RESEARCH ARTICLE

Text Extraction From Big Data Images

Vaishnavi Ganesh [1], Dr. L.G.Malik[2]

1. PG Student Computer science and Engineering Department G.H.R.C.E, Nagpur,India
2. HOD Computer science and Engineering Department G.H.R.C.E, Nagpur,India

Manuscript Info	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. In this paper various streetview images have been taken, that are provided online in the Google API, which acts as the Big data in this paper. The proposed framework in this paper consists of Color based partition method and a mesh concept, which is used to obtain the pixel color values of various locations of an image including the text portion and the non text portion. Now this information collected from various images is used to train the neural network classifier and the classifier is used to detect the text from any new given image. Experiments have been carried on various streetview images and text has been extracted successfully from 75% of the images.

INTRODUCTION

Big data can be seen as a collection of <u>data sets</u> that are very large and complex which becomes 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 <u>capturing</u> , curating, managing, and processing the data within a tolerable elapsed time.

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

Text detection and localization in big data images is required for content-based image analysis. This problem is very 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 have been increasing in number text-based image analysis techniques are receiving high importance in past few years. Out of all the contents in images text information has drawn lot of interest 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 classified into two groups: region-based and connected component (CC)-based. Region-based methods effort to detect and extract text regions by texture analysis. Usually, a feature vector extracted from each local region is fed into a classifier which estimates the likelihood of text. Then merging of neighbouring text regions is carried out to generate text blocks. On the other hand, direct segmentation of candidate text components by edge detection or color clustering is carried out in CC-based methods. Then pruning of non-text components with heuristic rules or classifiers is carried out.

I. 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, in the text binarization method [2] a bounding box has been assigned first to the boundary of each candidate character in the edge image and then text characters have been detected based on the boundary model. In the text detection using structural features [3] method ridge points were calculated in different scales for describing text skeletons at the level of higher resolution and text orientations at the level of lower resolution. In text segmentation using stroke filter method [4], a stroke filter is used to extract the stroke-like structures. In the Text extraction from colored book covers [5] method , a top–bottom analysis based on color variations in each row was combined with the column with a bottom–top analysis based on region growing by color similarity. In the Morphological text extraction method [6] a robust morphological processing was designed. In the Text localization enhancement and binarization method[7] , Otsu's method for binarization of text regions from background was improved, after which there was a set of morphological operations for reducing noise and correct classification errors. For grouping together text characters and removing out false positives, some conditions were employed by these algorithms 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 failed to remove the background noise resulting from wired mesh, atmospheric distortion, or other background objects. For reducing background noise, splitting of images to blocks is done by the algorithms in the connected component method and then merging of the blocks is done verified by the features of text characters. In the Edge based technique from video frames [8] different edge detectors are applied for searching the blocks containing the most apparent edges of text characters. In the Caption localization method [9] a fusion strategy is used which combined color detectors, detectors in texture, contour, and temporal invariance, respectively. In the Sign detection with conditional random fields [10] method, a group of filters is used to analyze texture features in each block and joint texture distributions between adjacent blocks by using conditional random field. One drawback here is that partitioning of images has been done here without any content in it and image has been divided spatially into blocks of equal size before grouping is performed. Image partitioning with no contents in it will usually break up text characters or text strings into fragments which fail to satisfy the texture constraints. Thus, in Laplacian method for text detection [11] line-by-line scans are performed in edge images for combining rows and columns with high density of edge pixels into text regions. In Adaptive algorithm for text detection [12] heuristic grouping is performed and layout analysis has been done for clustering 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 has been taken as the foundation where a variety of streetview images are provided which act as the Big Data here. The Google API is having millions of streetview images and other images which play the role of Big data in this paper. Actually here two main areas come into portrait. The Data mining field which is acting as the Big data here and the Image processing field which is the text extraction part.

After the images have been obtained, filters have been applied to the images, which remove noise from the images and give more clear picture, which becomes simple for text extraction. Study about various types of filters has been done in this paper.

After this the Color Based Partition Method is applied and a mesh is used to capture the color values of pixels at various locations of an image, including the text portion and the non text portion. This information is used to train a classifier about the text pattern and the non text pattern. Then the classifier will be able to extract text successfully from any image.

