An improved edge profile based method for text detection in images of natural scenes

Andrej Ikica, Peter Peer
Computer Vision Laboratory
Faculty of Computer and Information Science
Tržaška 25, Ljubljana, Slovenia
{andrej.ikica, peter.peer}@fri.uni-lj.si

***Abstract*—Text detection in natural images has gained much attention in the last years as it is a primary step towards fully autonomous text recognition. Understanding the visual text content is of a vital importance in many applicative areas from the internet search engines to the PDA signboard translators. Images of natural scenes, however, pose numerous difficulties compared to the traditional scanned documents. They mainly contain diverse complex text of different sizes, styles and colors with complex backgrounds. Furthermore, such images are captured under variable lighting conditions and are often affected by the skew distortion and perspective projections. In this article an improved edge profile based text detection method is presented. It uses a set of heuristic rules to eliminate detection of non-text areas. The method is evaluated on CVL OCR DB, an annotated image database of text in natural scenes.**

Keywords-text detection; natural scenes; edge profiles; computer vision

I. INTRODUCTION

Text detection is a preliminary step in automatic text recognition. It needs to be fast, efficient and robust in order to feed an OCR classifier with the correct input. In other words, segmented regions must correspond to the actual text. Due to its enormous applicative potential - from querying text in visual content to signboard translation using PDA – text detection in natural scenes has been given a lot of attention over the last decade and even earlier. As opposed to the traditional documents with black ink on a white paper, images of natural scenes introduce difficulties that are far from being completely solved, such as complex backgrounds, uneven illumination, complex symbols that can be easily misclassified as text, perspective projections and complex text styles (Fig. 1).

Several text detection methods have been proposed based on edge detection, binarization, spatial-frequency image analysis and mathematical morphology [1]. Generally text detection methods can be classified as either edge-based, connected-component based and texture-based methods [2]. According to [1] the best results were achieved using edge-based text detection. It obtained top overall performance among 4 methods including mathematical morphology and color-based character extraction. Edge-based text detection has also been used in combination with edge profiles. Park et al. [2, 3] use them for automatic detection and recognition of Korean text in outdoor signboard images. However, they assume that a single text sign is located around the center line of the image. Edge profiles have also been used for detecting text in video data. Shivakumara et al. [4, 5] use edge profiles in combination with additional edge features to eliminate false positives selection.

Due to their simplicity and efficiency edge profiles are used in the proposed implementation as well. To eliminate some of the non-text candidate regions, edge profiles are combined with a set of heuristic rules. The proposed method is evaluated on CVL OCR DB, an annotated image database of text in natural scenes. It is assumed in this article that multiple text regions of different text styles and sizes can appear in the image.

The article is organized as follows. Edge profile based text detection is described in Section 2. In Section 3 the proposed text detection method is presented. Section 4 contains a brief overview of CVL OCR DB and evaluation results. Finally, article is concluded in Section 5.

II. EDGE PROFILE BASED TEXT DETECTION

Horizontal edge profile (HP) of an edge-map (EM), where *w* and *h* correspond to the image width and height respectively, is an *h*-dimensional vector. A certain HP component corresponds to the total number of edge pixels in a corresponding edge-map row. Thus *j*-th HP component is defined as:

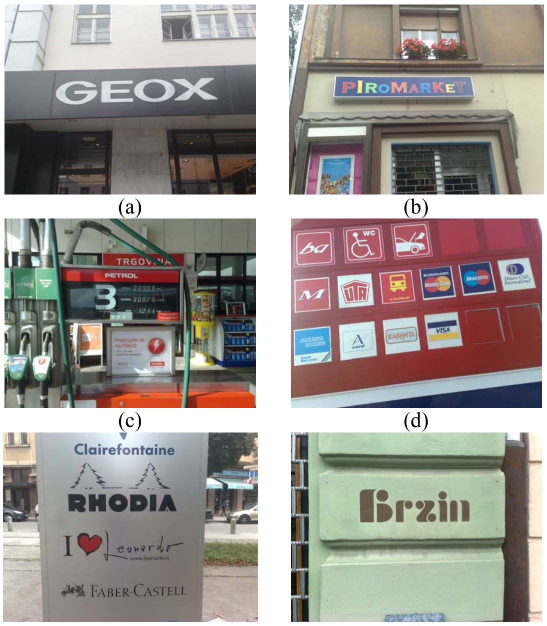
$$HP(j)=\sum_{i=1}^{w} EM(i,j)\,. \tag{1}$$

