ICDAR 2011 Robust Reading Competition
Challenge 2: Reading Text in Scene Images

Asif Shahab, Faisal Shafait, Andreas Dengel
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
{*asif.shahab, faisal.shafait, andreas.dengel*}*@dfki.de*

***Abstract*—Recognition of text in natural scene images is becoming a prominent research area due to the widespread availablity of imaging devices in low-cost consumer products like mobile phones. To evaluate the performance of recent algorithms in detecting and recognizing text from complex images, the ICDAR 2011 Robust Reading Competition was organized. Challenge 2 of the competition dealt specifically with detecting/recognizing text in natural scene images. This paper presents an overview of the approaches that the participants used, the evaluation measure, and the dataset used in the Challenge 2 of the contest. We also report the performance of all participating methods for text localization and word recognition tasks and compare their results using standard methods of area precision/recall and edit distance.**

***Keywords*-Scene text detection, natural images, text recognition**

I. Introduction

Research in document analysis and recognition has traditionally focused on processing and analyzing scanned documents. The advent of low cost consumer-end digital cameras has opened new challenges for the document analysis community. On one hand, documents captured with a camera have different kinds of degradations (perspective deformation, page curl, non-uniform illumination, blur, etc.). On the other hand, document analysis domain is not limited to documents any more – one can have photographs of vehicle number plates, street names, gas/electricity meters, and so on where automatic recognition of text is desired. The latter set of challenges constitute the area of scene text regonition requiring the researchers to go beyond the traditional techniques for document image analysis to solve them.

Research on scene text detection and recognition already started in mid-90s [1], [2] and since then has seen a substantial amount of growth with a large number of approaches published in the last decade [3], [4], [5], [6], [7]. While approaches for scene text detection/recognition tackle a much broader set of problems (non-planar surfaces, unkown layout, blur, varying distance to camera, much broader resolution ranges), synergies exist with other methods handling a subset of these problems. Examples are extracting text in color images (book or journal cover pages) [8], locating text in videos [9], or segmenting text in web images [10].

To benchmark performance of different scene text segmentation and recognition algorithms, a Robust Reading Competition was organized at ICDAR 2003 [11], [12]. The robust reading problem was divided into three sub-problems, and competitions were run for each stage. The sub-problems chosen were text locating, character recognition, and word recognition. The text locating problem received five entries, whereas no one participated in the character recognition and word recognition problem. A follow-up competition was held at ICDAR 2005 [13] using the same dataset as that of the previous competition. Again, only text locating problem received a sufficient number of entries.

To track progress in the field in the meanwhile, a Robust Reading Competition was held at ICDAR 2011 dealing with two challenges. The first challenge deals with born-digital images (web, email) whereas the second challenge deals with scene text recognition. A summary of the results of Challenge 1 is presented in [14]. This paper summarizes results of Challenge 2 on scene text recognition. The concept of scene text recognition challenge is similar to ICDAR 2003 and ICDAR 2005 Robust Reading competitions. Some problems were reported about the dataset (slightly larger bounding boxes, inconsistent definition of a ”word” for instance whether a hyphen breaks the words or not) as well as the evaluation scheme (handling of one-to-many and many-to-one matches) used in the previous competitions [15]. We created ground-truth of the ICDAR 2003 Robust Reading competition dataset from scratch and adapted a new evaluation scheme [15] to resolve these issues. We hope the new ground-truth will make the dataset even more widely used in the community. Besides, we extended the dataset with more images containing a variety of outdoor and indoor scenes.

The rest of this paper is organized as follows. Section II introduces the challenge format, Section II-A describes our collected dataset as well as the new ground-truth of ICDAR 2003 Robust Reading Competition dataset. The performance evaluation scheme employed in this contest is outlined in Section II-B. A summary of participating methods is provided in Section II-C. Section III shows the results of the competing methods followed by a conclusion in Section IV.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.296

II. The Challenge

Reading text in scene images challenge consisted of two tasks:

1) **Text localization task**: The target of text localization task was to identify text regions in scene images and mark their location with axis-aligned rectangular bounding boxes.
2) **Word recognition task**: The target of word recognition task was to recognize cropped word images of scene text. Cropping was done based on ground-truth word bounding boxes to evaluate recognition performance independently from text localization accuracy.

Owing to independent input for both tasks, it was allowed to participate in both tasks or in either of them. In contrast to the previous ICDAR Robust Reading Contests [12], [13], both challenges in ICDAR 2011 contest were organized in an open mode. The training and test datasets were made available at specified dates and the participants were given a time frame of three days after the release of the test data to submit their results. We rely on the academic integrity of the participants when reporting final results.

