International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Novel Framework for Text Recognition in Street View Images

Mehmet Serdar Guzel[*1]

Accepted : 13/07/2017 Published: 30/09/2017

Abstract: This paper addresses a new text recognition solution, which is mainly used for the detection of street view images. This paper employs two different approaches to detect text-based regions and recognise corresponding text fields. The first approach utilises maximally stable extremal regions (MSER), whereas the second approach relies on the class specific extremal regions (CSER) algorithm. Two separate frameworks, designed with respect to the aforementioned methods, are applied to the street view images so as to extract text-based regions. Numerous experiments were performed to evaluate and compare both approaches. Results obtained from the CSER-based approach are especially quite encouraging and verify the system's ability to detect text-based regions and recognise corresponding text fields.

Keywords: *Text recognition, MSER, CSER, signboard detection, street view images*

1. Introduction

The analysis and evaluation of signboards is a critical issue, especially in developing countries in terms of environmental transformation and renewal processes. Furthermore, the taxation polices for signboard detection is also an important problem. Accordingly, many government resources are wasted to detect signboards and recognise text fields located on each signboard [1-3]. It is clear that the text in the image provides very useful information about the image, which essentially provides appropriate clues for a wide variety of applications.

In our daily life, we encounter signboards and signboard posts. The autonomous detection of signboards and recognition of posts will open the door to many different areas. For instance, the location of shops can be determined using signboards and GPS. Within this approach, a map of the environment can be easily obtained only by using the text fields written on signboards [4,5].

Figure 1: An example street view images including signboards.

An example of a complex street view image, including numerous signboards is shown in Figure 1.

Direct extraction of text regions from street view images is a challenging task due to the texts having different formats and variant background interventions [6]. Signboard detection tasks mainly involve two sub-processes, namely, text detection from images and recognition of characters, which may have different fonts and sizes. As opposed to other natural view images, street view images have quite complex contents due to the issues aforementioned. In the literature, there exist several different classification approaches for analysing text detection and recognition methodologies, which have been summarized in the following review papers [7,8]. One of those approaches categorises them into two groups, namely, stepwise and integrated methodologies. The algorithms using the stepwise methodology primarily employ localization, validation, segmentation and recognition steps respectively. The frameworks relying on this methodology primarily follow a coarse-to-fine strategy, which first estimates position of text candidates, and then perform validation, segmentation, and recognition steps respectively [7,9,10]. Integrated methodologies, on the other hand, first utilize a character classification module, which is the most critical step and the results of this procedure are then employed by detection and recognition modules [11]. This methodology not only tends to extract characters from background but also from each other. This is a challenging problem, requiring a reliable feature detector and a robust classifier. For instance, well-known approaches, considered in this category, employ HOG algorithm to extract features and a nearest neighbour or SVM as classifier [12]. An interesting study, alternatively, proposes a multi-layer CNN based design to overcome both detection and recognition phases of text recognition problems [13]. It is clear that the advent of deep learning will lead text recognition capabilities of integrated methodologies in a more advanced level [14,15].

Besides, text detection based methods can also be classified into two categories, namely, texture based approach and connected component based approaches (CC) [7,8]

[1]*Ankara University, Computer Engineering Department – Ankara/Turkey*
** Corresponding Author:Email: mguzel@ankara.edu.tr*

Text detection-based methods can be classified into two categories, namely, a texture-based approach and connected component-based approaches (CC). Texture-based approaches consider the text as a special pattern that is particularly different from the background. Typically, features are extracted over a certain region using well-known feature extractors, and then a classifier is utilised to identify the existence of text [16]. Alternatively, CC-based methods extract regions from the image and employ different geometric parameters or statistical approaches to exclude non-text fields. One recent study employs this approach to an image with stroke width transformed [17]. In another study,K-mean clustering is employed to detect connected components in which straightness and edge density parameters are used to eliminate false positives [18].

This paper proposes two solutions for signboard detection and text recognition problems. One employs the MSER-based text recognition approach, which is basically intensity-based blob detection; and similarly, the second approach is an extremal region (ER)-based scene text detection system, which is also a CC-based method and estimates connected components whose intensity is higher or lower than its nearby pixels. This method is called CSER (class specific extremal regions), which is a generalisation of an MSER detector possessed within learning capacity. Overall, section 2 addresses the design of both approaches, whereas section 3 includes the experimental section. The paper is concluded in section 4.

