ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Recognizing Students Handwriting using OCR System: A Comprehensive Survey

Rahul V.Chaugule, Prof. Sachin Godse

M.E Student, Dept. of Computer Engineering, Sinhgad Academy Of Engineering, Pune, Maharashtra, India

Assistant Professor, Dept. of Computer Engineering, Sinhgad Academy Of Engineering, Pune, Maharashtra, India

ABSTRACT: Manual assessment of exam answer papers is always a tedious job which takes lots of time and effort. This paper has suggested different algorithms forrecognizing students handwriting for developing an automatic, fast and reliable system whichevaluate the marks of particular student from theory exam papers. There are different algorithms are used to recognize students handwriting and then that particular letters or words are converted to editable format like .doc file. Basically to convert non-editable files to editable file we are using optical character recognition (OCR system).Automation in assessment of theory exam papers, implementing Optical Character Recognition (OCR) system will help teachers save a lot of time in checking of each students answer sheet. OCR system uses different algorithms for conversion. In this paper we discuss some of the important algorithms used in OCR System and compare each one to find out which algorithm is better for conversion.

KEYWORDS: Feature Extraction, Grayscale conversion, Hidden Markov Models (HMMs) and Neural Networks (NNs),I-A novel Algorithm, Optical character Recognition (OCR system), OCR algorithms, Pattern Recognition.

I. INTRODUCTION

Almost every college and university needs to check the exam answer sheet as fast as possible and display the result as soon as possible.Manual handling of such tasks often leads to irregularities and discrepancies. An automated system capable of performing these tasks in a quicker, more efficient manner is discussed in this paper. For this purpose there is Optical character recognition(OCR) system and by using different algorithms of OCR system we can easily convert the image file to text file so that we can search andedit everything from that file.Here the main task is to recognize the handwriting ofdifferent students and then convert that handwriting into editable format like .doc file .for this purpose the optical character recognition (OCR)system is used. And afterconversion we are just comparing that editable file with learned set i.e. machine copy. In this paper Firstly we will discuss the exact concept of OCR after that we will studydifferent algorithms used in OCR and then we will compare all this algorithm and select one which is fast and reliable.

II. RELATED WORK

OCR is a method in which the input characters are recognized from optical data obtained by digital data. Optical character recognition is the process of electronic conversion of images like typed image, handwritten or printed text into editable text.Means In short OCR system takes any image file like JPG,PNG as a input and then convert that image file into editable or searchable format. So anyone can easily modify and search from that file. This OCR system mostly used in data entry application for example printed paper data records and passport documents, bank statements, computerized receipts. It is a most important methodology of digitizing printed texts so that it can be easily edited and searched. OCR is a area of research inartificial intelligenceand pattern recognition. There are so many algorithms have been developed for this purpose using different approaches and ideas.In short we can say that Optical Character Recognition (OCR) is a process of converting alphabets in images to computer readable coded text.

In general, the process of OCR can be divided into four steps as follows: We will discuss this steps one by one as follows:

DOI: 10.15680/IJIRCCE.2015. 0311178

(1) Pre-processing: In pre-processing OCR system convert original image into more specific or readable format means In short Pre-processing technique removes the noise. The recognition accuracy is depend on the quality of the actual text and the presence of noise. The text quality may be affected by some of these factors:

(a) The copies of documents are very hard to read because the text may change its thickness and that's why more noise dots may comes..

(b) The print quality may differ from printer to printer, e.g.colour or monochrome.

(2) Feature extraction: Any letters from the input file or image is divided into loops, branches, dots, zigzags, etc. These features are more useful and more commonly used for text recognition than statistical features, which include the characteristics such as invariant moments, pixel densities, Fourier descriptors etc.

(3) Recognition: Here we actually recognize the letters from the file. Recognition algorithms may be applied on individual letters or on complete words. For this recognition there are two most-useful techniques known as Hidden Markov Models (HMMs) and Neural Networks. HMMs technique is used where learning properties are not as intuitive. Usually, the HMMs model single dimensional sequences of data, and are composed of states and transition probabilities (based on 'observations') among the states. For text recognition application, the observations may be based on pixels, and the states could represent different parts of letters. NNs are composed of different layers of interconnected, computing nodes. Different sets of input values when passed through the layers produce different outputs. For text recognition, the information entering the NNs would be some type of word/letter representation (structural and/or statistical), such as pixels, invariant moments, etc.

(4) Post-processing: In this step we check the additional features like spell-checkers. In case of Arabic, the pre-fixes and suffixes may have to be removed before the words are looked up in a lexicon.

