World Applied Sciences Journal 33 (8): 1355-1361, 2015
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2015.33.08.244

Agile Text Detection for Active Real World Visual Representation

S. Pavithra, A. Sathiyavani and A. Rengarajan

Department of Computer Science and Engineering, Veltech Multitech Engineering College, India

Abstract: In recent years, vision system of Text detection in natural scene images is an important requirement for much content-based retrieval in image analysis tasks. Most of the existing methods have only focused on detecting horizontal texts. In this paper, we define a system which detects text of arbitrary orientation in natural scene images and also an effective pruning algorithm is designed to extract as character candidates using the strategy of minimizing regularized variations for better estimation of our algorithm compare it with other algorithms, we produce a new data-set, which includes different texts in real-world scenarios. The proposed experimental result shows that our algorithm produces better and enhanced performance on texts of arbitrary orientations in complex natural scenes.

Key words: Candidate Construction · Classification · MSER · Pruning

INTRODUCTION

The great success of smart phones and large demands in content-based image search/understanding has made text detection a crucial task in human computer interaction. It is desirable to build practical systems that are robust and fast enough to deal with natural scenes of various conditions; as shown in Fig. 1, we want to detect texts of large variations in language, font, color, scale and orientation in complex scenes. Although text detection has been studied extensively in the past [5, 6], the problem remains unsolved. The difficulties mainly come from two aspects: (1) the diversity of the texts and (2) the complexity of the backgrounds. On one hand, text is a high level concept but better defined than the generic objects [7]; on the other hand, repeated patterns (such s windows and barriers) and random clutters (such as grasses and leaves) may be similar to texts and thus lead to potential false positives.

A number of algorithms to extract caption texts from still images and video have been published in recent years [1, 4, 6]. These methods utilize the following properties of text:

- Characters are bounded in size;
- A text line always contains a cluster of characters which are aligned horizontally; and
- Texts usually has a good contrast from the background;

In recent years, visual search systems have been developed for applications such as product recognition [8, 9] and landmark recognition [10]. In these systems, local image features [11, 12] are extracted from images taken with a camera-phone and are matched to a large database using visual word indexing techniques [13, 14]. Although current visual search technologies have reached a certain level of maturity, they have largely ignored a class of informative features often observed in images: text. In fact, text is particularly interesting because it provides contextual clues for the object appearing inside an image. Given the vast number of text-based search engines, retrieving an image using the embedded text offers an efficient supplement to the visual search systems. As an essential prerequisite for text-based image search, text within images has to be robustly located. However, this is a challenging task due to the 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 numerous methods are reported in the literature [15, 16, 17, 18, 19]. These techniques can be classified into two categories: texture-based and connected component (CC)-based. Although widely used in the community, the ICDAR dataset [2, 3, 5, 20, 21] has two major drawbacks. First, most of the text lines (or single characters) in the ICDAR dataset are horizontal. In real scenarios, however, text may appear in any orientation. The second drawback is that all the text

Corresponding Author: S. Pavithra, Department of Computer Science and Engineering, Veltech Multitech Engineering College, India.

