IJDAR (2015) 18:169–182
DOI 10.1007/s10032-015-0240-4

ORIGINAL PAPER

Fast and accurate scene text understanding with image binarization and off-the-shelf OCR

Sergey Milyaev · Olga Barinova · Tatiana Novikova · Pushmeet Kohli · Victor Lempitsky

Received: 3 February 2014 / Revised: 19 August 2014 / Accepted: 12 January 2015 / Published online: 4 February 2015

Abstract While modern off-the-shelf OCR engines show particularly high accuracy on scanned text, text detection and recognition in natural images still remain a challenging problem. Here, we demonstrate that OCR engines can still perform well on this harder task as long as an appropriate image binarization is applied to input photographs. We propose a new binarization algorithm that is particularly suitable for scene text and systematically evaluate its performance along with 12 existing binarization methods. While most existing binarization techniques are designed specifically either for text detection or for recognition of localized text, our method shows very similar results for both large images and localized text regions. Therefore, it can be applied to large images directly with no need for re-binarization of localized text regions. We also propose the real-time variant of this method based on linear-time bilateral filtering. Evaluation across different metrics on established natural image text recognition benchmarks (ICDAR 2003 and ICDAR 2011) shows that our simple and fast image binarization method combined with off-the-shelf OCR engine achieves state-of-the-art performance for end-to-end text understanding in natural images and outperforms recent fancy methods.

Keywords Natural scene binarization · Scene text localization · Document binarization

S. Milyaev (✉) · O. Barinova · T. Novikova
Lomonosov Moscow State University, 1-52 Leninskiye Gory,
119991 Moscow, Russia
e-mail: smilyaev@graphics.cs.msu.ru

O. Barinova
e-mail: obarinova@graphics.cs.msu.ru

T. Novikova
e-mail: tnovikova@graphics.cs.msu.ru

P. Kohli
Microsoft Research, Cambridge, UK
e-mail: pkohli@microsoft.com

V. Lempitsky
Skolkovo Institute of Science and Technology, Moscow, Russia
e-mail: lempitsky@skoltech.ru

1 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 [25]. There are two directions of research related to binarization of natural scene test. The first one [27,35,38] uses image binarization as a part of text detection pipeline for detection of letter candidates that are further classified and grouped into text lines. Another direction of research studies binarization of cropped word images assuming that text localization is done at the previous step of a pipeline [20,30]. These two use-cases require different properties from binary maps and therefore result in different binarization methods.

In this work, we propose a new binarization method that is particularly suitable for text in natural images. The method does not require any information about the position and size of the text in an image and shows very similar results when applied to large images for text detection and when applied to cropped text regions for text recognition; therefore, only one round of image binarization is required for the end-to-end scene text understanding pipeline.

We describe two variants of our method. The first variant embeds local binarization into a global optimization framework. The global optimization problem is formulated in terms of a Markov random field (MRF) model, which can be solved using graph-cut inference [4]. The second variant of the method is based on linear-time bilateral filtering and therefore enables real-time processing of video sequences.

We systematically evaluate the performance of our method and compare it to several previous image binarization techniques 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. The comparative analysis includes two main use-cases of scene text binarization: for text localization and for cropped text recognition.

For the comparison of segmentation accuracy, we have manually created a new pixel-wise annotation of ICDAR 2003 dataset. An example of this new ground truth annotation is shown in Fig. 1b. Crucially, such re-annotation allows a principled comparison of different binarization approaches. As a result of this evaluation, we select the top methods for text localization and for cropped word recognition and compare them within the most interesting end-to-end text detection and recognition scenario.

As we demonstrate, our method shows superior results in terms of the OCR accuracy compared to existing binarization methods and demonstrates even more competitive performance w.r.t. recently presented pipelines for text understanding. The use of linear-time bilateral filtering allows real-time performance with just a little loss of accuracy compared to global optimization.

2 Related work

We first provide a very brief review of existing binarization methods that we have considered. These methods can be roughly divided into two groups: The first group uses a fixed threshold for a given image [16,26], while the second group (*local binarization*) uses local thresholds [24,29]. 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 parameters (e.g., the characteristic scale). In particular, optimal parameter values may differ for text of different sizes even within a single image. For this reason, some text detection and recognition pipelines [27] precede local binarization with the local text scale estimation step.