IV. METHODOLOGY USED

A. Big Data Analysis

Here the Google's API(Application Programming Interface) has been included. There are quite a few 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 are taken from Google Street view API. There are millions of pictures taken in all directions of many streets of different cities, which act as the Big Data here, which is provided by the Google API. After the registration process is complete in Google Street view API, the user receives the key. By correctly specifying the required parameters and optional parameters in the URL, the user can get the different street view images, which are found 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 either be a text string or a lat/lng value) , sensor(sensor indicates whether the request came from a device using a location sensor (e.g. a GPS) or not, to determine the location sent in this request)

The optional parameters include heading (heading indicates the camera's compass heading), 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 that are obtained are a part of big data of the streets of any city. From these street view images the text will be highlighted and exposed. The text could be from shops or from banner or the street names located from the street view images. Those text will be tinted and shown. This would be helpful for nighttime driving purposes. This could be a real time application of this project.

B. Applying Filters

There are three different types of filters.They are:-

i. Averaging filters(low pass filter)

This filfter will be consistind of a 3*3 or 5*5 mask and it should be placed on the image starting from top left corner. Then it should be moved rightwards. Here average of pixels of the original image matrix will be taken and the centre value of original image is altered. This process is recurred for all pixels. The corner pixels of actual image remain the same.

ii. Median filter(low pass filter)

The averaging filter eliminates the noise by blurring it, till it is no longer seen. But it also blurres the edges. Bigger the size of the averaging mask, more is the blurring. If the image contains salt and pepper noise and if we use the averaging filter, to remove it, it will blur the noise, but then it will also damage the edges. Hence in order to

eliminate the salt and pepper noise, we work with a non linear filter, known as median filter, or it is also called as a order statistic filter. Because their response is based on the ordering of pixels, contained inside the mask.

iii. Adaptive filter(Weiner filter)

It eliminates the low frequency regions at the same time enhancing the high frequency components.

Figure 2a. Image without applying filter

Figure 2b.Image after applying filter

C. Now the image processing stage consists of the following steps:-

Color based Partition Method

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

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

Figure 3a. RGB Image

Figure 3b.Output Image of Canny edge detector

Now the scatter plot of non edge pixels which is obtained as output is shown below. There are three axes in the scatter plot. One is for red, another is for green and another is for blue color pixels. Now the non edge pixels of the actual image has been distributed in the scatter plot according to their red, green and blue composition.
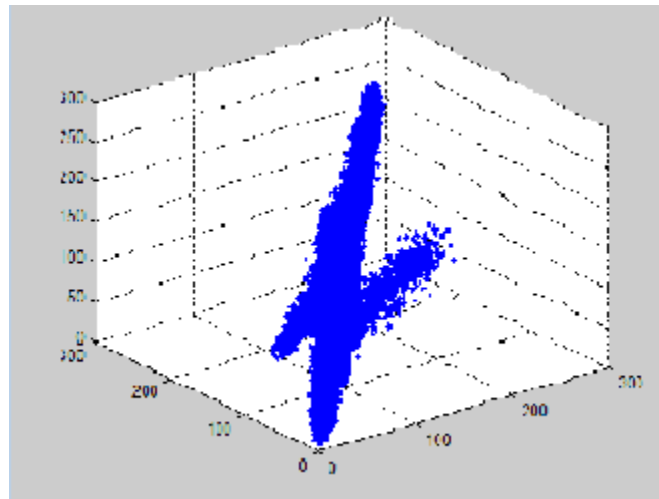

Figure 4.Scatter plot of non edge pixels

After this process the k-means clustering has been performed on the non edge pixels and they are placed into clusters. Now the k-means clustering method computes the distance among the pixels. Here for color based partition, the distances between the color values are computed for deciding which pixels fall under one cluster. Distance calculating methods that are commonly used are Euclidean distance method and city block distance method.

