MATEC Web of Conferences **57**, 01010 (2016) DOI: 10.1051/matecconf/20165701010
ICAET-2016

Text Character Extraction Implementation from Captured Handwritten Image to Text Conversionusing Template Matching Technique

Seema Barate[1], Chaitrali Kamthe[2], Shweta Phadtare[3], Rupali Jagtap[4], *M.R.M Veeramanickam*[5]

[1, 2, 3, 4]*Under Graduate Students of Information Tech. Dept., TCOER, Pune*

[5]*Asst. Prof., TCOER, Pune*

Abstract. Images contain various types of useful information that should be extracted whenever required. A various algorithms and methods are proposed to extract text from the given image, and by using that user will be able to access the text from any image. Variations in text may occur because of differences in size, style,orientation, alignment of text, and low image contrast, composite backgrounds make the problem during extraction of text. If we develop an application that extracts and recognizes those texts accurately in real time, then it can be applied to many important applications like document analysis, vehicle license plate extraction, text- based image indexing, etc and many applications have become realities in recent years. To overcome the above problems we develop such application that will convert the image into text by using algorithms, such as bounding box, HSV model, blob analysis,template matching, template generation.

1 INTRODUCTION

Handwritten text from image has been a subject of study and research for many years. There are different methods currently employed to improve the visibility of text in text documents. We use different algorithms to improve the accuracy of the existing system. In this method to improve readability of text documents through manually enhancing text strokes is proposed [1]. We develop such application that will convert the image into text by using algorithms, such as bounding box, HSV model,blob analysis, template matching, and template generation.

To develop an interface to train the application for understands the difficult and unreadable handwritingCollect handwritten characters, symbols as an input data and train the application. Give handwritten character input as an image and extract the character. Generate the document file from the given image.

We assume the input as A-Z or a-z. To simplify the recognition process, certain pairs of similar upper and lower case characters are combined into singleClasses like Cc, Kk, 00, Ii, Pp, Ss, Uu, Vv, Ww, Xx, Yy, and Zz etc [2]. The database that is used for these experiments consists of about 10,000 isolated hand printed characters. Connected component analysis was used to locate blobs that were about the size of characters.The characters were manually identified and Stored in the database [2].

In case of offline character recognition the handwritten character is typically scanned in form of a paper document and made available in the form of a gray scaleimage to the recognition algorithms [3].

2 Literature Study

To improve the text of historical document through the multi-resolution Gaussian and to improve the readability of the text two algorithms are used LU algorithm and Gaussian filter. Advantage is to detect text of different scale. Disadvantage is work for image of small resolution (Oliver A. Nina) [1].

Text images are recognized in the text file using a character based bi-gram language model for that bi-gram language model is used .Advantage is to improve the recognition performance. Disadvantage is the presented approach will be verified on text images only (Yanwei Wang,Xiaoqing Ding,Fellow,IEEE and Changsong Liu) [2].

To translate images of typewritten or handwritten characters into electronically editable format by preserving font properties OCR algorithm is used. But it can't recognize multiple font size. In this paper used algorithm is Optical Character Recognition (OCR).Advantage is decrease some possible human errors and high speed of recognition. Disadvantage is multiple font and size characters and handwritten characters are not recognize (Faisal Mohammad,Jyoti Anarase,Milan Shingote and Pratik Ghanwat) [3].

A handwritten Arabic text written by each writer can be seen as a having a specific texture and characterization of texture image is based on run length of image gray level. In this paper two methods are used Gray Level Run Length (GLRL), Gray Level Co-occurrence Matrices

[2] Corresponding author: chaitukamthe029@gmail.com

(GLCM). Advantage is uses IFN/ENIT database which is unique Arabic publicly available dataset. Disadvantage is only extract global feature (DJEDDI Chawki and SOUICI-MESLATI Labiba) [4].

Various techniques are used to extract the text from image but it can be use only detect the text area, two algorithms are used OCR and Edge detection Advantage is any types of images content extracts accurately. Disadvantage is only text area is detected (A.J.Jadhav ,Vaibhav Kolhe and Sagar Peshwe)[5].

Image is captured by the camera, that image contains many text formats. This text information can be extracted from the image by using edge detection, collected component and texture based algorithms, two algorithms are used Compare Prewitt edge detector and Gaussian edge detector. Advantage is only character extraction removes non text image. Disadvantage is work only for natural scene image (Prof. Amit Choksi ,Nihar Desai,Ajay Chauhan,Vishal Revdiwala and Kaushal Patel) [6].

