ISSN 2395-1621

DETECTION & EXTRACTION OF CHARACTER FROM PRINTED TEXT AND IMAGE

[#1]Gaikwad Divya, [#2]Pharande Kajal, [#3]Sakshi Bindroo, [#4]Prof .S.K.Bhatia

[1]divyagai30@gmail.com

[#1234]Department of Electronics & Telecommunication
Jspm's Imperial College Of Engineering & Research
Wagholi, Pune- 412207

ABSTRACT

Character extraction is the most critical pre-processing step for any off-line text recognition system because the characters are the smallest unit of any language script. We have proposed an approach to segment character images from the text containing images and computer printed words. This segmentation approach is based on a set of properties for each connected component (object) in the whole binary image of the machine printed or handwritten text containing some other images. These words which are printed along with some images are of different lengths and are printed by different cursive fonts of different sizes.

ARTICLE INFO

Article History

Received: 25th May 2017

Received in revised form :

25th May 2017

Accepted: 31st May 2017

Published online :

1st June 2017

I. INTRODUCTION

A character being the smallest unit of any language script and therefore the segmentation of characters is the most crucial step for any OCR (Optical Character Recognition) System. this technique interprets scanned or printed image of the document into a text document which will be altered. the choice of segmentation algorithmic program being employed is that the key factor in deciding the accuracy of the OCR system. If there's an honest segmentation of characters, the recognition accuracy also will be high. Segmentation of words into characters becomes terribly tough because of the cursive and free nature of the handwritten script. Image processing could be a terribly different field of technology. Human race is highly desirous to train their system to know the traditional human legible format to computer readable format. Since several decades researchers are engaged on this aim and through their rigorous effort they have finally discovered OCR as a really booming application to convert human legible text to computer readable. The techniques like neural networks, structural and applied mathematics pattern are in the market for recognition of text. However the most important disadvantage of neural network is massive training information that takes a lot of| time to create this makes the neural network more sophisticated for a naive user to know leading it to less user friendly.

Character segmentation and recognition has been a vigorous field of analysis for several years. It still remains an open drawback within the field of pattern recognition and image process. There are primarily 3 phases of a character recognition system specifically preprocessing, segmentation and recognition. Preprocessing aims at eliminating the variability that is inherent in written words. The preprocessing techniques like background signal removal, scaling, cutting skew removal etc. are used by numerous researchers in an endeavor to extend the performance of the segmentation and recognition process; the role of segmentation is to seek out correct letter boundaries. Segmentation precedes character recognition, which suggests that the output of segmentation becomes the input to the character recognition module. Segmentation of off-line cursive words into characters is one amongst the foremost tough and vital method in handwriting recognition because it directly affects the results of recognition method Most scientific and engineering publications contain mathematical symbols and expressions. Recognition of written mathematics wouldn't solely need less effort in writing technical documents however may even be used to transfer existing written documents into electronic format and between machines once required. thus written arithmetic recognition is one in all the key forces that drive the knowledge transformation between human and machine. written arithmetic recognition has been studied for over thirty years. As mathematical expressions seem in sizable amount of scientific documents, no doubt transferring such documents into electronic format needs utilities for

recognition of mathematical content. Handwriting input provides traditional and appropriate means of inputting mathematical text into pc for storage or sharing with others, another time underlining the requirement of effective mathematical recognition software system.

Text extraction is a vital innovate document image analysis and it doesn't have a universal accepted answer. so as to segment text from a page document it's necessary to discover all the attainable manuscript text regions. Text Image segmentation is usually accustomed find objects and limits (lines, curves, etc.) in pictures. It implies a labeling method that assigns a similar label to spatially align units i.e. pixel, connected parts or characteristic points such a bunch of pixels with the similar label share specific visual options. The results of image segmentation could be a set of segments that jointly cover the whole image, or a group of contours extracted from the image (edge detection). Every of the pixels in a very region is comparable with regard to some characteristic properties, like color, intensity, or texture. Adjacent regions are considerably totally different with regard to a similar characteristic.

