature extraction is to capture the characteristics of each character, and it is generally accepted that this is one of the most difficult problems of pattern recognition. Techniques for extracting such characteristics are often divided into three main groups, where the resources are found from the distribution points, and the processing series expansions and structural analysis.


stats

106x1 struct with 23 fields

	Area	Centroid	BoundingBox	Subarrayldx	MajorAxisLength	MinorAxisLength	Eccentricity	Orientation	ConvexHull	ConvexImage
1	20404	[270.0066,2...	[0.5000,0.5000,51...	1x2 cell	586.8295	520.0937	0.4632	-12.1549	68x2 double	512x512 logical
2	15	[14,180.8667]	[11.5000,174.500...	1x2 cell	15.0727	3.7587	0.9684	-89.7131	9x2 double	12x5 logical
3	5	[29,69]	[28.5000,66.5000,...	1x2 cell	5.7735	1.1547	0.9798	90	13x2 double	[1; 1; 1; 1; 1]
4	4	[30.7500,54....	[29.5000,53.5000,...	1x2 cell	4.6188	1.8257	0.9186	45	11x2 double	[0, 1, 1; 1, 1, 0; 1, ...
5	7	[35.4286,174]	[34.5000,170.500...	1x2 cell	8.4152	2.0835	0.9689	73.3111	14x2 double	7x3 logical
6	10	[38.5000,18...	[36.5000,176.500...	1x2 cell	10.8129	2.2599	0.9779	75.9107	11x2 double	9x4 logical
7	16	[42.1875,37...	[40.5000,371.500...	1x2 cell	16.2645	3.0835	0.9819	88.1593	10x2 double	14x4 logical
8	11	[49.1818,271]	[47.5000,265.500...	1x2 cell	13.1121	2.1809	0.9861	75.4199	14x2 double	11x4 logical
9	13	[52.2308,28...	[50.5000,276.500...	1x2 cell	14.2389	2.5866	0.9834	83.6396	12x2 double	12x4 logical
10	4	[55,358.5000]	[54.5000,356.500...	1x2 cell	4.6188	1.1547	0.9682	90	11x2 double	[1; 1; 1; 1]
11	20	[58,484.4000]	[56.5000,474.500...	1x2 cell	22.4939	1.7123	0.9971	89.9089	8x2 double	19x3 logical
12	3	[60.3333,264]	[59.5000,262.500...	1x2 cell	3.8541	1.4264	0.9290	-61.8450	9x2 double	[1, 0; 1, 1; 0, 1]
13	13	[66.6923,75]	[64.5000,68.5000,...	1x2 cell	15.0390	2.5231	0.9858	86.4620	21x2 double	13x3 logical
14	19	[68,401]	[67.5000,391.500...	1x2 cell	21.9393	1.1547	0.9996	90	41x2 double	19x1 logical
15	15	[68.7333,464]	[67.5000,456.500...	1x2 cell	17.3245	2.0793	0.9928	88.7600	20x2 double	15x2 logical
16	31	[70.0968,302]	[68.5000,286.500...	1x2 cell	35.7957	2.4143	0.9977	90	16x2 double	31x3 logical

Fig. 5. Feature Extraction output of all characters present in the Input Handwritten Image.

e. Classification

The classification is the process of identify each character and assign it to a corresponding character class. The two most important categories of classification methods for OCR are Decision of theoretical methods and structural method. The main approaches for the recognition of decision theory are: Minimum distance; excellent statistical classifiers; neural networks.

f. Post-processing

The result of the character recognition is a set of individual symbols. Therefore, the post-processing phase is done grouping the characters, combining words and numbers forming a text. The grouping of characters is based on location of these in the document, and grouped the characters that are sufficiently close. Furthermore, this stage is also made one error detection and correction

B. Self Organizing Maps

Self-Organizing Maps, or SOM, were invented in 1982 by Teuvo Kohonen, a professor at the Academy of Finland, and provide a way to represent multidimensional data (vectors) in spaces of a smaller dimension, usually 2D. This process of reducing vector dimensionality is a technique of data compression known as Vector Quantization. In addition, the Kohonen technique creates a network that stores information so that the topological relationships of the training set are maintained. A common example used to show how SOM works is based on the projection of colors (associated with 3D vectors formed from, for example, their 3 RGB components) in a 2D space.

Network architecture in general, the SOM algorithm consider a 2-layer architecture: on the one hand we have a layer of learning nodes, of which we care about the topological relation between them, and that will be the ones that will finally contain the information about the representation resulting, along with a layer of input nodes, where the original vectors will be represented during the training process. In addition, all elements of the first layer are connected to all elements of the second layer. The following figure shows a possible 2D architecture for a SOM training, the learning network is represented by the red nodes, and the training vectors are represented in green.

IV. RELATED WORK AND DISCUSSION

A. System Architecture

Fig6, shows the architecture diagram of our proposed application. Initially, scanned answer sheets are converted to text using OCR operation, then keyword score mapping is done to calculate final score.

