Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 2, February 2017, pg.95 – 100

A Brief Review on Optical Character Recognition Techniques

Maninder Kaur
Department of Electronics & Communication Engineering, Rayat Bahra University, Mohali, India
Maninderkr90@gmail.com

Miss. Manjeet Kaur
Assistant Professor, Department of Electronics & Communication Engineering, Rayat Bahra University, Mohali, India
Manjeetkaurlehal1989@gmail.com

Abstract

In Image process, segmentation has important phase in OCR and numerous articles are revealed on completely different segmentation strategies like cutting, bar chart etc. for various scripts during previous couple of years. Usually there's not work done on Overlapped and touching scripts. Typically as a result of poor handwriting, the author left some gap between diacritics and character or between diacritics and header line as a result of that little text blocks gets created that ends up in improper text line segmentation and thus ends up in wrong results and overlapping. As a result accuracy of the algorithmic program degrades. In planned work reconciling SVM would be accustomed improve accuracy of the system.

I. Introduction

Image process is a sort of signal process that image is enter as input, resembling a photograph and image process output could also be whether or not a picture or, a sequence of characteristics or parameters associated to image. Image Segmentation is that the methodology within which digital image is split into range of regions and set of pixels. Image partitioning is of various texture and objects. In alternative words, we are able to say that it's outcomes of set of regions that cover the whole image along and a collection of contours extracted from the image. All of the pixels in an exceedingly neighborhood are equivalent with relevance some characteristics similar to color, depth, or texture.

Segmentation conjointly contains three major steps like line segmentation, word segmentation and character segmentation. Handwritten character recognition is difficult task compared to machine written character recognition within the space of Optical Character Recognition. In character recognition field, OCR could be a

crucial space. it's a method to scan document and create the document editable. OCR field is extremely standard in business world and banking industries. OCR system converts the text image in machine-encoded kind that reduces the house needed for storage. OCR is that the one amongst the strategies accustomed digitalize the written documents that create transmission of documents simple. OCR method consists of three major sub-phases like Pre-processing, Segmentation and Recognition. Various Applications of an OCR are: Header line and base line detection, Grouping approach Word Segmentation, Character Segmentation, Overlapping and Touching of Characters, Gap finding between diacritics and header line. Now this paper will describe the things which are organised in following way. Next section II will describe the various techniques. Section III describes the related work about the paper and at last section IV contains conclusion.

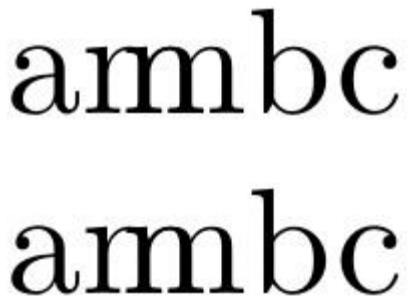

Fig. 1 Overlapping of Character

II. Techniques on Image Processing

There are various techniques for image processing in character readings:

Pre-processing

OCR software system typically "pre-processes" pictures to boost the probabilities of undefeated recognition. Segmentation of fixed-pitch fonts is accomplished comparatively just by orientating the image to an identical grid supported wherever vertical grid lines can least typically ran into black areas.

Character recognition

There are two basic kinds of core OCR formula, which can manufacture a hierarchal list of candidate characters. Matrix matching involves comparison a picture to a hold on glyptic art on a pixel-by-pixel basis; it's conjointly called "pattern matching", "pattern recognition", or "image correlation". This depends on the input glyptic art being properly isolated from the remainder of the image, and on the hold on glyptic art being in a very similar font and at constant scale. this system works best with written text and doesn't work well once new fonts square measure encountered. this is often the technique the first physical photocell-based OCR enforced, rather directly.

Feature extraction decomposes glyphs into "features" like lines, closed loops, line direction, and line intersections. These are compared with an abstract vector-like illustration of a personality, which could reduce to at least one or a lot of glyptic art prototypes. General techniques of feature detection in laptop vision are

applicable to the current form of OCR, that is usually seen in "intelligent" handwriting recognition and so newest OCR software system. Nearest neighbor classifiers like the k-nearest neighbors formula are accustomed compare image options with hold on glyptography options and opt for the closest match.

Post-processing

OCR accuracy are often inflated if the output is unnatural by a lexicon – a listing of words that square measure allowed to occur in a very document. This may be, as an example, all the words within the West Germanic language, or a lot of technical lexicon for a selected field. This system is often problematic if the document contains words not within the lexicon, like correct nouns.

Application-specific optimizations

In recent years, the foremost OCR technology suppliers began to tweak OCR systems to raised agitate specific varieties of input. on the far side AN application-specific lexicon, taking into consideration business rules, normal expression, or rich data contained in color pictures will have higher performance. This strategy is termed "Application-Oriented OCR" or "Customized OCR", and has been applied to OCR of license plates, invoices, screenshots, ID cards, driver licenses, and automobile producing.

III. Related Work

Bansal et al. (2010) [1]: This paper elaborates the segmentation of various irregular text words written in Gurumukhi script. This paper deals with the segmentation of words containing skewed, broken, irregular headline, touching and overlapped characters. Some of the new techniques like counter tracing methods are used along with horizontal and vertical projections.

Garg N. et.al (2011) [2]: Character recognition is an important stage of any text recognition system. In Optical Character Recognition (OCR) system, the presence of half characters decreases the recognition rate. Due to touching of half character with full characters, the determination of presence of half character is very challenging task. In this paper, they have proposed new algorithm based on structural properties of text to segment the half characters in handwritten Hindi text. The results are shown for both handwritten Hindi text as well as for printed Hindi text. The proposed algorithm achieves the segmentation accuracy as 83.02% for half characters in handwritten text and 87.5% in printed text.

