ISSN: 2321-7782 (Online)
Volume 2, Issue 3, March 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Paper / Case Study
Available online at: www.ijarcsms.com

Extraction of Text from Images of Big Data

Vaishnavi Ganesh[1]
PG student
CSE Dept
G. H. Raisoni College of Engineering
Nagpur – India

Dr. L. G. Malik[2]
H.O.D
CSE Dept
G. H. Raisoni College of Engineering
Nagpur – India

Abstract: ***Collection of data sets very large and complex that becomes difficult to be processed using on-hand database management tools or traditional data processing applications is called Big Data. Text information in images of big data serves as important clues for many image-based applications. However, locating text from a complex background with multiple colors is a difficult task. The proposed framework in this paper consists of two steps:-1.Color based partition method 2.Text line grouping method. Trained classifiers will be used after first step. Canny edge detector is used in first step and text line grouping makes use of Hough transform.***

Keywords: Big Data, Color based partition, Canny edge detector, Hough transform, Text line grouping.

I. INTRODUCTION

Big data can be understood as a collection of data sets that are very large and complex which is difficult to be processed using on-hand database management tools or traditional data processing applications. The data sets that are included here are with sizes beyond the ability of commonly used software tools for capturing, curating, managing, and processing the data within a tolerable elapsed time.

The big data target is important due to constant improvement in traditional DBMS technology and new databases like NoSQL and their ability to handle larger amounts of data. Big data is a huge volume and huge variety of information that require new forms of processing which helps in making decisions and helps in optimization of processes.

Text detection and localization in big data images is necessary for content-based image analysis. This problem is difficult due to the complications in the background, the non-uniform brightness of the image, the varying text font,and their sizes.

As the digital image capturing devices, such as digital cameras, mobile phones are increasing in number text-based image analysis techniques are receiving huge importance in past few years. Out of all the contents in images text information has drawn lot of attention as it is easily understood by humans and computer. It finds wide scope such as numeric detection in the license plate, sign detection alphanumeric detection on streetview images and so on.

The existing methods of text detection and extraction can be roughly categorized into two groups: region-based and connected component (CC)-based. Region-based methods attempt to detect and extract text regions by texture analysis. Usually, a feature vector extracted from each local region is fed into a classifier for estimating the likelihood of text. Then merging of neighbouring text regions takes place to generate text blocks. On the other hand, direct segmentation of candidate text components by edge detection or color clustering takes place in CC-based methods. Then pruning of non-text components with heuristic rules or classifiers takes place.

II. RELATED WORK

Previous work on text detection and extraction can be classified into two categories. The first category is focused on text region initialization and extension by using distinct features of text characters.

For extraction of candidates of text regions, the text binarization method [2] first assigned a bounding box to the boundary of each candidate character in the edge image and then detected text characters based on the boundary model. Text detection using structural features [3] method calculated ridge points in different scales to describe text skeletons at the level of higher resolution and text orientations at the level of low resolution. Text Segmentation using stroke filter method [4] used a stroke filter to extract the stroke-like structures. Text extraction from colored book covers [5] method combined a top–bottom analysis based on color variations in each row and column with a bottom–top analysis based on region growing by color similarity. Morphological text extraction method [6] designed robust morphological processing. Text localization enhancement and binarization method [7] improved Otsu's method for binarization of text regions from background, after which was a set of morphological operations to reduce noise and correct classification errors. For grouping together text characters and removing out false positives, these algorithms employed some conditions involved in character, such as the character should have a minimum size x and a maximum size y, brightness between character strokes and background. But, these algorithms usually fail to remove the background noise resulting from wire mesh,atmospheric distortion, or other background objects. For reducing background noise, the algorithms in the connected component method first do splitting of images to blocks and then merge the blocks verified by the features of text characters. Edge based technique from video frames [8] applied different edge detectors for searching of blocks containing the most apparent edges of text characters. Caption localization method [9] used a fusion strategy which combined color detectors, detectors in texture, contour, and temporal invariance, respectively. Sign detection with conditional random fields [10] method used a group of filters to analyze texture features in each block and joint texture distributions between adjacent blocks by using conditional random field. One drawback here is that they have been partitioning images without any content in it and dividing the image spatially into blocks of equal size before grouping is performed. Noncontent- based image partition will usually break up text characters or text strings into fragments which fail to satisfy the texture constraints. Thus, Laplacian method for text detection [11] performed line-by-line scans in edge images to combine rows and columns with high density of edge pixels into text regions. Adaptive algorithm for text detection [12] performed heuristic grouping and did layout analysis to cluster edges of objects in the images having same color, co-ordinates and size into text regions. However, these algorithms are not comfortable with slanted text lines.