A. Dataset

We extended the dataset used in earlier Robust Reading Competitions organized in ICDAR 2003 and 2005 [12], [11], [13]. The dataset used in these competitions along with word and character level ground truth in XML format is publicly available at http://algoval.essex.ac.uk/icdar/Datasets.html. We carefully analyzed the dataset and observed following shortcomings:

- Missing ground truth information for some of the files and text elements within some images.
- Mixed interpretation of punctuation and special characters as part of words.
- Bounding boxes around words are not tight.

Due to these reasons, we decided to prepare the ground truth of these files from scratch. Additionally, we captured around 100 images with digital camera using auto focus and natural lighting. The final dataset consisted of 485 images containing text in a variety of colors and fonts on many different backgrounds and in various orientations.

The ground truth is prepared in two phases. In the first phase, we prepared text location ground truth using kolourpaint[1]. We converted all images to gray level and used colored bounding boxes to mark the word location and save them as 24 bit PNG image. We took special care of the following:

- Space character is consistently used as word separation. All punctuation marks and special characters are considered as part of the word as long as there is no space character separating them.
- The bounding boxes are tight so they touch most of the boundary pixels of a word.

In the second phase, we prepared word recognition ground truth. We prepared a simple ground truth GUI to annotate words in an image. The GUI allows users to draw rough bounding boxes around words and label them with ASCII string.

We generated our ground truth automatically using the colored image files and labels generated using our GUI by evaluating bounding boxes overlap for a given image file. The ground truth consisted of bounding box co-ordinates which are stored in a separate text file for each of the image files.

The same method is used to extract word images and its associated ground truth. Our word recognition dataset consisted of 1564 word images. These word images are actually cropped from images in the text localization dataset using word bounding box ground truth. Each word is stored in a separate file and the ground truth transcription for these words is provided in a line separated file.

We partitioned our dataset for Text Localization and Word Recognition task into training and test set. The training set was published online on April 20, 2011, giving participants ample time to train their algoirthms. The test set was published on June 01, 2011 and participants were requested to submit the results by June 03, 2011.

B. Performance Evaluation

1) Text Localization Task: The task of text localization can be evaluated using any standard methodology for evaluating page segmentation performance [16], [17] that takes into account different categories of segmentation errors (over-, under-, and missed-segmentation). The main question when choosing a method for scene text detection particularly is how to deal with under and over segmentation errors. In this competition, we employ the method by Wolf et al. [15] that is specifically designed to evaluate scene text detection approaches. We used the DetEval [2] evaluation software with default parameters for evaluating the competing methods.

2) Word Recognition Task: To evaluate the word recognition accuracy, we simply use the edit distance with equal cost of deletions, substitutions, and insertions. We normalize the edit distance by the number of characters in ground truth word.

C. Participating Methods

Each of the participants provided a summary of their methods along with their results. These summaries are presented here for the completeness of this paper.

[1] Open source program shipped with KDE: http://kolourpaint.sourceforge.net/

[2] http://liris.cnrs.fr/christian.wolf/software/deteval/index.html

1) "Yi's Method": [3] To localize the text regions, adjacent character grouping [18] is performed to obtain a set of candidate image patches. Then, Haar features are extracted from gradient maps and stroke orientation maps by the block patterns presented in [19]. The features are input into Adaboost learning model to train a classifier of text to determine whether the candidate patches are text regions or not. The image patches classified as text regions are merged into rectangle boxes.

2) "Kim's Method": [4] We propose a new text detection algorithm for localizing text region in a mobile phone. First, blobs in an image are extracted by using the maximally stable extremal regions (MSER) approach and the horizontally neighboring blobs are merged when their sizes and colors are similar. The merged region is regarded as a candidate text region. The candidate region detection enables fast and robust text localization, while it also detects a huge amount of non-text regions as candidate regions. In order to minimize the false positives, gradient features obtained from oriented gradient images are used, where the feature value is computed by summing the pixel values inside a rectangle in an oriented gradient image. The Adaboost learning method is used for deciding the location and size of the rectangle in the oriented gradient image. A cascade classifier is used to discriminate text from non-text, where there are several stages for various features. The proposed detector provides very robust performance and runs in real-time on Nexus One phone.

