ISSN:0975-9646

Segmentation of Text From Image Document

Ankush Gautam
Department of Information Technology
Graphic Era University, Dehradun

Abstract- Segmentation of text from image documents has many important applications such as document retrieving, object identification, detection of vehicle license plate, etc. It is very popular research field in recent years. In this paper, we employ Symlet wavelet and 2-mean classification for segmentation of text from image document. We have used morphology operation like as dilation and erosion in post processing. Proposed method for text segmentation from image document has been implemented in MATLAB.

Keywords- Segmentation, Extraction Symlet wavelet, Erosion, Dilation etc.

I. INTRODUCTION

Image segmentation is a task of fundamental importance in digital image analysis. It is the process of partitioning a digital image into multiple segments.

Documents in which text is embedded in picture are increasingly common today for example, in magazines, advertisements and web pages. Robust detection of text from these documents receives a growing attention owing to their number of applications content based indexing, Text searches in images and Archiving documents.

In the face of the very important mass of information exchanged between different organizations, the need for systems allowing the recognition, the indexation, the information retrieval and the automatic classification of complex multi-lingual and multi-script document images has grown continuously. However, most works of backward-conversion of printed document images are limited to textual block recognition without handling complex documents such as letters of information, forms, all types of application sheet, etc. In practice, these documents can be noised, skewed, deformed, multi-lingual, multi-script with irregular textures and may contain several heterogeneous blocks such as annotations, machine print and/or handwritten script, graphics, pictures, logos, photographs, tabular structures. This situation makes it difficult to analyze and recognize document images.

II. BACKGROUND AND LITERATURE SURVEY

There are various approaches for text and picture segmentation in image document namely region based approach and texture based approach.

In the region based approach, we consider each pixel in the image and assign it to a particular region or object. This approach is basically divided into two subcategories: edge based [2, 3, 12] and connected component based.

The text-regions in a document image can be detected either by region based and texture-based methods. They are relatively independent of changes in text size and orientation, but having difficulties with complex images with non-uniform backgrounds, for example, if a text string touches a graphical object in the original image, they may form one connected component in the resultant binary image.

Basically the idea behind the edge-based algorithms is that the edges of text symbols are typically stronger than those of noise, textured-background and other graphical items [1, 3, and 4]. In these top-down techniques, a binary edge image is first generated using an edge detector, and then adjacent edges are connected by applying morphological operations or other algorithms such as run-length smoothing [3]. Connected components of the resultant image are the candidate text areas, as each one represents either several merged lines or a graphical item. Then, each component is decomposed into smaller regions by analyzing its vertical and horizontal projection profiles, and finally each of the small regions satisfying certain heuristic constraints is labeled as text. Edge-based methods are fast and can detect text in complex backgrounds but are restrictive to detect only horizontally or vertically aligned text strings.

In texture-based approach the input image is usually considered as a composite of text and non-text or text, picture and background texture classes. Many segmentation algorithms employ a classification window of a certain size in the hope that all or majority of pixels in the window belong to the same class [5]. Thereafter, a classification algorithm can be used to label each window in the feature space. For example, in [6] the number of classes is two, and a 2-means classification is used to classify each block of the image as text or non-text according to its local energy in the wavelet transform domain. By using a 3-means clustering in each image pixel is labeled as text, picture or background according to a 9-D feature vector based on Gaussian filtering. A large number of statistical and geometrical features have been proposed for texture segmentation. The wavelet transform has become a very effective tool in texture segmentation and classification due to its multi-resolution properties therefore wavelet based features are matter of interest. It provides a powerful transform domain for modeling images that are well characterized by their edges.

The literature on text segmentation is broad in scope but there appears to be very little literature on using machine learning techniques on this subject. Text segmentation algorithm should have adaptation and learning capability, but a learner usually needs much time and training data to achieve satisfactory results, which restricts its practicality. To overcome these problems, M. M. Haji, S. D. Katebi [7] give a simple procedure for generating training data from manually segmented images, then applying a Naive Bayes Classifier (NBC), which is fast both in training and application phase.

