Character Recognition: An Optimal Feature Based Approach

Kaiser J. Giri, Rumaan Bashir, Javaid Iqbal Bhat
Islamic University of Science and Technology, Awantipora, J&K
kaiser.giri@islamicuniversity.edu.in, rumaan.bashir@islamicuniversity.edu.in,
javaidonnet@gmail.com

Abstract

This paper presents an optimal feature based recognition technique for English characters. In order to improve the efficiency of the recognition system, the entire process is divided into two phases. In the first phase, the features necessary for recognition are extracted from the characters under consideration in an optimal manner and a feature matrix is generated so as to create a knowledge base. In second phase, a time efficient searching algorithm is used to recognize the characters using the already created feature matrix. The technique has been successfully tested on various fronts and satisfactory results have been obtained.

Keywords: *Character, Recognition, Feature Extraction, Node Matrix, Decision Tree, Active Nodes, Passive Nodes.*

1. INTRODUCTION

Character recognition is the process of extracting characters from a document image and converting them to their equivalent text [1]. Character recognition plays an important role in converting an electronic document stored as document image to its textual form which can later be edited and processed by the computers. Character recognition comprises of a series of steps such as digitization, preprocessing, segmentation, extraction of features, their classification and knowledgebase creation. Character recognition being an essential activity in document image processing & analysis, has undergone through research over past few decades [2- 7]. A number of character recognition techniques related to various languages have been developed so far, however the development of methods to read the text with same capability as those of humans is still an open area for researchers.

Among many other languages, English has comparatively a wide range of audience and is therefore being most widely used across the world. Recognition of characters belonging to English language has therefore resulted into an active area of research and many techniques have been developed accordingly [3, 8]. The traditional character recognition systems suffer from low recognition accuracy due

to variety complexities that arise during different stages of recognition process. These complexities include methods of capture, noise, variety of notations, illumination variance, methods of drawing, complex & dirty background etc. [9-10]. Since the feature selection and their extraction play an important role and affect the overall performance of the recognition process [11]. In this paper a cognitive approach based on selection of an optimal number of features has been proposed for recognition of English alphabets. It has been observed that overall processing time has been reduced substantially. The proposed method moreover has extensively reduced the storage requirements.

2. RELATED WORK

The automatic recognition of characters is considered to be an important step in document image analysis. The field on character recognition has gained significant importance due to availability of huge collection of document images, which need processed by computers.

The studies on character recognition can be traced back to 1900. The earliest studies were focused on mechanical devices instead of computers. The studies on character recognition were initially started by Russian scientist Tyurin in the year 1900 [12].

Sheppard in 1951 invented a reading and robot GISMO capable of reading musical notations and words on printed pages. This works is considered to be as the maiden work on modern OCR [13]. The work on character recognition during earlier days however was more focused towards either on printed text, well distinguished hand written text or small set of symbols. The algorithms developed were mostly for Latin characters and numerals. However some studies on Chinese, Japanese, Hebrew, Cyrillic, Indian, Greek and Arabic characters and numerals have also been reported [14-16].

Over the past few decades a substantial amount of research work has been done on character recognition which has resulted into emergence document image analysis and various multi-lingual & omni-font OCR's [17].

Bindo & Goutam [18] presented a Hidden Markov Model based character recognition technique. The proposed technique uses a novel feature extraction method. The technique has been test over 13000 samples from different writers and an accuracy rate of 94% has been reported.

Majida & Hamid [19] proposed a model of an OCR and discussed its various components. The proposed technique uses Particle Swarm Optimization Approach to achieve better recognition results.

Giri et al., [20-21] presented two character recognition techniques for offline printed character. The first technique is based on the concept cognition where a minimum number of parameters have been taken into consideration for purpose of recognition. In second technique, the recognition is done on the basis of structural analysis of the characters under consideration. Both the techniques have shown improved results in terms of space & time complexities.

Lipi et al., [22] presented a linear recognition based OCR technique for recognition of handwritten & printed Guajarati script. The proposed technique uses linear recognition for error detection and correction.

