A Hybrid Model to Automatically Extract Text

H.Thaheera[1] A. Abdul Samathum[2]

[1]Professor [2]M.Phil

Abstract— Text mining also referred to as text data mining, which is equivalent to text analytics, refers to the process of deriving high-quality information from text and such information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. This Text mining usually involves the process of structuring the input text like parsing, along with the addition of some derived language based features and the removal of others thereby deriving patterns within the structured data, and finally evaluation and interpretation of the output. In this paper a hybrid approach to robustly detect and localize texts in natural scene images is proposed. The variations of text like font, size and line orientation. A text region detector is designed to estimate the text existing confidence and scale information in image pyramid, which helps by segmenting the text components using local binarization. To efficiently filter out the non-text components, a conditional random field (CRF) model considering unary component properties and binary contextual component relationships with supervised parameter learning is proposed. Finally the extracted text components are grouped into text lines/words with a learning-based energy minimization tool and stored in an external file.

Key words: Text Mining, Text Extraction from Natural Scene Images, text detection, text localization - CRF

I. Introduction

A variety of approaches to text information extraction (TIE) from images have been proposed for specific applications. Some applications includeimage page segmentation, address block location, license plate location, and content-based image indexing. In spite of such extensive studies, it is still not easy to design a general-purposetext extraction system. This is because there are so many possible sources of variation when extracting text from a shaded or textured background, from low-contrast or complex images, or from images having variations in font size, style, color, orientation, and alignment. These variations make the problem of automatic TIE extremely difficult.

Content-based image indexing refers to the process of attaching labels to images based on their content. Image content can be divided into two main categories: *perceptual content* and *semantic content*. Perceptual content includes attributes such as color, intensity, shape, texture, and their temporal changes, whereas semantic content means objects, events, and their relations. A number of studies on the use of relatively low-level perceptual content for image and video indexing have already been reported. Studies on semantic image content in the form of text, face, vehicle, and human action have also attracted some recent interest. Among them, text within an image is of particular interest as (1) it is very useful for describing the contents of an image; (2) it can be easily extracted compared to other semantic contents, and (3) it enables various applications like keyword based image search, video logging, and text based image tagging. High quality in text mining refers to some combination of relevance, novelty, and quality interest. Typical text mining tasks include text categorization, clustering of text, words, concepts or entity extraction, production of taxonomies which are granular in nature, some sentiment analysis, document summarization, and of course entity relation modeling. To refer to in detail text analysis involves some information retrieval, lexical analysis in order to study word frequency distributions. This in turn is used to find patterns, tagging, annotations, information extraction, data mining techniques including link and association analysis, natural visualization, and predictive analytics. The goal however isto turn the text into meaningful data for analysis, via an application of natural language processing (NLP) and analytical methods. Text detection and localization in natural scene images is important for content-based image analysis. The afore mentioned problem is a challenging one due to the complex background and the non-uniform illumination.

Text extraction from images may exhibit many variations with respect to the following properties:

A. Dimensions:

Size is a major factor. Although the text size can vary a lot, assumptions can be made depending on the application domain. Alignment is the second major factor. The characters in the caption text appear in clusters and usually lie horizontally. But sometimes they may also appear as non-planar texts as a result of special effects. This will however not apply to scene text, which can have various perspective distortions, because scene text can be aligned in any direction and can have geometric distortions as well. The distance between characters in a text line also have an impact in text information extraction.

B. Foreground and Background Color:

The characters in a single text line always tend to have the same or similar colors. This is the reason that makes it possible to use a connected component-based approach for text detection. Images of color documents can contain text strings with more than two colors (polychrome) for effective visualization or different colors within one word.

C. Outline Edge Detection:

Most image caption and scene text are designed to be easily read and have strong coloured edges. This automatically results in strong edges at the boundaries of the text and also at the background.

D. Format:

Many digital images are stored and then transferred. This is done by processing in a compressed format. Thus, a faster text extraction system can be achieved if one can extract text without decompression.

The above methods result in noise, distortion and other vagaries which result in wrong or inconsistent detection of images. Besides segmentation and size which is known as the dimension problem also occurs, unlike as in region based methods, the speed is relatively slow. The performance is sensitive to text alignment orientation. The

CC-based methods cannot segment text components accurately without prior knowledge of text position and scale. So designing fast and reliable connected component analyzer is difficult since there are many non-text components. These are easily confused with texts when analyzed individually.

