Optical Character Recognition Technique using Intro Sort

Yusuf Khan
Department of computer science & Engineering
Goel institute of Technology & Management

Kapil Kumar Gupta
Department of computer science & Engineering
Goel institute of Technology & Management

Namrata Dhanda, PhD
Department of computer science & Engineering
Goel institute of Technology & Management

ABSTRACT

In this study, we propose an optical character recognition technique using Intro Sort. Main feature of this proposed technique is that we segment images using intro sort. It reduces the comparison time for matching the pixels of an image. It reflects reduction in OCR time. Intro sort algorithm begins with quick sort and when recursion depth exceeds a level it switches to heap sort, based on the number of pixels being sorted. This approach also has advantage of recognizing number plates and text documents in very nominal time. Our approach is able to extract characters of different font sizes. Our technique is performed well in noisy images too.

General Terms

Optical character recognition

Keywords

OCR, Intro sort, Image segmentation, Feature extraction, Digital image processing.

1. INTRODUCTION

The idea behind Optical Character Recognition (OCR) is to identify optical patterns (often contained in a digital image) corresponding to alpha-numeric or other characters. The process of OCR contains several steps including segmentation, feature extraction, and classification. Optical character recognition is a technique of computer recognition of optically scanned and digitized character images to produce an electronic text document automatically. That knowledge or data can be used to find the characters in digital images. OCR [1] is becoming a necessary part of modern research based computer applications. Especially with the advent of Unicode and support of complex scripts on personal computers, the importance of this application has increased. OCR is worldwide used to convert books and documents into electronic files (Sarkar 2006), to automate record-keeping in an office (Doucet et al 2011)

The present study is focused on exploration of possible techniques to develop an OCR [2] system for English language when noise is present in the image. A detailed analysis of English writing system has been done in order to understand the core challenges. Existing OCR systems are also studied to know the latest research going on in this field. The emphasis was on finding workable segmentation technique and diacritic handling for English strings, and built a recognition module for these ligatures. The complete ideology is proposed to develop an OCR system for English and a testing application is also made.

2. PREVIOUS WORK

In the past decades the trend is to digitize (ancient) paper based documents such as articles, textbooks and newspapers has emerged. Claudiu et al. (2011) [1] has test the process using simple training data set. Georgios et al. (2010) [2] has presented a system (approach) for off-line handwritten character recognition. A novel recognition approach has been presented by Sankaran et al. (2012) [4] that results in a 15% decrease in character error rate on heavily diminished document images of Indian language. He addressed these problems by proposing to recognize character n-gram images, which are simply groupings of consecutive character or component segments. Jawahar et al. (2012) [5] has proposed a recognition theory for the Indian script of Devanagari. Recognition accuracy of Devanagari script is not yet comparable to its Roman counterparts. This is mainly due to the complexity of the script, style of writing etc. Zhang et al. (2012) [6] has discussed the misty, winter day's foggy weather, or hazy weather conditions lead to image color degradation and reduce the resolution and the contrast of the observed object in outdoor scene acquisition. Badawy, W. et al. (2012) [7] has presented the Automatic license plate recognition (ALPR) is the extraction of vehicle license plate information from an image or a series of images. The extracted information can be used with or without a database in many applications, such as electronic payment systems (toll payment, parking fee payment), and freeway and arterial monitoring systems for traffic surveillance. A novel adaptive binarization method has proposed by Yang et al. (2012) [8] based on wavelet filter, which shows comparable performance to other similar methods and processes faster, so that it is more suitable for real-time processing and applicable for wireless devices.

3. PROPOSED WORK

In our proposed work we use optical character recognition using Intro sort technique. The process starts from scanning the image to find the output as a readable and editable format. After scanning of image we first convert image into pixels and it passes for the next step, which is image segmentation. Image segmentation is performed by making cluster of pixels based on slightly changes in the pixels. Clustered pixel is being sorted by using Intro sort technique. After that we extract feature, based on similarity and dissimilarity of pixels. Now it is passed for post processing step, where we get output of readable text in .txt format. Following figure shows implementation of proposed work.

