Optical Character Recognition

Tushar Gupta\textsuperscript{1}, Chahat Ahuja\textsuperscript{2}, Sanjukta Aich\textsuperscript{3}

\textsuperscript{1,2,3} School of Electronics Engineering, VIT University, Vellore, 632 014, India

Abstract—Recognition of characters relates to emblematic identity with the image of character. Majority of the OCR systems input characters are first converted to digital form by an optical scanner. Every character is first located and segmented, and the resulting character image is fed into a preprocessor for noise reduction and normalization. Certain characteristics are the extracted from the character for classification. Numerous techniques exist for feature extraction each one having its own merits and de-merits. After classification and training the identified characters are grouped to reconstruct the original symbol strings, and various algorithms may then be applied to detect and correct errors.

Keywords—character recognition, corr2, feature extraction, pixels, segmentation.

I. INTRODUCTION

Computer reproduction of human functions, like reading, has been a dream from a very long time. With the advancement in technology over the past fifty year’s computer based reading has become a reality. In the field of pattern recognition and machine learning OCR or optical based character recognition is one of the most researched and flourishing applications of technology. Majority of the current popular systems for evaluation for OCR exist for a vast number of purpose and applications. However computers are still at that level which is very far away when compared to human reading capabilities. Optical Character recognition, usually referred in short as OCR, is the automated or electronic transformation of images of handwritten, typewritten or printed text into a computer readable text. It can be related to recognizing of characters which are processed using optical methods such as scanners. Although the academic research is under progress, the main target on character recognition has been deviated to implementation of confirmed or proven techniques\cite{3}. For majority of the tasks which required document-input, character recognition is the most widely used method since it’s a cost-effective as well as easier to perform solution\cite{4}. Classification of different character is also one of the most testing ground for new ideas in recognition of specific pattern \cite{1}. However the experiments are operated on secluded set of characters and the outcome of the experiment may not be important to OCR.

As per the current trend OCR software is often an add-on to majority of the desktop scanners that are almost the same cost as the printer. The purpose of this paper will be to implement OCR on a Linux based system using a set of libraries which will be able to detect English language. Apart from detection, it will also train itself so when the time progresses, the recognition of character will be more accurate. This type of system can be applied to track number plates of a vehicle.

II. COMPONENTS OF AN OCR SYSTEM

Majority of the OCR system dwell of numerous of components. In figure 1 a typical setup of an OCR system is demonstrated. In the process the first step is to convert the document into a digital form with the help of an optical scanner. The next step is to locate the regions which contain text. At the same time noise has to be eliminated and the each symbol is extracted. The extracted symbols is pre-processed, to promote the extraction of features carried out in the next step. The integrity of each symbol or character is established by correlating the features extracted with depiction of the symbol classes which were obtained through a previous learning phase. In the end dependent information is used to reconstruct the words and numbers of the original text.

Fig. 1. Components of OCR system

A. Optical Scanning

In OCR based optical scanners, which generally comprises of a transit structure along a device usually a sensory that converts intensity of the light into grey scale levels. Majority of the printed documents usually consists of a black ink on a white sheet.
Therefore during OCR, it’s a common trend to convert the multilevel image such as RGB into a bi-level image comprising only of black and white pixels. This process if known as thresholding, and is usually performed on the scanner to save memory space as well as calculation effort by the computer, thereby increasing the speed of pre-processing. The thresholding process holds an important place as the outcome of the recognition is totally dependent on the quality of the grey scale image. However on the scanner only very basic thresholding is performed. For a high-contrast document with uniform background, fixed threshold is sufficient to get the required quality. A fixed threshold is said to be black and levels above are said to be white.

B. Location and Segmentation
Segmentation is the segregation of characters words, or symbols. Most of optical character recognition algorithms usually segment the words into separate characters which are recognized individually [3]. This segmentation is performed by isolating each connected neighbour component that is each connected black area. Although technique is very easy to implement, but the problem occur if the characters are touching each other. The main problems in segmentation may be divided into four groups.

C. Pre-processing
The output image from the scanner will contain noise. Majority of techniques are applies to eliminate the existing noise before the image if send for feature extraction. Majority of morphological operation takes place that includes both filling and thinning. Filling reduces or completely eliminates small breaks, gaps and holes in the characters, while thinning is used to reduce the width of the line. The most common smoothing algorithm is the "rectangular" or "unweighted sliding-average smooth". This method changes every single point in the image with the average of "k" neighbouring points, where "k" is a always positive integer called the "smooth width". The triangular smooth is like the rectangular smooth except that it implements a weighted smoothing function. The normalization is applied to obtain characters of uniform size, slant and rotation. The text which is rotated thorough a certain angle, the angle of rotation has to be calculated. This is normally done by using different variant of Hough transform which are also commonly used for detecting skew.