III. EXISTING ALGORITHMS FOR OCR

To recognize the student's handwriting there are multiple algorithms. Some of the existing algorithms are listed below:

- **A Stock Pattern Recognition Algorithm**

Currently, there are mainly two kinds of stockprice pattern recognition algorithms: the algorithmbased on rule-matching and the algorithm based ontemplate-matching. However, both of the twoalgorithms highly require the participation of domainexperts. In this algorithm we are providing a learned set to the machine. Means In short firstly we train to the machine and after that we pass out input file. This algorithm compares input characters with learned set characters and displays appropriate result. So the result and accuracy of this algorithm is totally depend on the learned set.[1]

- **Optical Character Recognition ImplementationUsing Pattern Matching**

Following diagram shows the actual flow of this algorithm. We will discuss this steps in brief as follows:

Step 1: Grayscale the image

Grayscale images have many shades of gray. Grayscale images is result of measuring intensity of each pixel. For achieving accuracy input document should be grayscaled. To convert a colour from a colourspace based on an RGB colour model to a grayscale representation following function is used

$Y = 0.2126R+0.7152G+0.0722B$

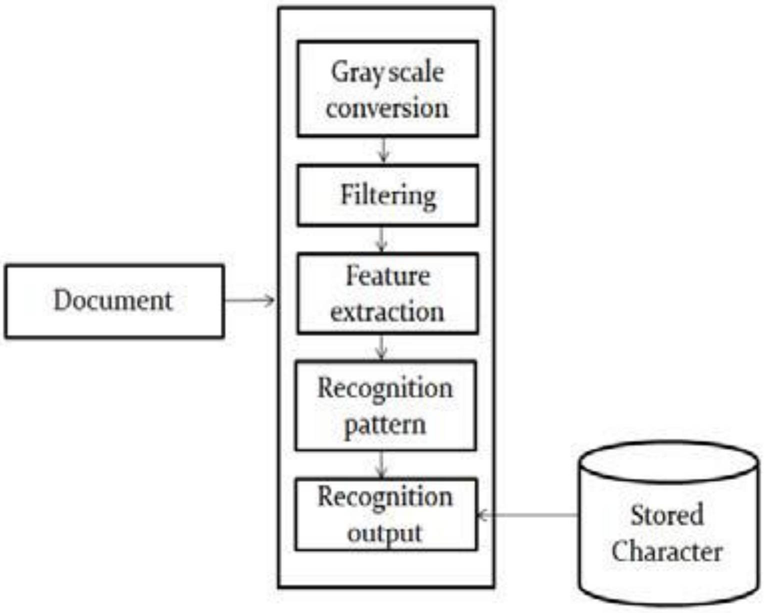

Fig.1 System Block Diagram

Step 2: Feature Extraction

Feature extraction is the process of getting information about an object or a group of object in order to facilitate classification. This is an important part in our system.

Step 3: Recognition of Pattern

Pattern based recognition require matching of generated binary format with the existing template for this purpose the binary has been divided into 5 tracks and each tracksubdivided into 8 sectors. A corresponding track-sector matrix is to be generated, identifying number of pixels in each region.

Step 4: Recognition of Output

The track-sector matrix generated above is then matched with existing template. The existing template consist of each track-sector intersection value, each track value and each sector value. If all these parameters are found to match with the template values then the resultant is the character identified. The resultant matrix contain unique value for each font and thus makes it easy to identify each font separately

- **Contour Analysis Algorithm**

A contour analysis method has been fully developed for digit recognition and is being modified for character recognition. This method calculates the curvature at every point along the inner and outer contours of a binary image. The features are similar to those described in . Three features are used for concave curvature, and five for convex curvature. Each feature is also associated with its direction and location. The feature string extracted from an unknown character is matched against a rule base to achieve recognition.

DRAWBACKS OF EXISTING ALGORITHM:

Following are the some of the drawbacks of the above OCR algorithms:

1. Above mentioned Algorithms use extensive mathematicaloperations and involve several calculations to deduce thewavelets

2. It Uses matrix multiplications, summationsand essentiallyit requires time to take different inputoutput pairs and calculate the relation between them.
3. Requires Complex Computation.
4. Large Processing Time.
5. Approaches using artificial neural networks also use trainingmechanism for OCR.

IV. THE PROPOSED ALGORITHM

When we compare all this existing algorithm then we are at the point that all this algorithms are having a drawbacks. To overcome that drawbacks we are going to propose the following algorithms**:**

A) “ i ” - A novel algorithm for Optical Character Recognition (OCR)

B) Text Recognition using Invariant Moments.