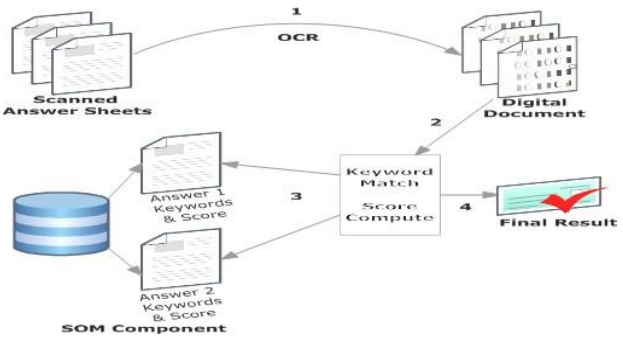


Fig. 6. System Architecture of Answer Sheet Evaluation Process

B. System Design

System design is done to illustrate the overall system and flow of information. It is a simple graphical representation to show various operations performed by various entities in the system. Various models are developed based on the system requirements specifications and implementations are carried out further.

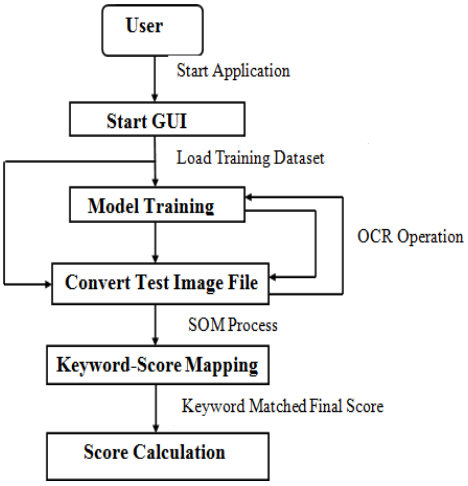


Fig7: sequence diagram of answer sheet evaluation.

Fig7, shows the sequence of operations from user initially, user starts application and a GUI appears, then loading of training data set is performed to train the model, then test images are converted to text using OCR operation, then keyword score mapping is done to calculated final score.

C. Use Case Diagram

Fig8, use case diagram shows action that can be performed by the users. The actions which can be performed are, load training dataset, define keyword and its score, train the OCR Model, Upload answer sheet, convert answer sheet to digital file, SOM keyword and answer mapping and final score calculation.

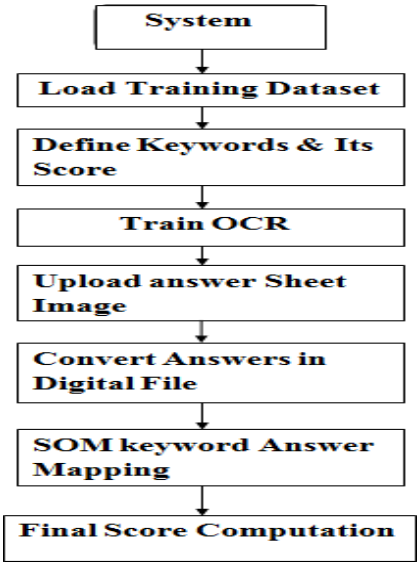


Fig. 8. Use Case Diagram of Descriptive Type Answer Script Evaluation Steps.

V. RESULTS

Fig9, represents, in the region of interest (ROI) which ever words are present in the input handwritten image that are all detects and extracted and later it displayed the result in figure window.

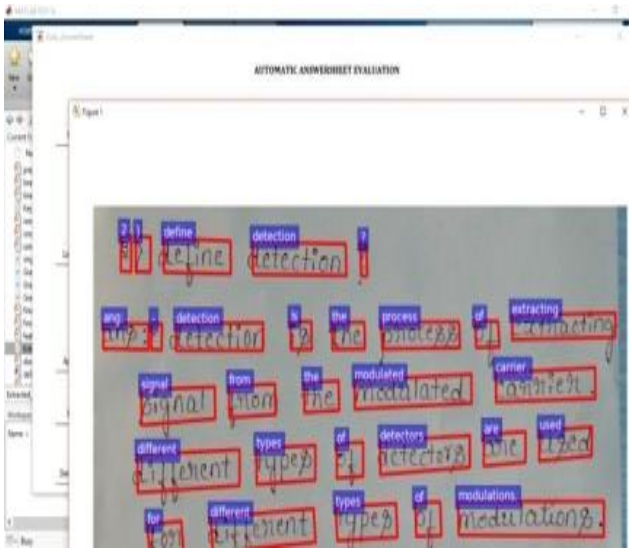


Fig. 9. Output of Detect and Extract Texts process

Fig10, shows the final marks awarded for answer sheet, this process is done by the comparing of standard keywords and answer sheet keywords matched and final marks is displayed on the figure window.

