International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)

Optical Character Recognition

Jagruti Chandarana[1], Mayank Kapadia[2]
[1]Department of Electronics and Communication Engineering, UKA TARSADIA University
[2]Assistant Professor, Department of Electronics and Communication Engineering, UKA TARSADIA University.

***Abstract*— Optical character recognition (OCR) is very popular research field since 1950's. Character recognition techniques associate a symbolic identity with the image of character. In a typical OCR systems input characters are digitized by an optical scanner. Each character is then located and segmented, and the resulting character image is fed into a pre-processor for noise reduction and normalization. Certain characteristics are the extracted from the character for classification. The feature extraction is critical and many different techniques exist, each having its strengths and weaknesses. After classification the identified characters are grouped to reconstruct the original symbol strings, and context may then be applied to detect and correct errors.**

***Keywords*— Feature extraction, Segmentation, Template Matching and Correlation.**

I. INTRODUCTION

Optical character recognition has become one of the most successful applications of technology in the field of pattern recognition and artificial intelligence. Many commercial systems for performing OCR exist for a variety of applications, although the machines are still not able to compete with human reading capabilities. Optical Character Recognition deals with the problem of recognizing optically processed characters. Optical recognition is performed off-line after the writing or printing has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are drawn. Both hand printed and printed characters may be recognized, but the performance is directly dependent upon the quality of the input documents. In the last decade the acceptance rates of form readers on hand-printed digits and constrained alphanumeric fields have raised significantly (form readers usually run at a high reject/error ratio). Many researchers now view off-line and on-line cursive writing as the next challenge or turn to multi-lingual recognition in a variety of scripts. Character classification is also a favorite testing ground for new ideas in pattern recognition, but since most of the resulting experiments are conducted on isolated characters, the results are not necessarily immediately relevant to OCR.

II. BLOCK DIAGRAM OF PROPOSED METHOD

The process of character recognition consists of a series of stages, with each stage passing its results on to the next in pipeline fashion as shown in fig.2.1. There is no feedback loop that would permit an earlier stage to make use of knowledge gained at a later point in the process.

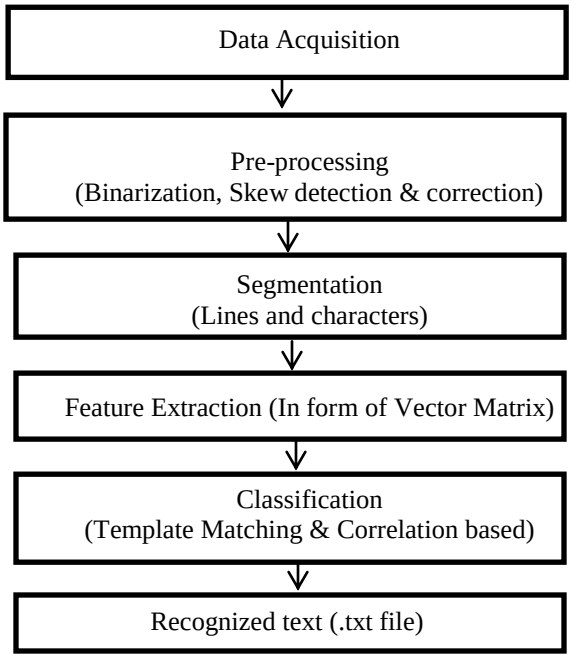

Figure 2.1: Block diagram of character recognition

Optical Character recognition is a system which loads a character (text) image, preprocesses the image, extracts proper image features, classify the characters based on the extracted image features (in the form of vector matrix) and the known features are stored in the image model library, and recognizes the image according to the degree of similarity between the loaded image and the image models. To recognize character firstly, the input images are acquired containing English text as an input image. Images are then stored in some picture file such as BMP, JPG etc. This image subsequently passes through preprocessing, segmentation, feature extraction and classification steps.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)

Preprocessing operations include image processing, binarization, noise reduction and skew detection & correction of a digital image so that subsequent algorithms along the road to final classification can be made simple and more accurate. Segmentation includes line segmentation-extract lines from a paragraph, and character segmentation-extract character from a line. After completing preprocessing and segmentation some features are extracted from the character image. Various feature extraction algorithms are there according to the behavior of the character.

The algorithm proposes an approach for extracting features in context of English character recognition. The techniques for extraction of such features are often divided into three main groups, where the features are found from:

• The distribution of points.

• Transformations and series expansions.

• Structural analysis.