To extract text lines and words from handwritten document Segmentation Algorithm and Viterbi Algorithm are presented . The line segmentation algorithm is based on locating the optimal succession of text and gap areas within vertical zones by applying Viterbi algorithm. Word segmentation is based on a gap metric that exploits the objective function of a soft-margin linear SVM that separates successive connected components. Advantage is it provides 97 percent accuracy. Disadvantage is it does not provide generalize well to variations encountered in handwritten documents (Vassilis Papavassilioua,Themos Stafylakisa,Vassilis Katsourosa andn George Carayannisa)[7].

3 Proposed system : Text extraction using template matching

In proposed system,we are using different algorithms to improve the accuracy of the text document or text image[16].It uses the image as input can be captured by laptop camera. After that image is matched using the template matching. It uses the filter to filtering image means to remove the noise, low frequency then it takes only the high frequency images. Then the image is cropped by using bounding box. Then convert image into Gray scale using HSV model, HSV means Hue, Saturation and Value .Extract the character from the image using the blob algorithm. Blob means separate the word into character and each character is called blob. Then matches the character with the trained dataset using template matching algorithm .Then generate the character sentence by using template generator. Then generate document file(.doc file).

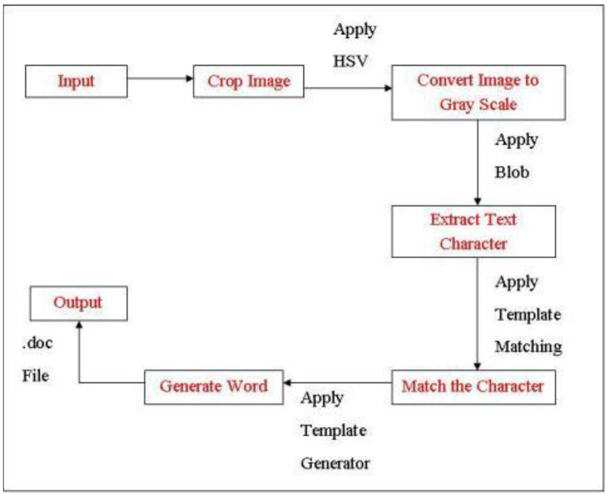


Figure 1 Architectural Diagram

4 MODULES

4.1 Template Matching

Template matching is a technique for finding areas of an image that match to a template image.In template matching, individual image pixels are used as features.Each comparison results in a similarity measure between the input character and the template. Template matching forcharacter recognition is reliable .The template matching works on two primary components:

Source image (I): match captured image to the template image

Template image (T): The patch image which will be compared to the template image

a) To detecting the highest matching area:../../../../../_images/Template_Matching_Template_Theory_Summary.jpg
b) To identify the matching area, we have to compare the template image against the source image by sliding it:../../../../../_images/Template_Matching_Template_Theory_Sliding.jpg

By sliding, means moving the patch one pixel at a time i.e left to right, up to down. At each location, a metric is calculated and it represents how “good” or “bad” the match at that location.

For each location of T over I, you store the metric in the result matrix (R). Each location (x,y) in R contains the match metric:../../../../../_images/Template_Matching_Template_Theory_Result.jpg

the image above is the result R of sliding the image with a metric TM_CCORR_NORMED. The brightest locations indicate the highest matches. If you can see, the location marked by the red circle is probably the one with the highest value, so that location of the rectangle formed by that point as a corner and width and height equal to the patch image is considered the match.we use the function minMaxLoc to locate the high value or low value, depending of the type of matching method in the R matrix.