2. Text Recognition Frameworks

This section will detail both solutions for estimation of text from street view images. As previously mentioned, two different methods were employed to overcome this critical computer vision problem. Essentially, two different frameworks were built relying on two comprehensive methods. Section 2.1 details the framework using the MSER method, whereas section 2.2 addresses the ER-based method.

2.1. MSER-based text recognition system

MSER is one of the most popular and efficient blob detectors due to its robustness against scale changes and lighting conditions. It is in essence a natural choice for text detection problems [16]. MSER is an intensity-based algorithm whose size remains unchanged over a range of thresholds. The methods work well but have problems especially on blurry or low contrast images [19].According to the MSER algorithm, first, a series of threshold values (sweeping) from black to white is applied, and afterwards, connected components are extracted. A threshold value within the maximally stable region (1) is estimated. Finally, each region is considered as a feature and may be approximated to each region with an ellipse. Extremal denotes that all pixels inside the MSER regions have higher or lower intensity values than all the pixels on its outer boundary.

$$R_1 * = \text{argmin}|R_{i+\Delta} \backslash R_{i-\Delta}|/|R_i| \quad (1)$$

where, $R_1 R_2, \ldots . R_i$ are nested extremal regions and "Δ" is a parameter.

R_{i^*} is an MSER and produces a local minimum on the nested chain $R_1, R_2, \ldots, R_{\max}$ along the threshold variable. The extremal regions are rejected if they are too small, large or similar to its parent MSER. Details can be seen in [16, 19]. MSER works with images having homogeneous regions with distinctive boundaries aswell as with small regions, whereas it cannot tolerate motion blur.

The MSER framework is illustrated in Figure 2. As illustrated in the figure, the proposed framework detects signboards from street view images and then recognises text-based regions. According to the framework, MSER is employed to detect ROI, which may cover text fields. Afterwards, a canny edge detector is applied to detect corner pixels in a more efficient manner. The connected component analysis algorithm is then applied to estimate transitions between meaningful pixel blocks, and then the most stable pixel blocks are obtained. In the final step, stroke width transformation is an image operator that computes per pixel width most likely stroke containing the pixel are applied to remove pixel based defects [17]. Next, a reliable OCR library is employed to recognise characters, and a dictionary module is employed to remove both unlisted characters and complete missing words. Figure 3 depicts the results of an example scenario using the MSER-based signboard detection system obtained from street view images.

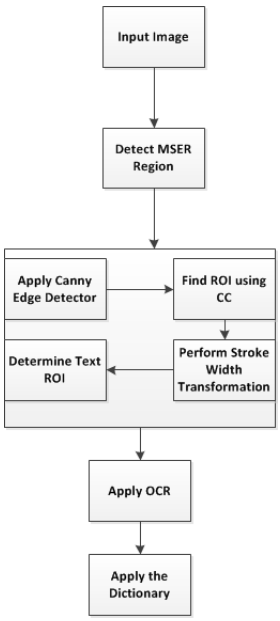

Figure 2: MSER-based signboard detection system.

Figure 3: MSER-based framework is applied to the scenario obtained from street view images based signboard detection.

2.2. CSER-based text recognition system

The class specific extremal regions (CSER) algorithm is similar to the MSER algorithm, where appropriate extremal regions are calculated using the intensity-based approach. However, the main difference is that the CSER algorithm relies on a sequential

classifier trained for character recognition, which drops the stability requirements of MSER but selects class-specific regions [20]. The CSER-based text recognition algorithm first checks the probability of extremal regions (ERs) having characters. ERs within local maximum values pass to the second stage. The classification is supported by employing computationally expensive features. Finally, an exhaustive search using a feedback mechanism is applied to groups so as to extract probable character regions, and then an OCR module is applied to recognise characters. The details of the algorithm can be seen in [20] and the pseudocode of the algorithm is also shown in algorithm 1.

