Detecting Text in Natural Scenes with Connected Component Clustering and Nontext Filtering

Mr. Mule S. S.
PG Student, Department of CSE
TPCT'S College of Engineering Osmanabad, Maharashtra, India
ssmule137@gmail.com

Mr. Holambe S. N.
Asst. Prof. Department of CSE
TPCT'S College of Engineering Osmanabad, Maharashtra, India
snholambe@yahoo.com

Abstract - *Text detection and recognition is an emerging topic for researchers in the field of image processing, pattern recognition and multimedia. Several methods have been developed for text detection and extraction that achieve reasonable accuracy for natural scene text as well as multi-oriented text. To improve the text detection accuracy most of the methods use classifier and large number training samples. An efficient method of scene text using two machine learning classifiers: one for generating candidate word regions and the other for the classification of text or nontext components. At first we extract connected components with the help of maximally stable extremal regions algorithm. The resulting components are partitioned into clusters with help of an adaboost classifier based on adjacency relationship. After that we extract features for classification from the clusters. In the existing system, the character detection in the image is done by connected component (CC) based approach. CC based approach is computational complexity, based on the pixel difference between the text and background of the image text is detected. The proposed system will extract and recognize the characters too. This process is going to do with the help of training the fonts of alphabets and numbers which are stored in datasets.*

General Terms

Computer vision

Key words: Connected component, Maximally Stable Extremal Region, CC clustering, nontext filtering

1. INTRODUCTION

Recognizing text in natural scene images is becoming a popular research area due to the wide spread availability of image capturing devices like digital cameras, mobile phones in low-cost. Various scene text detection and recognition have received much attention for the last decades. Among them, text detection and recognition in camera based images have been considered as very important problems in computer-vision community [1], [2].In this paper, we presents an innovative scene text detection algorithm with the help of two machine learning classifiers one for candidate generation and the other which filters out non text ones. Texts present in an image or video can be classified as scene text and caption text. In such cases, scene text exists in the image naturally while caption texts refer to those texts which are added manually by the user. Scene texts overlap with the background therefore scene text detection and extraction are difficult as compared to the detection of caption text. Text detection and extraction from images have a lot of valuable and useful application. For the text-based image search, text which is appeared within images has to be robustly located. This is a challenging task because the presence of wide variety of text appearances, such as variations in font and style, geometric and photometric distortions, partial occlusions and different lighting conditions. Text detection has been considered in many recent studies and lots of different methods are reported in [1, 3, 4, 5, 6, 7, 8].

With the increasing popularity of computer vision systems and mobile phones, text detection in natural scenes becomes a critical challenging task. To extract scene text information from camera captured images many algorithms and optical character recognition (OCR) systems have been developed [1]. It is impossible to recognize text in natural scenes directly because the OCR software cannot handle complex background interferences and non orienting text lines. Document analysis is not limited to texts but also concerned with photographs of vehicle number plates, street names, gas/electricity meters and so on where automatic recognition of scene text is desired. Examples of scene texts include signs on streets, display boards on shops, texts on vehicles, advertisement boards etc. Fig 1 shows examples of text in natural scene images. The challenges constitute the area of scene text detection requiring the researchers to go beyond the traditional techniques for document image analysis to solve them. So we developed a system for the scene text detection using connected component clustering and nontext filtering.

Fig 1: Examples of natural images with scene text

2. RELATED WORK

Most of the text detection algorithms can be classified into two categories: texture-based and connected component (CC) – based [1], [2]. Texture-based approaches view text as a special texture that is distinguishable from the document image background. Generally, here the features are extracted over a certain region and a classifier is employed to identify the existence of text. As compared to texture-based methods, the CC-based approach extracts regions from the image and uses geometric constraints to filter out non-text candidates. More recently, Maximally Stable Extremal Regions (MSERs) [9] based methods, which can be categorized as connected component based method. Connected component analysis method is used to define the binary image that consists of text regions. After the CC extraction, CC-based approaches filter out non-text part.

The region-based methods [5], [6] have focused on binary classification that consists text verses non text part of a small image patch. They have focused on the following problem: 1) Problem (A): to determine whether a given patch is a part of a text region. In this problem researchers addressed problem by adopting cascade structure. This structure enables efficient text detection but problem (A) is challenging. It is not straightforward for human to determine the image patch when we do not have knowledge of text properties such as scale, skew and color. Many experimental results shows that this region based approach is efficient, however, it yields worse performance compared with CC-based approaches. CC-based methods begin with CC extraction and normalize text regions by processing only CC-level information. Therefore, they have focused on the following problems: 2) Problem (B): to extract text-like CCs, 3) Problem (C): to filter out non text CCs, 4) Problem (D): to infer text blocks from CCs.