Template-matching and correlation techniques are different from the others in that no features are actually extracted. Instead the matrix containing the image of the input character is directly matched with a set of prototype characters representing each possible class. The distance between the pattern and each prototype is computed, and the class of the prototype giving the best match is assigned to the pattern. The technique is simple and easy to implement in hardware and has been used in many commercial OCR machines. However, this technique is sensitive to noise and style variations and has no way of handling rotated characters.

III. Design And Implementation Steps

The steps of proposed algorithm for Character Recognition are implemented in MATLAB 7.6 version as per the above block diagram shown in fig 2.1.

3.1 Database Creation:

Initially, we have created a database of all character images having upper-case letters, lower-case letters and numeral digits of English scripts from A-Z, a-z and 0-9 of pixels 42×26.

3.2 Data Acquisition:

Through the scanning process a digital image of the original document is captured. Scanned images are then stored in some picture file such as BMP, JPG etc.

Figure 3.2: Sample input images (RGB images) with Skew (-8 degree & +5 degree)

3.3 RGB to gray conversion:

In the pre-processing 1st stage is to convert the input RGB image into gray scale image.

3.4 Binarization:

Binarization is the process of converting a gray scale image (0 to 255 pixel values) into binary image (0 and 1 pixel values) by selecting a threshold value in between 0 to 255 (here threshold value is 128).

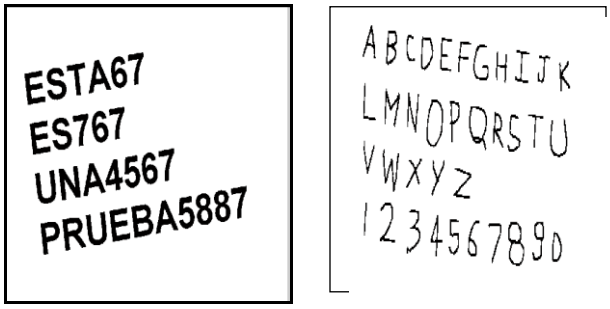

Figure 3.3: RGB images converted to Binary images

3.5 Skew Detection & Correction:

While scanning the image, if the paper/source document is not aligned properly, it may cause the components to be tilted. This could lead to erroneous behavior of the OCR system as shown in figure 3.2. To prevent this, Radon Transform (Skew detection & Correction method) has been devised, which detect & remove the skew from the image and later the boundaries of particular images are adjusted as shown in figure 3.5 so that image looks like an original image [16].

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)

Radon transforms maps Cartesian rectangular co-ordinates to the polar co-ordinates. Radon transform is a function of Rho and θ for each matching points calculated [16].

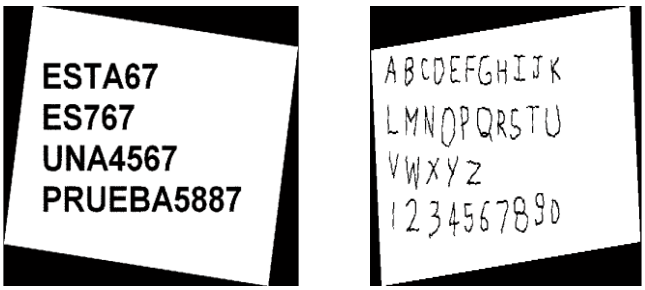

Figure 3.4: Skew Corrected Image (-8 degrees & 5 degrees)

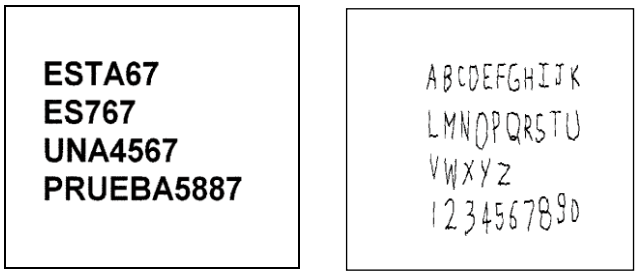

Figure 3.5: Boundary Adjusted Images

3.6 Segmentation:

It is an operation that seeks to decompose an image of sequence of characters into sub images of individual symbols. Character segmentation is a key requirement that determines the utility of conventional Character Recognition systems. It includes line, word and character segmentation. Different methods used can be classified based on the type of text and strategy being followed like recognition-based segmentation.

3.6.1 Line segmentation

In a printed script, the text lines are almost of same height, provided that the script is written in a specific font size. If the script is composed by a type-machine, surely the font size will be uniform everywhere. Between two text lines, there is a narrow horizontal band with either no pixel or very few pixels. Hence, by checking break-points through them and storing them will be useful for detecting the valleys in it, text line bands can be retrieved.