Algorithm 1: CSER-based text recognition system

Input: *Thresholds T on Image I*
Pixels p of the Image I
Output: *CSER regions*
While *ERs are updated*
If *unconnected pixel is < "T"*
Create a new region
Elseif *pixel lies on the border and < "T"*
Append pixel
Elseif *pixel if two regions are connected via p*
Merge Regions
endWhile
Recalculate features for updated ERs
Employ classifier to decide whether region belongs to CSER or not

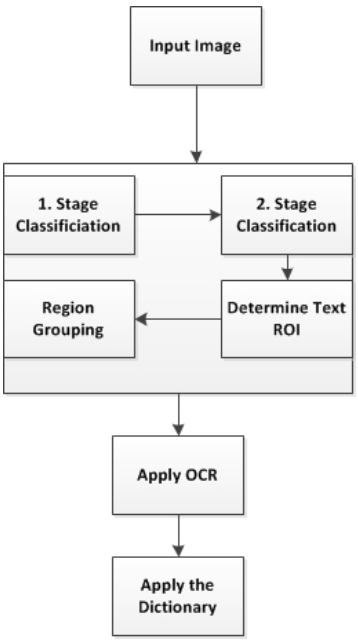

Figure 4: CSER-based signboard detection system.

Figure 4 also illustrates the CSER-based signboard detection system used for street view images. The CSER algorithm has a cascade structure (sequential classifier) with two stages. In the first stage, the following descriptors are employed, namely, 'area', 'bounding box', 'perimeter' and 'Euler number'. Afterwards, a real AdaBoost classifier using decision trees was employed with those features [21]. In the second stage, an SVM classifier additionally employs further parameters such as 'hole area ratio' and 'convex hull ratio'. For the grouping step, an efficient and pruned exhaustive search-based approach is employed, which searches character sequence space in real time. Details of this search can be seen in [22]. Afterwards, a reliable OCR library is utilised to identify characters, and a dictionary module is employed to remove both unknown characters and complete missing words.

Figure 5: CSER-based framework is applied to the scenario obtained from street view images based signboard detection.

Figure 5 shows the results of an example scenario using the CSER-based signboard detection system obtained from street view images.

3. Experimental Section

This section compares and details the experimental result of the proposed MSER-based and CSER-based frameworks for signboard detection and text recognition problems using street view images. The experiments are run on an Intel Core i7 2.2 GHz with 8 GB ram computer. The frameworks were developed using OpenCV 3.2 with the Windows operating system. As aforementioned, the main motivation lies behind this study to develop signboard recognition to be used in cluttered images obtained from street view images, especially in Turkey. Consequently, instead of utilizing well-known benchmark dataset, which cannot meet the requirements of commercial applications,a data set including 400 images was obtained.

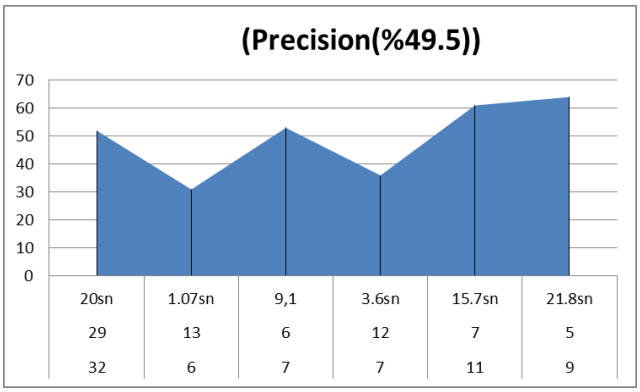

Figure 6: A randomly selected dataset; from top to bottom: process time, false positive (FP) and true positive are shown (MSER).

This dataset was obtained by employing open source mapping and imaging services; the dataset includes images from different municipalities all over Turkey. As previously mentioned, this dataset consists of images obtained from several municipalities located in Turkey. Also, the open source Tesseract OCR library is employed for the recognition library. For this experimental part, a small dataset is obtained from the given image corpus, and the

precision parameter (2) is employed to compare both architectures' accuracy over the given dataset.

$$Precision = TP/(TP + FP) \quad (2)$$

where, *TP is true positives and FP is false positives.*
While TP depicts correctly identified samples, FP depicts incorrectly identified ones.