In connected component based methods, at first an image is divided and candidate text components are extracted. After that non text elements are eliminated through different recent techniques. Connected component based methods make use of geometrical properties. Some of the works contributed to development of the present work are

Epshtein et al [8] describe a method that makes use of stroke width for the extraction of text components. A stroke is a contiguous part in an image that forms a band of approximately constant width. Constant stroke width is one of the important feature that separate texts from other components of a scene. In this method they make use of a logical operator together with geometrical reasoning that identifies the place having same stroke width for the identification of regions having text.

Yi et al [10] describes a method that use gradient features and color homogeneity of character components for the extraction of candidate text regions. After that character candidate grouping is performed to detect text strings. This is performed on the basis of structural features of characters in text string such as differences in character size, distance between neighboring characters and alignment of characters.

Gatos et al [11] described a methodology for text detection from natural scene image is based on an efficient binarization and enhancement technique followed by a connected component analysis procedure. Starting from the original image, the method produces a binary image and an inverted binary image. Then connected components are extracted from complementary images. Further, the text verification is conducted at character level and word level on the candidate connected components. Finally, text regions localized in two images are refined and merge in post-processing.

More recently, Maximally Stable Extremal Regions (MSER) have become one of the commonly used region detector because of their high repeatability and also partly because they are somewhat complementary to many other commonly used detectors. The original MSER detector finds regions that are stable over a wide range of thresholdings of a gray scale image. MSER-based methods have demonstrated very promising complementary performance in many real projects. However, current MSER-based methods still have some key limitations i.e. they may suffer from detecting of repeating components and also insufficient text candidates construction algorithms. The main advantage of MSER-based methods over traditional connected component based method is able to detect most characters even when the image is in low quality.

3. PROBLEM DOMAIN

Although many research efforts have been made to detect text regions from natural scene images, more robust and effective methods are expected to handle variations of scale, orientation and clutter background. CC-based approaches have shown better performance than region-based ones, they usually suffer from computational complexity. It is because their performance depends on the quality of CCs and they

adopted sophisticated CC extraction and filtering method.

The proposed framework is able to effectively detect text strings in arbitrary locations, sizes, orientations, colors and slight variations of illuminations or shape of attachment surface. Compared with the existing methods which focus on independent analysis of single character, the text string structure is more robust to distinguish background interferences from text information. Also used to determine whether the connected components belong to text characters or unexpected noises.

4. PROPOSED METHOD

The proposed method focus on scene text detection from different document images and camera captured images. After the detection the obtained scene text image which is free from complex background, variation in font, size and orientation, color distortions and noise. The proposed method consists of three steps: candidate generation, candidate normalization and nontext filtering. The figure shows the block diagram of our proposed method.

Our candidate generation method is based on popular CC based approaches which consists of a MSER-based CC extraction block and an AdaBoost-based CC-clustering block. The maximally stable extremal region (MSER) algorithm is invariant to scales changes and affine intensity changes and other blocks in our method are also designed to be invariant to these changes. In our method, both problems ((D) and (A)) are addressed based on machine learning techniques, so that our method is largely free from heuristics. We have trained a classifier that determines adjacency relationship between CCs for problem(D) and we generate candidates by identifying adjacent pairs. In training, we have selected efficient features and trained the classifier with the AdaBoost algorithm [9]. We apply the CC clustering method to raw CC sets that usually contains many nontext CCs and some candidate may correspond to nontext clusters.

Candidate normalization means not only geometric but binarization is also involved. We can localize character candidate regions with CC-level information and this localization allows us to build a simple but reliable text/nontext classifier. From that normalized image we built a binary image. Filtering means nothing but a text-nontext classifier is developed and rejects non-texts from normalized images.