After the clusters have been formed, they are plotted using the silhouette plot method, which is shown below.
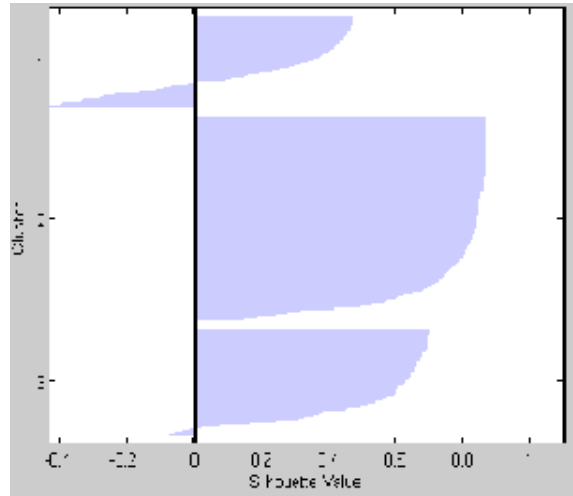
Figure 5.Sihoutte plot of clusters of non edge pixels

Now in this plot three clusters are shown. The values greater than one on the positive side represent those pixels that are well separated from other clusters. And the pixels which are having exact zero value represents those pixels on the boundary of two clusters. And the negative value pixels symbolize those pixels that are wrongly represented in other clusters.

Intentions of color based partition:-

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

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

Now the repainted image is shown below:-

Figure 6. Repainted Image

Here it can be seen that the original image has been repainted with three basic colors formed by the K-means clustering and the colors used are white,black and grey.

Now a mesh is taken.The mesh is actually a fine grid like structure.It has many horizontal and vertical lines in it.The length of mesh is taken as average length of all texts in many images. Similarly is the width of the mesh.

Now this mesh is used to capture the text patterns of various images. Basically the text and prominent borders in the image are shown by white outline in a canny edge detected output and other background are shown in black color. A white color pixel is stored as a one and a black color pixel is stored as a zero. So there exists a regular pattern of zeros and ones, if we observe the text region along with its background in canny edge detected output. So this idea is basically used for capturing the ordered pattern of zeros and ones of the text portions, through a mesh as shown below.
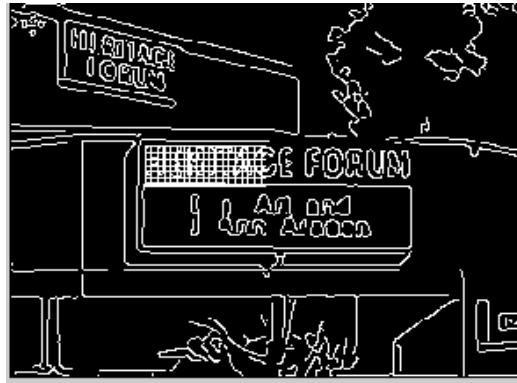

Figure 7 Mesh placed on text to capture its ordered pattern of zeros and ones

Like this the mesh is placed on different texts of different images and each ordered pattern of zeros and ones , captured by the mesh is used to train the neural network classifier1 .

Now the main idea behind the mesh concept is based on the following points:-

a.   The spatial frequency i.e the distance between the characters of a text is constant.

b.   The height of all characters of a text, almost fall inside one bounding box.

c.   The width of characters of a text is the same.

d.   The color of characters of one single text are usually the same.

The above mentioned points are inspired by the Adjacent character grouping method.

Now the non text regions which also have regular pattern of zeros and ones in them are captured as shown below and they are used to train the neural network classifier1.
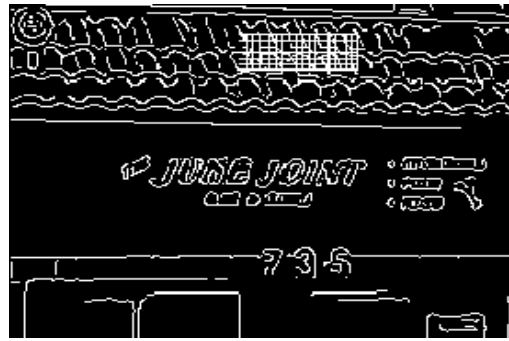
FIGURE 8 MESH PLACED ON NON TEXT REGION

Like this many non text regions were captured by the mesh and it is used for training the neural network classifier1, about the non text regions.

