Image binarization for end-to-end text understanding in natural images

Sergey Milyaev*, Olga Barinova*, Tatiana Novikova*, Pushmeet Kohli†, Victor Lempitsky‡
*Lomonosov Moscow State University, Moscow, Russia
†Microsoft Research, Cambridge, UK
‡Skolkovo Institute of Science and Technology, Moscow, Russia
{smilyaev,obarinova,tnovikova}@graphics.cs.msu.ru, pkohli@microsoft.com, lempitsky@skoltech.ru

***Abstract*—While modern off-the-shelf OCR engines show particularly high accuracy on scanned text, text detection and recognition in natural images still remains a challenging problem. Here, we demonstrate that OCR engines can still perform well on this harder task as long as appropriate image binarization is applied to input photographs. For such binarization, we systematically evaluate the performance of 12 binarization methods as well as of a new binarization algorithm that we propose here. Our evaluation includes different metrics and uses established natural image text recognition benchmarks (ICDAR 2003 and ICDAR 2011). Our main finding is thus the fact that image binarization methods combined with additional filtering of generated connected components and off-the-shelf OCR engines can achieve state-of-the-art performance for end-to-end text understanding in natural images.**

***Keywords*-natural scene binarization; text localization**

I. Introduction

Natural image text understanding, which includes localization and recognition of text in the photographs of indoor and outdoor environments, is a task that is gaining increasing importance due to the proliferation of mobile devices, robotics systems and Internet image search. This task remains a challenging one due to such factors as varying text orientation, font, color and lighting as well as the abundance of structured clutter in many photographs. At the same time, a related task of *optical character recognition* (OCR) for scanned document images can be considered a mature technology that efficiently combines information about text appearance, semantics and language, and achieves high accuracy and computational efficiency. Reusing the OCR technology to natural image text understanding is a subject of this work.

Most OCR engines use image binarization (segmenting the text from background) as a first step in their pipelines. Thereby, the simplest way to employ OCR for natural scenes would be to perform image binarization and pass the result to an off-the-shelf OCR module. Perhaps surprisingly, such a simple approach has not been investigated in much detail, despite the fact that text binarization of scanned documents is well-studied [1]. Several recent papers [2], [3] propose new methods for binarization of natural scene text in cropped word images assuming that text localization is done at the previous step of a pipeline (which, in practice, is highly non-trivial). Image binarization has been also used as a part of different text detection and recognition pipelines [4], [5], [6]. However its contribution to the overall performance of the system as well as the intuition behind the choice of each particular binarization method was not detailed.

The goal of this work is to evaluate image binarization techniques in the context of end-to-end text understanding. First and foremost, we systematically evaluate the performance of several well-known image binarization methods on established ICDAR benchmarks across different metrics, including segmentation accuracy and the final word recognition accuracy demonstrated by an OCR engine applied to the binarization result. As a result of this comparison we select the top methods and compare them within the most interesting end-to-end text detection and recognition scenario. We find that even a standard binarization method such as non-linear Niblack [7] in combination with an off-the-shelf OCR module show performance competitive to fancier state-of-the-art text understanding methods.

Encouraged by this finding, we have also designed a new binarization method that is particularly suitable for text in natural images. The method embeds local binarization into a global optimization framework. It does not require any information about the position and size of the text in an image and it can be used for text localization as well as for recognition of the cropped text. As we demonstrate, this new method shows superior results in terms of the OCR accuracy compared to existing binarization methods and demonstrates even more competitive performance w.r.t. recent methods for text understanding.

