International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 4 Issue 5, May 2017

ISSN: 2394-4404

Extraction Technology For Images Text Line And Keyword From Document

Tejaswini Vinod Ramekar

Dept. of CSE, Prof Rammeghe Institute of Technology & Research
College of Engg Amravati, India

Abstract: ***The extraction of text in an image is a classical problem in the computer vision. Extraction involves detection, localization, tracking, extraction, enhancement and recognition of the text from the given image. However variation of text due to difference in size, style, orientation, alignment, low image contrast and complex background make the problem of automatic text extraction extremely challenging. Text extraction requires binarization which leads to loss of significant information contained in gray scale images. The images may contain noise and have complex structure which makes the extraction more difficult. This paper proposes an algorithm which is insensitive to noise, skew and text orientation. It is free from artifacts that are usually introduced by thresholding using morphological operators. Examples are presented to illustrate the performance of proposed method. The text extraction system has been attempted over a corpus of three kinds of images and promising precision has been obtained.***

Keywords: ***Mathematical Morphology, Morphological Operators, Edge Detection, Localization, Connected Component***

I. INTRODUCTION

Text extraction from images and video sequences finds many useful applications in document processing [1],detection of vehicle license plate, analysis of technical papers with tables, maps, charts, and electric circuits [2],identification of parts in industrial automation [3], and content- based image/video retrieval from image/video databases [4], [5]. Educational and training video and TV programs such as news contain mixed text-picture-graphics regions. Region classification is helpful in object-based compression, manipulation and accessibility. Also, text regions may carry useful information about the visual content.

However due to the variety of fonts, sizes, styles, orientations, alignment effects of uncontrolled illuminations, reflections, shadows, the distortion due to perspective projection as well as the complexity of image background, automatic localizing and extracting text is a challenging problem.

Characters in a text are of different shapes and structures. Text extraction may employ binarization [7], [9]–[11] or directly process the original image [8], [12], [13]. In [5], a survey of existing techniques for page layout an analysis is presented. Mathematical morphology is a topological and geometrical based approach for image analysis. It provides powerful tools for extracting geometrical structures and representing shapes in many applications. Morphological feature extraction techniques have been efficiently applied to character recognition and document analysis, especially if dedicated hardware is used. In this paper, we propose an algorithm for text extraction based on morphological operations. The paper is organized as follows. In Section II, the proposed morphological text extraction technique is described. Examples and comparison with existing text extraction algorithms are presented in Section III. Conclusion is given in Section IV.

II. METHODOLOGY

Method has considered the fact that edges are reliable features of text regardless of color or intensity, layout, orientation etc. The edge detection operation is performed using the basic operators of mathematical morphology.

Using the edges the algorithm has tried to find out text candidate connected components. These components have been labeled to identify different components of the image. Once the components have been identified, the variance is found for each connected component considering the gray levels of those components. Then the text is extracted by selecting those connected components whose variance is less than some threshold value. The complete process of text extraction is given in the form of flow chart in Figure 1.

A. EDGE EXTRACTION

For the given input image an efficient morphological edge detection scheme is applied to find the edges of the image.

STEP 1: Apply non-linear filter to the given input image to remove noise.

- ✓ In this step apply open operation to the input image.
- ✓ In this step apply close operation to the input image.
- ✓ Now find the average of the above two steps, and the resultant image is a blurred image.

Algorithm: Non-linear_Filter (y)
Input: y (original image)
Output: ybl (blurred image){
% Apply open filter on image y
$y \circ B = \delta B(\varepsilon B(y))$
% δB = dilation with B on y
% εB = erosion with B on y
% Apply close filter on image y
$y \bullet B = \varepsilon B(\delta B(y))$
% Average of the two filtered images is the blurred image
ybl = ((y ○ B) + (y ● B))/2 ;
}

STEP 2: The blurred image obtained from the filtered image is taken as the input and we find the morphological gradient of this image.

Algorithm: Morphological gradient (ybl)
Input: ybl (blurred image)
Output: es (gradient image) {
es = δB(ybl) - εB(ybl) ;
% B: 8 connected structuring element
% δB = dilation with B on ybl
% εB = erosion with B on ybl
}

STEP 3: Threshold is used to obtain binary image from the grayscale gradient image.

The threshold value gamma is obtained using the algorithm given below.

Algorithm: Thresholding(es)
Input: es (gradient image)
Output: e (threshold image) {
g1 = [-1 0 1]
g2 = transpose(g1);
$y=\frac{-\Sigma(es.s)}{\Sigma s}$
s = max(|g1**es| , |g2**es|)
% •denote pixel wise multiplication
% ** denotes two dimensional convolution
}

B. TEXT CANDIDATE REGION FORMATION

From the threshold image the text candidate regions are obtained as follows. In text candidate region formation close operation is applied to connect all the edges.