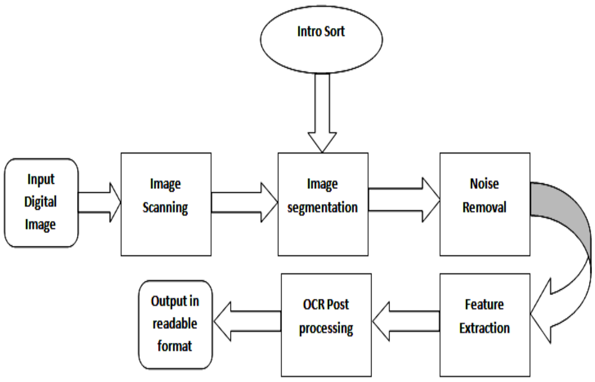


Fig 1: Components of Intro sort OCR System

Detail process is listed below:

3.1 Scanning of documents

Through the scanning process a digital image of the original document is captured. In OCR optical scanners are used, which generally consist of a transport mechanism plus a sensing device that converts light intensity into gray-levels.

3.2 Image segmentation using intro sort

The input image I_{PxQ} is, at first, clustered into m number of blocks B_i, i=1, 2,…, m such that $B_i \cap B_j = \emptyset$ and A block B_i is a set of pixels. Block B_i is clustered by using Euclidean distance between pixels by using

$$||x - y||_2 = \sqrt{\sum_n((x_n - y_n)^2)}. \qquad (1)$$

Clustered data is now sorted by using Intro Sort technique.

It this sorting technique, sorting of pixel start with quick sort but as number of pixel increases it switches to Heap sort. Switching of algorithm depends when recursion depth exceeds a level based on the number of pixels being sorted. It is the best of both worlds, with a worst-case O (n log n) runtime and practical performance comparable to Quick sort on typical pixel sets.

Complexity of Intro sort = O (n log n)

Where n is the number of pixels.

3.3 Noise removal step

The image resulting from the scanning process may contain a certain amount of noise. After Image segmentation step noisy pixel still present. To remove this noise we apply image smoothening algorithm. The smoothing implies both filling and thinning. Filling eliminates small breaks, gaps and holes in the digitized characters, while thinning reduces the width of the line. The most common techniques for smoothing, moves a window across the pixel set of the character, applying certain rules to the contents of the window. We compare central pixel to 8 neighbours and apply smoothening process.[9]

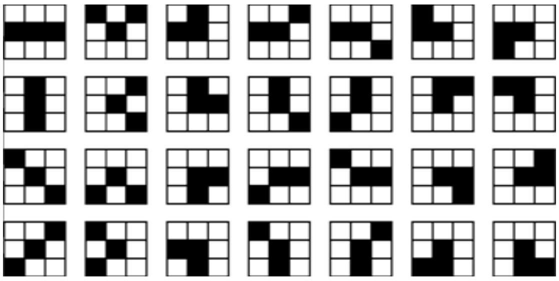


Fig 2: Comparisons of central pixel to its two neighbor's pixels in 3×3 Window

3.4 Feature extraction step

In this step we extract certain features that still characterize the symbols, but leaves out the unimportant attributes. The techniques for extraction of such features are often divided into three main groups. First is The distribution of points, transformations and series expansions and structural analysis of characters.

3.5 Post processing step

In the post processing step we group characters to form string. The process of performing this association of symbols into strings is commonly referred to as grouping. The grouping of the symbols into strings is based on the symbols' location in the document. Symbols that are found to be sufficiently close are grouped together.

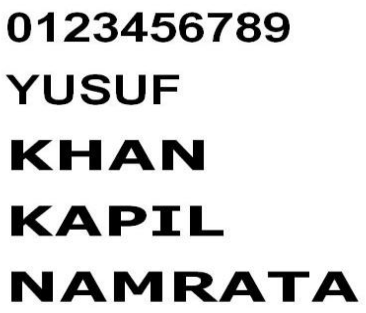
4. EXPERIMENTAL RESULT

4.1 Platform for evaluation

The platform utilized to evaluate the proposed approach includes a dual core CPU, the Intel Core 2 Duo with clock rate 2.4 GHz and memory 1 GB DDR2 667. The display card has GPU GeForce 7600 GT of NVIDIA Inc. and 1 GB memory. MATLAB is used for simulation of code and verifying result.