II. Binarization methods

Related work. We first provide a very brief review of existing binarization method that we have considered. These methods can be roughly divided into two groups: the first group uses a fixed threshold for a given image (Otsu [8], Kittler [9]), while the second group (*local binarization*) uses local thresholds (Sauvola [10], Niblack [11]). In general, methods that use a global threshold typically work well when the text occupies a large part of the picture and is well contrasted from background. On the other hand, local binarization techniques can handle uneven illumination and text color variations better, yet they are more sensitive to the choice of the parameters (e.g. the charactersitic scale). In particular, optimal parameter values may differ for text of different sizes even within a single

Fig. 1. A comparison of cropped image binarization results of methods with top OCR accuracy (labels flipped where appropriate). From top to bottom line: (1) original image, (2) Niblack, (3) Non-linear Niblack, (4) Proposed.

Fig. 2. The steps of the proposed binarization method. (Top-left - input image, top-right - local binarization for dark text on light background (the candidate text regions are shown in blue), bottom-left - the seeds resulting from incorporating local binarization and the Laplacian of the image intensity, bottom-right - the binarization after global optimization for dark text on light background. The candidate text regions are shown in blue.

image and some text detection and recognition pipelines [4] precede local binarization with the local text scale estimation.

Several methods for text binarization in natural images have been proposed more recently. For instance, Zhu et al. [7] suggested using the ordered statistics filter for estimating thresholds in the non-linear Nilblack decomposition. Howe [12] proposed to use the Laplacian of the image intensity for scanned document binarization within a Markov Random Field model (which is an algorithmic setup most similar to the one we propose below). Gatos et al. [13] used two binarized images by Sauvola's method for original gray-scale and inverted images for rough estimation of background and thresholded the difference between original and binarized images. Ezaki et al. [14] proposed generating connected components by combination of mathematical morphology operations, edge extraction and Otsu thresholding of image color channels. Epshtein [15] suggested using a new image operator (*Stroke Width Transform*) to segment letters. Minetto et al. [16] proposed using *toggle mapping* for character segmentation in a multiresolutional way since natural scene images have large character size variations and strong background clutter.

Other recent works [2], [3] focus on the binarization of cropped text assuming that the text is correctly localized in the preceeding steps of the pipeline. In this scenario, a bounding box of the text area is given and the boundary of the box is assumed to belong to background. Under this assumption, Mishra et al. [2] proposed a method for text binarization using iterated graph cut. Wakahara et al. [3] proposed a method based on k-means clustering and letter candidates classification for a similar cropped image scenario.

Proposed method. Apart from evaluating existing binarization methods, we propose a new binarization algorithm that consists of the following steps: 1) local binarization producing *seed pixels*, 2) seed pixel strength estimation and 3) global optimization. At the first step we use Niblack binarization. In particular, we perform local binarization with a rather small window size, since using large window size inside Niblack usually causes small letters to merge and we want to avoid this effect. Due to a deliberately small size of Niblack window, the result of the first step is a local binarization containing noise and holes but with a high "recall" for all characters including small ones (Figure 2). At the second step, the normalized absolute value of Laplacian of image intensity is computed at each pixel. The result of the Laplacian operator tends to have large absolute values near edges, where the local binarization with small window provides correct labels. Within the interior part of the letters the values of the Laplacian are usually close to zero. In this way, we can use values of the Laplacian as a confidence in initial labeling of the local binarization and then perform global optimization which accounts for pixel similarity for correcting errors of initial labeling. Figure 2 illustrates the steps of our algorithm.

For global optimization we construct an energy function $E(\mathbf{f}|I,\mathbf{n}) = E_{\text{local}}(\mathbf{f}|I,\mathbf{n}) + E_{\text{smooth}}(\mathbf{f}|I)$, where $\mathbf{f} = \{f_1, f_2, \ldots, f_N\}$ is the binary vector denoting the binarization result for pixels, $\mathbf{n} = \{n_1, n_2, \ldots, n_N\}$ is an initial labeling produced by the first two stages, and I is the input image. $E_{\text{local}}(f)$ is the unary term that measures the disagreement between f and the local binarization result, while E_{smooth} is a pairwise term that measures the smoothness of the binarization.