III. PROPOSED PLAN

In this paper, the google API is taken as the source which provides various streetview images which act as the Big Data here. The Google api will be having millions of streetview images and other images which play the part of Big data in this paper. Actually here two main areas come into picture. The Data mining field which acts as the Big data here and the Image processing field that is the text extraction part.

After the images are obtained, filters are applied to the images, which remove noise from the images and give clearer picture, which becomes easy for text extraction. Study of various types of filters and various types of noise has been done in this paper.

After this basically the image processing stage consists of two steps:-

1. Color Based partition method which makes use of Canny edge detection and K-means clustering.
2. Text grouping method which makes use of Hough transform.

And then trained classifiers will be used.

IV. METHODOLOGY USED

A. Big Data Analysis

Here the Google's API (Application Programming Interface) will be included. There are several Google APIs here. The Google Maps API, Google Places API, Google Street view API, Google Earth API and so on.

The images related to this paper will be taken from Google Street view API. There are millions of photos taken in various directions of various streets of various cities, which acts as a Big Data here, that is provided by the Google API. After the registration process in Google Street view API, the user gets the key. By properly specifying the required parameters and optional parameters in the URL, the user can obtain the different street view images, which are formed at different combinations of latitude and longitude values.

The required parameters are size (size specifies the output size of the image in pixels.), loc (location can be either a text string or a lat/lng value), sensor (sensor indicates whether or not the request came from a device using a location sensor (e.g. a GPS) to determine the location sent in this request).

The optional parameters include heading (heading indicates the compass heading of the camera), fov (default is 90) determines the horizontal field of view of the image, pitch (default is 0) specifies the up or down angle of the camera relative to the Street View vehicle, key (optional) identifies the user's application for quota purposes.

Figure1. Streetview Images containing text

Now the various street view images obtained are a part of big data of the streets of any city. From these street view images the text will be highlighted and shown. The text could be from shops or from banners or the street names located from the street view images. That text will be highlighted and shown. This would be helpful for night driving purposes. This could be a real time application of this project.

B. Applying Filters

There are three types of filters:-

i. Averaging filters(low pass filter)

This filfter consists of a 3*3 or 5*5 mask and it is placed on the image starting from top left corner. Then it is moved rightwards. Here average of pixels of the original image matrix is taken and the centre value of original image is changed. This process is repeated for all pixels. The corner pixels of original image remain the same.

ii. Median filter (low pass filter)

The averaging filter removes the noise by blurring it, till it is no longer seen. But it also blurres the edges. Bigger the averaging mask more is the blurring. When the image contains salt and pepper noise and if we use the averaging filter, to remove the same, it will blur the noise, but it would also damage the edges. Hence we need to eliminate the salt and pepper noise, we work with a non linear filter, known as median filter, or it is called a order statistic filter. Because their response is based on the ordering of pixels, contained within the mask.

iii. Adaptive filter (Weiner filter)

It eliminates the low frequency regions while enhancing the high frequency components.

Figure 2a. Image without applying filter

Figure 2b.Image after applying filter

C. Now the image processing consists of the following stages:-

1. Color based Partition Method

In this method, partitioning of the image has been done on the basis of color. That is all pixels having same color or pixels having small variations in their color component will be placed in one cluster.

For this canny edge detector has been used, which detects all the edges of the image. For this first the RGB image consisting of three dimensions is converted to grayscale image which consists of two dimensions. The three dimensions of the RGB image consist of the row, column and the color dimension for red, green and blue color pixels. The binary image is of two dimensions. Here there are only two color pixels that are black and white. A one value represents a black color pixel and a zero value represents a white color pixel. Now the canny edge detector is applied to this RGB image and a binary image consisting of edges in white color and non edge pixels in black color is obtained.

