Scene text segmentation based on redundant representation of character candidates

Matko Saric
University of Split
Rudjera Boskovica 32
21000, Split, Croatia
msaric@fesb.hr

ABSTRACT

Text segmentation is important step in extraction of textual information from natural scene images. This paper proposes novel method for generation of character candidate regions based on redundant representation of subpaths in extremal regions (ER) tree. These subpaths are constructed using area variation and pruned using their length: each sufficiently long subpath is character candidate which is represented by subset of regions contained in the subpath. Mean SVM probability score of regions in subset is used to filter out non character components. Proposed approach for character candidates generation is followed by character grouping and restoration steps. Experimental results obtained on the ICDAR 2013 dataset shows that the proposed text segmentation method obtains second highest precision and competitive recall rate.

Keywords
Scene text segmentation, extremal regions, SVM classification

1 INTRODUCTION

Extraction of textual information from scene images is challenging problem that is essential for range of different applications like text translation, content based image indexing, reading aid systems for visually impaired people etc. In contrast to the text in document images, recognition of natural scene text is more complex task. This is primarily caused by imaging conditions (perspective distortion, uneven illumination, shadows, blur, sensor noise) and specific properties of scene text (complex background, variable color, font line orientation). Extraction of textual information from scene images consists of three steps:localization, segmentation and recognition [Kar15]. Text localization methods can be classified into 3 groups: sliding-window based methods, methods based on connected components (CCs) and hybrid methods. Methods in first group [Jad14], [Wan12] use window that moves across the image and detect characters using local features. This approach is robust to background noise, but main drawback is computational complexity caused by high number of patches needed to cover text with different scales, aspect ratios and orientations. These methods also don't segment characters from background. Techniques in second group assume that characters are connected components. In this way text segmentation is also provided what can be exploited in recognition step. Maximally Stable Extremal Regions (MSER) region detector is widely used in literature achieving promising performance. Its main drawback is high number of repeating components that affects character grouping algorithm. In [Yin14] character candidates are extracted using MSER segmentation. Pruning of non-character CCs is performed by minimizing regularized variations and exploiting parent-children relations in MSER tree. Character candidates grouping is performed using single-link clustering algorithm with trained distance function. Neumann and Matas [Neu15] proposed extremal regions (ERs) based method for real time localization and recognition. Probability of each ER being a character is determined through two stage classification. First stage employs Adaboost classifier with incrementally computed descriptors, while in second stage remaining candidates are classified using SVM and more complex features. ERs are grouped into lines with highly efficient clustering algorithm followed by OCR module. Sung et al. [Sun15] presented text detection algorithm where firstly ER tree subpaths are constructed using similarity between parent and child regions. Each sufficiently long subpath correspond to characters. Finally, region with lowest normalized variation in subpath is chosen as character candidate. Hybrid text localization methods combines sliding-window and connected components based approaches. Huang et al. [Hua14] introduced text detection method combining MSER and convolutional neural network (CNN) classifier. MSER segmentation reduces the number of search windows, while sliding-window CNN classifier handles characters merged or missed by MSER. Lu et al. [Lu15] propose scene text extraction method where three text-specific edge features are introduced to detect candidate text boundaries. Character candidates are segmented based on Niblack thresholding of areas found in previous step. Support vector regression (SVR) and bag-of-words (BoW) model are employed to confirm true words after text-line construction. Text segmentation method presented in [Mil15] includes Niblack binarization, calculation of initial labeling confidence using Laplacian of pixel intensities and global optimization. Combined with of-the-shelf OCR software proposed method gives performance comparable to state-of-the-art text reading methods. In this paper novel method for text segmentation is proposed. Main contribution is character candidates generation method based on redundant representation of character candidates. ER tree subpaths are extracted based on area variation. Each sufficiently long subpath is character candidate which is represented by subset of subpath regions. SVM classification is performed on each ER in this subset and mean probability score of the subpath is calculated in order to accept or reject character candidate. In this way redundant representation of character candidates and multiple classifications are exploited to compensate SVM classification errors and obtain more relevant estimation of character probability. Compared to recent work, proposed approach achieves second highest precision and competitive recall for scene text segmentation step.