3. PROPOSED METHOD

The proposed technique uses a very simple and time efficient feature selection mechanism. The character under consideration is initially converted into a binary image and bounded in a hypothetical reference rectangle. The coordinates of the reference rectangle are later used for selection of features necessary for creation of knowledgebase used during recognition of the characters. The proposed technique takes a single character into consideration from which noise has already been removed. The particular character is scanned from all sides till it strikes the boundary of the character and there is a transaction due to change in the intensity values and a hypothetical reference rectangle is drawn around the particular character based on this information as shown in figure 1(a). The coordinates of the hypothetical rectangle are later used to draw a hypothetical 3X3 grid around the character under consideration irrespective of the size of the character as shown in figure 1 (b). The intersections of the grid are termed as nodes and labeled as node 1 to node9 as shown in figure 1(c). The intensity at each node (intersection of the grid is used to create the knowledgebase for recognition of the characters. If the intensity at a particular intersection (node) is 0, the corresponding node is designated as an active node, otherwise it is designated as a passive node as shown in figure 1(d). A list of active nodes is generated accordingly as shown in figure 1(e). The procedure is repeated for all characters and a corresponding list of active nodes is generated for each of them as shown in figure 2. The node list is finally used to recognition a particular character using a decision tree as shown in figure 3.

The characters having same list of active nodes (G & O and H & N) are further processed by applying the same procedure in one of their grid sections till a unique node list is obtained. It is evident from the search tree that the maximum time search time is 7 iterations and minimum search time is 2 iterations irrespective of the size of the character which is significantly less compared to some of the known techniques such as template matching etc.

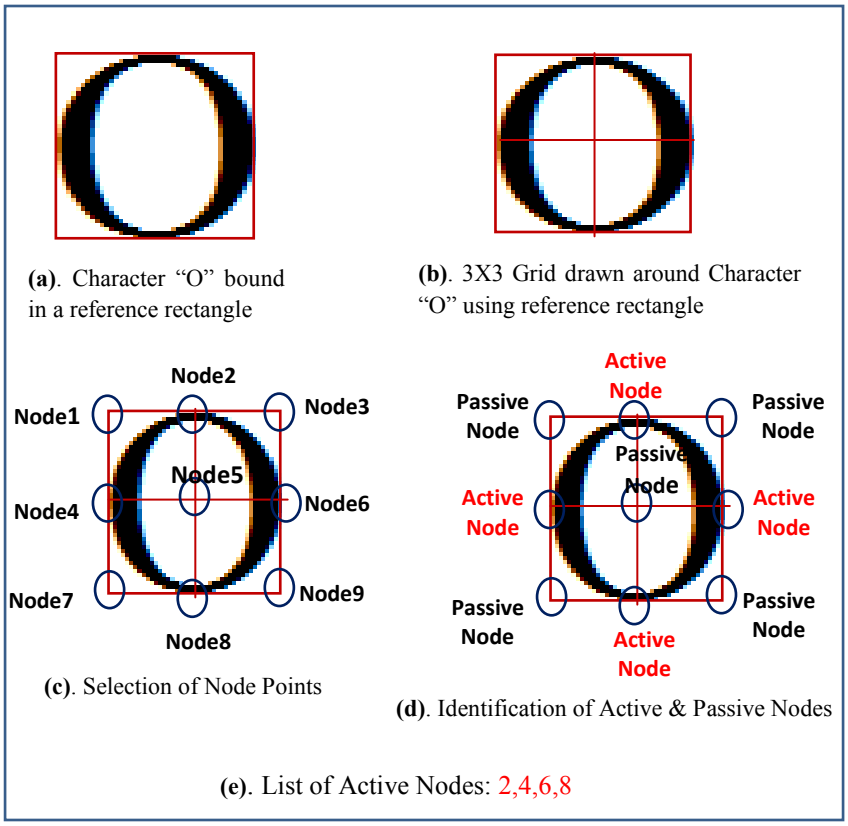

Figure 1: Illustration for Generation of active node list for Character "o"