OpenCV implements Template matching in the function matchTemplate. There are four methods :

a) method=CV_TM_SQDIFFR(x,y)= \sum _{x',y'} (T(x',y')-I(x+x',y+y'))^2

b) method=CV_TM_SQDIFF_NORMEDR(x,y)= \frac{\sum_{x',y'}(T(x',y')I(x+x',y+y'))^2}{\sqrt {\sum_{x',y'}T(x',y')^2\cdot\sum_{x',y'} I(x+x',y+y')^2}}

c) method=CV_TM_CCORRR(x,y)= \sum _{x',y'} (T(x',y') \cdot I(x+x',y+y'))

d) method=CV_TM_CCORR_NORMEDR(x,y)= \frac{\sum_{x',y'}(T(x',y')\cdot I(x+x',y+y'))}{\sqrt{\sum_{x',y'}T(x',y')^2 \cdot \sum_{x',y'} I(x+x',y+y')^2}}

Source:http://docs.opencv.org/2.4/doc/tutorials/imgproc/histograms/template_matching/template_matching.html

4.2 Bounding Box

The "**bounding box**" of a limited geometric object is the box with minimum area or minimumvolume, that contains a given geometric object. For any collection of linear objects like points segments ,polygon ,lines ,etc their bounding box is given by the minimum and maximum coordinate values for the point set **S** of all the object's *n* vertices[15].Suppose finding the area of a rectangle, area means the "level ground, an open space,"The number of square units it takes to completely fill a rectangle.
Formula: Width × Height
Try this Drag the orange dots to move and resize the rectangle. The area of a rectangle is given by multiplying the width times the height. As a formula: where
w is the width
h is the height so the formula is w*h.

Source:http://www.mathopenref.com/coordtriangleareabox.html

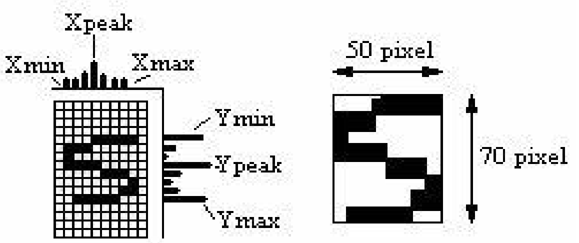


Figure 2: Example of Bounding Box
Source:Nafiz Arica and Fatos T. Yarman-Vural,"An Overview of Character Recognition Focused on Off-Line Handwriting".

4.3 Blob analysis

Blob Analysis is a fundamental technique of machine vision based on analysis of consistent image regions. The numbers of words are divided into characters, the performance of the algorithm both in terms of accuracy and efficiency is increases. Hence the number of words should be decided based on scanned image.

4.4 HSV model

The HSV color model, also called HSB (Hue, Saturation, Brightness), defines a color space in terms of three constituent components:
- Hue is the color type (such as red, magenta, blue, cyan, green or yellow). Hue ranges from 0-360 deg.
- Saturation refers to the intensity of specific hue. Saturation ranges are from 0 to 100%. In this work saturation is presenting in range 0-255.
- Value refers to the brightness of the colour. Saturation ranges are from 0 to 100%. Value ranges are from 0-100%. In this work saturation and value arepresenting in range 0-255. RGB to HSV conversion formula ,The R,G,B values are divided by 255 to change the range from 0..255 to 0..1:
R' = R/255,G' = G/255,B' = B/255
Cmax = max(R', G', B')
Cmin = min(R', G', B')
Δ = Cmax - Cmin
Value calculation:V = Cmax

Table1 RGB to HSV colour

Colourname	Hex value	(R,G,B)	(H,S,V)
Black	#000000	(0,0,0)	(0º,0%,0%)
White	#FFFFFF	(255,255,255)	(0º,0%,100%)
Red	#FF0000	(255,0,0)	(0º,100%,100%)
Blue	#0000FF	(0,0,255)	(240º,100%,100%)
Yellow	#FFFF00	(255,255,0)	(60º,100%,100%)
Green	#008000	(0,128,0)	(120º,100%,50%)
Cyan	#00FFFF	(0,255,255)	(180º,100%,100%)
Magenta	#FF00FF	(255,0,255)	(300º,100%,100%)

Source:http://math.stackexchange.com/questions/556341/rgb-to-hsv-color-conversion-algorithm

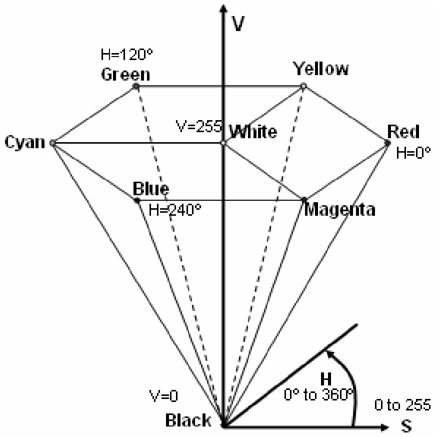


Figure 3 HSV Color Model

Source: Lidiya Georgieva, Tatyana Dimitrova, Nicola Angelov," RGB and HSV colour models in colour identification of digital traumasImages", (ICCST).