3.6.2 Character segmentation

After the line segmentation, consider each and every line which is segmented before going through the process of character segmentation. Each line is segmented in its individual characters (isolated) for further operation.

3.7 Feature Extraction

After character segmentation, features from each segmented character are extracted which is in the form of Matrix as shown in figure 3.6 [8].

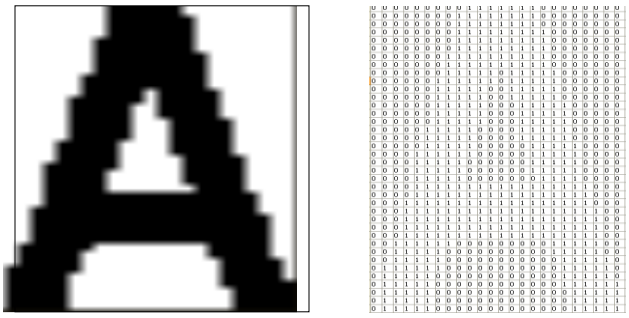

Figure 3.6: Character extraction in form of Matrix [8]

3.8 Template-matching and correlation method:

These techniques are different from the others in that no features are actually extracted. Instead the matrix containing the image of the input character is directly matched with a set of prototype characters representing each possible class. The distance between the pattern and each prototype is computed, and the class of the prototype giving the best match is assigned to the pattern [8]. The technique is simple and easy to implement in hardware and has been used in many commercial OCR machines.

$$\% \text{ Accuracy} = \frac{\text{No of characters found correctly}}{\text{Total no. of patterns}}$$

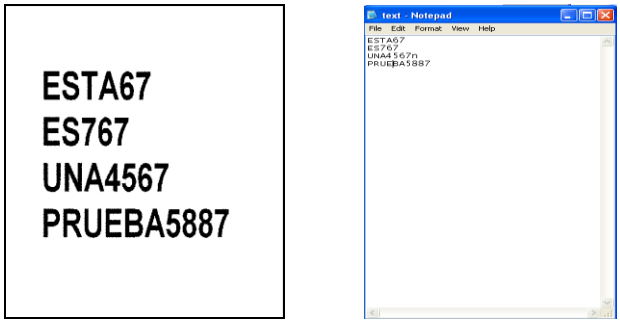

Figure 3.7: Text image sample and its output

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)

Figure 3.8: Handwritten Sample and its output

The text image samples and its output which is observed in notepad file are shown in figures 3.7 and 3.8.

IV. OCR PERFORMANCE EVALUATION

No standardized test sets exist for character recognition, and as the performance of an OCR system is highly dependent on the quality of the input, this makes it difficult to evaluate and compare different systems. Still, recognition rates are often given, and usually presented as the percentage of characters correctly classified. However, this does not say anything about the errors committed. Therefore in evaluation of OCR system, three different performance rates should be investigated:

4.1 Recognition rate:

The proportion of correctly classified characters.

4.2 Rejection rate:

The proportion of characters which the systems were unable to recognize. Rejected characters can be flagged by the OCR-system, and are therefore easily retraceable for manual correction [8].

4.3 Error rate:

The proportion of characters erroneously classified. Misclassified characters go by undetected by the system, and manual inspection of the recognized text is necessary to detect and correct these errors [8].

V. RESULTS AND INTERPRETATION

To illustrate the accuracy of English handwritten and different sample text images of different fonts of different sizes have been tested under OCR algorithm by using MATLAB (R2010.a/64-bit) and then performance was measured using the samples.

When a document is fed to a scanner, a few degrees of tilt (skew) are unavoidable [16]. Skew angle is the angle that the lines of text in the digital image make with the horizontal direction.

The algorithm approaches two methods for Skew detection and correction for the images and the comparison for % of accuracy for two methods is shown in table I from which it is observed that Radon Transform gives optimal solution with 100% accuracy.