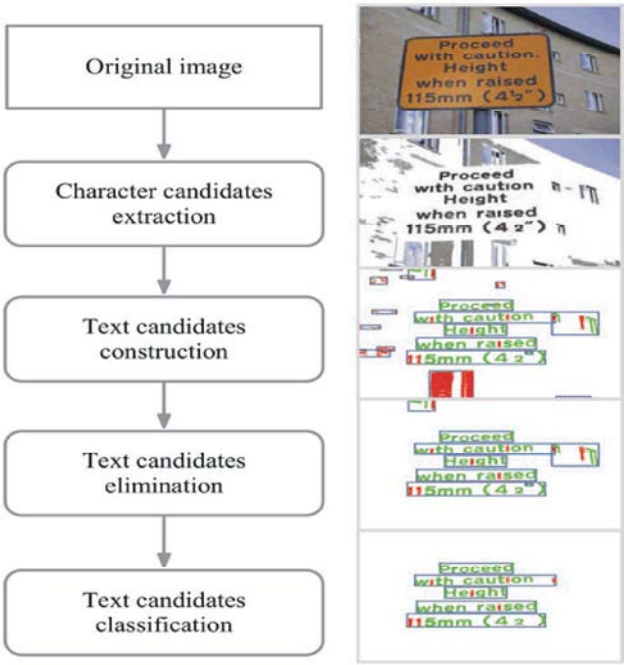

Fig. 1: Proposed architecture

lines or characters in this dataset are in English. These two shortcomings are also pointed out in [22, 23]. In this work, we generate a new multilingual image dataset with horizontal as well as skewed and slant texts. We name this dataset. Ground truth generation and overlap ratio calculation. (a) Human annotations. The annotators are required to bound each text line using a four-vertex polygon (red dots and yellow lines). (b) Ground truth rectangles (green). The ground truth rectangle is generated automatically by fitting a minimum area rectangle using the polygon. (c) Calculation of overlap ratio between detection rectangle and ground truth rectangle. MSRA Text Detection 500 Database (MSRA-TD500)2, because it contains 500 natural images in total. These images are taken from indoor (office and mall) and outdoor (street) scenes using a packet camera. The indoor images are mainly signs, doorplates and caution plates while the outdoor images are mostly guide boards and billboards in complex background. The resolutions of the images vary from 1296 × 864 to 1920 × 1280. Some typical images from this dataset are shown in Fig. 8 (a). This dataset is very challenging because of both the diversity of the texts and the complexity of the backgrounds in the images. The texts may be in different languages (Chinese, English or mixture of both), fonts, sizes, colors and orientations. The backgrounds may contain vegetation (e.g. trees and grasses) and repeated patterns (e.g. windows and bricks), which are not so distinguishable from text.

Related Work: Existing approaches to text detection can be roughly divided into three categories: texture-based methods, regionbased methods and hybrid methods. Texture-based methods [16, 7] treat texts as a special type of texture and make use of their properties, such as local intensities, filter responses and wavelet coefficients. These methods are computation demanding as all locations and scales are exhaustively scanned. Moreover, these algorithms mostly detect horizontal texts. Region-based methods [14, 22] first extract candidate text regions through edge detection or clustering and then eliminate non-text regions using various heuristic rules. The third category, hybridmethods [24], is a mixture of texture-based and region-based methods. Most existing algorithms, e.g. [24, 7], have focused on detecting horizontal texts. In this paper, we address the problem of detecting texts of large variations in natural images, which has great practical importance but has not been well studied. In [25], methods that can detect text strings of arbitrary directions are proposed but they have a large set of rules and parameters; how general and applicable they are in dealing with scenes of large variation is unclear. The dataset is divided into two parts: training set and test set. The training set contains 300 images randomly selected from the original dataset and the rest 200 images constitute the test set. All the images in this dataset are fully annotated. The basic unit in this dataset is text line rather than word, which is used in the ICDAR dataset.

MATERIALS AND METHODS

In this section, we present the details of the proposed algorithm. Specifically, the pipeline of the algorithm will be presented in Sec. 2 and the details of the features will be described in Fig. 1 shows the proposed system and results after each step of the sample. 1) Text candidates are labelled by blue bounding rectangles; character candidates identified as characters are colored green and others red. 2) Text candidates construction. Distance weights and clustering threshold are learned simultaneously using the proposed metric learning algorithm; character candidates are clustered into text candidates by the single-link clustering algorithm using the learned parameters. 3) Text candidates elimination. The posterior probabilities of text candidates corresponding to non-texts are estimated using the character classifier and text candidates with high non-text probabilities are removed.4) Text candidates classification. Text candidates corresponding to true texts are identified by the text classifier. An AdaBoost classifier is trained to decide whether a text candidate corresponding to the true text or not.

Algorithm Chain Process:

Step 1: Initially user will upload the photos from which text has to be extracted.
Step2: Character candidates are extracted by OCR engine.
Step 3: Extracted character candidates are then clustered into text candidates.
Step 4: Then non-text s are estimated and eliminated.
Step 5: Text candidates corresponding to true texts are identified by the text classifier.
Step 6: Results.

MSER Processing: The MSER (Maximally stable extremal region) extraction implements the following steps:

- Sweep threshold of intensity from black to white, performing a simple luminance thresholding of the image
- Extract connected components ("Extremal Regions")
- Find a threshold when an extremal region is "Maximally Stable", i.e. local minimum of the relative growth of its square.