Kumar et al. (2014) [3]: This paper presents the segmentation of handwritten Gurumukhi characters is carried out defining the whole process for segmentation including digitization process and pre-processed techniques. Water Reservoir method is applied for identification and segmentation of touching characters.

Kumar et al. (2010) [4]: This paper proposed a technique in which the segmentation of the scanned document image is done. In which the whole image is consider as a one large window. The large window is split into less large windows as giving lines and once the lines are recognized then each window consisting of a line is used to recognize a word that is present in a line and at the end character is recognized. This paper uses the concept of variable sized window.

Kumar and Singh (2011) [5]: It was tested on different documents, the results obtained were encouraging were detected with a great accuracy. The lines, which were having some characters in the lower zone, were interpreted almost correctly. To get the character, the coordinates of detected lines and words are used. For

character segmentation process was divided in two part, (i) to get the segmented region R (ii) to check, if R has a meaningful symbol or not. This can be a reverse approach to ensure correct segmentation, i.e. if R does not have a meaningful symbol then R is readjusted. As per the data shown in the Table 3, there is certain incorrect segmentation too. After close analysis, we found that this is due to the shapes of characters. Certain characters in Gurumukhi script are combined in nature. But overall, results were good and encouraging.

Literature Table

Sr. No.	Year	Authors	Proposed Work	Advantage	Disadvantage
1	**2010**	**Bansal G., Sharma D**	**Segmentation of words containing skewed, broken, irregular headline, touching and overlapped characters.**	**It is working upon handwritings**	**Accuracy is very low**
2	**2011**	**Garg N.K., Kaur L., Jindal M.K**	**Authors have proposed algorithm based on structural properties of text to segment the half characters in handwritten Hindi text.**	**Half character segmentation For hindi text**	**Limited availability, could be enhanced for other languages also**
3	**2014**	**Kumar D., Koshti, Govilkar S**	**Authors have carried out the research with defining the whole process for segmentation**	**Handwritten characters could be recognized**	**Better approach could be used**

			including digitization process and pre-processed techniques		
4	**2010**	**Kumar M., Jindal M.K., Sharma R.K**	**Authors have worked upon segmentation of the scanned image**	**Isolated and Touching characters of scanned documents**	**Model must be tested on large dataset**
5	**2011**	**Kumar R., Singh A**	**Authors have provided an algorithm which is used to segment the scanned document image as a lines, words and characters.**	**Concept used for find co-ordinates of character caused high accuracy**	**Used only for Punjabi language**

IV. Conclusion

Due to overlapping and touching of characters, there remains no significant gap between the text lines and hence two or more text lines comes in a same text block which leads to wrong results. In the existing system, overlapped character Recognition work was done on grey scale images by using SVM. In proposed system, Adaptive SVM scheme will be applied using weight based scheme. Documentation image would be taken and convert into histogram. To improve accuracy SVM algorithm would be enhanced in this approach. The main focus in this research project is to experiment deeply with, and find alternative solutions to the image segmentation and character recognition problems.

References

[1]. Bansal G., Sharma D., "Isolated Handwritten Words Segmentation Techniques in Gurmukhi Script", International Journal of Computer Applications, Vol. 1, No. 24, pp. 104-111, 2010.

[2]. Garg N.K., Kaur L., Jindal M.K. "The segmentation of half characters in Handwritten Hindi Text", SpringerVerlag Berlin Heidelberg, pp. 48-53, 2011.

[3]. Kumar D., Koshti, Govilkar S., "Segmentation of Touching Characters in Handwritten Devanagri Script", International Journal of Computer Science and its Applications, Vol. 2, Issue 2, pp. 83-87.

[4]. Kumar M., Jindal M.K., Sharma R.K., "Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition", International Journal Information Technology and Computer Science, pp. 58- 63, Feb, 2014.

[5]. Kumar R., Singh A., "Algorithm to Detect and Segment Gurmukhi Handwritten Text into Lines, Words and Characters", IACSIT International Journal of Engineering and Technology, Vol.3, No.4, 2011.

[6]. Kumar R., Singh A., "Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text" Institute of Electrical and Electronics Engineers (IEEE), pp. 353-356, 2010.

[7]. Mangla P., Kaur H., "An End Detection Algorithm for segmentation of Broken and Touching characters in Gurumukhi Word", Handwritten Institute of Electrical and Electronics Engineers (IEEE) , pp.1-4, 2014.

[8]. Mehta B., Rani S., "Segmentation of Broken Characters of handwritten Gurmukhi Script", International Journal of Engineering Sciences, Vol. 3, pp. 95-105, 2014.

[9]. Kumar R., Singh A., "Challenges in Segmentation of Text in Handwritten Gurmukhi Script" Proceedings in BAIP 2010, CCIS 70, Springer-Verlag Berlin Heidelberg, pp. 388-392, 2010

[10]. Binny Thakral, Manoj Kumar, "Devanagari Handwritten Text Segmentation for Overlapping and Conjunct Characters- A Proficient Technique". 978-1-4799-6896-1,IEEE 2014.

[11]. Naunita, Taneja A., Chawla M., "Segmentation of Touching Characters in Handwritten Gurumukhi Script", International Journal of Engineering Sciences, Vol. 3, pp. 90-94, 2014.