A Study of Image Segmentation and Edge Detection Techniques has been proposed by Punam Thakare. This paper discussed about some image segmentation techniques like edge based, region based and integrated techniques. The results show that the recognition rate depends on the type of the image and their ground Truths [10]

A method of extraction of textual information is proposed by Danial Md Nor, Rosli Omar , M. Zarar M.Jenu. This paper present a solution to the problem of extraction of textual information in presentation scene images. The proposed approach of extraction of textual information is composed of three steps: image segmentation, text localization and extraction, and Optical Character Recognition. The results are very dependent on the quality of OCR images or documents [11]

A methodology for extracting text from images such as document images, scene images etc by Neha Gupta, V .K. Banga. This paper employs discrete wavelet transform (DWT) for extracting text information from complex images. The input image may be a colour image or a grayscale image. If the image is colour image, then preprocessing is required. For extracting text edges, the sobel edge detector is applied on each sub image. The resultant edges

so obtained are used to form an edge map. Morphological operations are applied on the processed edge map and further thresholding is applied to improve the performance [12]

III. Proposed Algorithm For Segmentation Of Text

A. Image Acquisition

The Image Acquisition is the process of collection of images for text and picture segmentation in image document. We have used scanned image for text and picture segmentation in image document.

B. Image preprocessing

The aim of preprocessing is to improve image data so that it removes undesired distortions and/or it enhances image features that are relevant for further processing.

The analysis of a picture using image processing that can identify shades, colors and relationships that cannot be perceived by the human eye. Image processing is used to solve identification problems, such as in forensic medicine or in creating weather maps from satellite pictures. It deals with images in bitmapped graphics format that have been scanned in or captured with digital cameras. The color images are then converted to grey level images by using the following formula

grey (i, j) = 0.59 green(i, j) + 0.30 red(i, j) + 0.11 blue(i, j)

C. Symlet Wavelet

The wavelet transform provides a multi-resolution representation of an image that has become quite popular in recent years owing to their huge number of applications in various fields, such as, for example, telecommunications, geophysics, astrophysics and in computer vision field to enable to detection, analysis and recognition of image features and properties over varying ranges of scale.

Symlet wavelets are a family of wavelets. They are a modified version of Daubechies wavelets with increased symmetry.

Symlet Wavelet[n] is defined for any positive integer n. The scaling function (ϕ) and wavelet function (ψ) have compact support length of $2n$. The scaling function has n vanishing moments.

Symlet wavelet can be used with functions as Discrete Wavelet Transform. The discrete wavelet transform is a mathematical tool for signal analysis and image processing. By the wavelet transform, an image can be decomposed into multiresolution frame in which every portion has distinct frequency and spatial properties

D. Block processing

Images can either be too large to load into memory, or else they can be loaded into memory but then be too large to process therefore block processing is more useful to automatically divide the input image into blocks of the user- specified size, processes each block individually and then reassembles each block results into the output image.

If we want to divide an image into blocks and process each block individually, the function blkproc is used that allow to process distinct blocks.

E. K-means Clustering

The K-mean algorithm is an iterative technique that is used to partition an image into K clusters. The basic algorithm is:

1. Pick K cluster centers, either randomly or based on some heuristic.
2. Assign each pixel in the image to the cluster that minimizes the distance between the pixel and the cluster center
3. Re-compute the cluster centers by averaging all of the pixels in the cluster
4. Repeat steps 2 and 3 until convergence is attained (e.g. no pixels change clusters)

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. We used 2-means classification in our implementation. One group of white pixel and second group of black pixel are used in 2-means classification.

F. Post Processing

Post processing attempts to increase the quality of a mask image. Post processing is performed with the help of Morphology. The morphological operations are dilation and erosion. Dilation adds pixels to the boundaries of objects in an image, while erosion removes pixels on object boundaries. The number of pixels added or removed from the objects in an image depends on the size and shape of the structuring element used to process the image.

IV. Flow Work Diagram

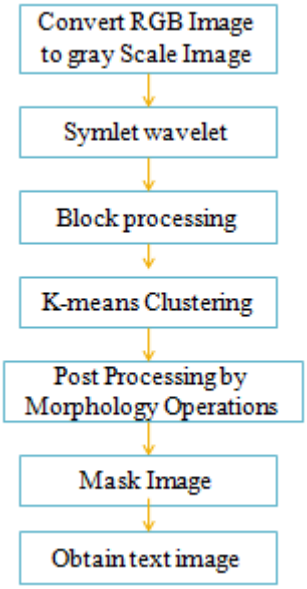

Figure 1: Flow work diagram for text segmentation

V. Experimental Results

We have developed segmentation of text and picture from image document with the help of Symlet wavelet and 2-mean classification using MATLAB R2009a. Here we have considered some images and illustrate the segmentation of text and picture.