Figure 3a. RGB Image

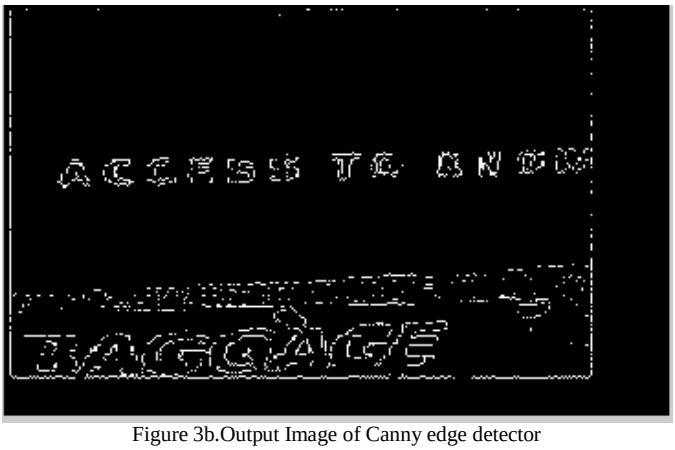

Figure 3b.Output Image of Canny edge detector

Now the scatter plot of non edge pixels is shown below. In the scatter plot there are three axes .One for red, another for green and another for blue. Now the non edge pixels of the original image are distributed in the scatter plot according to their red, green and blue composition.

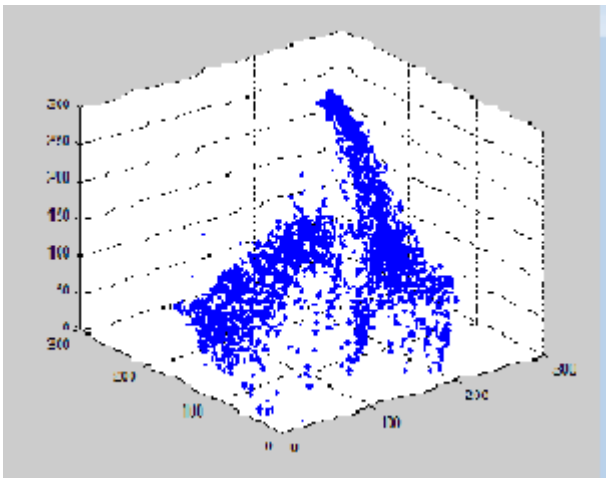

Figure 4.Scatter plot of non edge pixels

After this the k-means clustering is performed on the non edge pixels and they are placed into clusters. Now the k-means clustering method calculates the distance among the pixels. Here for color based partition, the distances between the color values are calculated to decide which pixels are in one cluster. Distance calculating methods used are Euclidean distance method and city block distance method.

After the clusters are formed, they are plotted using the silhouette plot method, which is shown below.

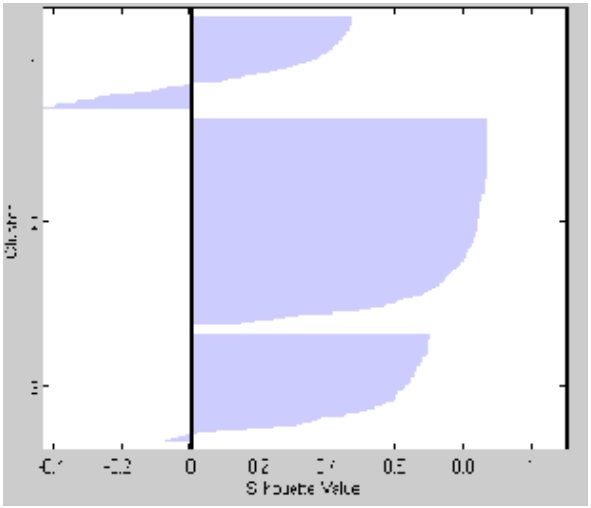

Figure 5.Sihoutte plot of clusters of non edge pixels

Now this plot shows three clusters. The values above one on the positive side shows those pixels which are well separated from other clusters. And the pixels which are having zero value represent those pixels on the boundary of two clusters. And the negative value pixels represent those pixels which are wrongly represented in other clusters.

Purpose of color based partition:-

a. K-means clustering represents all the pixels of the image, in terms of red, green and blue colors. Means by doing this, the total number of colors in the original image is reduced or similar color pixels are brought together into one cluster.

b. The pixels having slight variations in their color are brought into one cluster. So original image is simplified by repainting it with colors formed by the clusters.

After color based partition classifiers will be applied. The classifiers are used to detect whether the above partitions contain text in them or not. Because of classifiers it is sure that certain partitions will contain text definitely and by applying classifiers time required may be more. But then grouping method can be applied to those only, which contain text in them. So here efficiency is improved.

So at the cost of time, efficiency may be improved. Time and efficiency go hand in hand.