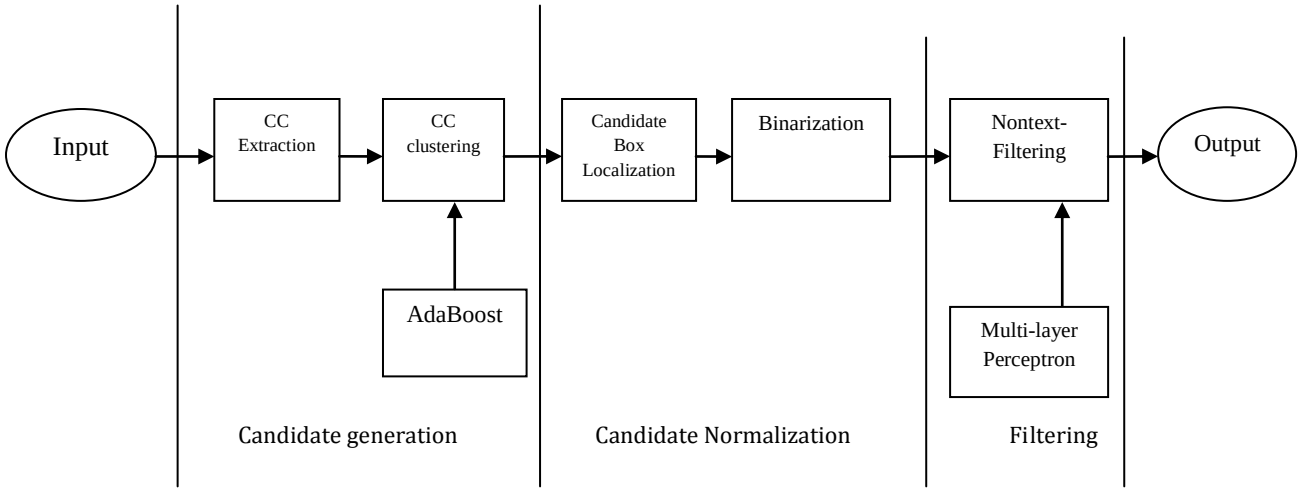

Fig 2: Block diagram of our proposed method

4.1 Candidate Generation

In the proposed method for the generating a candidates, extract CCs in images and partition the extracted CCs into set of clusters, using clustering algorithm is based on an adjacency relation classifier. In this section first explain the CC extraction method. Then explain the approaches (i) building a training samples, (ii) to train the classifier and (iii) using that classifier in CC clustering method.

Fig 3: Input Image

4.1.1 CCs Extraction

The MSER algorithm [9] is the most efficient in CC extraction because it shows a good performance with a small computation cost only the MSER algorithm could provide the stable binary results and also help us find most of the text components.

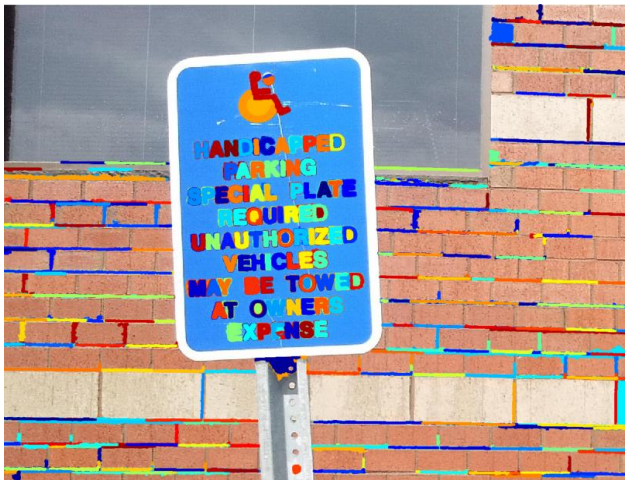

Fig 4: Detect MSER Regions

4.1.2 Building Training Sets

Our classifier is based on pair wise relations between CCs and CC pairs. Therefore, rather than focusing on this difficult problem, we address a relatively simple problem by adopting an idea of region-based approaches. If we have ci ~ cj it will yield a character candidate components consisting of non text CCs and this candidate will be rejected at the non text rejection step. Also we will perform word segmentation as a post processing step and based on these observations we build training sets. Specifically, we first obtain sets of CCs by applying the MSER algorithm to a training set released.

4.1.3 Adaboost Learning and CC Clustering

Using the collected samples we train an AdaBoost classifier that tells us whether it is adjacent or not. For these operations we shall define some local properties of CCs. We have used 6-dimensional feature vectors consisting of five geometrical features and one color-based feature. All of geometric features are designed to be invariant to the scale of an input image and the color feature is given by the color distance between two CCs in RGB space. All of these features are informative and consider each feature as a weak classifier. From these weak classifiers, build a strong classifier with the AdaBoost learning algorithm. The AdaBoost is easy to implement and known to show good performance in many applications.

4.2 Candidate Normalization

After CC clustering there are set of clusters. In this section, we will normalize corresponding regions for the reliable text/nontext classification.

4.2.1 Geometric Normalization

Here we first localize its corresponding region. Eventhough text boxes can experience perspective distortions, then approximate the shapes of text boxes with parallelograms whose left and right sides are parallel to y-axis. This approximation alleviates difficulties in estimating text boxes having a high degree of freedom (DOF). The normalization method is used only to find a skew and four boundary supporting points. To estimate the scale and skew of a given word candidate wk , we build two sets as:

Tk = {t(ci)|ci € wk }

Bk ={b(ci)|ci € wk }

Where t(ci) and b(ci) are the top-center point and the bottom center point of a bounding box of ci respectively. Then, perform geometric normalization by applying an affine transform mapping that transforms that corresponds region to a rectangle. During the transformation, use a constant target height and preserve the aspect ratio of the box.