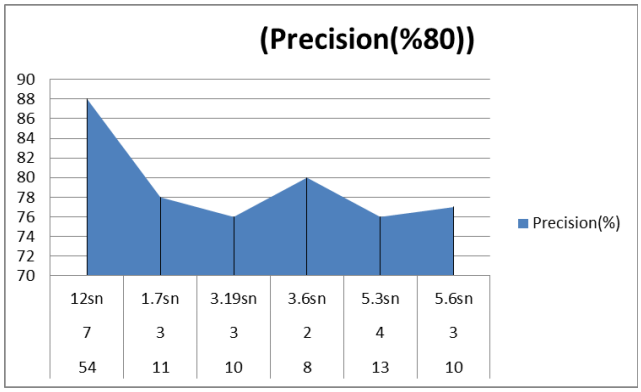

Figure 7: A randomly selected dataset (CSER) process time, false positive (FP) and true positive (CSER).

Figure 6 illustrates the results of the MSER-based architecture over a randomly selected dataset, which includes six high contrast and detailed images. However, this approach can achieve only 50% accuracy on signboard detection tasks, unexpectedly.

Figure 8: Recognition of text on signboards (MSER).

Figure 9: MSER-based signboard detection.

Alternatively, Figure 7 represents the results of the CSER-based architecture over the same dataset, which, however surprisingly, achieves 80% accuracy on the signboard detection task. Results of both methods highly depend on the quality of acquired images, as expected. However, within the given corpus overall performance advantage of the CSER-based approach is almost 30% better than the MSER-based approach.

Figure 10: Recognition of text on signboards (CSER).

Figure 11: CSER-based signboard detection.

An example scenario using the image shown in Figure 1 is used to reveal the text recognition skills of both systems. Figures 8–11 include the results of both approaches; Figures 8 and 9 illustrate the text recognition and signboard detection results of MSER method. Furthermore, Figures 10 and 11 illustrate the identified characters from CSER method, respectively. As mentioned previously, both architectures employ the same OCR and dictionary modules. Therefore, a critical comparison can be made considering the segmentation and signboard detection of both architectures that the CSER-based one achieves far more than the MSER based architecture.In order to have a better comparison, SVT (Street View Text Dataset), public and benchmark dataset, was also employed to compare both approaches. One of the recent and leading papers compares End-To-End Text detection performances of comprehensive text recognition algorithms [7]. Results reveal that despite MSER based detection approach achieves a solid performance; the integrated systems using an AI based learning phase results in better recognition performance. [7]. With respect to end-to-end text detection, as illustrated in Table 1. MSER based approches performance a low precison rate especially, for SVT dataset, however CSER based framework results in better detection performance especially in RSD dataset.

Table 1: End-to-End Text Detection Performance.

	DATASET	PRECISION (%)
MSER	***SVT***	%37.5
CSER	***SVT***	%67.3
MSER	***RSD***	%49.5
CSER	***RSD***	%80

This study mainly aims to propose a solution for detection of signboards, especially used in Turkey. In that regards, two different frameworks are tested. Consequently, results prove that CSER based framework performs high precision results, especially in dataset (RSD), gathered from different municipalities all over the Turkey.

4. Conclusion

The detection of signboards from street view images is a challenging task and requires obtaining a region of interest (ROI), including texts and characters. Accordingly, two different architectures were designed, based on two different segmentation algorithms. Both architectures are supported by a powerful OCR library and a dictionary module. The first architecture is mainly designed based on a well-known and efficient segmentation algorithm, namely, maximally stable extremal regions (MSER). A corresponding system and segmentation algorithm narrows the searching field and increases the overall possibility of correctly detecting signboards obtained from street view images. However, street images may include different fonts that reduce the overall performance of the first approach, which can have, at most, a 50% precision value at the detection of signboards. The second method, class specific extremal regions (CSER), on the other hand, employs trained data to detect the ROI that the trained set produces using the rotation and orientation models of each character. Therefore, CSER detects text-based regions in a more robust and efficient manner. Furthermore, the CSER-based system employs an advanced grouping method that achieves better performance in detecting text-based regions. A series of experiments were conducted to evaluate both approaches in detecting signboards from street view images. The results reveal that the CSER-based approach is superior to the MSER-based approach and can be efficiently used to detect text-based regions, even in cluttered images.