Figure 2: Original image

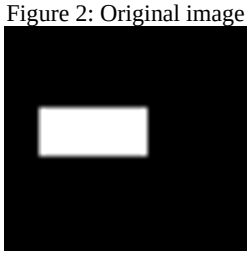

Figure 3: Output mask image of figure 2

Figure 4: Segmented image of figure 2

We have done experiment for another image by our method.

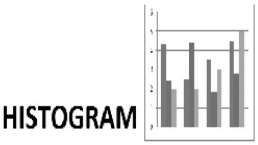

Figure 5: Original image

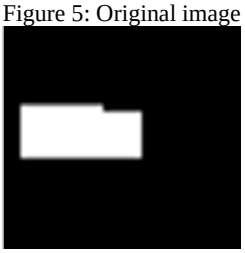

Figure 6: Output Mask image

Figure 7: Segmented image of figure 5

Figure 8: Original image

Figure 9: Output mask image of figure 8

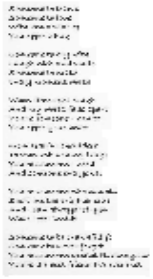

Figure 10: Segmented image of figure 8

CONCLUSION

The development of efficient method for detecting all types of graphics and text in any orientation from real life documents is a challenging process. Our Proposed method for segmentation of text in image document extract text from images efficiently by applying Symlet wavelet.

In the future, research efforts must be devoted to Multi-oriented annotations detection, improved separation of graphic linked to text and segmentation rate will be improved.

REFERENCES

[1.] D. Chen, H. Bourlard and J. Thiran, "Text Identification in Complex Backgrounds Using SVM", Proc. of the International Conf. On Computer Vision and Pattern Recognition, Chen Bourlard, H. Thiran ,pp. 621-626, 8-14 Dec. 2001.

[2.] M. Pietikäinen and O. Okun, "Text Extraction from Grey Scale Page Images by Simple Edge Detectors", Proc. of the 12th Scandinavian Conf. On Image Analysis, Bergen, Norway, pp. 628-635, 11-14 June 2001.

[3.] Jie Xi, Xian-Sheng Hua, Xiang-Rong Chen, et al., "A Video Text Detection and Recognition System", Proc. of ICME 2001, Waseda University, Japan, pp. 1080-1083, August 2001.

[4.] Q. Yuan and C. L. Tan, "Page Segmentation and Text Extraction from Grey-Scale Images in Micro Film Format", SPIE Proc. on Document Recognition and Retrieval, vol. 4307, pp.323-332, 2000.

[5.] H. Choi and R. G. Baraniuk, "Multiscale Image Segmentation Using Wavelet-Domain Hidden Markov Models", IEEE Transactions on Image Processing, vol. 10(9), pp. 1309-1321, Sep. 2001.

[6.] Shulan Deng and Shahram Latifi, "Fast Text Segmentation Using Wavelet for Document Processing", Proc. of the 4th WAC, ISSCI, IFMIP, Maui, Hawaii, USA, pp. 739-744, 11-15 June 2000.

[7.] M. M. Haji, S. D. Katebi, "An Efficient Text Segmentation Technique Based on Naive Bayes Classifier", GVIP Journal, Volume 5, Issue 7, July 2005

[8.] M. M. Haji, S. D. Katebi, "Machine Learning Approaches to Text Segmentation", Scientia Iranica, Vol. 13, No. 4, pp 395-403, October 2006.

[9.] S.Audithan, R.M. Chandrasekaran," Document Text Extraction from Document Images Using Haar Discrete Wavelet Transform" , European Journal of Scientific Research ISSN 1450-216X ,Vol.36, No.4, pp.502-512, June 2009.

[10.] PunamThakare, "A Study of Image Segmentation and Edge Detection Techniques", International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397 Vol. 3 No. 2 Feb 2011

[11.] Danial Md Nor1, Rosli Omar2 , M. Zarar M.Jenu, Jean Marc Ogier, "Image Segmentation and text Extraction: Application to the Extraction of Textual Information in Scene Images", ISASM 2011

[12.] Neha Gupta, V.K. Banga, "Image Segmentation for Text Extraction", ICEECE'2012 April 28-19, 2012

[13.] en.wikipedia.org/wiki/Segmentation_(image_processing)