The unary term $E_{\text{local}}(\mathbf{f}|I,\mathbf{n}) = \sum_i e_{\text{local}}(i)$, where

$$e_{\text{local}}(i) = \begin{cases} 1 - (0.5 + \nabla^2 I'_i/2), f_i = n_i \\ 0.5 + \nabla^2 I'_i/2, f_i \neq n_i \end{cases} \quad (1)$$

where $\nabla^2 I'_i$ denotes the absolute value of Laplacian of the image intensity normalized to its maximum value.

We use a conventional pairwise term traditional to graph cut segmentation [17]: $E_{\text{smooth}}(\mathbf{f}|I) = \lambda \sum_{(i,j)\in\mathbf{N}} e_{\text{smooth}}(i,j)$, defined by pixel similarity:

$$e_{\text{smooth}}(i,j) = \begin{cases} \exp(-\frac{\|x_i - x_j\|^2}{2\sigma_g^2} - \frac{\|c_i - c_j\|^2}{2\sigma_c^2}), f_i \neq f_j \\ 0, f_i = f_j \end{cases} \quad (2)$$

where $\mathbf{N}$ denotes a neighborhood system (we use 8-connected neighborhood in experiments), x denotes pixel coordinates, c means RGB color, σ_g and σ_c are normalization constants, λ determines the degree of smoothness. The pairwise term thus imposes a cost for the boundaries in the binarization result according to the local color contrast in the input image.

The global minimum of this energy can be found efficiently using the graph cut inference [18]. As long as text in natural

images can be either darker than background or lighter than background, we construct energy function for both cases and perform optimization twice, hence obtaining two binary maps. Both maps should then be submitted to the OCR engine.

III. Performance evaluation

A. Text binarization

Evaluated Methods. We now present the results of our evaluation. We selected 12 different binarization methods for evaluation[1]. We have included methods commonly used for document images, namely Otsu [8], Kittler [9], Niblack [11] and Sauvola [10]. We have also included several recent methods for document binarization, namely Wolf [19] [2], Howe [12], and Lu [20][3], the last one being a runner-up at ICDAR 2011 Document Image Binarization Contest (DIBCO 2011) [21]. We have also included methods developed for natural images: Ezaki [14], Gatos [13], Minetto [16] and non-linear Niblack decomposition [7]. Finally, we have also included the method based on stroke width transform from [15] implemented in text localization system[4].

Datasets. As the ultimate goal is end-to-end text detection and recognition, we applied these methods to whole uncropped images. We have looked at the accuracy of an OCR engine when applied to the binarization results as well as at the segmentation accuracy achieved by the methods. In the first set of experiments, we restricted our analysis to the interiors of the ground truth word bounding boxes (the methods were still applied to uncropped images). To be able to measure the segmentation accuracy, we have performed a pixel level annotation for ICDAR 2003 dataset[5]. Some of the compared methods assume dark text on light background, so we applied them to both the original and the inverted images. For these methods the result corresponding to higher F-score (separately for each cropped region) is reported.

We have validated the parameters of all local binarization methods on the training part of ICDAR 2003 dataset in order to achieve the maximum OCR accuracy. The parameters of Niblack method were set as suggested in [4]. The parameters for the Sauvola method were set as suggested in [22]. For [14], [13], [16], [7] we used parameters suggested by the authors. For the proposed method we set k to 0.4 as in [4] and $w = 21$ in order to obtain finer segmentation for small letters. Other parameters of our method ($\lambda = 2$, $\sigma_g = 12$ and $\sigma_c = 0.02$) were set by validation.