3) "Text Hunter": [5] The proposed approach to localize text in scenes and digital images is composed of two steps: text detection and fine localization of text regions. The text detector is a cascade of boosted classifiers trained using AdaBoost algorithm. In contrary to current research in object detection, we proposed to use heterogeneous (belonging to different families) weak classifiers in boosting paradigm. The weak classifiers used in this work belong to: generative and discriminant, linear and nonlinear, parametric and nonparametric families of classifiers. To encode textual information, we have proposed two set of features. One feature set is based on the contrast between foreground and background while the second feature set encodes shape and appearance of characters in a text region. These features are computed on an image (detection) window of small size which is further divided into 16 cells or blocks. Detection window size varies form 32×16 pixels to 288×144 pixels in 9 steps. Weak classifiers are trained in a feature space containing single and pair-wise features. The output of text detector is a set of detections at various scales along with confidence level of the detector. More details can be found in [20].

In the fine localization step, the detections in (quasi) horizontal direction are grouped and their confidence levels are added resulting in candidate text regions. Then, connected components are extracted by applying morphological operations on the canny edge map of each candidate text regions. Next, connected components are validated using simple thresholds on features such as confidence level, edge density, height, width and aspect ratio. Later, validated connected components are grouped to form text lines and/or words. The thresholds on features used for validation of connected components and grouping are learnt on training database using the genetic algorithm where the objective is to maximize the F-measure on the given training database. In case of scene text (Challenge 2), features based on gradient information of connected components are also taken into consideration during validation.

4) "KAIST AIPR System": [6] Our algorithm for text localization task is not yet published. The description is as follows.

We used hierarchical framework to segment scene text regions in images. Our framework utilizes a hierarchical model of text in which dependencies among levels at different scales are considered coherently. The main components of the hierarchical structure of scene text are "the text line (composed of characters)," "the characters," "the superpixels," and "pixels."

A superpixel is a coherent local region that preserves most of the object boundaries. These superpixels contain fine level details along object boundaries well but they cannot present geometric shape information of object regions. A segment is an enlarged region which has one or more neighboring superpixels together. Therefore, we can encode both properties of the text region well: distinctiveness from the background and homogeneity of color inside the character region.

We extened classical one-layer Conditional Random Field (CRF) to multilayer CRF with up to second order cliques. With this model, we could efficiently solve the scene text extraction problem by integrating local and global information and obtain binarized text image as result.

From the binarized text image, we group the connected components(CCs) into text lines. Then we seperated the text lines into words and applied a boundary extension algorithm for each word to find missing letters. After the missing CCs are recovered, all CCs are grouped into text lines again and seperated into words for final output.

After obtaining the binarized text region, we used an OCR engine made by INZI soft [21] to recognize the text.

5) "Neumann's Method": [7] The method [22], [23] is based on an exhaustive effectively pruned search of the space of all character sequences, where individual characters are detected as Maximally Stable Extremal Regions (MSERs).

[3] Chucai Yi and Yingli Tian – City University of New York, U.S.

[4] Chunghoon Kim – Qualcomm Korea R&D Center, South Korea

[5] Muhammad Shehzad Hanif – UET Lahore, Pakistan and Lionel Prevost – ISIR, Université Paris 06, France

[6] Seonghun Lee, Kyungmin Min, Jae-Hyun Seok and Jin Hyung Kim – Department of Computer Science, KAIST, South Korea

[7] Lukas Neumann and Jiri Matas – Faculty of Electrical Engineering, Czech Technical University, Czech Republic

The method exploits higher-order features and feedback loops to compensate errors in text detection.

Text regonition is based on contour-based features and a multi-class SVM classifier, which is trained using synthetic data. The method takes into account multiple hypotheses in both text localization and recognition stages and selects the best hypothesis in the final stage using a simple language model.

6) "LIP6-Retin": [8] We propose a robust and accurate multi-resolution approach to detect and classify text regions in such scenarios. Based on generation/validation paradigm, we first segment images to detect character regions with a multi-resolution algorithm able to manage large character size variations. The segmented regions are then filtered out using shape-based classification, and neighboring characters are merged to generate text hypotheses. A validation step computes a region signature based on texture analysis to reject wrong region hypotheses.

7) "TDM_IACAS": [9] Firstly, zoom the input image to $(s1 \times \text{width}) \times (s1 \times \text{height})$, $(s2 \times \text{width}) \times (s2 \times \text{height})$ and $(s3 \times \text{width}) \times (s3 \times \text{height})$ respectively. Then detect text regions on the three scaled image respectively. At last, put all text regions extracted in the three scales together, if there are two region $r1$ and $r2$, and the corresponding areas are $r1_{\text{area}}$ and $r2_{\text{area}}$, the intersecting area between $r1$ and $r2$ is $r1 \cap r2$, and if

$$r1 \cap r2 / \min(r1_{\text{area}}, r2_{\text{area}}) > T_{\text{area}}$$

then discard the region whose area is smaller.