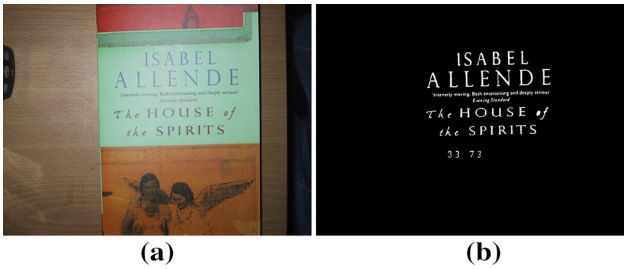
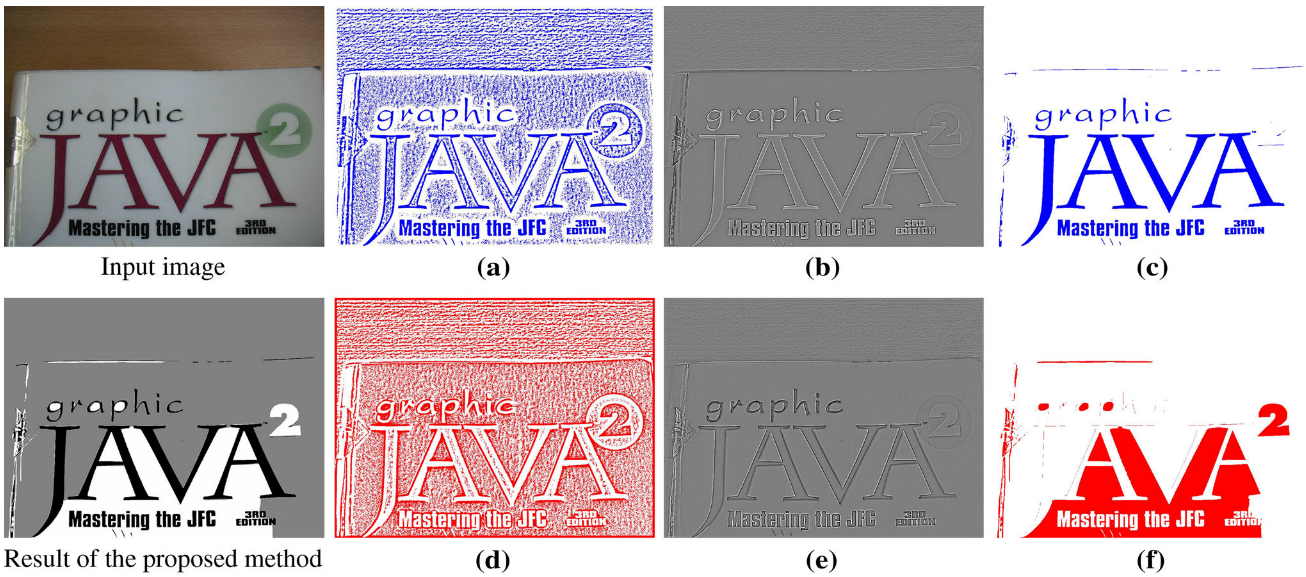


Fig. 1 **a** Example image from ICDAR dataset. **b** Image binarization result (pixel-wise ground truth annotation)

Several methods for text binarization in natural images have been proposed more recently. For instance, Zhu et al. [40] suggested using the ordered statistics filter for estimating thresholds in the nonlinear Nilblack decomposition. Gatos et al. [11] 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 [8] proposed generating connected components by combination of mathematical morphology operations, edge extraction and Otsu thresholding of image color channels. Epshtein et al. [7] suggested using a new image operator (*Stroke Width Transform*) to segment letters. Minetto et al. [19] proposed using *toggle mapping* for character segmentation in a multiresolutional way since natural scene images have large character size variations and strong background clutter.

Howe [13] proposed to use the Laplacian of the image intensity for scanned document binarization within a Markov random field model. This algorithmic setup is similar to the one we propose below. While Howe works only with edge detection result, we use color and distance information for nonlocal aggregation of image evidence in our method. This allows us to effectively process images containing background clutter, in contrast to Howe method which is very sensitive to the presence of outlier edges in the image.