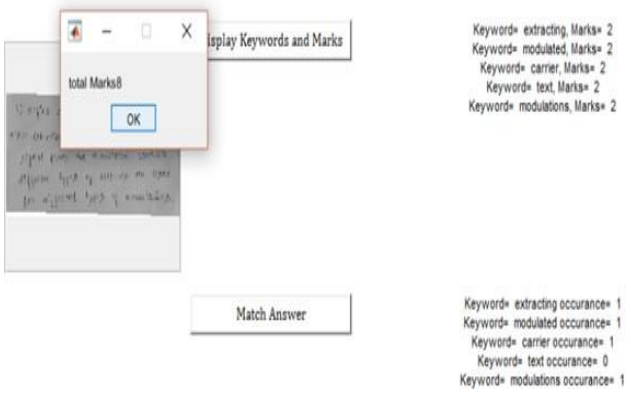


Fig. 10. Marks Awarded for Evaluation of Answer Sheet

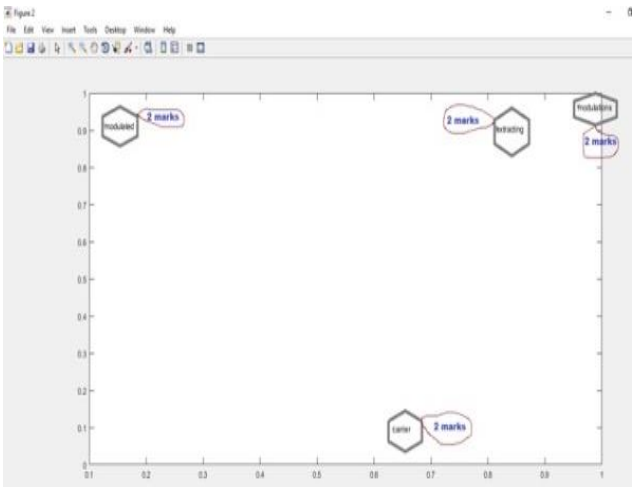


Fig11: clustering of matched keywords

Fig11, shows the figure window of Clustering of matched keywords. Finally, a clustering graph is presented to illustrate the occurrence of keywords in processed answer sheets. Keyword Matching from Answers to SOM Indexed Keywords. Score Calculation based on matching Keyword Count and Final Cumulative Score. Score of each keyword matched is stored and aggregated as per the answer numbering and finally all the answer scores are cumulatively aggregated as final result score.

VI. CONCLUSION

In recent years, due to availability and use of information and communication technologies for some uses of paper have been replaced by digital and electronic media. OCR is widely used for automatic form processing where huge amount of data is accessible in printed format. This application enables supervisors to assess answer sheets without much human efforts. Results prove the efficacy of proposed system. Finally, a clustering graph is presented to illustrate the occurrence of keywords in processed answer sheets.

This application is a prototype where single answers are processed for score evaluation. In future a system keeping this work as a base can be developed to process multiple answers at a time and calculate cumulative score. Also, keywords of multiple answers can be saved to make multiple assessments.

REFERENCES

1. Shen, H., Coughlan, J.M, 2012, Towards A Real Time System for Finding and Reading Signs for Visually Impaired Users. Computers Helping People with Special Needs. Linz, Austria: Springer International Publishing.
2. Bhavani, S., Thanushkodi, K, 2010, A Survey On Coding Algorithms In Medical Image Compression. International Journal on Computer Science and Engineering, 2(5), 1429-1434.
3. I. Bazzi, R. Schwartz, and J.Makhoul, Anomn ifontopen- vocabulary ocr system for english and arabic", Pattern Analysis and Machine Intelligence, IEEE Transactions on,vol.21,pp.495-504,June1999.
4. J. Hu, S. G. Lim, and M. K. Brown, Writer independent on-line handwriting recognition using an hmm approach", Pattern Recognition, vol.33, no.1, pp. 133-147,2000.
5. Bozinovic, Radmilo M., and Sargur N. Srihari. "Off-line cursive script word recognition." Pattern Analysis and Machine Intelligence, IEEE Transactions on 11, no. 1(1989):68-83.
6. Bunke, Horst, Markus Roth, and Ernst GnterSchukat- Talamazzini. "Off-line cursive handwriting recognition using hidden Markov models." Pattern recognition 28,no. 9 (1995):1399-1413.
7. ARICA, NAFIZ. "An off-line character recognition sys- tem for free style handwriting." PhD diss., MIDDLE EAST TECHNICAL UNIVERSITY,1998.
8. Tay, Yong Haur, Pierre-Michel Lallican, Marzuki Khalid, Christian Viard-Gaudin, and S. Kneer. "An offline cur- sive handwritten word recognition system." In TENCON 2001. Proceedings of IEEE Region 10 International Con- ference on Electrical and Electronic Technology, vol. 2, pp. 519-524. IEEE,2001.
9. Gupta, Anshul, Manisha Srivastava, and Chitralekha Mahanta "Offline handwritten character recognition using neural network." In Computer Applications and Industrial Electronics (ICCAIE), 2011 IEEE International Confer- enceon, pp.102-107.IEEE,2011.
10. Patel, D. K., T. Som, and M. K. Singh. "Improving the Recognition of Handwritten Characters using Neural Network through Multiresolution Technique and Euclidean Distance Metric." International Journal of Computer Applications45, no.6(2012):38-50.
11. Blumenstein, Michael, Brijesh Verma, and Hasan Basli." A novel feature extraction technique for the recognition of segmented handwritten characters." In Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on, pp. 137-147. IEEE, 2003.