Similarly vertical edge profile (VP) of an edge-map is a *w*-dimensional vector where a certain VP component corresponds to the total number of edge pixels in a corresponding edge-map column. In the same way as in HP, *i*-th VP component is defined as:

$$VP(i)=\sum_{j=1}^{h} EM(i,j)\,. \tag{2}$$

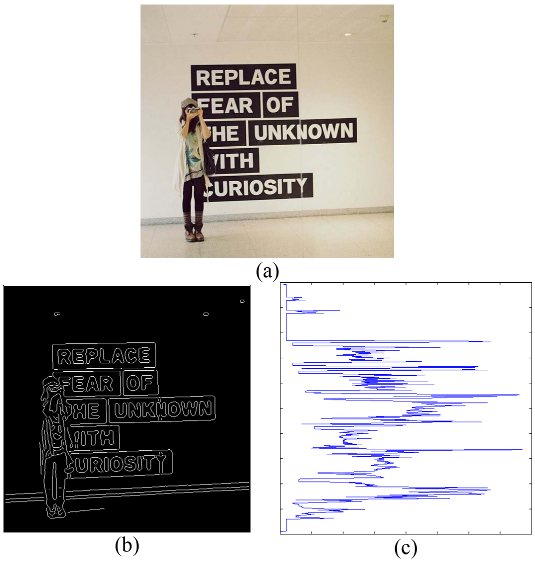
HP-based text detection assumes that the text is always aligned horizontally. Otherwise skew detection and correction must be applied. Candidate text regions correspond to the peaks in the HP (Fig. 2c). Further, vertical profile is computed on the extracted candidate text regions. By detecting peaks in the VP single characters or clusters of connected characters are obtained.

Edge profile based text detection is simple, fast and efficient. However, due to its edge-orientated behavior it lacks of being enough text-aware. Therefore non-text regions are often extracted as well. To avoid this drawback, we propose some heuristic rules to eliminate false positives detection.



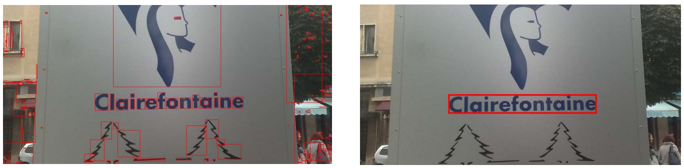
(a) (b) (c) (d) (e) (f)

Figure 1. (a) A typical image of text in natural scenes, (b) different character colors, (c) very complex background, (d) skewed image, (e) multiple text styles and (f) complex text style.



(a) (b) (c)

Figure 2. (a) Original image, (b) detected edges and (c) horizontal edge profile.



(a) (b)

Figure 3. (a) False positives and (b) final candidate text region.

III. TEXT DETECTION METHOD

After obtaining the Canny edge-map, connected component labeling is performed and edge pixels are grouped into connected components (CCs) ie. candidate text regions. All regions that are smaller than 5 pixels are ignored in this step. Since images of text in natural scenes contain very complex backgrounds, most of the candidate text regions correspond to the non-text areas as shown in Fig. 3a.

Many heuristic rules have been proposed to solve this problem. However, most of them rely on region sizes and aspect ratios controlled by the constant values such as region size thresholds. These heuristic rules therefore often eliminate the true text regions and instead leave the non-text regions intact. Since the goal is to be able to detect multiple texts of different sizes and since heuristic thresholds are hard to be tuned perfectly, a slightly modified strategy is proposed.

First, the horizontal edge profile is computed. HP represents a useful clue where the text might globally appear in the image. Since the text is assumed to be aligned horizontally and characters expose typical vertical characteristics, HP is computed on the vertical edge-map. Peak regions in HP more or less correspond to the true vertical text boundaries (VTB). They, however, often cut out the parts of the characters that extend over or below the vertical boundary. Such an example is letter ‘j’ or the capital letter which appears at the beginning of the word. In order not to eliminate these parts, original connected components that overlap with a particular VTB are used, not only the parts inside the VTB.

Next, the following heuristic rules are applied. All the connected components that do not overlap with any of the VTBs are eliminated. Since the text characters always appear close and are more or less of the same height, all the overlapping regions that extend far above or below a particular VTB violate this assumption and are thus eliminated. Similarly the overlapping regions that are too small compared to the VTB height are eliminated as well. VTB based heuristic rules are illustrated in the Fig. 4.

Finally, the heuristic rules similar to [1] are used to further eliminate non-text candidate regions, such as the aspect ratio rule and closeness rule. At the end all the candidate text regions that appear close enough are merged into a single area. This facilitates scenarios, where a certain character is eliminated from the word. Merging stage reconstructs the total word area.

The text detection algorithm is presented in the Fig. 5.

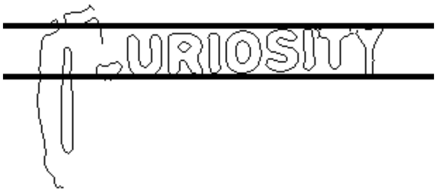


Figure 4. The edge structure on the left extends far above and below the overlapped VTB and is therefore ignored. The small edge structure on the left of the letter "U" is too small compared to the VTB height and is therefore ignored as well. All other characters are left intact.