Other recent works [20,30] focus on the binarization of cropped text assuming that the text is correctly localized in the preceding 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. [20] proposed a method for text binarization using iterated graph cut. Wakahara and Kita [30] proposed



Input image (a) (b) (c)

Result of the proposed method (d) (e) (f)

Fig. 2 The steps of the proposed binarization method. **a**, **d** Local binarization for *dark text on light background* and *light text on dark background*, respectively. The candidate text regions are shown in *blue* and *red*. **b**, **e** The seeds resulting from incorporating local binarization and the Laplacian of the image intensity that allows to detect *dark text on light background* and *light text on dark background*, respectively. **c**, **f** The binarization after global optimization for *dark text on light background* and *light text on dark background*, respectively. The candidate text regions are shown in *blue* and *red*. Notice how the global optimization significantly refines the initial local binarization by filling holes inside the letters and reducing number of generated connected components (color figure online)

a method based on k-means clustering and letter candidates classification for a similar cropped image scenario.

The preliminary version of our approach was presented in [18]. The main contributions of this paper compared to [18] are twofold. Firstly, we propose a real-time variation of our method with the use of linear-time recursive bilateral filter. Second, we provide evaluation of the binarization methods in the context of recognition of localized text.

3 Proposed method

We propose a new binarization algorithm that consists of the following steps:

1. local binarization producing *seed pixels*,
2. seed pixel strength estimation,
3. construction of binary maps by aggregation of nonlocal evidence and
4. trimap construction

In particular, at the first step we perform local binarization with a rather small window size (Niblack in our experiments), since using large window size inside local binarization 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 (Fig. 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 aggregate the evidence nonlocally, thus accounting for pixel similarity and correcting errors of initial labeling. Figure 2 illustrates the steps of our algorithm.

As long as the text in natural images can be either darker than background or lighter than background, we apply nonlocal aggregation for both cases, hence obtaining two binary maps. These two maps are combined into a single trimap in the following way. A pixel is assigned a "dark text" label if it is marked as foreground in the binary map for dark text and marked as background in the binary map for light text; it is assigned a "light text" label if it is marked as foreground in the binary map for light text and marked as background in the binary map for dark text. Otherwise, it is assigned a "background" label.

3.1 Global optimization

For global optimization, we construct an energy function

$$E(\mathbf{f}|I,\mathbf{n}) = E_{\text{local}}(\mathbf{f}|I,\mathbf{n}) + E_{\text{smooth}}(\mathbf{f}|I), \qquad (1)$$

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.

In more detail. the unary term is defined as:

$$E_{\text{local}}(\mathbf{f}|I, \mathbf{n}) = \sum_{i=1}^{N} e_{\text{local}}(i), \tag{2}$$

where

$$e_{\text{local}}(i) = \begin{cases} \nabla^2 I'_i, & f_i = n_i \\ 1 - \nabla^2 I'_i, & f_i \neq n_i \end{cases}. \tag{3}$$

Here, $\nabla^2 I'_i$ denotes Laplacian of the image intensity normalized to its maximum value. This form of unary potential is related to the flux optimization [5,15]

We use a conventional contrast-sensitive pairwise term traditional to graph-cut segmentation [3]:

$$E_{\text{smooth}}(\mathbf{f}|I) = \lambda \sum_{(i,j)\in\mathbf{N}} e_{\text{smooth}}(i, j), \tag{4}$$

defined by pixel similarity:

$$e_{\text{smooth}}(i, j) = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_g^2} - \frac{\|c_i - c_j\|^2}{2\sigma_c^2}\right), & f_i \neq f_j \\ 0, & f_i = f_j \end{cases}, \tag{5}$$

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 [4].

3.2 Real-time variant with linear-time bilateral filtering

For many practical applications, the graph-cut inference described above can be prohibitively slow. Within our method, this inference is used to refine the results of local binarization by nonlocal aggregation of the evidence across the image. An alternative approach to aggregation of nonlocal evidence based on Gestalt grouping principles of similarity and proximity has been proposed for the task of stereo reconstruction [39]. The idea of this approach is to refine the local estimates by computing the weighted sum of the local results for neighboring pixels. The support weight of each neighboring pixel is computed according to the strength of grouping by similarity and proximity. The more similar the color of a pixel, the larger is its support weight. In addition, the closer the pixel is, the larger the support weight is. The former is related to the grouping by similarity, and the latter is related to the grouping by proximity. This idea is also related to *guided filter* [12].