II. LITERATURE SURVEY

1.Samrajya P, Lakshmi M, Hanmandlu, Swaroop A. projected that, segmentation is a crucial step within the longhand recognition method. a better recognition rate is achieved if the characters of a word are properly isolated. Hypergraph model to section a cursive handwritten word image into isolated characters. Hypergraph model treats an image as packets of pixels. Authors claim that by recombining these packets of various sizes a given word image may be divided into characters if a minimum of one amongst the mixtures provided an accurate segmentation. However, neither segmentation results are bestowed for comparison nor the technique appears to yield successful results for touching characters. The technique is tested on CEDAR benchmark information containing cursive written words of various persons. the greatest advantage of this methodology is its simplicity in separating the characters of the word without losing any data, once the segmentation points are situated. there's no need to write separate program to seek out the segmentation methods.

2.Dawoud A. projected that, the repetitive cross section sequence graph (ICSSG) is an algorithmic program for written character segmentation. It expands the cross section sequence graph idea by applying it iteratively at equally spaced thresholds. repetitive cross section sequence graph (ICSSG) for the character phase at particle. ICSSG tracks the characters growth at equally spaced thresholds. The repetitive thresholding reduces the impact of knowledge loss related to image binarization. However, the experiments are performed on written digits uniquely.

3.Lee H, Verma B. has proposed that, a new segmentation algorithmic rule for off-line cursive handwriting recognition. Initially, word pictures are compound heuristically supported picture element density between higher and lower baselines. every section expeirenced multiple skilled primarily based validation processes to determine valid character boundaries. an over-segmentation algorithmic rule is introduced to dissect the words from handwritten text supported the picture element density between higher and lower baselines. each and every section from the over-segmentation is passed to a multiple expert-based validation method. first professional compares the overall foreground picture element of the segmentation point to a threshold value. the threshold is set and calculated before the segmentation by scanning the stroke parts within the word. Second professional checks for closed areas like holes. Third professional validates segmentation points employing a neural voting approach that is trained on metameric characters before validation method starts. during this system, a unique segmentation paradigm for off-line handwritten text recognition has been planned and investigated. The segmentation paradigm contains a baseline pixel-based over-segmented, hole detection, section foreground picture element comparison and a neural voting primarily based validation. Also, outsized segment analysis is performed before manufacturing final segmentation points. The new segmentation paradigm has been tested on cursive handwritten text. The planned segmentation approach produced lowest errors compared to existing approaches..

4 Rehman A, Dzulkifli M. has proposed a new, easy and quick approach for character segmentation of unconstrained handwritten words. The developed segmentation algorithmic rule over-segments in some cases due to the inherent nature of the cursive words. but the over segmentation is minimum. to extend the potential of the algorithmic rule an artificial Neural Network is trained with vital quantity of valid segmentation points for cursive words manually. Segmentation is that the vital step of analytical approaches utilized to handwritten word recognition. therefore it's the bottom of latest approaches. it's admitted undeniable fact that no segmentation methodology will directly find character location accurately without an intelligent methodology. during this paper, proposed segmentation algorithmic rule is integrated with neural network using standard back propagation.

5. Namrata Dave, has propsed that, written Character Recognition is space of analysis since many years. Automation of existing manual system is need of most industries as well as government areas. Recognition of hand written characters is a demand for several fields
. during this paper they have mentioned our approach for hand written character segmentation. This paper discusses varied methodologies to segment a text primarily based image at varied levels of segmentation. This paper is a guide for individuals acting on the text primarily based image segmentation space of computer Vision. First, the necessity for segmentation is justified within the context of text primarily based data retrieval. Then, the various factors poignant the segmentation method are mentioned. Followed by the amount of text segmentation are explored.

III. BLOCK DIAGRAM

Following Fig shows the proposed system block diagram. This system implement for character recognition from

printed image. The Matlab software is used to implement a system. The image preprocessing steps are shown in figure.