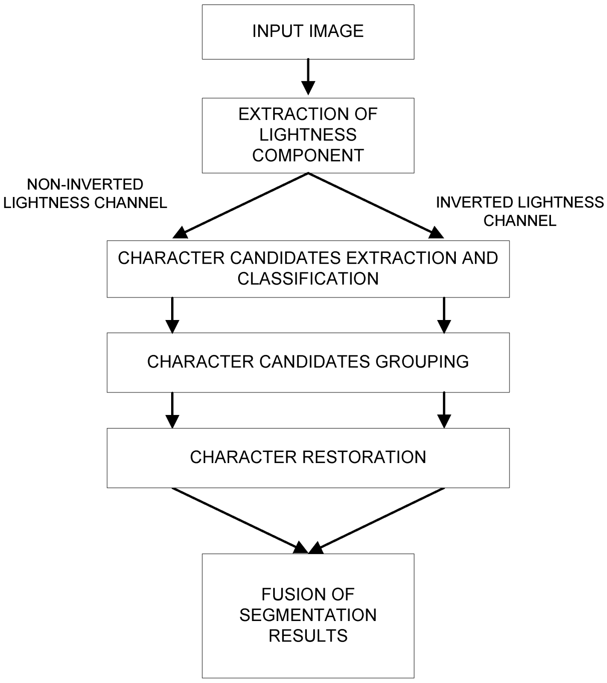
INPUT IMAGE
EXTRACTION OF LIGHTNESS COMPONENT
NON-INVERTED LIGHTNESS CHANNEL
INVERTED LIGHTNESS CHANNEL
CHARACTER CANDIDATES EXTRACTION AND CLASSIFICATION
CHARACTER CANDIDATES GROUPING
CHARACTER RESTORATION
FUSION OF SEGMENTATION RESULTS

Figure 1: Overview of the proposed method

2 PROPOSED METHOD

An illustration of proposed system is showed in Figure 1. First step is extraction of lightness component from the input RGB image. Next step is ER-based character candidates generation. After that classification of character candidate regions is performed using geometrical features thresholding and SVM classifier. This is followed by character grouping and restoration steps which aims to correct errors from classification step.

2.1 Character candidates generation and classification

Drawback of MSER-based character extraction is high number of repeating components. It should also be noted that maximum stability of the region doesn't guarantee correct character extraction. In order to address these problems a novel method for character candidates selection is proposed in this paper. Firstly, ER tree subpaths are extracted based on ER variation values. Each sufficiently long subpath is considered as character candidate. Extremal region [Mat04] is connected component with pixels having lower or higher values than boundary pixels. Binary threshold image $B(p)$ is obtained from gray level image I as

$$B(p) = \begin{cases} 1, & \text{if } I(p) > t \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where t is threshold and p is pixel position in image I. An extremal region R_t at threshold t is defined with

$$\forall p \in R_t, \forall q \in boundary(R_t) \Rightarrow B_t(p) \geq B_t(q) \tag{2}$$

where p and q are pixel positions. The variation of extremal region R_t is defined as

$$\Psi(R_t) = \frac{(|R_{t+\Delta}| - |R_t|)}{|R_t|} \tag{3}$$

where $|R_t|$ denotes number of pixels in region R_t and Δ is a parameter. Relation between extremal regions is usually represented by ER tree. It is component tree where each node represents extremal region. Edge represents inclusion relation between parent and child region. Parent region R_i and child region R_j satisfy next condition:

$$\bigvee p \in R_i \rightarrow p \in R_j \tag{4}$$

By moving to the root node, value of threshold t decreases what results with larger regions (Figure 2). The root is region covering the whole image. In this paper