In our system $T_{\text{area}} = 0.6$, $s1$, $s2$ and $s3$ are decided according to the input image's width and height:

$s1$=0.5, $s2$ = 1, $s3$=2,	if $\max$(width,height) > 1600
$s1$=1, $s2$=2, $s3$=4,	else if $\max$(width,height) > 800
$s1$=1, $s2$=2, $s3$=6	otherwise

The detection method on each scaled image includes the following steps [24]:

- Get the And-Valley image, I_{AV} and the And-Ridged image, I_{AR}
- Classify connected components in I_{AV} and I_{AR} as a character or non-character connect component. 8-direction gradient direction features are extracted for character classification which is different to [24]
- Group all character connected components to text regions.

8) "TH-TextLoc, TH-OCR System": [10] In TH-TextLoc system, we adopt a region-based method to localize text in scene images. It mainly consists of three stages as follows:

In text detection stage, steps of text candidates generation and text candidates selection are contained.

We first extract all the text candidates components using adaptive local binarization method in gray level. For text selection, a coarse-to-fine theme is employed to selecte text candidates. In coarse text candidates selection, neighboring connected components analysis based on heuristic knowledge is used to filter out apparently noise CCs. The main idea for neighboring components analysis comes from [25]. Text characters generally don't appear alone, but together with other characters of similar properties and are usually regularly placed in horizontal string. We take several relative loose rules to eliminate the distinct false candidates roughly, especially for the isolated or bar-shaped components [26]. In the fine text candidates selection, we train a SVM classifier using several connected components features. It includes geometric features, shape related features and strokes width features.The discriminative classifier SVM is employed to classify the remaining text candidates into text/non-text ones.

Finally, we group these text candidates into text regions using projection histogram analysis. To refine these text regions, we employ a region shrinking and extension approach by analyzing the candidate's CCs located inside and outside of the region's boundary. In addition, text lines are broken into separate words using heuristic rules.

In word recognition system, TH-OCR, we mainly focus on the text binarization stage rather than core algorithms like character segmentation and recognition in OCR engine to improve the performance. The system mainly consists of three stages as follows:

In pre-processing stage, all the images have been normalized to the same height using bi-cubic interpolation method; we set the height to 100 pixels.

In text binarization stage, the foreground text pixels have been segmented from background pixels. We first classify the text polarity (black or white text pixels) of each image using connected component analysis method, and reverse the gray images in which text pixels are white. Then, we utilize adaptive local binarization method to generate coarse text components. A morphological opening operation is employed on the binarized image to separate the consecutive connected characters. Then post-processing using connected components analysis is used to remove the non-text regions, such as noise points and bar-style background blocks.

In word recognition stage, the binarized image and the gray image were fed into the OCR engine. We adopt the OCR engine TH-OCR 2007, which is an Asian multi-languages recognition software mainly developed by our group from Tsinghua University in China. In the engine, characters are segmented guided by the input binary image and recognized using grayscale features. It should be noted that the recognition result come without using any language models. For more information about TH-OCR engine, one can refer to [27].

[8] Jonathan Guyomard, Frederic Precioso, Nicolas Thome, Matthieu Cord and Rodrigo Minetto – LIP6, UPMC-Sorbonne Universités, France

[9] Yunxue Shao, Chunheng Wang and Yang Zhang – Institute of Automation, Chinese Academy of Sciences, Beijing, China

[10] Cheng Yang, Changsong Liu and Xiaoqing Ding – Department of Electronics Engineering, Tsinghua University, Beijing, China.

9) "ECNU-CCG Method": [11] Firstly, Sobel edge detectors are applied to introduce an edge map which contains the strong edge information of the original image. The edge map is obtained by combing four individual edge maps with typical directions, viz., horizontal, vertical, up-right slanting, and up-left slanting.

Then, the connected component analysis (CCA) is employed to detect potential text areas of the image from the edge map. Some connected components, closed to each other and with similar width or height, may be merged.