2. Text Line Grouping Method

This method uses the fact that all the charactes of one text are in one single line and they are placed at equal distances from one another. It calculates the centroids of each character and finds whether the centroids of those characters fall in one single straight line or not. For this it uses the hough transform method.

V. CONCLUSION

In this paper, first Big Data analysis has been done by using Google Apis. Then filters have been applied to the images for obtaining noiseless images. Then two methods have been applied in the image processing stage. The first method is Color based Partition. Here canny edge detector and k-means clustering have been performed. After this the trained classifiers are applied on the partitions to determine whether each cluster contains text or not. Then the text grouping method uses Hough transform to detect the text. By applying trained classifiers either time or efficiency will be improved.

The future scope will be to develop a project which will be a real time application that can be used in vehicles which will highlight the text from streetview images captured online, that will be especially useful during night driving.

References

1. Chucai Yi and YingLi Tian,"Text String Detection From Natural Scenes by Structure –Based Partition and Grouping",IEEE Trans on Image Processing, Vol.20,No 9,September 2011.
2. T. Kasar, J. Kumar, and A. G. Ramakrishnan, "Font and background color independent text binarization," in Proc. 2nd Int. Workshop Camera-Based Document Anal. Recognit., 2007, pp. 3–9.
3. H. Tran, A. Lux, H. L. Nguyen, and A. Boucher, "A novel approach for text detection in images using structural features," in Proc. 3rd Int. Conf. Adv. Pattern Recognit., 2005, pp. 627–635.
4. Q. Liu, C. Jung, and Y. Moon, "Text segmentation based on stroke filter," in Proc. Int. Conf. Multimedia, 2006, pp. 129–132.
5. K. Sobottka, H. Kronenberg, T. Perroud, and H. Bunke, "Text extraction from colored book and journal covers," in Proc. 10th Int. Conf. Document Anal. Recognit., 1999, no. 4, pp. 163–176.
6. Y. M. Y. Hasan and L. J. Karam, "Morphological text extraction from images," IEEE Trans. Image Process., vol. 9, no. 11, pp. 1978–1983, Nov. 2000.
7. C. Wolf, J. M. Jolion, and F. Chassaing, "Text localization, enhancement and binarization in multimedia documents," in Proc. Int. Conf. Pattern Recognit., 2002, vol. 4, pp. 1037–1040.
8. P. Shivakumara, W. Huang, and C. L. Tan, "An efficient edge based technique for text detection in video frames," in The Eighth IAPRWorkshop on Document Analysis Systems, 2008.
9. S. Lefevre and N. Vincent, "Caption localisation invideo sequences by fusion of multiple detectors," in Proc. 8th Int. Conf. Document Anal. Recognit., 2005, pp. 106–110.

10. J. Weinman, A. Hanson, and A. McCallum, "Sign detection in natural images with conditional random fields," in Proc. IEEE Int. Workshop Mach. Learning Signal Process., 2004, pp. 549–558.
11. T. Phan, P. Shivakumara, and C. L. Tan, "A Laplacian method for video text detection," in Proc. 10th Int. Conf. Document Anal. Recognit., 2009, pp. 66–70.
12. J. Gao and J. Yang, "An adaptive algorithm for text detection from natural scenes," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,2001, vol. 2, pp. 84–89.
13. B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in nature scenes with stroke width transform," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2010, pp. 2963–2970.
14. N. Nikolaou and N. Papamarkos, "Color reduction for complex document images," Int. J. Imaging Syst. Technol., vol. 19, pp. 14–26,2009.
15. Yi-Feng Pan,Xinwen Hou, and Cheng-Lin Liu ,"A Hybrid Approach to Detect and Localize Texts in Natural Scene Images",IEEE Trans on Image Processing,Vol.20,No 3,March 2011.
16. Y.-F. Pan, X. W. Hou, and C.-L. Liu, "A robust system to detect and localize texts in natural scene images," in Proc. 8th IAPR Workshop on Document Analysis Syetems (DAS'08), Nara, Japan, 2008, pp. 35–42.
17. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition(CVPR'05), San Diego, CA, 2005, pp. 886–893.
18. J. Sochman and J. Matas, "WaldBoost – Learning for time constrained sequential detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, 2005, pp. 150–156.
19. J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proc. 18th Int. Conf. Machine Learning (ICML'01), San Francisco, CA, 2001, pp. 282–289.
20. F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in Proc. Conf. North American Chapter Assoc. Computational Linguistics on Human Language Technology (NAACL'03), Morristown, NJ, 2003, pp. 134–141.