Algorithm: Region_Formation(e)
Input: e threshold image
Output: ec text candidate region images {
% dilation(e);
ec1 = δ s(e);
% erosion(ec1);
ec = εs(ec1);
% s is 8 connected structuring elements;
}

C. LABELLING OF TEXT CANDIDATE REGIONS

Apply labelling on the text candidate region as follows

- ✓ To the above obtained text candidate region each candidate is uniquely labelled.
- ✓ Re-labeled the text candidate regions by sequentially assigning unique values to the same component.

Algorithm: Labeling_of_Region(ec)
Input: ec text candidate region image
Output: bw labelled image {
bw= bwlabel(ec);
% calculates all the connected components with nsequential label;
}

D. ELIMINATION OF NON TEXT REGION

From the labelled image which contains text and non text regions eliminate the non text regions using variance operation as follows

Algorithm: Region_Elimination(bw)
Input: bw labelled image
Output: ext image containing text {
var= $e^x = \frac{1}{N}\sum_{i=0}^{n}(x_i - x)2$
% Finds variance on the original Gray scale image on the labelled components
% Select the regions whose variance values are less than threshold.
% Threshold is calculated by taking average of all gray level values
ext = union(var(images) < threshold));
}

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. EXPERIMENTATION SETUP

This algorithm has been tested over a corpus of 60 text images of three type including caption text, scene text and document images in which text has different font size, color,

orientation, alignments. These images are analyzed to demonstrate the performance of the proposed algorithm. Performance is verified with the oriented text in horizontal & vertical direction with different languages (English & Hindi). Various metrics have been evaluated from the tested results.

B. PERFORMANCE ANALYSIS

Metrics used to evaluate the performance of the system are Precision, Recall and F-Score. Precision and Recall rates have been computed based on the number of Correctly Detected

Characters (CDC) in an image, in order to evaluate the efficiency and robustness of the algorithm. The metrics are as follows:

Definition 1: False Positives (FP) / False alarms are those regions in the image which are actually not characters of a text, but have been detected by the algorithm as text.

Definition 2: False Negatives (FN)/ Misses are those regions in the image which are actually text characters, but have not been detected by the algorithm.

Definition 3: Precision rate (P) is defined as the ratio of correctly detected characters to the sum of correctly detected characters plus false positives as represented in equation below.

$$P=\frac{CDC\,(True\ Positive)}{CDC+FP}*100\%$$

Definition 4: Recall rate (R) is defined as the ratio of the correctly detected characters to sum of correctly detected characters plus false negatives as represented in equation below.

Definition 5: F-score is the harmonic mean of recall and precision rate as represents in equation below.

C. RESULTS AND DISCUSSION

The algorithms is tested on the oriented text in Horizontal and vertical direction. The output image of purposed algorithms in fig 2,3 and 4 Consist of detected text for caption text document image and scene text respectively .

This images extract from text using the OCR system to recognized the contained information. The results obtained on varied set of images are compared with respect to precision and recall rates. Promising results have been obtained on a number of images in which almost all text lines can be retrieved from the graphics and figure regions. The images have been tested using various threshold techniques.

Figure 2: Text extraction for caption text images

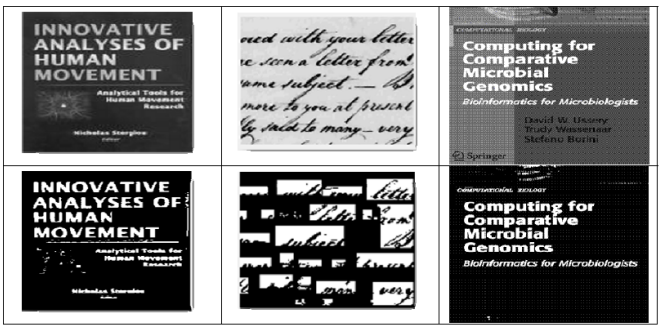

Figure 3: Text extraction for Document images

Figure 4: Text extraction for Scene images

D. COMPARISON WITH OTHER TEXT EXTRACTION TECHNIQUES

To give an average estimate of the performance of the text extraction the results have been compared against two existing algorithms [14] and [15]. Both the methods have used the aspect ratio to identify the text and non text regions within an image. The first method has used the complicated procedure of finding inner, outer and inner-outer corners.

The second procedure has identified edge at different orientation i.e. 0, 45, 90, and 135 degrees and grouping these strokes at different heights, text is extracted. This increases the complexity of algorithm to identify the edges at different orientations. The new Connected Component Variance (CCV) approach has solved the above problems.