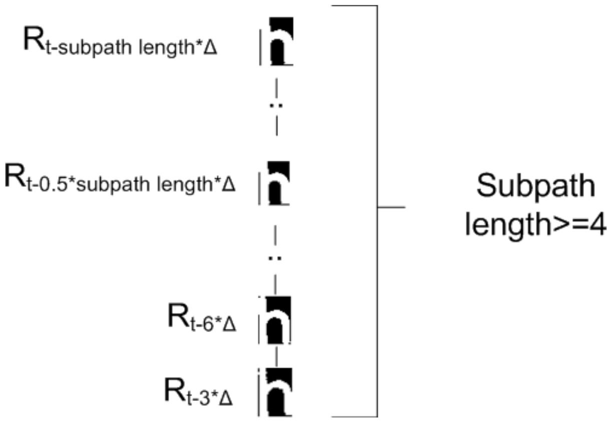

Figure 2: Characters candidate subpath represented by subset of ERs

ER tree subpaths are constructed using variation values. ERs R_t and its child $R_{t+\Delta}$ are split into two subpaths if

$$|\Psi(R_t) - \Psi(R_{t+\Delta})| \geq t_{variation} \tag{5}$$

where $t_{variation}$ is variation threshold that is set to value 0.5 according to [Yin14] and Δ is set to value 4. In this way only subset of regions R_t with $t = k * \Delta, k \in [0, 255/\Delta]$ is considered. Characters correspond to extremal regions with lower variations what is manifested in longer ER subpath.

Therefore, subpath $S = \{R_{t-\Delta}, R_{t-2*\Delta}, ..., R_{t-slen*\Delta}\}$, where $slen$ denotes subpath length, is considered as character candidate if $slen >= 4$. Each subpath is represented with subset of regions: $R_{t-j*\Delta}$, where $j = 3 * n, n \in [1, subpathlength/3]$ (Figure 2). In this way subpath regions are subsampled by factor 3 to lower the number of classifications needed for one subpath. Regions from this subset are further sent to classification step in order to calculate their SVM character probability scores. Final score is calculated as mean of these probabilities:

$$character\ probability = mean(p_{t-j*\Delta}) \tag{6}$$

where $p_{t-j*\Delta}$ represents SVM probability score of region $R_{t-j*\Delta}$. If $character\ probability \geq 0.5$, subpath is accepted as valid character and in the further steps it is represented by region $R_{t-0.5*subpathlength*\Delta}$ which corresponds to the middle of the subpath. This choice is motivated by fact that regions on the subpath bottom tend to be eroded version of character, while top of the subpath often contains dilated components. Proposed approach exploits redundant subpath representation taking mean of multiple probability scores of the same character candidate in order to compensate SVM classifier errors.

Classification into character and non-character ERs is performed in two steps. Obviously non character objects are eliminated by thresholding the values of area, height, aspect ratio and number of holes. These regions are discarded from the processing pipeline and can't be recovered in character restoration step. Further, SVM classifier is employed to distinguish character ERs from non-character ERs. It was trained on training set consisting of 3198 positive and 4558 negative samples. Positive samples are extracted from ICDAR 2011 training [Sha11], KAIST [Kai17] and Char74k [Cam09] datasets, while negative samples are generated from ICDAR 2011 training and BSD 300 [Mar01] datasets followed by extracting non-character CCs. RBF kernel was used with $\gamma = 0.69$. Following features used in [Gon13] and [Neu15] forms input vector for SVM classifier:

Aspect ratio

$$Aspect\ ratio = \frac{width}{height} \tag{7}$$

where *width* and *height* denote width and height of region bounding box.

Occupy rate

$$Occupy\ rate = \frac{area}{height * width} \tag{8}$$

where *area* is region area.

Compactness

$$Compactness(CC) = area/perimeter^2 \tag{9}$$

where *perimeter* is number of contour pixels.

