IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 9, September 2017

Extraction Techniques Of Image, Text-Line and Keywords from Document Image

Tejaswini V. Ramekar[1], Prof. Dr. A. S. Alvi[2]

M.E., Computer Science & Engg., Prof. Ram Meghe Institute Of Technology & Research, Amravati, India[1]

Professor, Computer Science & Engg., Prof. Ram Meghe Institute Of Technology & Research, Amravati, India[2]

Abstract: In computer vision the extraction of text in an image is a classical problem . Extraction process includes detection, localization, tracking, extraction, enhancement and recognition of the text from the given image. The problem of automatic text extraction extremely challenging because of variation of text due to difference in size, style, orientation, alignment, low image contrast and complex background make. Text extraction needs binarization which leads to loss of important information contained in gray scale images. Extraction becomes more difficult because the images may contain noise and have complex structure. This paper involves an algorithm which is insensitive to noise, skew and text orientation. It does not contain artifacts that are generally introduced by thresholding using morphological operators. There are many examples that presents the performance of proposed method.

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

I. INTRODUCTION

Text extraction from images and video sequences are many useful applications in document processing, detection of vehicle license plate, maps, charts, and electric circuits, verification of parts in industrial automation , analysis of technical papers with tables and content- based image/video retrieval from image/video databases. Educational and training video and TV programs such as news contain mixed text-picture-graphics regions. In object-based compression, manipulation and accessibility, region classification is helpful. However due to the variety of fonts, sizes, reflections, shadows, styles, orientations, , the distortion due to perspective projection as well as the complexity of image background, alignment effects, automatic localizing and extracting text is a challenging problem.

Characters in a text are of various shapes and structures. Text extraction may imply binarization or directly process the original image. In a survey it is analysed that, for image analysis mathematical morphology is a topological and geometrical based approach. It gives powerful tools for extracting geometrical structures and representing shapes in many applications. Especially if dedicated hardware is used, morphological feature extraction techniques have been efficiently applied to character recognition and document analysis. This paper focuses on 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 associated fact that edges are reliable features of text regardless of color, orientation, intensity, layout, etc. By using the basic operators of mathematical morphology. the edge detection operation is performed. The algorithm has tried to find out text candidate connected components by using the edges. To identify different components of the image, these components have been labelled. Once the components have been labelled, the variety is found for each connected component considering the gray levels of those components. Then the text is extracted by choosing those connected components whose variance is less than some threshold value

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.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 9, September 2017

Algorithm: Non-linear_Filter (y)

Input: y (original image)
Output: ybl (blurred image){
% Apply open filter on image y B(y))
% δB = dilation with B on y
% εB = erosion with B on y
% Apply close filter on image y
y ● B = ε B (δB(y))
% Average of the two filtered images is the blurred image
ybl = ((y ○ B) + (y ● B))/2 ;
}
Step 2: Input is taken as the blurred image obtained from the filtered image 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=
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
- ✓ From the above obtained text candidate region each candidate is uniquely labelled.
- ✓ Re-labelled the text candidate regions by serially assigning unique values to the same component.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 9, September 2017

Algorithm: Labelling of Region (ec)
Input: ec text candidate region image
Output: bw labelled image
{
bw = bw label(ec);
% calculates all the connected components with n sequential label; }

D. Elimination of non text region

From the labelled image which contains text and non text regions drop out the non text regions using variance operation as follows -
Algorithm: Region_ Elimination(bw)
Input: bw labelled image
Output: ext image containing text
{
var
% 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 invole caption text, scene text and document images in which text has different font size, color, orientation, alignments. These images are analyzed to display the performance of the proposed algorithm. Performance is identified with the oriented text in horizontal & vertical direction with different languages (English & Hindi). Various metrics have been calculated from the tested results.

B. Performance analysis

Metrics used to calculate the performance of the system are Precision, Recall and F-Score. Precision and Recall rates have been evaluated based on the number of Correctly Detected Characters (CDC) in an image, in order to evaluate the effectiveness and robustness of the algorithm. The metrics are as follows:
Definition 1: False Negatives (FN)/ Misses are those regions in the image which are actually text characters, but have not been detected by the algorithm.
Definition 2: False Positives (FP) / False alarms are those regions in the image which are actually not characters of a text, but have been identified by the algorithm as text.
Definition 3: Precision rate (P) is defined as the ratio of correctly identified characters to the sum of correctly identified characters plus false positives as represented in equation below.
P=*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 proposed algorithms in fig 2, 3 and 4 involves detected text for caption text document image and scene text respectively.
This images extract from text using the OCR system to recall the involved information. The results obtained on various set of images are compared with precision and recall rates. Promising results have been obtained on a number of images in which almost all text lines can be regain from the graphics and figure regions.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 9, September 2017