A) Overview of I-A novel Algorithm

Many methodologies and algorithms have been developed for this purpose using different approaches. Here we present one such approach for OCR named “ i . “ i ”-A novel algorithm has a some extra features as compared to other algorithms like it aims at a high speed, simple to implement , font independent and size independent OCR system based on a unique segment extraction technique.[9] The I-novel algorithm can be used as a heart for single alphabet detection within a complete OCR solution system without the need for any complex mathematical operations so it reduces the time required to recognize the characters. The highlight of this algorithm is that, it does not use any libraries or databases of image matrices to recognizealphabets, but it has a unique algorithm to recognize alphabets instead. This algorithm which works on basis of the feature recognition based OCR. The i-novel algorithm extracts details about the lines or curves which the alphabet is made of, and then based on the extracted details, it takes a decision about the alphabet in the image. Here we use no databases or libraries to compare with, but only logic based on patterns and shapes which decides the alphabet. What is really interesting about the algorithm is that it is font independent while having no database, training or SVM mechanisms. The algorithm is efficient in the terms of small footprint and better efficiency.[9]

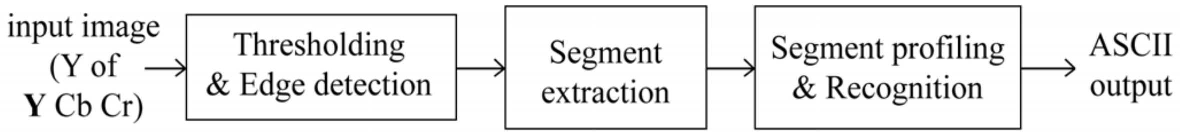
input image (Y of Y Cb Cr)
Thresholding & Edge detection
Segment extraction
Segment profiling & Recognition
ASCII output

Fig.2Flow of I-Novel Algorithm

First, an image is acquired through any of the standard image acquisition techniques. The input image is assumed to be in the *Y Cb Cr* color format. The algorithm works on the *Y* partof the input image, which is a gray level image. Using an appropriate thresholding algorithm, the gray level image is then thresholded to obtain a binary image which quantizes alphabets and background to black and white colours respectively. The obtained binary image is then passed on to a specific edge detection process. The edge detection algorithm is performed such that only the right sided edges of each alphabet are obtained and the other edges are eliminated. After edge detection, the image is then segmented and feature extraction is performed. A segment is defined as a continuous path of black pixels in the edge detected image, for this algorithm. In this step, different details of the segments, which are required for further processing, are stored. The next step is to profile stored line segments. Profiling of segments is the process of categorizing them into different types of segments such as short, long, line or curve, etc.

- **Why I-novel Algorithm:**

Accuracy, flexibility and speed are the main features that characterise a good OCR system. Several algorithms for character recognition have been developed based on feature selection**.** Some of them have been found commercially viable and have gone into production like OmniPage,Wordscan, TypeReader etc. The performance of the systems have been constrained by the dependence on font, size and orientation. The recognition rate in these algorithms depends on the choice of features. Most of the existing algorithms involve extensive processing on the image before the features are extracted that results in increased computational time. In this paper, we discuss a pattern matching based method for character recognition that would effectively reduce the image processing time while maintaining efficiency and versatility.[9]

The above mentioned algorithms use extensive mathematical operations and involve several calculations to deduce the wavelets. These in turn use matrix multiplications, summations and the SVM essentially requires time to take different input outputpairs and calculate the relation between them. Approaches using artificial neural networks also use training mechanism for OCR [6]. These approaches are equallyexhausting in terms of mathematical processing as the previous technique discussed above. There are also systems which do not need training and memory based training or recognition. Some use fuzzy logic and histogram type area-weight detection of the areas of alphabet [7]. The main disadvantage of these methods would be the complex computations and the processing time taken. Yet another methodology used for OCR is feature recognition based OCR. In feature recognition type [8] OCRs, different features are extracted from an alphabet present in the input image. This is a highly advantageous approach in terms of memory utilization and computations, since only a certain set of features of the character are sufficient to identify the letter.

B) Text Recognition using Invariant Moments.

Optical character recognition is anvery important image processingtask. Its aim is to enable computers to recognise graphic characters withouthuman supervision. The process of optical symbol recognition is divided intotwo stages. First, certain features of the character undergoing recognitionare extracted, and second, a match to them is searched for in the library ofmodels. This algorithm looks at Hu invariant moments, a well established set ofimage features and the performance of this method in optical character recognition.One approach to using Hu invariant moments in pattern recognitionis using a metric function to find the pattern in the library of models, thatis of the same class as the pattern considered. In this paper a new classificationmethod is proposed that performs better than the classic method ofmetric function.[11]

- **Hu Invariant Moments**

Central and Normalised Moments

One of the most common ways to describe an image pattern is to compute itsgeometrical moments [10]. The summation of pixel intensities multiplied by pixel coordinates x raised to thepower of p and y raised to the power of q is over all the pixels in the image.Central moments di_er from regular moments in that the values of pixel coordinatesx and y are displaced by their mean values in the image. Using the centroidof an image results in the value of the (p,q) central moment of an image patternto be equal to the value of the (p,q) regular moment of the same pattern shifted sothat its centre coincides with its centroid. That makes central moments invariant to
image translation.