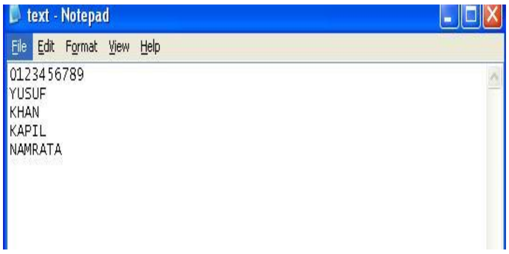
4.2 Test sample of images

In order to validate the proposed approach, we performed testing on 300 images. We test proposed approach on noisy as well as noiseless both training image set. Some results are listed below:



0123456789
YUSUF
KHAN
KAPIL
NAMRATA

Fig 3: Noiseless gray scale image



```
0123456789
YUSUF
KHAN
KAPIL
NAMRATA
```

Fig 4: Output received of image in Fig. 3 after passing through Intro Sort OCR

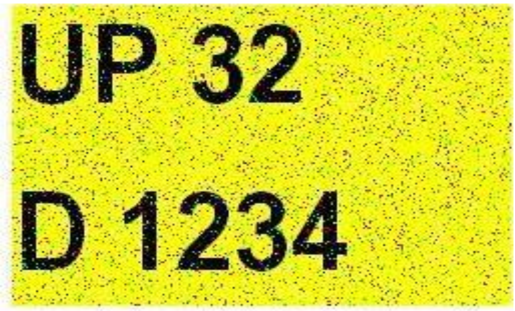
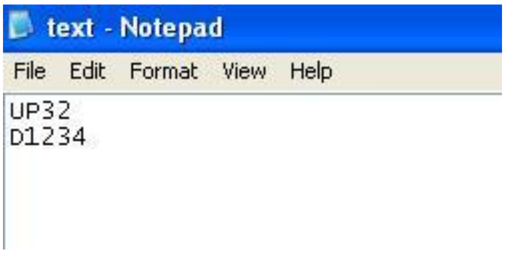


Fig 5: Noisy color image



```
UP32
D1234
```

Fig 6: Output received of image in Fig. 5 after passing through Intro Sort OCR

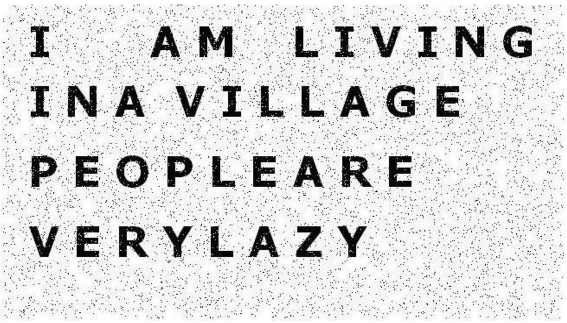
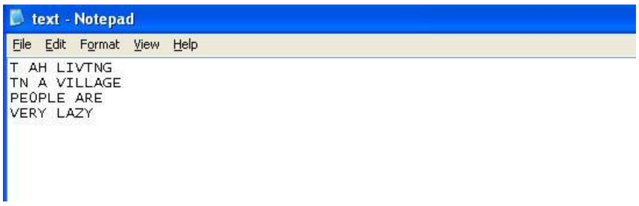


Fig 7: Noisy gray scale image



```
T AH LIVTNG
TN A VILLAGE
PEOPLE ARE
VERY LAZY
```

Fig 8: Output received of image in Fig. 7 after passing through Intro Sort OCR

5. COMPARISON

5.1 Comparison based on size of training data Vs. Elapsed time

We tested on 300 images for varying number of poses in training data. We conclude that as the image size increases, the character recognition time in Matlab also increases.