4.5 Template Generator

Template Generator is used for automatic generation of templates of numerals of a certain font. After applying template matching, template generator is used to form the word from the separated characters in blob analysis.

5 Implementation of text extraction

Template Matching

First image is captured with the help of camera or browse image. Then template matching algorithm is used to calculate the size of the image, if image size is greater than 15cm x 15cm then don't accept the image, otherwise accept it.

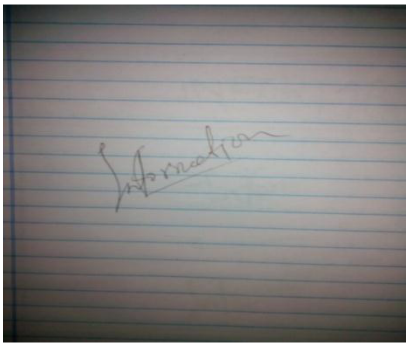
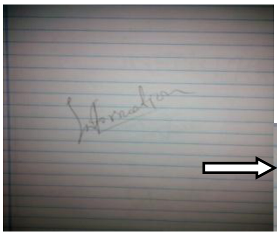


Figure 4 Input Image

Bounding Box

The output of Template Matching module i.e. 15cm x 15cm size image is taken as an input for Bounding Box.For discard the unwanted portion of the image use bounding box algorithm. By applying bounding box algorithm it take only text region from the image and remaining part from the image is discarded.



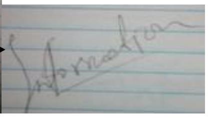
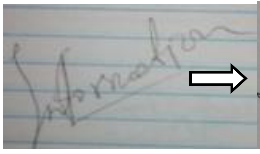


Figure 5 Before applying BoundingBox

Figure 6 Afterapplying Bounding Box

HSV Model

In HSV image is calculate Hue, Saturation and Value separately as a pixel. HSV model is used to convert the image into gray scale. HSV model takes the input from Bounding Box. Then combine all these pixel as output. In output image display the text by black color and make background as white.



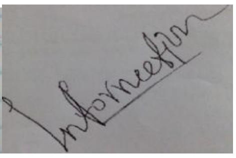


Figure 7 Beforeapplying HSV HSVModel

Figure 8 Afterapplying Model

Blob Analysis

After getting gray scale image separate the word into the character using blob analysis. Each character is called as a blob. Each character is match with the training database.

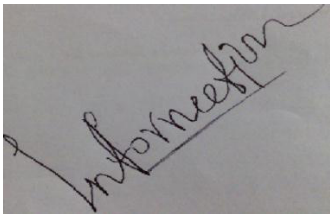


Figure 9 Before applying Blob Analysis

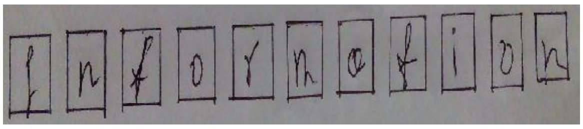


Figure 10 After applying Blob Analysis

Template Matching

Template matching algorithm is used for matching blob or character with training dataset. To match blob with training dataset feature extraction algorithm is used. If character is matched accurately then that character is given as input to template generator.

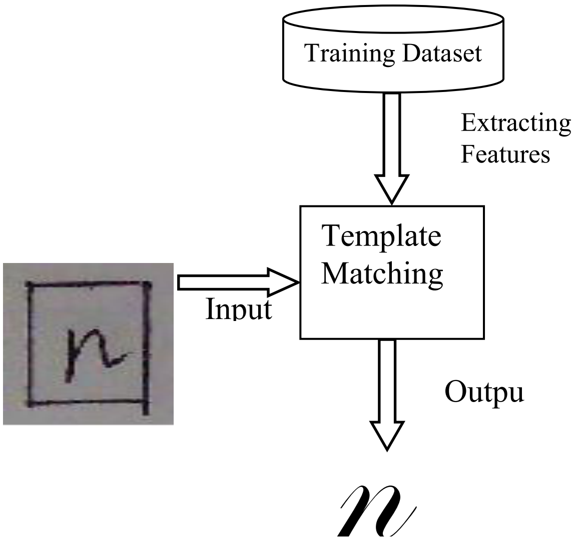
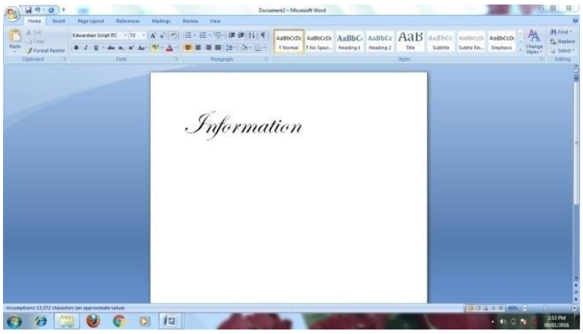