Metrics. The quantitative results are presented in Table I for the ICDAR 2003 database and in Table II for the ICDAR 2011 database. We perform detailed evaluation on ICDAR 2003 using the pixel-wise annotation. We report standard accuracy measures including precision, recall, F-score and peak signal to noise ratio (PSNR). Although pixel-wise metrics are widely used in comparative analysis of document binarization techniques (see [23], [21]), they do not describe morphological structure of the generated connected components, which is important for the accuracy of text recognition. Therefore we also report morphological metrics proposed in [24]. These metrics are based on classification of all connected components into *background*, *whole*, *fraction*, *multiple*, *fraction & multiple*, *mixed* classes using the notions of minimal and maximal coverage. To evaluate the text binarization we compute the fraction of segments of each of the mentioned types as suggested in [24]. Finally we have measured the accuracy of word recognition (in a case-sensitive manner) using different binarization methods. We used a popular commercial OCR software Omnipage Professional 18[6]. Examples of the cropped word recognition are shown in figure 1. For ICDAR 2011 we compare OCR accuracies for the methods that showed highest OCR accuracy on ICDAR 2003 dataset with the results of ICDAR 2011 Robust Reading Competition. One can see that even applied to uncropped images both non-linear Niblack and proposed method in combintion with OCR engine show higher accuracy than the winner of ICDAR 2011 competition.

Key Results and Observations. The most popular methods for document image binarization like Otsu [8], Kittler [9], Sauvola [10] show significantly degraded performance on natural scenes. In the cases when color and illumination variations are high, global thresholding methods (Otsu [8], Kittler [9]) are unable to divide natural images into text and background using a single threshold. We believe that the reasons of degraded performance of local binarization methods is the locality of their operation as well as their high sensitivity to the choice of parameters. E.g. the window size parameter in many of those methods should roughly correspond to the letter size, which is typically not known a priori and can vary through the same image.

It is interesting that the state-of-the-art document binarization of Lu et al. [20] showed low performance compared to other methods thus highlighting the gap between the text binarization in scanned document images and natural scene images. At the same time, a rather simple Niblack method as well as its widely used non-linear modification achieve high OCR accuracy. While the method of Howe [12] uses Laplacian-based unary terms similarly to our method, it shows significantly lower accuracy in the case of natural images with complex backgrounds, which we believe is due to better choice of unary and pairwise terms inside the global optimization in proposed method.

Interestingly, it can be seen that that pixel-wise metrics, such as precision, recall, F-score and PSNR do not demonstrate strong correlation with OCR accuracy. For example, Niblack method, which has the highest F-score, is on the fourth place in terms of OCR accuracy. And vice versa, non-linear Niblack method which has a mediocre pixel-level results shows very high recognition accuracy. As a consequence structured-output machine learning of binarization techniques based on the

[1]available at *http://graphics.cs.msu.ru/en/science/research/msr/text*
[2]available at *http://liris.cnrs.fr/christian.wolf/software/binarize/index.html*
[3]available at *http://www.comp.nus.edu.sg/%7Esubolan/*
[4]available at *https://sites.google.com/site/roboticssaurav/strokewidthnokia*
[5]available at *http://graphics.cs.msu.ru/en/science/research/msr/text*
[6]available at *http://www.nuance.com/*

TABLE I

COMPARISON OF THE BINARIZATION METHODS ACROSS A NUMBER OF ACCURACY MEASURES ON THE ICDAR 2003 DATASET. NUMBER OF SEGMENTS IS DIVIDED BY THE NUMBER OF GROUND TRUTH CHARACTERS IN DATASET. SEE THE TEXT FOR MORE DETAILS.