Fig.1.Text extraction for caption text images

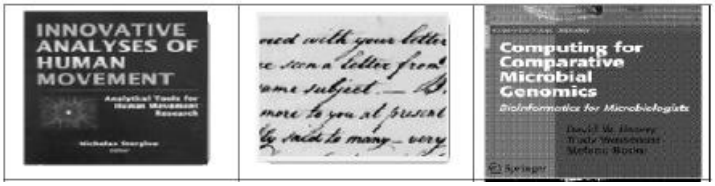

Fig.2 Text extraction for Document images

Fig.3 Text extraction for Scene images

D. Comparison with other text extraction techniques

To give an average prediction of the performance of the text extraction the results have been compared against two existing algorithms. Both the methods have used the concept of ratio to verify the text and non text regions within an image. The first method has used the complex procedure of finding inner, outer and inner-outer corners. The second procedure has verified edge at different orientation i.e. 0, 45, 90, and 135 degrees and grouping these all at different heights, text is extracted. This raises the complexity of algorithm to verify the edges at different orientations. The new Connected Component Variance (CCV) concept has solved the above problems. A expansion operation with varying structuring element is used, when we have to group the isolated characters to a meaning full word. 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 is

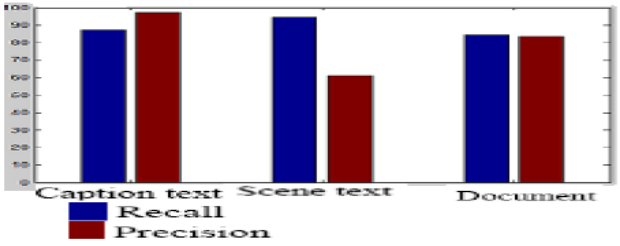

Fig.4 Precision vs Recall rate graph

Fig.5 Text extraction using TMO and TAG algorithms

The text extraction algorithm is associated with an OCR system to recognize the existing information. The main moto of the text extraction algorithm is to decrease the number of false text candidate that may be fed to the OCR. Different images with different lighting and contrast is used for text extraction. Further investigation on the threshold value to

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 9, September 2017

select the correct text candidate is being performed. These images are tested using the two algorithms Threshold with Average Gray values (TAG) and the other Threshold Using Morphological operators (TMO). Any number of characters in a text string under different lighting conditions is considered, to evaluate the performance of TAG and TMO 25 text images with different font size, perspective, alignments. When TMO is used, it has been observed that in many case the images gives good text extraction. , It has been observed that TMO is in sensitive to noise and introduce minimum noise on the removal of non text information, if various foregrounds and backgrounds are presented in the image.

IV. CONCLUSION

This paper focuses on a new text extraction algorithm from a text/graphics mixed document images. This algorithm is in sensitive to skew and text orientation. To recognize the information, the output of the text extraction algorithm is fed to an OCR system. The results obtained on various set of images which are compared with respect to precision and recall rates. If all text lines can be retrieved from the graphics and figure regions, promising results have been obtained on a number of images. The images have been tested using various threshold techniques. It has been observed that the threshold depends on a different parameters like the illumination condition, the scan point spread function and reflections. This approach has used morphological clearing of images which would help to decrease the number of false positive obtained. This cleaning of the image could result in a higher exact rate. Further investigation on the threshold value to select the correct text candidate is being performed. Using various threshold techniques, the images have been tested. When the images are taken in poor illumination, it has been found that TAG gives effective extraction comparative to TMO. The TAG gives noisy misshape binary images comparative to TMO. To reduce the number of false positive rate obtained, this approach has used morphological clearing of images which would help. The OCR algorithm with proposed morphological text extraction method yields a useful system for text analysis in images.

REFERENCES

[1] K. Jain, and Y. Zhong, “Page Segmentation using Texture Analysis”.
[2] 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.
[3] K. Jain, and B. Yu, “Automatic Text Location in Images and Video Frames”.
[4] D. Crandall, S. Antani, and R. Kasturi,“Robust Detection of Stylized Text Events in Digital Video”.
[5] 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.
[6] 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.
[7] Chitrakala Gopalan and D. Manjula, “Contourlet Based Approach for Text Identification and Extraction from Heterogeneous Textual Images”.
[8] K. Jain and B. Yu, “Document representation and its application to page decomposition,” Mar. 1998.
[9] J. Serra, Image Analysis and Mathematical Morphology. New York: Academic, 1982.
[10] R. Lienhart, “Indexing and retrieval of digital video sequences based on automatic text recognition Nov. 1996