Due to the discrete nature of the image, the region below / above may be coincident with the actual region, in which case the region is still deemed maximal. Approximate a region with an ellipse (this step is optional). Keep those regions descriptors as features. However, even if an extremal region is maximally stable, it might.

Pruning Algorithm: Repeating components is the major pitfall when the MSER algorithm is applied as a character segmentation algorithm. Considering the MSERs tree presented in Fig. 2(a), this figure shows that fourteen MSERs are detected for the word "PACT" but only four of them are really interested to us. The hierarchical structure of MSERs is quite useful for designing a pruning algorithm. As characters cannot "contain" or be "contained" by other characters in real world, it is safe to remove children once the parent is known to be a character and vice versa. If the MSERs tree is pruned by applying this kind of parent-children elimination operation recursively, we are still "safe" and all characters are preserved after the elimination. As an example, Fig. 2(e) shows that a set of disconnected nodes containing all the desired characters can be extracted by applying this algorithm to the MSERs tree in Fig. 2. However, it can be computationally expensive to identify characters, which usually entails the computations of complex features. Our proposed system for detecting text from images is a web application. So anyone can upload the photos and detect text in those images.

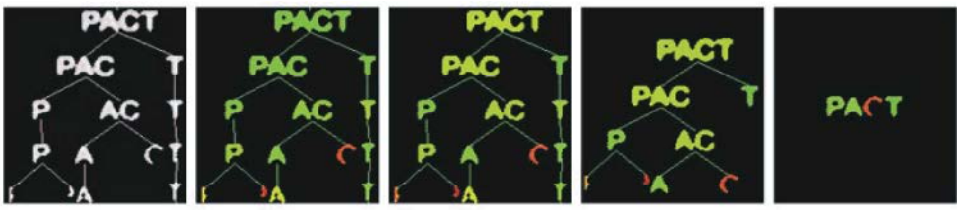

Fig. 2: The process of MSERs pruning. (a) MSERs tree of a text segment. (b) MSERs colored according to variations, as variations increase, MSERs are colored from green to yellow then to red. (c) MSERs colored according to regularized variations. (d) MSERs tree after linear reduction. (e) Character candidates after tree accumulation

Fig. 3: Text detection examples on the ICDAR 2011dataset. Detected text by our system are labeled using red rectangles. Notice the robustness against low contrast, complex background and font variations.

RESULTS

For evaluation we used a similar approach and the same quality measures as in the evaluation scheme of ICDAR 2013 competition. The following quality measures are used: precision, recall and f measure. They are defined as following:

$$\text{Recal l} = \frac{\sum_{i=1}^{|G|} match_G(G_i)}{|G|}$$

$$\text{PRECISION} = \frac{\sum_{i=1}^{|D|} match_D(D_i)}{|D|}$$

$$\text{F} = 2\,\frac{recall.precision}{recall + precision}$$

$$\text{Match}_G(G_i) = \max_{j=1..|D|} \frac{2.area(G_i \cap D_j)}{area(G_i)+area(D_j)} \quad (1)$$

$$\text{Match}_D(D_j) = \max_{i=1..|G|} \frac{2.area(D_j \cap G_i)}{area(D_j)+area(G_i)} \quad (2)$$

In [25], a dataset called Oriented Scene Text Database (OSTD), which contains texts of various orientations, is released.

This dataset contains 89 images of logos, indoor scenes and street views. We perform text detection on all the images in this dataset. The quantitative results are presented in Table 4. Our method outperforms [25] on the Oriented Scene Text Database (OSTD), with an improvement of 0.20 in F-measure.

Table 1: Performance (%) comparison of text detection algorithms on ICDAR 2011 Robust Reading Competition dataset.