II. PROPOSED MODEL

The proposed uses a hybrid system instead of any single approach and uses both CC and Region Based approaches taking each method advantages. First it designs a text region detector. This detects text position (location) and scale (size). Next a local binarization algorithm is used. To segment candidate text fields.

A hybrid approach to robustly detect texts in natural scene images is proposed by taking advantages of both region-based and Connected Component based methods. Since local region detection can robustly detect scene texts even in noisy images, we design a text region detector to estimate the probabilities of text position and scale, which help segment candidate text components with an efficient local binarization algorithm.

To combine unary component properties and binary contextual component relationships, a conditional random field (CRF) model with necessary supervised parameter learning is used. Finally, text components are grouped into text lines/words robustly with an energy minimization method. Experimental results on English and multilingual image datasets show that the proposed approach yields higher precision and recall performance compared with state-of-the-art methods.

III. ALGORITHMS AND TECHNIQUES USED LOCAL BINARIZATION ALGORITHM

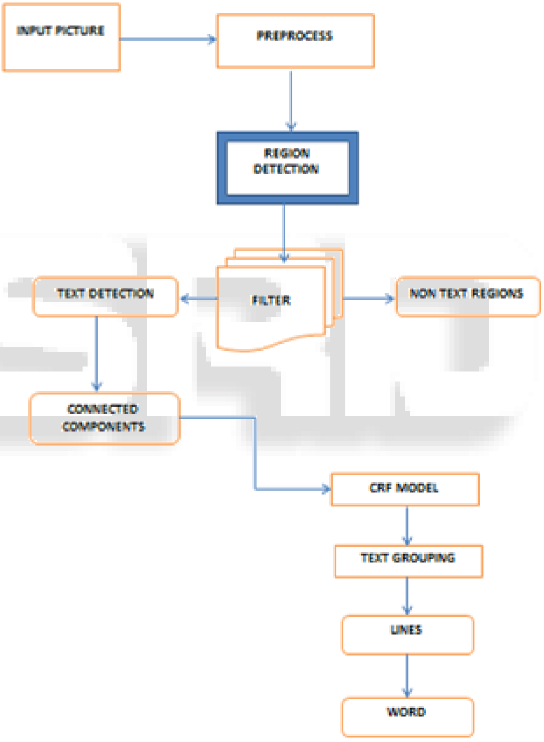
Binarization (thresholding) converts the grayscale document image to binary , by changing the foreground pixels (text characters) to black and background pixels to white

A. *Algorithm*

(1) Load the Image into a Matrix of pixels
(2) Read the value of each pixel
(3) For each pixel
(4) Get the radius of the window to be centered on that pixel
(5) Get the intensity mean
(6) Get the Standard Deviation
(7) Now check for the gray level.
(8) If 0 or 255 check for gray
(9) If 1 otherwise
(10) End if
(11) Store the gradient converted pixel into the temporary matrix
(12) Invert the results and check the values
(13) Display the temporary pixels matrix into a picture.
(14) Now we get the binarized results

IV. MINIMUM SPANNING TREE ALGORITHM

(1) Calculate the Distance metric for each point
(2) This is taken from the connected cluster component.
(3) Extract the features
(4) This is done by calculating the weights
(5) The weights are stored into vectors.
(6) The vectors are classified into mean square error criteria
(7) This gives the connected components.
(8) Given a connected graph G=(V,E) and a weight d:E->R+, find a minimum spanning tree T.
(9) Set i=1 and let E0={}
(10) Select an edge ei of minimum value not in Ei-1 such that Ti=<Ei-1 cup {ei} >is acyclic and define Ei=Ei-1 cup {ei}. If no such edge exists, Let T=<Ei>and stop.
(11) Replace i by i+1. Return to Step 2.
(12) The time required by Kruskals algorithm is O(|E|log|V|).



V. ARCHITECTURE

A. *Segmentation - Preprocessing*

The result of image segmentation is a set of segments that collectively cover the whole image, or a set of edge extracted contours from the image. Each of the pixels in a region are similar with respect to some characteristic property like color, intensity, or texture. The adjacent regions are significantly different with respect to the above identified characteristics.

B. *Identifying Text Regions – Text Detection*

The original color image is converted into a graylevel image, on which the image pyramid with scale is built up with nearest interpolation to capture text information of different scales. Based on our previous work, a text region detector is designed by integrating a widely used feature descriptor: histogram of oriented gradients (HOG) and a boosted cascade classifier: WaldBoost. Note that the aim of text region detector is not to find accurate text positions but to estimate probabilities of the text position and scale information.