In order to group the isolated characters to a meaning full word, a dilation operation with varying structuring element is used. The algorithm identifies number of components and calculates the variance of each component if the variance of each component is high then it is believed to a kind of symbol rather than a text. This algorithm is in sensitive to skew and text orientation, the output of

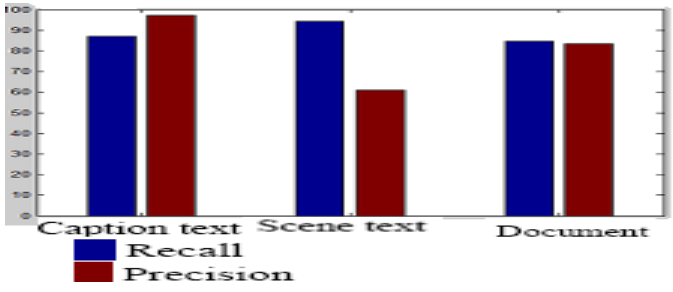

Figure 5: Precision vs Recall rate graph

Figure 6: Text extraction using TMO and TAG algorithms

The text extraction algorithm is fed to an OCR system to recognize the contained information. The main objective of the text extraction algorithm is to reduce the number of false text candidate that may be fed to the OCR. Further investigation on the threshold value to select the correct text candidate is being performed. Different images with different lighting and contrast is used for text extraction. These images are tested using the two algorithms Threshold with Average Gray values (TAG) and the other Threshold Using Morphological operators (TMO). To evaluate the performance of TAG and TMO 25 text images with different font size, perspective, alignments any number of characters in a text string under different lighting conditions is considered. It has been observed that in many case the images gives good text extraction when TMO is used, when the image has uniform background. If various foregrounds and backgrounds are presented in the image, it has been observed that TMO is in sensitive to noise and introduce minimum noise on the removal of non text information. The TAG algorithm introduces noises while removing non text characters but has an advantage of extracting those text characters whose gray levels are close to the back ground shown in the Figure 6.

IV. CONCLUSION

This paper proposes a new text extraction algorithm from a text/graphics mixed document images. This algorithm is in sensitive to skew and text orientation the output of the text extraction algorithm is fed to an OCR system to recognize the information. The results obtained on varied set of images are compared with respect to precision and recall rates. Promising results have been obtained on a number of images in which almost all text lines can be retrieved from the graphics and figure regions. The images have been tested using various threshold techniques. It has been observed that the threshold depends on a various parameters like the illumination condition, reflections and the scan point spread function. This approach has used morphological clearing of images which would help to reduce the number of false positive obtained. This cleaning of the image could result in a higher precision rate.

Further investigation on the threshold value to select the correct text candidate is being performed. The images have been tested using various threshold techniques. It has been found that TAG gives efficient extraction comparative to TMO when the images are taken in poor illumination. However the TAG gives noisy distorted binary images comparative to TMO. This approach has used morphological clearing of images which would help to reduce the number of false positive obtained. In corporately the OCR algorithm with proposed morphological text extraction method yields a useful system for text analysis in images.

REFERENCES

[1] K. Jain and B. Yu, "Document representation and its application to page decomposition," IEEE Trans. Pattern Anal. Machine Intell., vol. 20, pp. 294–308, Mar. 1998.
[2] K. Jain, and B. Yu, "Automatic Text Location in Images and Video Frames", Pattern Recognition, 31 (12) (1998) 2055-2076.
[3] K. Jain, and Y. Zhong, "Page Segmentation using Texture Analysis", Pattern Recognition, 29 (5) (1996) 743-770.
[4] Chitrakala Gopalan and D. Manjula, "Contourlet Based Approach for Text Identification and Extraction from Heterogeneous Textual Images", International Journal of Computer Science and Engineering (2008) pp.202-211.
[5] D. Crandall, S. Antani, and R. Kasturi,"Robust Detection of Stylized Text Events in Digital Video", Proceedings of International Conference on Document Analysis and Recognition, 2001, pp. 865-869.
[6] J. Serra, Image Analysis and Mathematical Morphology. New York: Academic, 1982.
[7] K. C. Fan, L. S. Wang, and Y. K. Wang, "Page segmentation and identification for intelligent signal processing," Signal Process., vol. 45, pp.329–346, 1995.
[8] R. Lienhart, "Indexing and retrieval of digital video sequences based on automatic text recognition," in Proc. ACM Int. Conf., Boston, MA, Nov. 1996.
[9] K. Jain and B. Yu, "Document representation and its application to page decomposition," IEEE Trans. Pattern Anal. Machine Intell., vol. 20, pp. 294–308, Mar. 1998.
[10] J. Ohya, A. Shio, and S. Akamatsu, "Recognizing characters in scene images," IEEE Trans. Pattern Anal. Machine Intell., vol. 16, pp. 215–220, Feb. 1994.
[11] H. M. Suen and J. F.Wang, "Text string extraction from images of colorpriented documents," Proc. Inst. Elect. Eng. Vis., Image, Signal Process., vol. 143, no. 4, pp. 210–216, 1996.
[12] L. Wang and T. Pavlidis, "Direct gray-scale extraction of features for character recognition," IEEE Trans. Pattern Anal. Machine Intell., vol. 15, pp. 1053–1067, Oct.1993.
[13] Y. Zhong, K. Karu, and A. K. Jain, "Locating text in complex color images," Pattern Recognit., vol. 28, no. 10, pp. 1523–1535, 1995.
[14] Jagath Samarabandu, Member, IEEE, and Xiaoqing Liu 2007 "An Edge-based Text Region Extraction Algorithm for Indoor Mobile Robot Navigation", International Journal of Signal Processing 3(4)2007.
[15] Yassin M. Y. Hasan and Lina J. Karam 2000 "Morphological Text Extraction from Images"