Methods	Recall	Precision	F-measure
Our Method	70.26	88.29	78.22
Yi's Method	58.09	67.22	62.32
Neumann's Method	52.54	68.93	59.63

Table 2: Performances of different text detection methods evaluated on the ICDAR test set

Algorithm	Precision	Recall	F-measure
TD-ICDAR	0.68	0.66	0.66
Epshtein *et al.* [3]	0.73	0.60	0.66
Yi *et al.* [25]	0.71	0.62	0.62
Becker *et al.* [20]	0.62	0.67	0.62
Chen *et al.* [19]	0.60	0.60	0.58

Table 3: Performances of different text detection methods evaluated on the Oriented Scene Text Database (OSTD)

Algorithm	Precision	Recall	F-measure
TD-Mixture	0.77	0.73	0.74
TD-ICDAR	0.71	0.69	0.68
Yi *et al.* [25]	0.56	0.64	0.55
Epshtein *et al.* [3]	0.37	0.32	0.32
Chen *et al.* [19]	0.07	0.06	0.06

Fig. 4: Upload a image

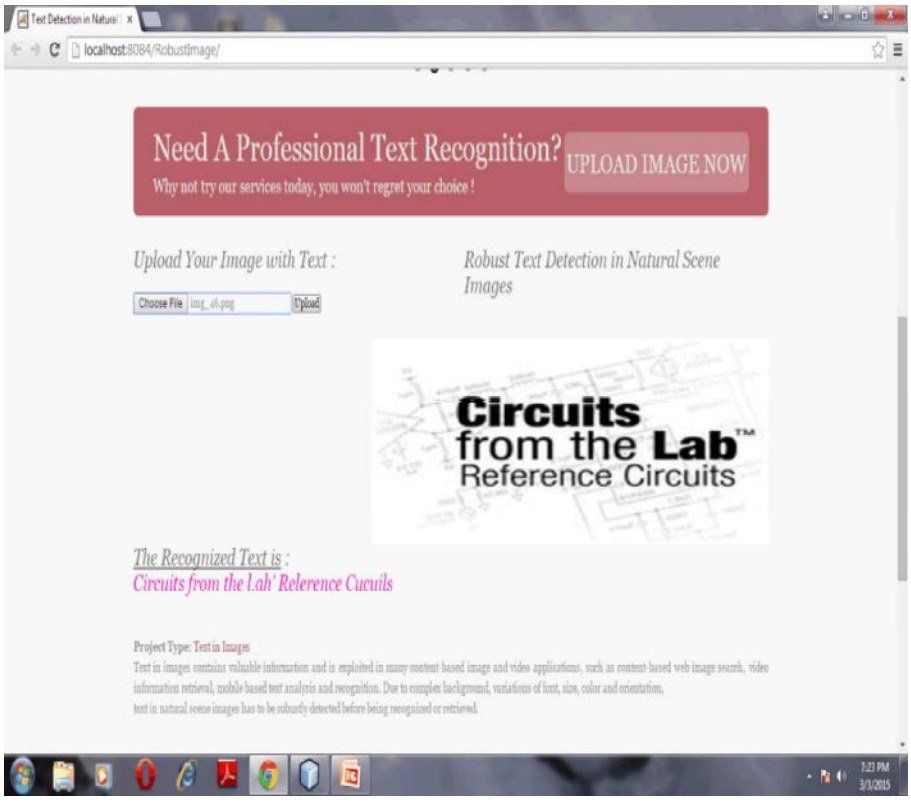

Fig. 5: Robust text detection in natural images

CONCLUSIONS

This paper presents a new novel text detection and pruning algorithm is designed . We propose a fast and Active MSERs pruning algorithm that enables us to detect most the candidate characters even when the image is in low quality, Our system significantly enhanced performance on texts in arbitrary orientations in complex real world scenes. The results show that the use of information can greatly enhance the quality of text clustering and classification, while maintaining a high level of effectiveness. we make a powerful text detection system that exhibited superior performance over state-of-the-art methods on both the ICDAR 2011 and a OSTD data set. For the further work it is planned to implement a complete algorithm that solves the problem of text detection and character recognition irrespective of a language.