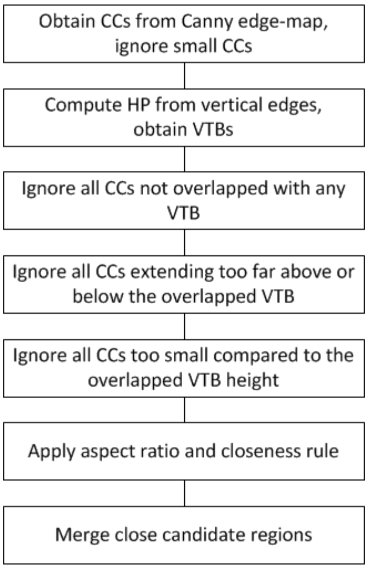

Obtain CCs from Canny edge-map, ignore small CCs
Compute HP from vertical edges, obtain VTBs
Ignore all CCs not overlapped with any VTB
Ignore all CCs extending too far above or below the overlapped VTB
Ignore all CCs too small compared to the overlapped VTB height
Apply aspect ratio and closeness rule
Merge close candidate regions

Figure 5. The text detection algorithm. See Chapter III for abbreviation explanations.

IV. RESULTS

A. *CVL OCR DB*

CVL OCR DB is a collection of images of text in natural scenes collected at the Computer Vision Laboratory at the Faculty of Computer and Information Science, University of Ljubljana. It includes images of various text scenes such as signboards, traffic signs, shop names etc. In order to provide a wide range of real life scenarios, images are captured with different compact digital cameras and mobile phones at different angles, positions and under variable lighting and weather conditions. The assembly of the CVL OCR DB is currently in its starting stage and contains 341 images, annotated with ground truth information. In the near future CVL OCR DB will be expanded with thousands of new, already captured images, and additional categories. The database will be publicly available to the research community.

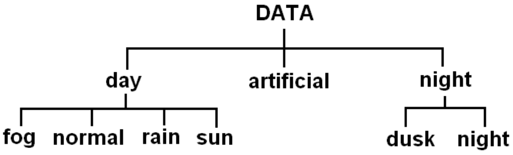


Figure 6. Organization of the CVL OCR DB.

CVL OCR dataset is organized into three main categories according to the capture time - *day*, *night* and *artificial*. The last category corresponds to the images captured under artificial lighting, eg. in shopping centers. *Day* and *night* are further divided into subcategories *fog*, *night*, *rain* & *sun* and *dusk* & *night*, respectively (Fig. 6).

All the image data is physically stored in a directory structure, where each category/subcategory corresponds to a particular directory/subdirectory. Each image is equipped with an XML file containing the ground truth data annotated by the human annotator. Ground truth data includes bounding rectangles of the actual text regions as well as the locations and UTF-8 codes of the individual characters appearing in the image. Thus, CVL OCR DB can be used to test both text detection and recognition methods. In order to train recognition classifier, cropped images of individual characters are stored in subdirectories and can be immediately used – no cropping is requested. To easily navigate the directory structure, all files are given a specific name, which uniquely identifies them. For instance a 110th file appearing in the *day/normal* category is marked as '*d_n_00110_full.jpg*' and it's corresponding XML file as '*d_n_00110.xml*'. Single characters are stored in '*d_n_00110*' subdirectory, and for instance an image of the character 'a' is marked as '*d_n_00110_char_a_lc_001.jpg*'.

B. *Evaluation results*

To measure the accuracy of the proposed method on the single image, two terms are used: precision and recall. Both terms were used in [1] and are similar to those in ICDAR 2003 competition [7]. According to [1] precision *p* is defined as the number of correct estimates *C* divided by the total number of estimates *E*:

$$p = \frac{C}{E}. \tag{3}$$

Recall *r* is defined as the number of correct estimates *C* divided by the total number of targets *T*:

$$r = \frac{C}{T}. \tag{4}$$

In other words, *T* corresponds to the total area of all the annotated bounding rectangles in the image, *E* corresponds to the total area of all the text regions detected in the image and *C* corresponds to the total area of the regions correctly detected ie. intersection between *E* and *T*. Both precision and recall are combined in the quality measure *f*:

$$f = \frac{1}{\alpha / p + (1-\alpha) / r} . \qquad (5)$$

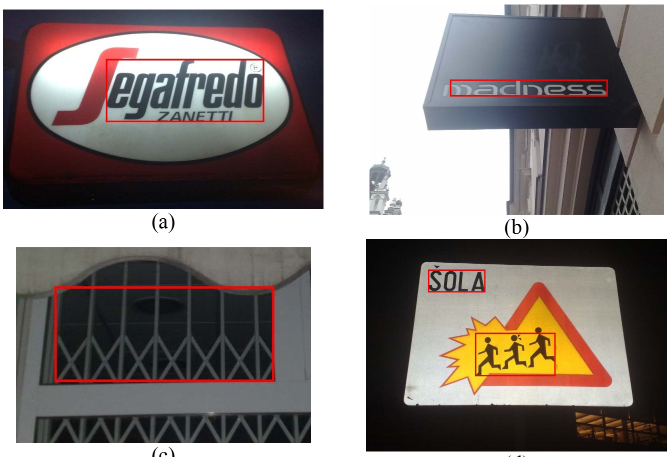
The parameter α is set to 0.5 in order to give equal weights to the precision and recall.