Table I
Comparison table for % of Accuracy for Skew detection and Correction method

Methodology	Skew angle (Scanned image)	% of Accuracy
Hough Transform	1 to 15 degrees	99.60%
Radon Transform	1 to15 degrees & -6 to +6 degrees	100%

Table II
Recognition Accuracy of sample input images using proposed method

Images of Different fonts	Skew detection & Correction	Output Formats	Accuracy (Proposed Method)
Arial_Black.png	+\-5	Arial_Black.txt	96.77%
Calibri.png	+\-8	Calibri.txt	95.16%
Times_New_Roman.png	+\-12	Times_New_Roman.txt	83.87%
Trebuchet_MS.png	+\-15	Trebuchet_MS.txt	80.64%
TEST_2.jpg	+\-8	text.txt	98.38%
Handwritten-img.jpg	+\-5	Handimg.txt	80.55%

The images were then filtered, binarized and resized. Lines of text were then extracted from the images. The font size was identified; segmentation was performed on each line to segment characters taking in consideration the characteristics of English Verdana font's templates. MATLAB (R2010.a/64-bit) is used to implement the proposed OCR algorithm. The Output is obtained in the form of notepad file (.txt file). The comparison table below shows the performance measurement accuracy for different font images. The templates of all Characters and numbers are of 42X26 pixels.

Table III
Comparison table between existing method and proposed method

Existing Method [8]	Proposed Method
85-90%	91.16%

From table III, it is observed that the recognition accuracy is 91.16% on an average for the proposed method when the number of images is tested under the experiment and which is greater as compared with the existing method [8].

VI. CONCLUSION

A survey of feature extraction and classification techniques for optical character recognition is studied. A lot of research has been done in this field. Still the work is going on to improve the accuracy of feature extraction and classification techniques. Due to algorithmic simplicity and higher degree of flexibility, template matching and Correlation method is easy to implement with the change of recognition target classes. Its recognition is strongest on monotype and different types of fonts considering the sample input images for example handwritten image and it takes shorter time and does not require sample training but one template is only capable of recognizing characters of the same size. The OCR algorithm which is implemented in MATLAB (R2010.a/64-bit) gives optimal accuracy on an average as 91.16% when compared with existing method and also the Radon transform applied for skew detection and correction gives better results as compared with Hough transform.

REFERENCES

[1] R. C. Gonzalez & R. E. Woods. Digital Image Processing.Addison-Wesley,1992.

[2] S. Mori C.Y. Suen & K. Yamamoto. "Historical Review of OCR research and Development". IEEE Proceedings, special issue on OCR, pp. 1029-1057, July 1992.

[3] V. K. Govindan & A.P. Shivaprasad. "Character Recognition - a Review". Pattern Recognition, Vol. 23, No , PP. 671-683, 1990.

[4] Mori S., Optical Character Recognition, Wiley- Interscience,1999.

[5] Rice S., Nagy G., Nartker T., Optical Character Recognition:An Illustrated Guide to the Frontier, Springer, 1999.

[6] O. D. Trier, A. K. Jain and T. Text, "Feature Extraction Methods For Character Recognition- A Survey", Pattern Recognition, Vol. 29, No. 4, pp. 641-662, 1996.

[7] M. Zahid Hossain_ M. Ashraful Amin, Hong Yan, "Rapid Feature Extraction for Optical Character Recognition", Manuscript Draft, June 4, 2012.

[8] Sandeep Tiwari, Shivangi Mishra, Priyank Bhatia, Praveen Km. Yadav "Optical Character Recognition using MATLAB" International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) ISSN: 2278 – 909X , Volume 2, Issue 5, May 2013.

[9] M.-K. Hu, Visual pattern recognition by moment invariants,IRE Transactions Information Theory 8, 179–187,1962.

[10] Pritpal Singh, Sumit Budhiraja, "Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script – A Survey"International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 ,Vol. 1, Issue 4, pp. 1736-1739.

[11] M. Bokser, "Omnifont technologies," Proc. IEEE, vol. 80, pp. 1066–1078, 1992.

[12] K. M. Mohiuddin and J. Mao, "A comparative study of different classifiers for handprinted character recognition," Pattern Recognition Practice IV, pp. 437– 448, 1994.

[13] Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin, " A Survey of OCR Applications" International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012.

[14] Jatin M Patil, Ashok P. Mane, "Multi Font and Size Optical Character Recognition Using Template Matching" International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 1, and January 2013.

[15] Priya Sharma, Randhir Singh, "Survey and Classification of Character Recognition System," International Journal of Engineering Trends and Technology- Volume4 Issue3- 2013.

[16] Prakash K Aithal, Rajesh G, U Dinesh Achary and Siddalingaswamy P. C, " A Fast and Novel Skew Estimation Approach using Radon Transform", International Journal of Computer Information Systems and Industrial Management Applications. ISSN 2150-7988 Volume 5 (2012) pp. 337-344.