Solidity (*convex area*):

$$Solidity(CC) = \frac{area}{convex\ area} \tag{10}$$

Occupy rate convex area:

$$Occupy\ rate\ convex\ area = \frac{convex\ area}{height * width} \tag{11}$$

Skeleton length to perimeter ratio

$$Skeleton\ length\ to\ perimeter\ ratio = \frac{skeleton\ length}{perimeter} \tag{12}$$

Stroke width features

$$Mean\ stroke\ width\ size\ ratio = \frac{mean(stroke\ width)}{max(height, width)} \tag{13}$$

$$Max\ stroke\ width\ size\ ratio = \frac{max(stroke\ width)}{max(height, width)} \tag{14}$$

$$Stroke\ width\ variance = \frac{\sigma^2_{stroke\ width}}{mean(stroke\ width} \tag{15}$$

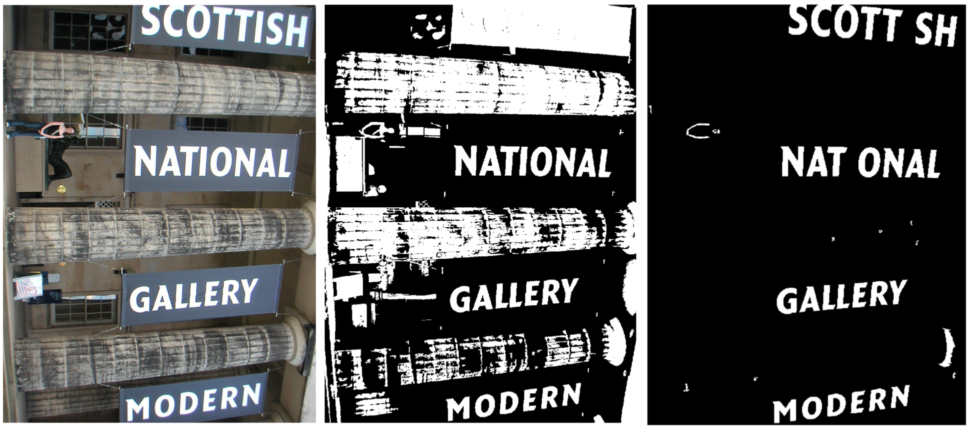

Figure 3: Characters candidates classification: input image (left), generated character candidates (middle) and character candidates after SVM classification (right)

Filled area ratio

$$Filled\ area\ ratio = \frac{abs(Filled\ area - Area)}{Area} \quad (16)$$

where *Filled area* is the number of pixels in region with all holes filled.

Number of inflection points is used as indicator of shape complexity. **Euler number** is the difference between the number of connected components and number of holes.

Regions with SVM probability score lower than 0.5 are temporarily discarded, but they are considered as candidates for character restoration in the last step. Example of character candidate classification is shown on Figure 3. In case of low aspect ratio ERs classifier can easily misinterpret letters like "i" or "l" as noise and vice versa. Therefore, these regions are simply discarded when its aspect ratio is lower than 0.3 assuming that low aspect ratio characters will be recovered in character restoration step.

2.2 Character candidates grouping and restoration

Grouping of character candidates and restoration of missed characters are performed with same method as in [Sar16]. Firstly, extremal regions classified as characters are grouped into text lines. In this step text line properties are exploited to filter out false positive classifications. Procedure consists of the following steps:
1. Regions are firstly sorted by centroid y coordinates. Transition between lines are found as abrupt changes in y coordinates between adjacent regions. In this way y coordinates of text lines are determined and each ER is joined to the closest line. Line break is detected if difference between y coordinates of region centroids exceeds $threshold_{line\ break}$:

$$threshold_{line\ break} = \begin{cases} 0,4 * max(height[i], height[i-1]), \\ if\ heightRatio > 2 \\ 0,4 * mean(height[i], height[i-1]), \\ if\ heightRatio \leq 2 \end{cases} \quad (17)$$

$$heightRatio = \frac{max(height[i], height[i-1])}{min(height[i], height[i-1])} \quad (18)$$

This threshold allows characters to deviate from horizontal orientation what enables to handle certain degree of diagonal text orientation.
2. Regions in each text line are sorted by centroid x coordinate. Relation of each region to its neighboring components is checked using distance function proposed in [Yin14] which includes features like width and height difference, vertical and horizontal alignment etc. If distance value to both neighbors is higher than threshold (also defined in [Yin14]), ER is discarded as non-character object.