Finally, an N-level scale space model is constructed and the spatial responses to the Laplacian-of-Gaussian operator are computed. In view of the time consumption, we use the Difference-of-Gaussian (DoG) operator to approximate Laplacian-of-Gaussian (LoG) operator. The scale at which the strongest spatial responses present may indicate the stroke width of the text characters. Therefore, we use the distribution of strongest response and the scale they appearing to identify the candidate text regions.

III. Results and Discussion

A. Task 1 - Text Localization

We received nine entries for the text localization task. The bounding boxes of text returned by different participants were compared with the ground-truth bounding boxes as outline in Section II-B1. Precision and recall are cumulated for all images in the test set and their harmonic mean is used as the ranking metric. The results of the evaluation are shown in Table I. The results show a wide range of performance achieved by participating algorithms. In almost all cases, precision was higher than recall. The method submited by Chunghoon Kim achieved highest performance on both precision and recall and therefore is a clear winner of the text localization task of the competition.

Table I
Text Localization Results (%)

Method	Recall	Precision	Harmonic Mean
Kim's Method	62.47	82.98	71.28
Yi's Method	58.09	67.22	62.32
TH-TextLoc System	57.68	66.97	61.98
Neumann's Method	52.54	68.93	59.63
TDM_IACS	53.52	63.52	58.09
LIP6-Retin	50.07	62.97	55.78
KAIST AIPR System	44.57	59.67	51.03
ECNU-CCG Method	38.32	35.01	36.59
Text Hunter	25.96	50.05	34.19

B. Task 2 - Word Recognition

We received three entries for word recognition task. We used normalized edit distance as evaluation metric for word recognition task as described in section II-B2. We report the results for the three participating algorithms in Table II. We report results in form of *Total Edit Distance* which is obtained by accumulating the normalized edit distance for each of the ground truth word and the corresponding recognition result.

Table II
Word Recognition Results

Method	Total Edit Distance	Correct Recognition (%)
TH-OCR System	176.23	41.2
KAIST AIPR System	318.46	35.6
Neumann's Method	429.75	33.11

The results of Table II shows TH-OCR System as clear winner which correctly recognizes 41.2% of the words. We also did further analysis of these correct detections which revealed that 13.41% of the words are correctly recognized by all of the participating algorithms. These are the images which are rather simple in a way that the word appears in a standard font and size on a uniform single colored background.

However, it is to be noted that majority of correct detections by these three algorithms donot overlap and 38.41% of the words are not correctly recognized by any of the participating systems.

IV. Conclusion

This paper summarizes results of the reading text in scene images challenge of ICDAR 2011 Robust Reading Competition. The dataset used in this challenge was an extended version of the ICDAR 2003 Robust Reading Competition dataset with more images and revised ground-truth. The challenge was very well received by the community and we received nine entries for text localization task and three entries for word recognition task. The results submitted by Chunghoon Kim from Qualcomm R&D center in Korea achieved the highest performance with about 62% recall and 83% precision for the text localization task. The results submitted by Cheng Yang from Tsinghua University, China achieved highest accuracy with about 41% correct detections for word recognition task. The results show that although current algorithms achieve decent accuracy on real-world scene text recognition problem, there is still a lot of room for improvement. Hence, we expect that reading text in natural scene images will remain a hot topic of research in the document analysis community in the future.

Acknowledgment

This work was partially funded by the BMBF (German Federal Ministry of Education and Research), projects INBEKI (13N10787) and Perspecting (01 IW 08002).

[11] Qiaoyu Sun – Dept. of Computer Science and Technology, East China Normal University, Shanghai, China and Yue Lu – Dept. of Electronic Engineering, Huaihai Institute of Technology, Jiangsu, China

References

[1] J. Ohya, A. Shio, and S. Akamatsu, "Recognizing characters in scene images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 214–220, Feb. 1994.

[2] Y. Zhong, K. Karu, and A. Jain, "Locating text in complex color images," *Pattern Recognition*, vol. 28, no. 10, pp. 1523–1535, Oct. 1995.

[3] P. Clark and M. Mirmehdi, "Recognising text in real scenes," *Int. Jour. on Document Analysis and Recognition*, vol. 4, no. 4, pp. 243–257, 2002.

[4] C. Mancas Thillou and B. Gosselin, "Color text extraction with selective metric-based clustering," *Computer Vision and Image Understanding*, vol. 107, no. 1-2, pp. 97–107, Jul. 2007.

[5] J. Weinman, E. Learned Miller, and A. Hanson, "Scene text recognition using similarity and a lexicon with sparse belief propagation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1733–1746, Oct. 2009.