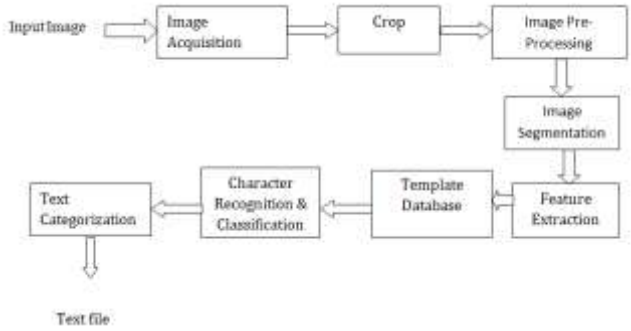

Fig 1. Block diagram

Cropping:

Cropping refers to the removal of the outer components of a picture to boost framing, intensify material or modification ratio. reckoning on the appliance, this could be performed on a physical photograph, design or film footage, or achieved digitally mistreatment image redaction computer code. The term is common to the film, broadcasting, photographic, graphic style and printing industries.

This is vital as a result of in input image there ar several different information's apart from our fascinating half like varied shapes, objects etc that isn't fascinating in optical character recognition system.

If we tend to don't do the cropping then the knowledge that isn't needed additionally gets processed which supplies failing result.

Fig 2. Cropping

Image Pre-processing:

Preprocessing is a crucial step of applying variety of procedures for smoothing, enhancing, filtering etc, for creating a picture usable by consequent formula so as to enhance their readability for optical character recognition code. The pre-processing could be a series of operations performed on scanned input image. It primarily enhances the image rendering it appropriate for segmentation. The role of pre-processing is to phase the fascinating pattern from the background. Generally, noise filtering, smoothing and normalisation ought to be drained this step. The pre-processing conjointly defines a compact illustration of the pattern. Binarization method converts a grey scale image into a binary image.

Image getting is that the activity of sick photos from supply, unremarkably associate instrumentation primarily based supply.

Image pre-processing is needed to get rid of unwanted distortions and enhance the image options. There area unit varied image representations and filtering techniques which will cut back the impact of lighting conditions and improve image quality. when acquisition of image some preprocessing is finished on noninheritable image. Pre-processing includes following processes,

Image Segmentation

Segmentation precedes character recognition, which implies that the output of segmentation becomes the input to the character recognition module. Segmentation of off-line cursive words into characters is one amongst the foremost tough and vital method in handwriting recognition because it directly affects the results of recognition method Segmentation step is one amongst the foremost vital and tough task in text extraction and recognition. Segmentation is that the method of decomposition of various objects by extracting their individual boundaries and also the text part is isolated from the background. during this step the input pre-processed image consisting of sequence of character is there by rotten into sub pictures. foremost the image is segmental line by line, then the road is segmental word by word and any the word is segmental into characters. The segmental character is provided as Associate in Nursing input for consequent step. Segmentation is one amongst the foremost vital and essential method that decides the success rate of character recognition system. Segmentation is that the method of partitioning a picture / document into disjoint and homogenous regions. This task is earned by finding the boundaries. There area unit many approaches for locating the character bounds. during this stage, a picture of sequence of characters is rotten into sub-images of individual character. The input image was copied vertically from the higher left corner and every one the connected parts were known primarily based o n their foreground space. All the valid connected parts has been extracted and enveloped in an exceedingly rectangular region with smallest doable space.

There area unit varied factors that hinder the method of text primarily based image segmentation.

A few area unit as follows:

Image Quality:

The quality of the image may be a important issue for text segmentation. Presence of noise within the image ends up in degradation of accuracy and potency.

Handwritten or written Document:

Most text line segmentation strategies area unit supported the assumptions that distance between neighboring text lines is precise likewise as that text lines area unit equitably straight. However,

These assumptions aren't characterised for written documents. just in case of written document, text image segmentation may be a leading challenge. The prior, is that the case of the written text document. For such a document segmentation is a simple task, attributable to the biradial nature of the document. The line, word and even character

spacing is outlined, that get rid of the challenges as Janus-faced with written documents.

Orientation of text content:

For written document if the individual lines don't seem to be straight or if there's a presence of skew then the quality for text extraction will increase.

Rough document:
Presence of texture, like pictures, patterns, et al. within the text document makes the task of Segmentation varied.