Character restoration step deals with problem of character regions erroneously discarded by SVM classifier. Candidates for restoration are ERs having low SVM probability score. For each region that already has been classified as character its adjacent restoration candidates are checked and restored if:

- Difference of centroid y coordinates between character region i and restoration candidate region j is lower than 50% of the character region height
- Horizontal distance between character region i and restoration candidate region j should satisfy:

$$\frac{H_{ij}}{mean(w_i, w_j)} < 2 \quad (19)$$

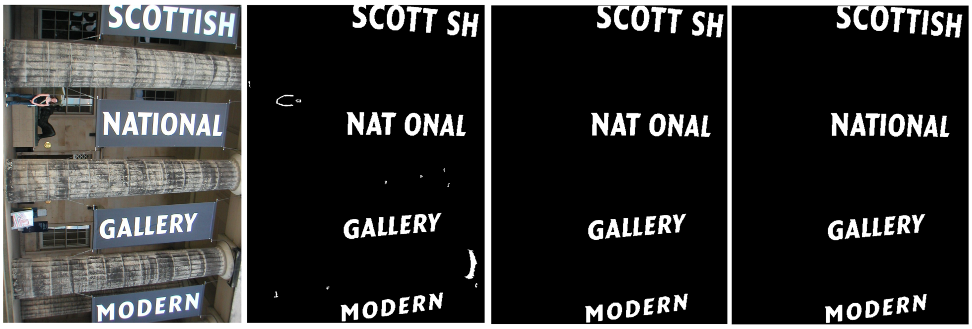

Figure 4: Character candidates grouping and restoration (from left to right): input image, image after SVM classification of character candidates, image after character candidates grouping and image after character candidates restoration

where H_{ij} is distance between right edge of left region and left edge of the right region, while w_i and w_j represent bounding box width for regions i and j.

- Restoration candidate is not contained inside character region
- Overlap of restoration candidate region with character region is lower than 20% of the character region area
- Value of distance function ([Yin14]) between restoration candidate region j and character region i is lower than threshold already used for character grouping

Example of character candidates grouping and restoration steps is shown on Figure 4.

3 EXPERIMENTAL RESULTS

Proposed method was evaluated on the dataset of ICDAR 2013 challenge "Reading text in scene images" [Kar13]. Same dataset is used in ICDAR 2015 challenge "Focused scene text" [Kar15]. Training set (229 images) and testing set (233 images) contain variety of scene text examples captured under challenging imaging conditions (low contrast, blur, uneven illumination, perspective distortion). Text segmentation performance is evaluated using precision and recall measures on pixel level and atom level. Since pixel-based measures don't reflect morphological properties of extracted region, atom-based metrics presented in [Cla10] is considered more relevant. It introduces minimal and maximal coverage criteria. The first criterion requires that extracted region should cover Tmin = 0.9 of a character skeleton. According to the second criterion, pixels outside the character area should be close to the character edge. More precisely, distance between pixel and edge should not exceed Tmax = max(5; 0.5*G), where G is the maximum stroke width of a character. Based on these criteria each region can be classified into following categories: Background,Whole, Fraction, Multiple,Fraction and Multiple,Mixed.

Table 1 compares performance of proposed text segmentation method with results presented in [Kar13] and [Lu15]. Proposed method achieves second highest precision and competitive recall. Higher precision value can be explained by influence of the proposed character generation method which uses redundant representation and multiple classifications of the same region. In this way elimination of non character candidates is more efficient. Proposed approach outperforms USTB_FuStar method which is based on algorithm described in [Yin14]. Proposed approach use simpler character candidates generation scheme without using maximum stability criterion for ER pruning. NSTsegmentator variant of method proposed in [Mil15] is significantly outperformed, while in comparison with NS-Textractor variant of [Mil15] slightly lower F-score is achieved. Figure 5 shows examples of proposed scene text segmentation method. Method successfully extracted text in third example characterized by reflections and fourth image where complex background and windows structures were eliminated. In the first image error is caused by windows structures that are misinterpreted as character regions, while in the second image character "L" is lost due to highlights. Average running time per image calculated on ICDAR 2015 training set is about 14 seconds (Intel Core 2 Duo 3 GHz, 5 GB RAM). Execution time can be significantly improved by more efficient implementation in C++ and more recent CPU.