These two rules can be treated as a single rule in an integrated manner and lead to a formula for the support weights similar to the one used in bilateral filtering. The support weight $w_i(j)$ of a pixel j for the pixel i is computed as

$$w_i(j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_g^2} - \frac{\|c_i - c_j\|^2}{2\sigma_c^2}\right), \tag{6}$$

where x denotes pixel coordinates, c means RGB color, σ_g and σ_c are normalization constants.

Thus, to obtain one binary map in step 3 of our method, we can apply such aggregation to the results of local binarization obtained in step 1 using the strengths computed in step 2:

$$h_i(c) = \sum_{j:n_j=c} w_i(j) e_{\text{local}}(j), \tag{7}$$

where n_j is the result of local binarization performed in step 1, and $e_{\text{local}}(j)$ is the value of the unary potential computed using Eq. (3). As a result, we obtain two refined maps of votes $h_i(c)$ for each class c (which can be 1 or 0 corresponding to dark/light text and background respectively), and then, each pixel is assigned a label that corresponds to the larger vote.

The summation in (7) includes all image pixels labeled as text by local the binarization. The brute-force implementation of (7) is in $O(n^3)$ time, where n is the number of pixels in the image, which is prohibitively high. In a practical implementation of the bilateral filter, only pixels within distance r from the pixel i are usually taken into account. The complexity then becomes $O(nr^2)$, which is still very high when the radius r is large.

A few $O(n)$ time algorithms for bilateral filtering based on histograms have been developed [1,36], which require a high quantization degree to achieve satisfactory speed, but at the expense of quality degradation. In [37], an exact recursive implementation of the bilateral filter with computational and memory complexity linear in the size of input image and its dimensionality has been proposed. It uses a new range filter kernel and adopts any spatial filter kernel that can be recursively implemented.

We have investigated the use of recursive bilateral filtering [37] as a faster substitute for graph-cut-based inference. As long as such variant of our algorithm requires just two passes of bilateral filter over the image and linear-time local binarization, it also has $O(n)$ complexity. For reference, the time complexity of nonlinear Niblack method that showed second best performance in our experiments is $O(nw^2)$, which is significantly larger given that the optimal value of the window size parameter w chosen in our experiments is equal to $1/16$ of image height.

The $O(n)$ computational time implies that the time complexity is independent of the kernel radius r, so we are free to use arbitrary kernel sizes. Therefore, in this work we used the whole image as a support for each image pixel.

4 Performance evaluation

4.1 Evaluation metrics

4.1.1 Pixel-wise accuracy

We manually produced pixel-level binary masks for the ground truth characters segmentation in the ICDAR 2003 database. With new pixel-level annotation, we evaluated standard accuracy measures including precision, recall, F-score and peak signal-to-noise ratio (PSNR). In all experiments, the pixel-wise metrics were computed over the cropped word images using database ground truth annotation for word bounding boxes. We compared the binarization result to the pixel-wise ground truth segmentation of the letters available in the annotation of ICDAR dataset. Precision and recall are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{8}$$

where $\text{TP}, \text{FP} and \text{FN}$ denote the number of true-positive, false-positive and false-negative segmented pixels. F-score is defined in following way:

$$F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{9}$$

The PSNR measure is defined as

$$PSNR = 10 \log \left(\frac{C^2}{\text{MSE}} \right), \tag{10}$$

where $C = 1$ and

$$MSE = \frac{\sum_{x=1}^{M} \sum_{y=1}^{N} (I(x, y) - I'(x, y))^2}{\text{MN}}, \tag{11}$$

where M and N are image width and height, respectively.