Rotation Iinvariant Moments

Normalised moments are translation and scale invariant. In [10] Hu described amethod that led to the construction of new moments that are invariant to rotation.Algebraic invariants were applied to the moment generating function under a rotationtransformation. The resulting new moments have been since then referred
to as Hu invariant moments. The first Hu moment is a linear combination of normalisedmoments, whereas the next Hu moments are non-linear combinations ofnormalised moments.[11]

OCR using Hu invariant moments

Hu invariant moments are combinations of (p,q) normalised moments. Thosenormalised moments are set together in such a way that the resulting sum is invariantto translation, scale and rotation. Characterrecognition based on Hu invariant moments seems trivial at first sight. What weneed to do is calculate Hu moments of the character considered and compare themwith the moments of characters obtained in the training process of the classifier.An important observation is, that although the seven Hu moments change only alittle with scale, they change to the extent that can lead to false results when the setof possible matches is the alphabet. If we comparecorresponding values for each character we will see that they di_er to someextent but the consecutive values occur in more or less the same proportions toeach other.Now the question is whether the change in Hu moment values with scale caninfluence the

OCR process or not. Let us consider an OCR task where the trainingdata were a set of Times New Roman characters of size sixty and we are torecognise characters sized forty. The library of model features is build from dataobtained from the training set; therefore the two terms – training set and libraryof model features are used interchangeably in this methods.[11]

V. CONCLUSION

In this paper, we have discussed different algorithms to recognize studentshandwriting from hardcopy of studentpapers. OCR algorithm has been discussed and we have compared this this algorithm to give accurate result of students handwriting OCR is one of the most emerging technologies and its reliability is continually improving. Soon OCR will become apowerful tool for data entry applications which will lead to automated data entry by OCR thus reducing labour. Incorporating OCR will be an attractive feature of any Data Entry System. However in past due to limited availability of a capital and short environment was restricting the growth of this technology, but today more and more enterprises are working on this technology and that will definitely lead to 100% accuracy in this technology thus making the dream of paperless world true. We have studied different existing algorithms and drawbacks of this algorithms and after that we have discussed i-A novel algorithm for OCR which gives more accuracy as compared to others as well as it requires less amount of time and memory to run. So in short in this paper we have studied this different algorithms for OCR and comparison between them.

REFERENCES

[1]XinyuGuo, Xun Liang, Xiang Li, "A Stock Pattern Recognition Algorithm Based on Neural Networks", Third International Conference on Natural Computation (ICNC 2007).

[2] Ramanathan. R. et al., "A Novel Technique for English Font Recognition Using Support Vector Machines", in Advances in Recent Technologies in Communication and Computing, Kottayam, Kerala, 2009, pp. 766 - 769.

[3] L Priit. (2011, November 1). How to extract text from images: a comparison of 10 free OCR tools [online]. Available: http://www.freewaregenius.com/2011/11/01/how-to-extracttext-from-images-a-comparison-of-free-ocr-tools.

[4] LineEikvil, "Optical Character Recognition", NorskRegnesentral , Oslo, Norway, Rep. 876, 1993.

[5] Yang Guang, "License Plate Character Recognition Based on WaveletKernel LS-SVM", inComputer Research and Development (ICCRD) 3rdInternational Conference, Shanghai, 2011, pp. 222 - 226.

[6] M UsmanRaza, et al., "Text Extraction Using Artificial Neural Networks", in Networked Computing and Advanced Information Management (NCM) 7th International Conference,,Gyeongju, North Gyeongsang, 2011, pp. 134 - 137.

[7] Fonseca, J.M., et al., "Optical Character Recognition Using Automatically Generated Fuzzy Classifiers", in Eighth International Conference on Fuzzy Systems and Knowledge Discovery, Shanghai, 2011, pp. 448 - 452.

[8] Kumar, M., et al., "k - Nearest Neighbor Based Offline Handwritten Gurmukhi Character Recognition", in International Conference on Image Information Processing, Himachal Pradesh, 2011, pp. 1 - 4.

[9]SushruthShastry, Gunasheela G, ThejusDutt, Vinay D S and SudhirRaoRupanagudi, " i " - A novel algorithm for Optical Character Recognition (OCR),IEEE-2013.

[10] Hu, M., Visual Pattern Recognition by Moment Invariants, IRE Transactionson Information Theory, 1962.

[11] MarcinKmie´c, "New Optical Character Recognition MethodBased on Hu Invariant MomentsandWeighted Voting", AGH University of Science and Technology, Institute of Automaticsal. A. Mickiewicza 30, 30-059 Kraków