V RESULTS AND DISCUSSIONS

While training the neural network, the training has been done in such a way that the text region will be outputed as 1 and the non text region will be outputted as 0.The output from the trained set have been compared with the output of the test set. So we obtained the following confusion matrix.
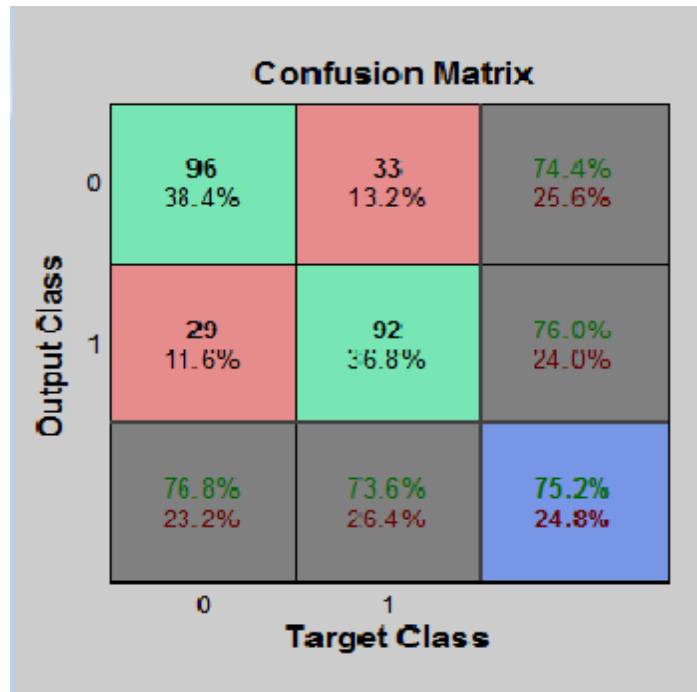

Figure 9 Confusion matrix showing the accuracy percentage

It can be seen that wherever the expected output of the text .i.e the target class is 1, for 92 cases we got 1 i.e the detected texts in actual output also. Similarly for 96 cases the target class and the output class, both have a 0. In brief, the proposed methodology has 75.2% accuracy.

Now if the classifier is given a new image, it will scan the image from top to bottom for finding the text and the output is as shown below:-

Figure 10a Text extracted

Figure 10b Text extracted from streetview image

The text is extracted from various locations, where actually there is text, which are shown by red rectangles.Also there are some false positives.

Now classifier2 will be taken to train about the color values of the pixels of the text region including its background , to show it the regular change in color values of the background pixels with that of the text, which will be taken from the repainted image. Owing to this the false positives will be reduced.

Pros and Cons :-

The pros are, if trained properly for other languages the text will be extracted in those languages also other than English language.

The cons are , if the text size is too small, then it will not be shown in the canny edge detected output and hence it cannot be extracted from the image. But this is a genuine limitation as even human eye cannot read very small texts.
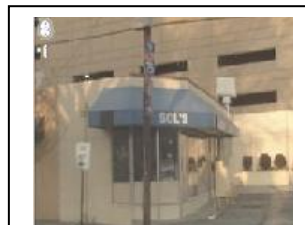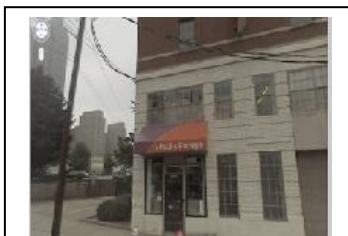
Figure 11 Type of text which cannot be extracted

VI 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 the color based partition has been applied on the image. Here we have applied the canny edge detector and the k-means clustering on the image. Then a mesh has been designed for capturing the ordered pattern of zeros and ones between the text region and its background, in a canny edge detected output. The collected information from many images have been used to train the classifier. When the trained classifier was compared with the test set, text was extracted successfully from 75% of the images. And also another classifier is trained regarding the ordered change in color values of the text pixels with that of its background pixels. Owing to this some false positives were reduced and more accuracy was obtained.

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 specially 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.