Figure 11. Working of Template Matching

Template Generator

In template generator input is taken from template matching, each generated character is combined into word and generate the.doc file.



Information

Figure 12. Final output using word doc

6 CONCLUSION

Though we have large no of algorithms and methods for text extraction from image butnone of them provide accurate result .Pattern Matching can be implemented successfully for English language text character recognition. The system has image processing modules for converting text image to generate .doc file. The expected experiment recognition rate will be good.

7 Future Scope

As it is restricted only for English language in future make this application other than English language. In future make this application for Android mobile phones as an App. And will be train for graphical symbols also.

References

1. Oliver A. Nina, “Interactive Enhancement of Handwritten Text through Multi-resolution Gaussian”, in 2012 International Conference on Frontiers in Handwriting Recognition.
2. Jonathan J. HULL, Alan COMMIKE and Tin-Kam HO,” Multiple Algorithms for Handwritten Character Recognition”.
3. Yanwei Wang, Xiaoqing Ding, Fellow, IEEE, and Changsong Liu,” Topic Language Model Adaption for Recognition of Homologous Offline Handwritten Chinese Text Image”, IEEE SIGNAL PROCESSING LETTERS, VOL. 21, NO. 5, MAY 2014.
4. Faisal Mohammad, Jyoti Anarase, Milan Shingote, Pratik Ghanwat “Optical Character Recognition Implementation Using Pattern Matching”,Faisal Mohammad et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2088-2090.
5. DJEDDI Chawki, SOUICI-MESLATI Labiba,” A texture based approach for Arabic Writer Identification and Verification”, 978-14244-8611-3/10/$26.00 ©2010 IEEE.
6. A.J.Jadhav Vaibhav Kolhe Sagar Peshwe, “ Text Extraction from Images: A Survey”, in Volume 3, Issue 3, March 2013, ISSN: 2277 128X.
7. Prof. Amit Choksi1, Nihar Desai2, Ajay Chauhan3, Vishal Revdiwala4, Prof. Kaushal Patel5, “Text Extraction from Natural Scene Images using Prewitt Edge Detection Method”, Volume 3, Issue12, December 2013, ISSN: 2277 128X.
8. Vassilis Papavassilioua,Themos Stafylakisa,Vassilis Katsourosa and George Carayannisa, "Handwritten document image segmentation into text lines and words",2010.
9. Debapratim Sarkar, Raghunath Ghosh ,A Bottom-Up Approach of Line Segmentation from Handwritten Text ,2009.
10. Bartosz Paszkowski, Wojciech Bieniecki and Szymon Grabowski Katedra Informatyki Stosowanej, Politechnika ALodzka, "Preprocessing for real-time handwritten character recognition".
11. Juan Chen,Chuanxiong Guo,"Online Detection and Prevention of Phishing Attacks".Mohammed Ali Qatran,"Template Matching Method For Recognition Musnad Characters Based On Correlation Analysis".
12. Mr.Danish Nadeem,Miss.Saleha Rizvi,"Character Recognition Using Template Matching".
13. Harmeet Kaur Kelda ,Prabhpreet Kaur,"A Review: Color Models in Image Processing".
14. Nafiz Arica and Fatos T. Yarman-Vural,"An Overview of Character Recognition Focused on Off-Line Handwriting".
15. Prof. Primož Potočnik ,Student: Žiga Zadnik,”Handwritten character Recognition: Training a Simple NN for classification using MATLAB”.
16. Lidiya Georgieva, Tatyana Dimitrova, Nicola Angelov,” RGB and HSV colour models in colour

identification of digital traumasImages”, International Conference on Computer Systems and Technologies - CompSysTech’ 2005.