Character	List of Active Nodes	Character	List of Active Nodes
A	2, 7, 9	N**	1, 3, 4, 5, 6, 7, 9
B	1, 2, 4, 5, 7, 8	O*	2, 4, 6, 8
C	2, 4, 8	P	1, 2, 4, 5, 7
D	1, 2, 4, 6, 7, 8	Q	2, 4
E	1, 2, 4, 5, 7, 8, 9	R	1, 2, 4, 5, 7, 9
F	1, 2, 3, 4, 5, 7	S	2, 4, 8
G *	2, 4, 6, 8	T	1, 2, 3, 5, 8
H**	1, 3, 4, 5, 6, 7, 9	U	1, 3, 4, 6, 8
I	1, 2, 3, 4, 5, 6, 7, 8, 9	V	1, 3, 8
J	3, 6, 8	W	1, 2, 3
K	1, 4, 5, 7, 9	X	5, 7, 9
L	1, 4, 7, 8, 9	Y	1, 3, 5, 8
M	1, 3, 4, 6, 7, 8, 9	Z	2, 5, 7, 8, 9

Figure 2: List of Active Nodes for each character

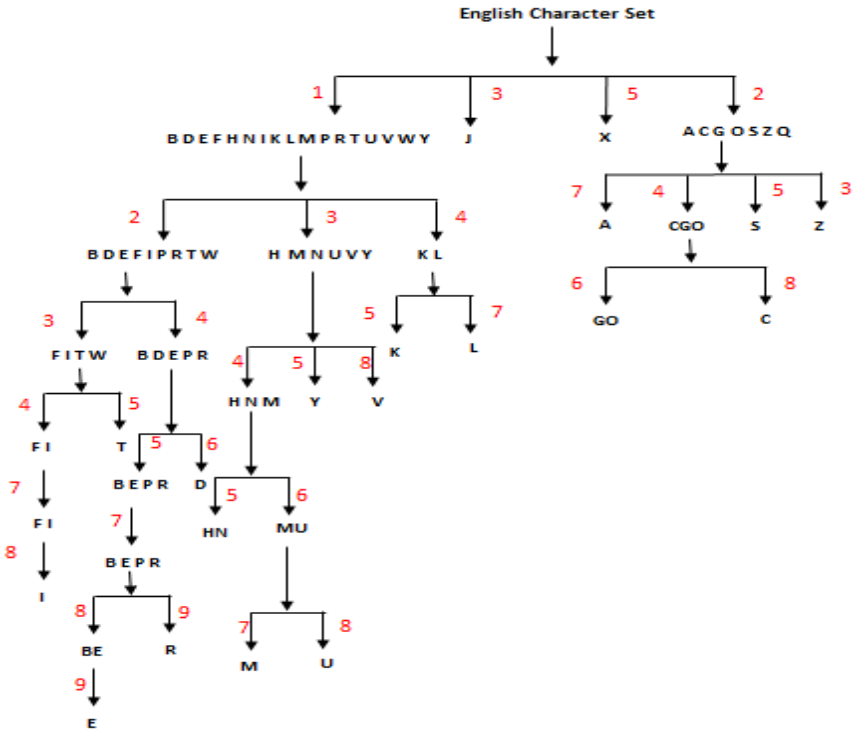

Figure 3: Search Tree for Recognition of a Particular Character

4. CONCLUSION

This paper presents an efficient recognition technique for machine printed offline characters. The number of parameters necessary for recognizing a particular character is considerably less which in turn helps to significantly reduce the time & storage requirements. The number of parameters moreover is independent of the size of the characters to be recognized. The technique has been tested on English characters in the first instance; however it can be improved to recognize the characters of other languages such as Arabic, Hindi, and Persian etc. The technique is very simple to implement and has promising results in terms of recognition of characters.