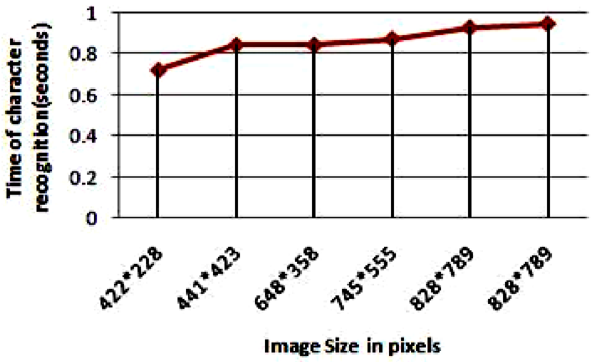
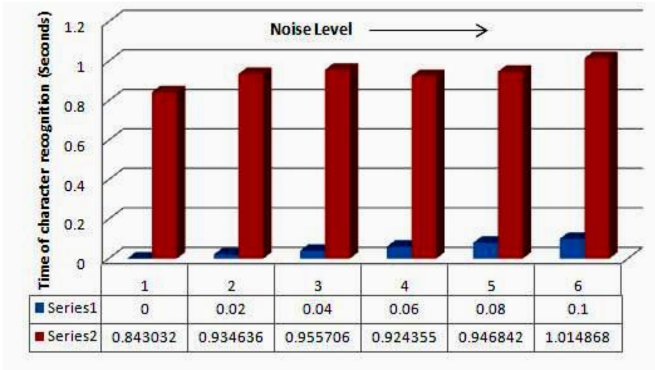


Fig 9: Comparison chart between image size and character recognition time

5.2 Comparison based on depth of noise level Vs. Elapsed time

We conclude that as the Noise level in an image increases, the character recognition time in Matlab also increases.



	1	2	3	4	5	6
■ Series1	0	0.02	0.04	0.06	0.08	0.1
■ Series2	0.843032	0.934636	0.955706	0.924355	0.946842	1.014868

Fig 10: Comparison chart between depth of noise level and character recognition time

6. CONCLUSION

It is found that proposed method is able to recognise character 100% in noiseless images. It is also able to recognise characters more than 99 % in noisy images. Proposed method is also significantly reduces the OCR time.

7. ACKNOWLEDGEMENTS

This work is supported by resources from Mr. Mahesh Goel, Managing Director, Goel Group of Institutions, Lucknow for promoting research activities in the Campus. They would also like to thank the anonymous reviewers for their significant and constructive critiques and suggestions, which substantially improved the quality of this paper.

8. REFERENCES

[1] Dan ClaudiuCires¸an and Ueli Meier and Luca Maria Gambardella and JurgenSchmidhuber, 2011 “Convolutional Neural Network Committees for Handwritten Character Classification”,International Conference on Document Analysis and Recognition, IEEE.

[2] Soumen Bag & Gaurav Harit, “A survey on optical character recognition for Bangla and Devanagari scripts”. Vol. 38, Part 1, February 2013, pp. 133–168. Indian Academy of Sciences

[3] GeorgiosVamvakas, Basilis Gatos, Stavros J. Perantonis, 2010 “Handwritten character recognition through two-stage foreground sub-sampling”, Pattern Recognition, Volume 43, Issue 8, August

[4] Shrey Dutta, Naveen Sankaran, PramodSankar K., C.V. Jawahar, 2012. “Robust Recognition of Degraded Documents Using Character N-Grams”, IEEE,

[5] Naveen Sankaran and C.V Jawahar, 2012. “Recognition of Printed Devanagari Text Using BLSTM Neural Network”, IEEE

[6] Yong-Qin Zhang, Yu Ding, Jin-Sheng Xiao, Jiaying Liu and Zongming Guo1, 2012. “Visibility enhancement using an image filtering approach”, Zhang et al. EURASIP Journal on Advances in Signal Processing.

[7] Badawy, W. (2012): 1-1. "Automatic License Plate Recognition (ALPR): A State of the Art Review."

[8] Yang, Jufeng, Kai Wang, Jiaofeng Li, Jiao Jiao, and Jing Xu. 2012. "A fast adaptive binarization method for complex scene images." In Image Processing (ICIP), 2012 19th IEEE International Conference on, pp. 1889-1892. IEEE

[9] Kapil Kumar Gupta, M. Rizwan Beg, Jitendra Kumar Niranjan.“A Novel Approach to Fast Image Filtering Algorithm of Infrared Images based on Intro Sort Algorithm”. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 1, November 2011 on pp. 235-241.

[10] Sukhpreet Singh, “Optical Character Recognition Techniques: A Survey”, Journal of Emerging Trends in Computing and Information Sciences, Vol. 4, No. 6 June 2013, on pp. 545-550.