For evaluation purposes the CVL OCR DB was used. The proposed method, as well as two other already mentioned text detection methods, to which we will refer to as Ezaki [1] and Shivakumara [4], were tested on 341 images of the dataset. Since none of the methods uses machine learning, the training stage was skipped and therefore all 341 images were actually used for evaluation. Furthermore, none of the 341 images were used in the algorithm development stage. Results in terms of average precision p, average recall r, and average quality measure f of all 341 images are shown in Table 1.

TABLE I. EVALUATION RESULTS FOR ALL THREE TEXT DETECTION METHODS

Text detection method	***p***	***r***	***f***
Shivakumara	58,7%	51,0%	54,6%
Ezaki	58,0%	53,1%	55,4%
Proposed method	70,9%	55,2%	62,1%

It is clear that the proposed method gained better results compared to the other two methods. Some of the text detection examples are shown in Fig. 7. By using a set of heuristic rules the proposed method achieves better precision. The proposed method gained approximately the same recall as the other two methods. This is due to the fact that the other two methods often find large connected regions (for example a complete signboard) due to the long edges and are not able to eliminate them. Since the annotated text region is always included in the large connected region, both methods often gain recall of 1. However, a human annotator would mark such detected area as a false positive.



(a)

(b)

(c)

(d)

Figure 7. (a) An example of detected text region, (b) left side of the text region ignored, (c) grid can be easily mistaken for ‘x’ characters, and (d) humans detected as text.

V. CONCLUSION

Edge profiles have proven to be an efficient preprocessing tool for the candidate text region extraction. Due to their simplicity and speed they are suitable for the preliminary segmentation tasks. This article shows that simple heuristic rules can improve the text detection performance. Results, however, indicate that the text detection task is far from being solved, especially if it is used in a commercial system. Therefore in the near future the method will be improved in terms of accuracy. Several possible improvements will be investigated. One of them is to eliminate non-text areas by using Gabor filtering. The other improvement is proposed in [6] and is based on analyzing candidate text region histograms and Fisher’s Discriminant Rate (FDR). Finally, text detection based on approximate - and therefore fast - OCR classifiers will be investigated as well. Simultaneously CVL OCR DB will be constantly expanded with the new images and new categories.

In this article the CVL OCR DB was used for the evaluation purposes. In the future research work the methods will also be evaluated on similar datasets such as ICDAR 2003 [7].

ACKNOWLEDGMENT

This research is supported by the Public Agency for Technology of the Republic of Slovenia (TIA) – operation partly financed by the European Union, European Social Fund.

REFERENCES

[1] N. Ezaki, M. Bulacu, L. Schomaker, “Text Detection from Natural Scene Images: Towards a System for Visually Impaired Persons”, *Int. Conf. on Pattern Recognition* (ICPR 2004), vol. II, pp. 683-686.

[2] J. Park, G. Lee, E. Kim, J. Lim, S. Kim, H. Yang, M. Lee, S. Hwang, “Automatic detection and recognition of Korean text in outdoor signboard images”, *Pattern Recognition Letters*, 2010.

[3] T. N. Dinh, J. Park, G. Lee, “Korean Text Detection and Binarization in Color Signboards”, *Proc. of The Seventh Int. Conf. on Advanced Language Processing and Web Information Technology* (ALPIT 2008), pp. 235-240.

[4] P. Shivakumara, W. Huang, C. L. Tan, “Efficient Video Text Detection using Edge Features”, *Int. Conf. on Pattern Recognition* (ICPR 2008), pp. 1-4.

[5] P. Shivakumara, T. Q. Phan, C. L. Tan, “Video text detection based on filters and edge features”, *Int. Conf. on Multimedia & Expo* (ICME 2009), pp. 514-517.

[6] N. Ezaki, K. Kiyota, B. T. Minh, M. Bulacu, L. Schomaker, “Improved Text-Detection Methods for a Camera-based Text Reading System for Blind Persons”, *Int. Conf. on Document Analysis and Recognition* (ICDAR 2005), pp. 257-261.

[7] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, “ICDAR 2003 Robust Reading Competitions”, *Proc. of the ICDAR 2003*, pp. 682-687.