[6] J. Park, G. Lee, E. Kim, J. Lim, S. Kim, H. Yang, M. Lee, and S. Hwang, "Automatic detection and recognition of korean text in outdoor signboard images," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1728–1739, Sep. 2010.

[7] Y. Pan, X. Hou, and C. Liu, "A hybrid approach to detect and localize texts in natural scene images," *IEEE Trans. on Image Processing*, vol. 20, no. 3, pp. 800–813, Mar. 2011.

[8] K. Sobottka, H. Kronenberg, T. Perroud, and H. Bunke, "Text extraction from colored book and journal covers," *Int. Jour. on Document Analysis and Recognition*, vol. 2, no. 4, pp. 163–176, 1999.

[9] X. Wang, L. Huang, and C. Liu, "A video text location method based on background classification," *Int. Jour. on Document Analysis and Recognition*, vol. 13, no. 3, pp. 187–207, Sep. 2010.

[10] D. Karatzs and A. Antonacopoulos, "Colour text segmentation in web images based on human perception," *Image and Vision Computing*, vol. 25, no. 5, pp. 564–577, May 2007.

[11] S. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *Proc. Int. Conf. on Document Analysis and Recognition*, Edinburgh, UK, Aug. 2003, pp. 682–687.

[12] S. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, H. Miyao, J. Zhu, W. Ou, C. Wolf, J. Jolion, L. Todoran, M. Worring, and X. Lin, "ICDAR 2003 robust reading competitions: Entries, results, and future directions," *Int. Jour. on Document Analysis and Recognition*, vol. 7, no. 2-3, pp. 105–122, Jul. 2005.

[13] S. Lucas, "ICDAR 2005 text locating competition results," in *Proc. Int. Conf. on Document Analysis and Recognition*, Seoul, Korea, Aug. 2005, pp. 80–84.

[14] D. Karatzas, S. R. Mestre, J. Mas, F. Nourbakhsh, and P. P. Roy, "ICDAR 2011 robust reading competition challenge 1: Reading text in born-digital images (web and email)," in *Proc. Int. Conf. on Document Analysis and Recognition*, Beijing, China, Sep. 2011.

[15] C. Wolf and J. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *Int. Jour. on Document Analysis and Recognition*, vol. 8, no. 4, pp. 280–296, Sep. 2006.

[16] F. Shafait, D. Keysers, and T. M. Breuel, "Performance evaluation and benchmarking of six page segmentation algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 941–954, 2008.

[17] A. Antonacopoulos, S. Pletschacher, D. Bridson, and C. Papadopoulos, "ICDAR 2009 page segmentation competition," in *Proc. Int. Conf. on Document Analysis and Recognition*, Barcelona, Spain, Jul. 2007, pp. 1370–1374.

[18] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," vol. 19, no. 12, 2011.

[19] X. Chen and A. Yuille, "Detecting and reading text in natural scenes," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, Jul. 2004, pp. 366–373.

[20] S. Muhammad Hanif and L. Prevost, "Text detection and localization in complex scene image using constrained adaboost algorithm," in *Int. Conf. on Document Analysis and Recognition*, Barcelona, Spain, Jul. 2009, pp. 1–5.

[21] "Inzisoft," Website, http://www.inzisoft.com.

[22] L. Neumann and J. Matas., "Text localization in real-world images using efficiently pruned exhaustive search," in *Int. Conf. on Document Analysis and Recognition*, Beijing, China, Sep. 2011.

[23] L.Neumann and J. Matas., "A method for text localization and recognition in real-world images," in *Proc. of the 10th Asian Conf. on Computer Vision*, Queenstown, New Zealand, Nov. 2010, pp. 770–783.

[24] Y. Shao, C. Wang, B. Xiao, Y. Zhang, L. Zhang, and L. Ma, "Text detection in natural images based on character classification," in *Proc. of the Advances in Multimedia Information Processing - PCM'10*, Shanghai, China, 2010, pp. 736–746.

[25] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Text detection in indoor/outdoor scene images," in *1st Int. Workshop on Camera-based Document Analysis and Recognition*, Seoul, Korea, Aug. 2005, pp. 127–132.

[26] N. Ezaki, M. Bulacu, and L. Schomaker, "Text detection from natural scene images: Towards a system for visually impaired persons," in *Int. Conf. on Pattern Recognition*, Cambridge, UK, Aug. 2004, pp. 683–686.

[27] H. Liu and X. Ding, "Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes," in *Int. Conf. on Document Analysis and Recognition*, Washington, DC, USA, 2005, pp. 19–25.