4 CONCLUSION

In this paper novel method for generation of character candidate regions is proposed. ER tree subpaths are constructed using area variation. Sufficiently long subpaths are taken as character candidates that are represented by subset of ERs from that subpath. Mean SVM probability score of ERs in this subset is used to decide whether subpath represent character or noise component. This step is followed by character grouping and restoration steps. Experimental results obtained on the ICDAR 2013 dataset show that the proposed

Table 1: Text segmentation results on ICDAR 2013 test set

	Pixel based results			Atom based results		
Method	Recall	Precision	F-Score	Recall	Precision	F-Score
Lu et al.[Lu15]	77.27 %	82.10 %	79.63 %	70.12 %	81.20 %	75.24 %
I2R_NUS_FAR	74.73 %	81.70 %	78.06 %	68.64 %	80.59 %	74.14 %
NSTextractor [Mil15]	60.71 %	76.28 %	67.61 %	63.38 %	83.57 %	72.09 %
Proposed method	60.32 %	79.93 %	68.76 %	64.43 %	81.21 %	71.85 %
USTB_FuStar [Yin14]	69.58 %	74.45 %	71.93 %	68.03 %	72.46 %	70.18 %
I2R_NUS	73.57 %	79.04 %	76.21 %	60.33 %	76.62 %	67.51 %
NSTsegmentator [Mil15]	68.41 %	63.95 %	66.10 %	68.00 %	54.35 %	60.41 %
Text Detection	64.74 %	76.20 %	70.01 %	62.03 %	57.43 %	59.64 %

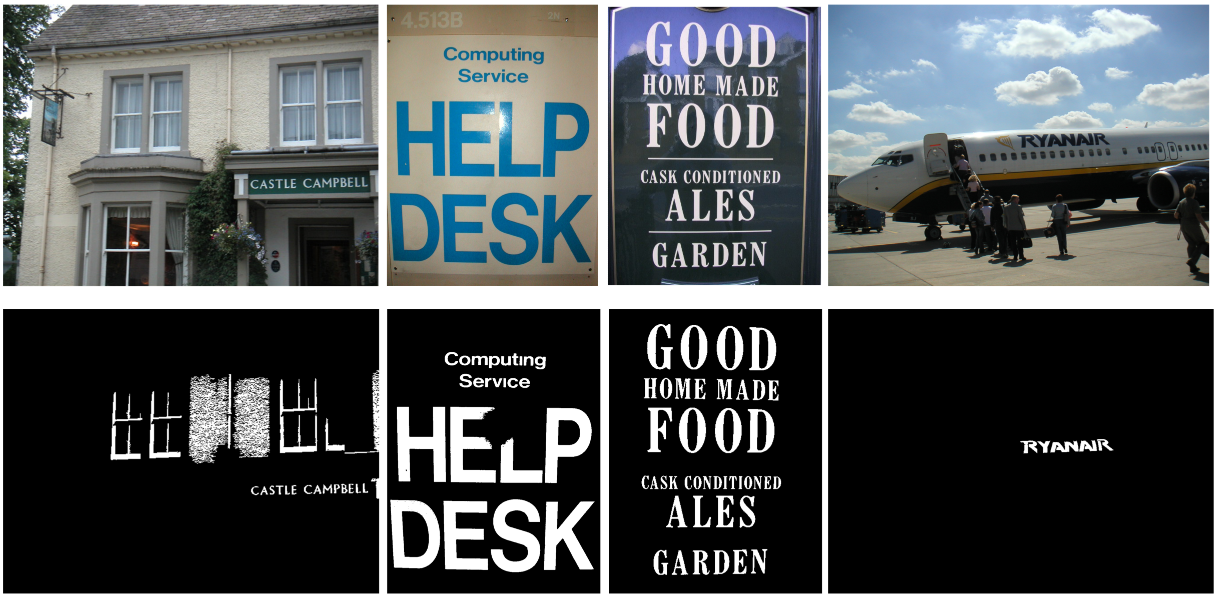

Figure 5: Scene text segmentation examples

approach efficiently eliminates non character regions yielding high precision value and competitive recall value. Future work will focus on the improval of recall rate through integration of the sliding window approach after character candidates grouping, that is in the regions of interest corresponding to extracted text lines. In this way number of scanned windows will be reduced and scale would be known in advance. This approach would lower computational complexity of standard sliding window methods which requires processing of the whole image on multiple scales.