5. REFERENCES

1. A. M. Al-Shatnawi, "A new method in image steganography with improved image quality," Applied Mathematical Sciences, vol. 6, no. 79, pp. 3907-3915, 2012.
2. Pereira e Silva, G. and Dueire Lins, R., "An Automatic Method for Enhancing Character Recognition in Degraded Historical Documents", in Proc. Int. Conf on Document Analysis and Recognition (ICDAR), 2011, pp: 553 – 557
3. Sulistiyo, M.D. , Saepudin, D. and Adiwijaya, "Optical Character Recognition using modified Direction Feature and Nested Multi Layer Perceptrons Network", in Proc. Int. Conf. On Computational Intelligence and Cybernetics (CyberneticsCom), 2012, pp: 30 – 34
4. Zaafouri, A., Sayadi, M. and Fnaiech, F., "Printed Arabic character recognition using local energy and structural features", in Proc. 2nd Int. Conf. on Communications, Computing and Control Applications (CCCA),2012,pp 1- 5
5. Mohammad Tanvir Parvez, Sabri A. Mahmoud, " Arabic handwriting recognition using structural and syntactic pattern attributes", Pattern Recognition 46(2013) 141-154
6. Chaivatan Sumetphong, Supachai Tangwongsan, "Modeling broken character recognition as a set-partitioning problem", Pattern Recognition Letters 33(2012), 2270-2279
7. Yang Yang, Xu Lijia & Cheng Chen, "English Character Recognition Based on Feature combination", Intl. Conf. on Advances in Engineering, Elsevier Engineering 24 (2011) 159–164
8. Marisa R. De Giusti, María Marta Vila, Gonzalo Luján Villarreal, Manuscript Character Recognition Overview of features for the Feature Vector, JCS&T Vol. 6 No. 2 October (2006)

9. Sushila aghav, Prof. S. S. Paygude, "Computer Assisted Character recognition in document based images", Elsevier Engineering 38 (2012) 3222 – 3227 ICMOC (2012)

10. Anju K Sadasivana & T. Senthilkumarb, "Automatic Character Recognition in Complex Images", Intl. Conf. on Communication Technology and System Design, Elsevier Engineering 30 (2012), 218– 225

11. Dar M. D., P. Naghabhushan, Chisti. N., "Document Image Classification: A Cognitive Approach", Intl. Conf. on computer vision and information technology, Dr. Baba Ambedkar University, Awrangabad, Nov. 15 to 18, (2007)

12. V. K. Govindan and A. P. Shivaprasad, Character recognition- A review, Pattern recognition J. vol. 23, no 7 pp. 671-683, 1990.

13. Satti, D.A., 2013, Offline Urdu Nastaliq OCR for Printed Text using Analytical Approach. MS thesis report Quaid-i-Azam University: Islamabad, Pakistan. p. 141.

14. Adisleven Lev and Miriam Furst, Recognition of Handwritten Hebrew One-Stroke Letters by Learning Syntactic Representations of Symbols, IEEE Trans. on Sys. Man and Cybernetics, vol: 19, no: 5, pp.1306-13131, 1991.

15. El. Sheikh and R. M. Guindi, Computer Recognition of Arabic Cursive Scripts, Pattern Recognition, vol. 21, no. 4, pp.293-302, 1988.

16. Shumji Mori, Kazuhiko Yamamoto and Michio Yasuda, Research on Machine Recognition of Handprinted Characters, IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) vol. 6 , no. 4, pp.386-404, 1984.

17. Mahmoud, S.A., & Al-Badr, B., 1995, Survey and bibliography of Arabic optical text recognition. Signal processing, 41(1), 49-77.

18. Binod Kumar Prasad, GoutamSanyal A model Approach to Off-line English Character Recognition International Journal of Scientific and Research Publications, Volume 2, Issue 6, June 2012

19. Majida Ali Abed Hamid Ali Abed Alasadi Simplifying Handwritten Characters Recognition Using a Particle Swarm Optimization Approach European Academic Research, Volume I, Issue 5/ August 2013

20. Kaiser J. Giri, Rumaan Bashir. (2013), "Design and Implementation of a Novel Cognitive Character Recognition Technique", IEEE 2013 International Conference on Signal Processing and Communication organized by Jaypee Institute of Information Technology, Nodia, India from 12-14 December, 2013

21. Kaiser J. Giri, Rumaan Bashir. (2013), "Character Recognition Based on Structural Analysis Using Code & Decision Matrix", IEEE International Conference on Machine Intelligence

and Research Advancement (ICMIRA-2013) organized by SMVDU, Katra (Jammu), India from 21-23, December 2013.

22. Lipi Shah, Ripal Patel, Shreyal Patel, Jay Maniar Skew Detection and Correction for Gujarati Printed and Handwritten Character using Linear Regression International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 1, January 2014