D. Template-matching and comparing methods
Instead of extracting the features in the image the matrix which contains the digitalized image of the input character is directly matched with a set of pre-existing characters representing each possible types of character [2]. The distance between the pattern and each sample is computed, and the depending upon the class of the pre-defined character the best possible match is assigned as output. Although the method is simple as well as easy to apply on hardware level it is very sensitive to noise and variations in different style of character. Moreover if the character are in rotated form or two consecutive characters are touching each other, this method cannot provide a match of the character.
III. OPENCV

OPENCV stands for Open Computer Vision. It’s a set of libraries that was initially developed by Intel but now is open source. Image processing is a very wide field to deal with. There are number of features we can detect such as colours, intensity, edges, texture or pattern in an image. Initial testing of the algorithm was done on MATLAB, however the speed of computation was slow as compared to opencv with the same algorithm. Moreover training algorithms are easier to implement on opencv then on MATLAB. Since the code was ported to a general set of libraries, it can be run on any Linux based system and hence can be embedded into the OS of the optical scanner. This will not only eliminate the need of a separate machine for OCR but will also be a cost-effective solution.

IV. PERFORMANCE

Despite number of developments in OCR field, there is still no general test that exists for character recognition. The performance of an OCR system entirely depends on the input quality image which makes it complicated to judge and contrast different systems. However, in spite of so many limitations established rates are often given, and usually referred as the percentage of symbols or words correctly recognized Therefore in evaluation of OCR system, three different performance rates are necessary:

A. Recognition Rate

Recognition rate is the percentage of correctly identified symbols.

B. Rejection rate

Rejection rate is the percentage of symbols which the system was unable to recognize but they are not errors. Rejected characters can be set as a flag by the system, and will require manual correction which will also act as feedback to system.

C. Error rate

The percentage of characters that were identified wrong is the error rate. The only ways to correct this error is with the help of manual inspection. There is usually a trade-off between the different recognition rates. A low error rate may lead to a higher rejection rate and a lower recognition rate. Since there is a time limitation on the error to be corrected, the error rate is the effective criteria which is used to judge whether the system is cost-effective or not. An example from barcode reading may be the best example to illustrate the trade-off.

A rejection while reading a bar-coded price tag will only lead to either re-scanning of code or manual entry of the code, however a wrongly decoded price tag will result in the customer being charged for the wrong amount. In the barcode industry the error rates are therefore as low as one in a million labels, while a rejection rate of one in a hundred is acceptable. In view of this, it is necessary that it is not enough to look entirely on the recognition rates of a system. A correct recognition rate of 99%, might imply an error rate of 1%. In the case of text recognition on a printed page, this on average contains about 2000 characters, an error rate of 1% means 20 undetected errors per page. In applications such as mail sorting, where an address contains about 50 characters, an error rate of 1% implies an error on every other piece of mail which is not an ideal-case.

V. RESULTS

To demonstrate the accuracy of efficiency of hand-written English text and sample text images obtained from the web performance of the system was measured for our OCR. Figure 3 shows the sample document scanned from Cannon scanner at 200 dpi. The images were then filtered, binarized, clipped and resized and also skewing was performed. Lines of text were obtained from the image iteratively based on the value of sum of pixels. In the end, the font size was detected; segmentation was performed on each line to segment symbols considering the characteristics of English Verdana font’s templates [1]. OpenCV 2.4.8 is used to implement the proposed OCR algorithm. The recognition accuracy was 60% to 70% in the initial phase but when the errors were rectified by manual inspection, they were given back to the training model which extracted features. After a large number of experiment, the accuracy rose to 90% which will keep on increasing as more experiments are performed.

Fig. 4. Expected sample and its output
VI. FUTURE SCOPE

Various new methods mainly describing or using character recognition are still more expected to appear as the computer technology is developing and decreasing computational restrictions which ultimately makes the way out for new approaches. There might be a potential in case for performing character recognition straight directly on the grey level images. However, the greatest promising ideas seems to lie within taking advantage of existing methods, by adding methodologies and making more value of context.

Recognition of joined and split characters can be improved by integration of segmentation and conceptual analysis. Also, higher level contextual examination which looks mainly at the explanation of entire sentences may come out to be useful. Generally there is more efficiency in using context to a larger amount than what is there at present. In addition to this, it acts as a compound of various independent feature sets and analyzers, where the weakness of one method is resituated by the strength of another, this may finally help in improving the recognition of individual characters. The people pursuing research within character recognition have now shifted towards the recognition of cursive script, which is handwritten connected or based on calligraphic characters [3]. Bright techniques within this area mainly deal with the recognition of entire words instead of individual characters.

VII. CONCLUSION

Today optical character recognition is mainly recognized for bound object that is documents are producing under some command [3]. However, it seems that the need for forced OCR will get reduced in the coming future years. The reason for this is that command of the construction process usually means that the document is manufactured from material already stored on a computer. Hence, if a computer readable form is already there, this means that data may be transferred electronically or stamped in a more computer readable version, for particular barcodes. The usage for future OCR systems will lie in the approach of forms where command over the manufactured process is impossible. These are the material where the recipient is out source from an electronic form and has no command of the production process or old material which at manufactured time could not be generated electronically. This means that in the coming future, OCR systems are mainly intended for reading form text must be up front. Another important area aspect for OCR is the identification of manually created documents [3]. Within postal uses in these cases, OCR must mainly focus on the reading of the addresses on letter or mail written or produced by people without any clear access to computer advanced technology. Already, it is not favorable for the companies etc., with a collective approach to make use of computer technology to mark the corresponding mail with barcodes. The main ultimate idea of recognition of handwritten text is therefore expected to maximize.

REFERENCES