5 ACKNOWLEDGMENT

This work has been fully supported by Croatian Science Foundation under the project UIP-2014-09-3875.

6 REFERENCES

[Kar15] Karatzas D., Gomez-Bigorda L., Nicolaou A., Ghosh, S., Bagdanov, A, Iwamura, M.,Matas, J., Neumann, L. Icdar 2015 competition on robust reading, in Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, 1156-1160,2015.

[Jad14] Jaderberg, M:, Vedaldi, A., and Zisserman, A. Deep features for text spotting, in Computer Vision–ECCV 2014, 512-528, 2014.

[Wan12] Wang, T., David, J. W., Coates, A., and Ng, A.Y. End-to-end text recognition with convolutional neural networks, in Pattern Recognition (ICPR), 2012 21st International Conference on, 3304-3308, 2012.

[Yin14] Yin, X-C., Yin, X., Huang, K., Hao, H-W. Robust text detection in natural scene images. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 970-983, 36, No. 5, 2014.

[Neu15] Neumann, L., Matas, J. Real-time Lexicon-free Scene Text Localization and Recognition.

Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1872-1885, No. 99, 2015.

[Sun15] Sung, M-C., Jun, B., Cho, H., and Kim, D. Scene text detection with robust character candidate extraction method. Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, 426-430, 2015.

[Hua14] Huang, W., Qiao, Y., and Tang, X. Robust scene text detection with convolution neural network induced mser trees, in Computer Vision–ECCV 2014, 497-511, 2014.

[Lu15] Lu, S., Chen, T., Tian, S., Lim, J-H., and Tan, C-L. Scene text extraction based on edges and support vector regression. International Journal on Document Analysis and Recognition (IJDAR), 125-135, 18, No. 2, 2015.

[Mil15] Milyaev, S., and Barinova, O., Novikova, T., and Kohli, P., and Lempitsky, Victor. Fast and accurate scene text understanding with image binarization and off-the-shelf OCR, International Journal on Document Analysis and Recognition (IJDAR), 169-182, 18, No. 2, 2015.

[Mat04] Matas, J., and Chum, O., and Urban, M., and Pajdla, T. Robust wide-baseline stereo from maximally stable extremal regions, Image and vision computing, 761-767, 22, No. 10, 2004.

[Gon13] González, Á., and Bergasa, L. M. A text reading algorithm for natural images. Image and Vision Computing, 255-274, 31, No. 5, 2013.

[Sar16] SAric, M. Scene text segmentation using low variation extremal regions and sorting based character grouping, submitted to Neurocomputing, 2016

[Kar13] Karatzas D., et al. ICDAR 2013 robust reading competition, in Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, 1484-1493, 2013.

[Cla10] Clavelli, A., Karatzas, D., and Lladós, J. A framework for the assessment of text extraction algorithms on complex colour images, in Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, 19-26, 2010.

[Sha11] Shahab, A., Shafait, F., Dengel, A. ICDAR 2011 robust reading competition, in Document Analysis and Recognition (ICDAR), 2011 11th International Conference on, 1491-1496, 2011.

[Kai17] KAIST Scene Text Database. http://www.iapr-tc11.org/mediawiki/index.php?title=KAIST_ Scene _ Text_ Database

[Cam09] De Campos, T.E., Babu, B.R., Varma, M. Character recognition in natural images, in VISAPP (2), pp. 273-280, 2009.

[Mar01] Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, Vol. 2, IEEE, pp. 416-423, 2001.