VI. Binarization

A. Adaptive Binarization

A novel adaptive binarization algorithm using ternary entropy-based approach is used. Given an input image, the contrast of intensity is first estimated by a grayscale morphological closing operator. A near double threshold is generated by the entropy-based method to classify pixels into text, near-text, and non-text regions. The pixels in the second region are relabeled by the local mean and the standard deviation values. The proposed method classifies noise into categories which are processed by binary operators namely shrink and swell filtersand finally filter using the graph searching strategy.

VII. Connected Component Classifier

The points at which image brightness changes sharply are normally organized into a set of curved line segments which are termed as the edges of the object. The problem of finding discontinuities in signals is known as step detection. The problem of finding signal discontinuities over time is known as the change detection. Edge detection tool in image processing uses the connected component classifier to identify the text areas and feature detection and finally in feature extraction.

A. CRF

Detection of skin color in color images is a very popular and useful technique for face detection. Many techniques have reported locating skin color regions in the input image. Normally any input color image is typically in the RGB format. The color components in the color spaceare in the HSV or YIQ formats. That is because RGB components are subject to the lighting conditions, thus the image or object detection may fail if the lighting condition changes. Of the many color spaces, this paper usesYCbCr components since it takes lesser computation time. In the YCbCr color space, the luminance information is contained in the Y component; and, the chrominance information is in Cb and Cr. This in turn makes the luminance information to be easily de-embedded. The RGB components were converted to the YcbCrformat.

B. Detection

In the skin color detection process, each pixel was classified as skin or non-skin based on its color components. The detection window for skin color was determined based on the mean and The color segmentation has been applied to a training image and its result is shown in the below figure. Some non-skin objects are inevitably observed in the result as their colors fall into the skin color space.

VIII. Conclusion

Thus the hybridapproach to localize scene texts by integrating region information into a robust CC-based method. The binary contextual component relationships which along with the unary component propertiesare both integrated in the hybrid CRF model, whose parameters are jointly optimized by supervised learning. Our experimental results demonstrated that the proposed method is effective in unconstrained scene text localization in several respects: 1) region-based information is very helpful for text component segmentation and analysis; 2) incorporating contextual component relationships, the CRF model differentiates text components from non-text components better than local classifiers; 3) joint optimization of base classifier parameters with CRF is beneficial; and 4) learning-based energy minimization method can group text components into text lines (words) robustly.

In future **t**he proposed model can be extended to detect hard-to-segment texts by taking into account color information. The speed of the proposed approach (about 1.2 s per 640 480 image) can be accelerated further. In addition, text recognition may be integrated with text localization to complete the need of text information extraction. Also the text recognition can also help reduce false positives of text extraction.

References

[1] C. M. Bishop, Neural Networks for Pattern Recognition. New York: Oxford University Press, 1995.

[2] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 11, pp. 1222–1239, 2001.

[3] D. T. Chen, J.-M. Odobez, and H. Bourlard, “Text detection and recognition in images and video frames,” Pattern Recogn., vol. 37, no. 5, pp.595–608, 2004.

[4] X. L. Chen, J. Yang, J. Zhang, and A.Waibel, “Automatic detection and recognition of signs from natural scenes,” IEEE Trans. Image Process., vol. 13, no. 1, pp. 87–99, Jan. 2004.

[5] X. R. Chen and A. L. Yuille, “Detecting and reading text in natural scenes,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR04), Washington, DC, 2004, pp. 366–373.

[6] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR05), San Diego, CA, 2005, pp. 886–893.

[7] S. L. Feng, R. Manmatha, and A. McCallum, “Exploring the use of conditional random field models and HMMs for historical handwritten document recognition,” in Proc. 2nd Int. Conf. Document Image Analysis for Libraries (DIAL06), Washington, DC, 2006, pp. 30–37.

[8] J. Gllavata, R. Ewerth, and B. Freisleben, “Text detection in images based on unsupervised classification of high-frequency wavelet coefficients,” in Proc. 17th Int. Conf. Pattern Recognition (ICPR04), Cambridge, U.K., 2004, pp. 425–428.

[9] J. M. Hammersley and P. Clifford, Markov Field on Finite Graphs and Lattices, 1971, unpublished.

[10] X.-B. Jin, C.-L. Liu, and X. Hou, “Regularized margin-based conditional log-likelihood loss for

prototype learning," Pattern Recogn., vol. 43, no. 7, pp. 2428–2438, 2010.

[11] Y. Jin and S. Geman, "Context and hierarchy in a probabilistic image model," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR06), New York, NY, 2006, pp. 2145–2152.