Method	Prec.	Rec.	F-sc.	PSNR	Backgr.	Whole	Fract.	Mult.	F. & M.	Mixed	OCR
Otsu [8]	.79	.85	78	8.85	1.79	.43	.33	.02	.01	.07	47.1%
Kittler [9]	.70	**.89**	72	7.36	.93	.32	**.25**	.03	**.01**	.01	35.1%
Niblack [11]	.90	.80	**84**	10.05	23.57	.60	1.48	.02	.02	.04	56.0%
Sauvola [10]	.90	.66	73	9.62	4.05	.47	.84	.02	.01	**.02**	53.8%
NL Niblack [7]	.93	.73	79	10.34	4.05	.47	.84	.02	.01	.02	59.3%
Howe [12]	.81	.66	71	9.01	**.61**	.46	.32	.01	.01	.03	53.2%
Gatos [13]	.90	.68	75	9.80	.88	.50	.56	.02	.01	.03	56.2%
Ezaki [14]	.85	.82	82	9.61	2.57	.43	.43	.03	.02	.05	47.6%
Minneto [16]	.87	.79	82	9.41	2.90	.50	.42	.02	.02	.05	47.3%
Epstein [15]	.81	.85	82	9.40	1.24	.44	.42	.01	.03	.12	47.6%
Wolf [19]	.88	.66	72	9.59	4.17	.48	.78	.02	.01	.02	53.4%
Lu [20]	.87	.66	73	8.80	1.92	.43	.63	**.01**	.01	.04	52.2%
Proposed	**.91**	.78	82	**10.44**	2.22	**.64**	.33	.02	.01	.03	**63.5%**

TABLE II

THE ACCURACY OF WORD RECOGNITION FOR THE ICDAR 2011 DATASET FOR IMAGE BINARIZATION METHODS FOLLOWED BY AN OCR ENGINE AS WELL AS FOR THE PARTICIPANTS OF THE ICDAR 2011 CHALLENGE.

NL Niblack	Proposed	TH - OCR	KAIST AIPR	Neumann
54.9%	**60.3%**	41.2%	35.6%	33.11%

pixel-level loss (e.g. Hamming) is unlikely to perform well.

At the same time morphological metrics correlate much stronger with the OCR accuracy. In particular, as can be expected, the increasing number of whole segmented characters leads to increasing OCR accuracy. The number of mixed connected components shows a negative correlation with the OCR accuracy. Intuitive explanation for this fact may be that mixed components, that contain both text and non-text parts, are problematic for OCR engine. On the other hand the presence of merged and broken segments seems to be not crucial for OCR accuracy since an OCR engine can cope with such errors.

TABLE III

END-TO-END TEXT UNDERSTANDING ACCURACY ON THE ICDAR 2003 AND ICDAR 2011 DATASETS. THE ABILITY TO CORRECTLY LOCALIZE AND RECOGNIZE WORDS IS EVALUATED. THE FIXED LEXICON COMPRISES ALL WORDS THAT OCCUR IN THE DATASETS.

ICDAR 2003 dataset			
Method	Prec.	Rec.	F-meas.
Wang [25] (no lexicon)	0.54	0.30	0.38
Neumann and Matas (no lexicon) [26]	0.42	0.41	0.41
NL Niblack (no lexicon)	0.63	0.41	0.50
Multiscale NL Niblack (no lexicon)	0.62	0.43	0.50
Proposed (no lexicon)	**0.66**	**0.48**	**0.55**
Wang [25] (fixed lexicon)	0.45	**0.54**	0.51
Wang [27] (fixed lexicon)	-	-	**0.67**
NL Niblack (fixed lexicon)	0.85	0.44	0.58
Multiscale NL Niblack (fixed lexicon)	0.81	0.47	0.60
Proposed (fixed lexicon)	**0.88**	0.50	0.63
ICDAR 2011 dataset			
Method	Prec.	Rec.	F-meas.
Neumann and Matas[28]	0.37	0.37	0.36
Proposed (no lexicon provided)	0.66	0.46	0.54
Proposed (fixed lexicon provided)	**0.89**	**0.49**	**0.64**

B. End-to-end text understanding

Implementation details for creating the full pipeline. In our final set of experiments, we performed end-to-end text localization and recognition that required constructing a more complex pipeline. In it, we consider the output of image binarization and treat each connected component as a letter candidate. We then apply an AdaBoost classifier trained for character/non-character classification (we have used our pixel-wise annotation of the ICDAR'2003 training set augmented with projective distortions to get positive examples). The classifier uses simple features computed with *regionprops* function from Matlab Image Processing Toolbox[7] (area, width, height, aspect ratio, length ratio, compactness, solidity, number of holes, occupy ratio, holes to area ratio, equivalent diameter, fitted ellipse axis ratio and orientation).