REFERENCE

1. Zhong, Y., H. Zhang and A. Jain, 2000. "Automatic caption localization in compressed video," *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(4): 385-392.
2. Karatzas D. *et al*., 2013. "ICDAR 2013 robust reading competition," in *Proc. ICDAR*, Washington, DC, USA, 2013, pp: 1115-1124.
3. Epshtein, B., E. Ofek and Y. Wexler, 2010. Detecting text in natural sceneswith stroke width transform. In *Proc. CVPR*.
4. Weinman, J., E. Learned-Miller and A. Hanson, 2009. "Scene text recognition using similarity and a lexicon with sparse belief propagation,"*IEEE Trans. Pattern Anal. Mach. Intell.*, 31(10): 1733-1746.
5. Yin, X.C., H.W. Hao, J. Sun and S. Naoi, 2011. "Robust vanishing point detection for mobile cam-based documents," in *Proc. ICDAR*, Beijing, China, 2011, pp: 136-140.
5. Liang, J., D. Doermann and H. Li., 2005. "Camera-based analysis of text and documents: a survey". *IJDAR*.
6. Jung, K., K. Kim and A. Jain. 2004. "Text information extraction in images and video: a survey." *PR*, 2004.
7. Everingham, M., L.V. Gool, C.K.I. Williams, J. Winn and A. Zisserman 2010. The pascal visual object classes (voc) challenge. *IJCV*.
8. Tsai, S.S. D. Chen, V. Chandrasekhar, G. Takacs, N.M. Cheung, R. Vedantham, R. Grzeszczuk and B. Girod, 2010. "Mobile product recognition," in Proc. ACM, Multimedia.
9. Chen, D., S.S. Tsai, C.H. Hsu, K. Kim, J.P. Singh and B. Girod, 2010. "Building book inventories using smartphones," in Proc. ACM Multimedia.
10. Takacs, G., Y. Xiong, R. Grzeszczuk, V. Chandrasekhar, W. Chen, L. Pulli N. Gelfand, T. Bismpigiannis and B. Girod, 2008. "Outdoors augmented reality on mobile phone using loxel-based visual feature organization," in Proc. ACM Multimedia Information Retrieval, pp: 427-434.
11. Bay, H., A. Ess, T. Tuytelaars and L. Van Gool, 2008. "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, 110(3): 346-359.
12. Chandrasekhar, V., G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk and B. Girod, 2009. "CHoG: Compressed histogram of gradients. a low bit-rate feature descriptor," in *CVPR*, pp: 2504-2511.
13. Nist´er D. and H. Stew´enius, 2006. "Scalable recognition with a vocabulary tree," in *CVPR*, pp: 2161-2168.
14. Chen, D.M., S.S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk and B. Girod, 2010. "Inverted Index Compression for Scalable Image Matching," in *Proc. of IEEE Data Compression Conference (DCC)*, Snowbird, Utah.
15. Liang, J., D. Doermann and H.P. Li, 2005. "Camera-based analysis of text and documents: a survey," *IJDAR*, 7(2-3): 84-104.
16. Jung, K., K.I. Kim and A.K. Jain, 2004. "Text information extraction in images and video: a survey," *Pattern Recognition*, 37(5): 977-997.
17. Zhong, Y., H. Zhang and A.K. Jain, 2000. "Automatic caption localization in compressed video," IEEE Trans. Pattern Anal. Mach. Intell., 22(4): 385-392.
18. Ye, Q., Q. Huang, W. Gao and D. Zhao, 2005. "Fast and robust text detection in images and video frames," *Image Vision Comput.*, 23: 565-576.
19. Chen X. and A.L. Yuille, 2004. "Detecting and reading text in natural scenes," in *CVPR*, 2: II–366–II–373.
20. Lucas, S.M. Icdar, 2005. text locating competition results. In *Proc. ICDAR*.
21. Lucas, S.M., A. Panaretos, L. Sosa, A. Tang, S. Wong and R. Young, 2003. Icdar 2003 robust reading competitions. In *Proc. ICDAR*.
22. Srivastav A. and J. Kumar, 2008. "Text detection in scene images using stroke width and nearest-neighbor constraints," in *TENCON 2008-2008 IEEE Region 10 Conference*, pp: 1-5.
23. Tsai, S.S., H. Chen, D.M. Chen, G. Schroth, R. Grzeszczuk and B. Girod, 2011. "Mobile visual search on papers using text and low bit-rate features," in ICIP.

24. Pan, Y., X. Hou and C. Liu, 2011. A hybrid approach to detect and localize texts in natural scene images. IEEE Trans. IP.

25. Yi, C. and Y. Tian, 2011. Text string detection from natural scenes by structure-based partition and grouping. IEEE Trans. IP.