Type of Text:
Cursive text provides further problem throughout character segmentation, thanks to the presence of ligatures.

Feature Extraction
In machine learning, pattern recognition and in image process, feature extraction starts from associate initial set of measured knowledge and builds derived values (features) meant to be informative and non-redundant, facilitating the following learning and generalization steps, and in some cases resulting in higher human interpretations. Feature extraction is said to spatiality reduction.
When the input file to associate algorithmic rule is just too giant to be processed and it's suspected to be redundant, then it are often remodeled into a reduced set of options. This method is named feature choice. the chosen options area unit expected to contain the relevant info from the input file, so the required task are often performed by mistreatment this reduced illustration rather than the entire initial knowledge.
Feature extraction involves reducing the number of resources needed to explain an oversized set of knowledge. once performing arts analysis of complicated information one in all the main issues stems from the amount of variables concerned. Analysis with an oversized range of variables typically needs an oversized quantity of memory and computation power, conjointly it's going to cause a classification algorithmic rule to over acceptable coaching samples and generalize poorly to new samples. Feature extraction could be a general term for ways of constructing mixtures of the variables to induce around these issues whereas still describing the info with adequate accuracy.

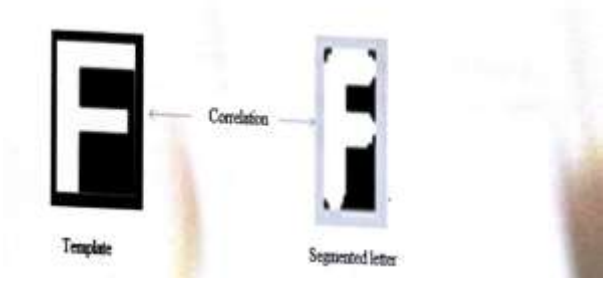

Fig 2. Correlation

Algorithm:
Get letter cropped image.
Correlate cropped letter with the pre-saved guide.
Store correlation result into a vector.
Notice & store the address of most price of correlation in vector.
Show the character that relates with the address calculated in previous step.
Move to step one.

Character Recognition:
Character recognition method depends upon range of things like varied font sizes, noise, broken lines or characters etc. and these factors influence the results of recognition system.

Algorithm:
Step 1: Load input color image.
Step 2: Crop or resize the input image.
Step3: Convert the grayscale image into black and white image.
Step4: Remove the objects that need less than 15 pixels.
Step 5: Separate out lines one by one.
Step 6: Separate out each letter one by one from every separated line.
Step7: Correlate separated letter with templates and store the result in vector.
Step8: Determine address of maximum value of correlation vector.
Step9: Show the detected letter as per address calculated in previous step.

IV. RESULT

Fig. 3 Color image

V. CONCLUSION

As an overall view of the system prototype, we conclude that this system prototype has been developed by using the technique that has been mentioned and elaborated which id the Template Matching approach to recognize the character image. Besides, the interface of the system prototype looks user friendly and makes the user of this system prototype easier to use it.
As a result, the recognition process of this system while recognizing the character. Even though this system prototype could give several advantages to the users, but this system prototype are facing a number of limitations like the system prototype has some limitation related to

performance and it works only with stored templates of alphabets and numbers with fixed size templates.

REFERENCES

[1] Dawoud A. "Iterative cross section sequence graph for handwritten character segmentation". IEEE Trans Image Process; 2007, 16(8) 2150 2154.

[2] Lee H, Verma B. "A novel multiple experts and fusion based segmentation algorithm for cursive handwriting recognition" In: ; 2008, 2994 2999.

[3] Rehman A, Dzulkifli M."A simple segmentation approach for unconstrained cursive handwritten words in conjunction with the neural network". Int J Image Process 2(3); 2008, 29 35.

[4] Saba T, Rehman A, Sulong G. "Cursive script segmentation with neural confidence". Int J Innov Comput Int Control; 2011, 7(7).

[5] Samrajya P, Lakshmi M, Hanmandlu, Swaroop A. "Segmentation of cursive handwritten words using Hypergraph", TENCON, IEEE region 10 Conference; 2006, 1-4.