During testing, we generated a graph on the candidate segments that pass the classifier using the following simple rules. The segments were connected with an edge if: 1) they were spatially close and had similar size, 2) they had labels of the same type ("dark text" or "light text"), 3) they had similar colors (differences of mean a and b values of *Lab* colorspace do not exceed 20). The connected components of the resulting graph were then considered as text line candidates. These text line candidates were then split into words based on the assumption that the distance between two subsequent characters in the same word can not exceed twice the median distance between characters in the same text line. Generated word candidates were passed to the OCR module for recognition. We filtered out the word candidates with the height smaller than 15 pixels, since the OCR engine is unable to process text below this size. For each word candidates that passed the filters we computed the average probabilistic classifier output for the segments that constitute this word (sigmoid transform [29] are considered to map the outputs of boosted classifier to probabilities). By varying the threshold on this output we generated the recall-precision curve.

Evaluated methods. Here we report results for three different binarization strategies: 1) single-scale non-linear Niblack, 2) multi-scale non-linear Niblack and 3) our binarization (other binarization methods showed clearly inferior performance). Non-linear Niblack has been used in several previous works (e.g. [6]) in multi-scale fashion in order to achieve higher recall. In our experiments, we used three scales inside the non-linear Niblack method with varying window size, and performed non-maxima suppression of word candidates that overlap by more than 50%. Among the overlapping candidates we chose the one with higher average probabilistic score. The results of this comparison are shown in figure 4.

We now compare the results of this pipeline with other end-to-end pipelines reported in the literature. In the first case, we did not use any lexicon, but fixed the alphabet (as in [26]) and pruned out the recognition results that contained non alpha-numeric characters. The results are presented in the table III. In the second case, we used the lexicon provided

[7]available at *http://www.mathworks.com/products/image/*

Fig. 3. End-to-end text localization and recognition results of proposed binarization method (without lexicon).

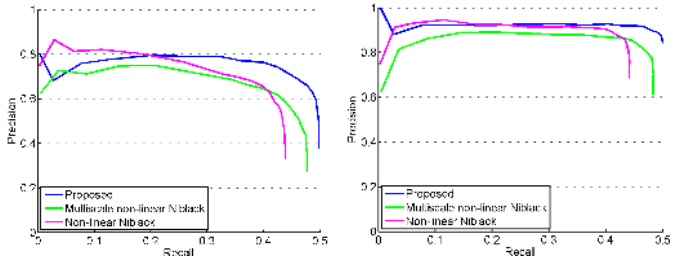

Fig. 4. Precision-recall curve for text localization and recognition on ICDAR 2003 test set. Left plot - end-to-end text recognition without lexicon, right plot - with fixed lexicon. .

with ground truth annotation to ICDAR 2003 dataset. The results are presented in table III. We can see that proposed binarization method significantly outperforms NL Niblack. So finally we selected our method and performed experiments on the ICDAR 2011 dataset with the results presented in table III comparing to the recent result of Neumann and Matas [28] (to the best of our knowledge, this is the only published result for end-to-end text understanding on this dataset).

Key Results and Observations. One can observe that, perhaps surprisingly, a pipeline based on image binarization and an off-the-shelf OCR achieves higher accuracy than some of the recent fancier methods. Non-linear Niblack and proposed method show better performance for text recognition without lexicon than existing methods, and the performance when using a lexicon is quite close to the very recent result in [27].

Conclusion. We have performed analysis of several image binarization techniques on the ICDAR 2003 and the ICDAR 2011 benchmarks. Overall, we have found that a pipeline consisting of an image binarization method and off-the-shelf OCR module was able to achieve state-of-the-art end-to-end text recognition on these challenging datasets.