4.2.2 Binarization

Given geometrically normalized images which build binary images. In many cases, MSER results can be considered as binarization results. However, it performs the binarization separately by estimating document text and background colors. It is because (i) the MSER results may miss some character components and/or yield noisy regions and (ii) it have to store the point information of all CCs for the MSER-based binarization.

4.3 Non Text Filtering

A text/nontext classifier that rejects nontext blocks among normalized images. The main challenge of the approach is the variable aspect ratio. One possible approach to solve this problem is to split the normalized images into patches covering one of the letters and develop a character/non-character classifier. However, character segmentation is not an easy problem so split a normalized block into overlapping squares and develops a classifier that assigns a textness value to each square block. Finally, the decision results for all square blocks are integrated so that the original block is classified.

4.3.1 Feature Extraction from a Square Block

Here each square blocks can be divided into four horizontal and vertical ones and extract the features. For a horizontal block Hi (i= 1, 2, 3, 4), we consider

1) The number of white pixels
2) The number of vertical white-black transitions
3) The number of vertical black-white transitions

as features and features for a vertical block is similarly defined.

4.3.2 Multilayer Perceptron Learning

For the training, we need normalized images for this goal, applied the method that is candidate generation and normalization algorithms to the training images. Then, manually classified them into text and nontext and also discarded some images showing poor binarization results. The text/ nontext images are divided into squares and have trained a multilayer perceptron for the classification of square patches. To help the learning, input features are normalized.

5. FUTURE WORK

Our future work will focus on developing learning based methods for text extraction from complex backgrounds and text localization for OCR recognition. We also attempt to improve the efficiency and transplant the algorithms into a navigation system prepared for the finding of visually impaired people.

6. CONCLUSION

We have presented in this paper an improved text string detection method which can effectively detect text from the document background. Due to the unpredictable text appearances and complex backgrounds, text detection in natural scene images is still an unsolved problem to locate text regions embedded in those images. In this paper, we have presented a novel scene text detection algorithm with the help of two machine learning classifiers: one for candidate generation and the other which filters out non text ones.

7. ACKNOWLEDGMENTS

The present work has been undertaken and completed with direct and indirect help from many people and I would like acknowledge all of them.

8. REFERENCES

[1] K. Jung, “Text information extraction in images and video: A survey,” Pattern Recognit., vol. 37, no. 5, pp. 977-997, May 2004.

[2] J. Zhang and R. Kasturi, “Extraction of text objects in video documents: Recent progress,” in Proc. 8th IAPR Int. Workshop Document Anal. Syst. , Sep. 2008, pp. 5-17.

[3] J. Liang, D. Doermann, and H. P. Li., “Camera-based analysis of text and documents: a survey,” IJDAR, vol. 7, no. 2-3, pp. 84-104, 2005.

[4] Y. Zhong, H. Zhang and A. K. Jain, “Automatic caption localization in compressed video,” IEEE Trans. Pattern Anal. March. Intell. Vol. 22, no. 4, pp. 385-392, 2000.

[5] X. Chen and A. L. Yuillle, “Detecting and reading text in natural scenes,” in CVPR, 2004, vol. 2,pp. II-366-II-373 Vol. 2.

[6] X. Chen and A. L. Yuillle, “A time-efficient cascade for real-time object detection: With applications for the visually impaired,” in CVPR-workshops, 2005, p. 28.

[7] S. M. Lucas, “ICDAR 2005 text locating competition results,” in ICDAR, 2005, pp. 80-84 Vol. 1.

[8] B. Epshtein, E. Ofekand Y. Wexler, “Detecting text in natural scenes with stroke width transform,” in CVPR, 2010, pp. 2963-2970.

[9] J. Matas, O. Chum, U. Martin and T. Pajdla, “Robust wide baseline stereo from maximally stable extremal regions,” in Proc. Brit. Mach. Vis. Conf., 2002, pp. 384-393.

[10] Yingli Tian and Chucai Yi, “Text string detection from natural scenes by structure based partition and grouping,” IEEE transaction on image processing, vol. 20, no. 9, pp. 2594-2605, 2011.

[11] Gatos, B. Pratikakis, I. Perantonis, S. J., “Towards text recognition in natural scene images,” in proceedings of Int. Conf. Automation and Technology, p. 354-359, 2005.