Acknowledgment

Some part of this study is supported by the Scientific and Technological Research Council of Turkey (TUBITAK-TEYDEP Project No: 3140566).The author is also grateful for support and collaboration of Netcad Software Inc.

References

[1] R. D. Brown, "Example-based machine translation in the pangloss system," in Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, DK, 1996, pp. 169-174.

[2] Y. Cui and Q. Huang, "Character Extraction of License Plates from Video," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, San Juan, USA, 1997, pp. 502-507.

[3] J. Gao and J. Yang, "An Adaptive Algorithm for Text Detection from Natural Scenes," in Proceedings of Computer Vision and Pattern Recognition, 2001, (CVPR 2001).

[4] J. W. Hutchins, "Machine Translation: Past, Present," Future, Ellis Horwood Limited, England, 1986.

[5] A. K. Jain and B. Yu, "Automatic text location in images and video frames," Pattern Recognition, vol. 31, no. 12, pp. 2055-2076, 1998.

[6] K. S. Lahari, "Text Detection from Natural Image using MSER and BOW,"IJEERT, vol.3, pp. 152-156, 2015.

[7] Q. Ye and D. Doermann, "Text Detection and Recognition in Imagery: A Survey," IEEE Transcactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 7, pp. 1480-1500, 2015.

[8] K. Wadhawan and E. Gajendran, "Automatic Recognition of Text in Images: A Survey", International Journal of Computer Applications, vol. 127, no. 15, pp. 15-19, 2015.

[9] K. Elagouni, C. Garcia, and P. Sbillot, "A comprehensive neuralbased approach for text recognition in videos using natural language processing," in Proc. ACM Conf. Multimedia Retrieval, 2011

[10] C. Yao, X. Zhang, X. Bail, W. Liu, Y. Ma and Z. Tu, "Detecting texts of arbitrary orientations in nautral images," in Proc. IEEE Int. Conf. Comput. Vis. Pattern Rec., pp. 1083-1090, 2012.

[11] W. Wu, D. Chen, and J. Yang, "Integrating co-training and recognition for text detection, "in Proc. IEEE Int. Conf. Multimedia Expo, pp. 1169-1169, 2005.

[12] L. Tang and J. R. Kender, "A unified text extraction method for instructional videos, " in Proc. IEEE Int. Conf. Image Process, pp. 1216-1219, 2005.

[13] T. Wang et al., "End-to-end text recognition with CNN, " in Proc. IEE Int. Conf. Pattern Recognition, pp 3304-3308, 2012.

[14] A. Bissacco et al., "PhotoOCR: Reading Text in Uncontrolled Conditions, " IEEE International Conferenceo n Computer Vision, Sydney, NSW, 2013, pp. 785-792, 2013.

[15] S. Yousfi et al., "Deep Learning and recurrent connectionist-based approaches for Arabic text recognition in videos," 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, pp. 1026-1030, 2015.

[16] G. Nagaraju et al., "Text Extraction From Images With Edge-Enhanced MSER And Hardware Interfacing Using Arduino", IJECS, vol.4, pp 11798-11803, 2015.

[17] B. Epshtein et al., "Detecting text in natural scenes with stroke width transform," in CVPR, CA, USA, pp. 2963-2970, 2010.

[18] P. Shivakumara et al., "A Laplacian approach to multi-oriented text detection in video," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, pp. 412-419, 2011.

[19] L. Neumann et al.,"Real-Time Scene Text Localization and Recognition," in 25th IEEE Conference on Computer Vision and Patter Rec. , RI, USA, 16-22 June, 2012.

[20] G. Li et al., "Scene text detection with extremal region based cascade filtering, " in IEEE ICP Conference, Phoenix , AZ, USA, pp. 2896-2900, 25-28 Sept, 2016.

[21] J. Mataset et al., "A new class of learnable detectors for categorization, "In Image Analysis, vol. 3540 of LNCS, pp. 541-550, 2005.

[22] K. R. Muller et al.," An introduction to kernel-based learning algorithms, " IEEE Trans. on NN., vol. 12, pp. 181-201, 2001.