REFERENCES

[1] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "An objective evaluation methodology for document image binarization techniques," in *DAS*, 2008, pp. 217–224.

[2] A. Mishra, K. Alahari, and C. V. Jawahar, "An MRF model for binarization of natural scene text," in *ICDAR*, 2011, pp. 11–16.

[3] T. Wakahara and K. Kita, "Binarization of color character strings in scene images using k-means clustering and support vector machines," in *ICDAR*, 2011, pp. 274–278.

[4] Y.-F. Pan, X. Hou, and C.-L. Liu, "Text localization in natural scene images based on conditional random field," in *ICDAR*, 2009, pp. 6–10.

[5] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1083–1090.

[6] T. Yamazoe, M. Etoh, T. Yoshimura, and K. Tsujino, "Hypothesis preservation approach to scene text recognition with weighted finite-state transducer," in *ICDAR*, 2011, pp. 359–363.

[7] K. Zhu, F. Qi, R. Jiang, L. Xu, M. Kimaci, Y. Wu, and T. Aizawa, "Using adaboost to detect and segment characters from natural scenes," in *Camera-Based Document Analysis and Recognition (CBDAR)*, 2005.

[8] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Trans. Systems, Man and Cybernetics*, vol. 9, pp. 62–66, 1979.

[9] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, vol. 19, pp. 41–47, 1986.

[10] J. Sauvola and M. Pietikinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, pp. 225–236, 2000.

[11] W. Niblack, "An introduction to digital image processing." Strandberg Publishing Company, 1985.

[12] N. Howe, "A laplacian energy for document binarization," in *ICDAR*, 2011, pp. 6–10.

[13] . Gatos, . Pratikakis, and P. S.J., "Text detection in indoor/outdoor scene images," in *CBDAR'05*, 2005, pp. 127–132.

[14] N. Ezaki, "Text detection from natural scene images: towards a system for visually impaired persons," in *In Int. Conf. on Pattern Recognition*, 2004, pp. 683–686.

[15] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *CVPR*, 2010.

[16] R. Minetto, N. Thome, M. Cord, J. Stolfi, F. Precioso, J. Guyomard, and N. J. Leite, "Text detection and recognition in urban scenes," in *ICCV Workshops*, 2011, pp. 227–234.

[17] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," in *ICCV*, 2001, pp. 105–112.

[18] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2004.

[19] C. Wolf and D. Doermann, "Binarization of low quality text using a markov random field model," in *Proc. Intl Conf. Pattern Recognition*, 2002, pp. 160–163.

[20] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," *IJDAR*, vol. 13, no. 4, pp. 303–314, 2010.

[21] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 document image binarization contest (DIBCO 2011)," in *ICDAR*, 2011, pp. 1506–1510.

[22] E. Badekas and N. Papamarkos, "Automatic evaluation of document binarization results," in *CIARP*, 2005, pp. 1005–1014.

[23] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," in *ICDAR*, 2009, pp. 1375–1382.

[24] A. Clavelli, D. Karatzas, and J. Lladós, "A framework for the assessment of text extraction algorithms on complex colour images," in *Document Analysis Systems*, 2010, pp. 19–26.

[25] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *IEEE International Conference on Computer Vision (ICCV)*, Barcelona, Spain, 2011.

[26] L. Neumann and J. Matas, "Estimating hidden parameters for text localization and recognition," in *Computer Vision Winter Workshop*, 2011.

[27] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Pattern Recognition (ICPR), 21st International Conference on*. IEEE, 2012, pp. 3304–3308.

[28] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *CVPR*, 2012, pp. 3538–3545.

[29] J. Friedman, T. Hastie, and R. Tibshirani, "Additive Logistic Regression: a Statistical View of Boosting," *The Annals of Statistics*, vol. 38, no. 2, 2000.