4.1.2 Morphological metrics

Although pixel-wise metrics are widely used in comparative analysis of document binarization techniques (see [10,28]), they do not describe morphological structure of the generated segment. Such morphological properties however drastically affect the accuracy of text recognition (see e.g., [6]). For example, in Fig. 3a), a character “c” may be segmented as bold “c” (Fig. 3b) by one segmentation method and as “o” (Fig. 3c) by another one. Pixel-wise metrics in these two cases can be very close, while recognition results would be completely different. Another example is when “c” character is segmented as eroded character (Fig. 3d) by one method and as part of character (Fig. 3e) by another method. Again while the pixel-wise metrics are similar, the recognition results would be different. Therefore, to better evaluate the segmentation accuracy, one needs quantitative metrics that take into account the morphological properties of the segmented characters (connected components).

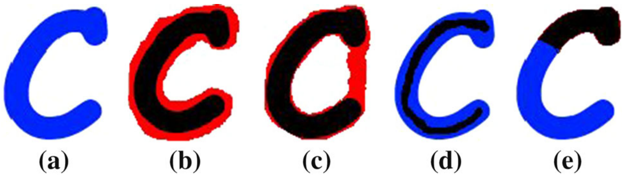


Fig. 3 An example of character segmentation. **a** Ground truth segmentation. **b** Dilated result. **c** Deformed result. **d** Eroded result. **e** Fractioned result. While traditional error measures based on pixel-wise metrics for (**b**)–(**c**) are similar, the effects on the recognition are very different

In our comparison, we therefore included the metrics proposed in [6] for the assessment of segmentation performance. These metrics require ground truth skeletons for each character. They consider *minimal coverage* and *maximal coverage criteria*. To meet the minimal coverage criterion, the connected component should cover more than $T_{\min} = 0.9$ of a character skeleton. To pass the maximal coverage criteria, all pixels of a connected component should lie close to the boundary of a character. Namely, the distance between the connected component and the character boundary should be less than $T_{\max} = \max(5, 0.5 * G)$, where G is the maximum stroke width of a character.

Using these criteria, each connected component can be classified into one of the following types:

- *Background* A connected component that does not overlap with any of the character skeletons.
- *Whole* A connected component that overlaps with a single character skeleton and satisfies the minimal coverage criterion and maximal coverage criterion.
- *Fraction* A connected component that overlaps with a single character skeleton and satisfies the maximal coverage criterion, but does not satisfy the minimal coverage criterion.
- *Multiple* A connected component that overlaps several character skeletons and satisfies minimal coverage criterion and maximal coverage criterion.

- *Fraction and Multiple* A connected component that overlaps several character skeletons and satisfies maximal coverage criterion but does not satisfy minimal coverage criterion.
- *Mixed* otherwise.

The percentage of connected components of each type characterize the quality of image binarization. Thus, large fraction of *Whole* components intuitively correspond to better binarization performance, while larger fraction of *Mixed* components correspond to worse text binarization. In the experiments, we also investigate the correlations of different morphological metrics with OCR accuracy.

4.1.3 OCR accuracy

For testing the accuracy of word recognition of binarization result, we used a commercial OCR system for document recognition *Omnipage Professional 18*[1] and the state-of-the-art open source OCR software *Tesseract*.[2] Again we cropped the word bounding boxes from binarization results using ground truth word annotation. Then, we selected two-label binary word image ("light-vs-all" and "dark-vs-all") with highest F-score of pixel-level accuracy for further OCR processing. This assumption implies that the segmentation process is guided by an ideal algorithm for determining text polarity.

4.2 Evaluated methods and datasets

We selected 12 different binarization methods for evaluation.[3] We have included methods commonly used for document images, namely Otsu [26], Kittler and Illingworth [16], Niblack [24] and Sauvola and Pietikinen [29]. We have also included several recent methods for document binarization, namely Wolf [33],[4] Howe [13] and Lu et al. [17],[5] the last one being a runner-up at ICDAR 2011 Document Image Binarization Contest (DIBCO 2011) [28]. We have also included methods developed for natural images: Ezaki [8], Gatos [11], Minetto et al. [19] and nonlinear Niblack decomposition [40]. Finally, we have also included the method based on *stroke width transform* (*SWT*) from [7] implemented in text localization system.[6]

We have validated the parameters of all local binarization methods on the training part of the ICDAR 2003 dataset in order to achieve the maximum OCR accuracy. The parameters of Niblack method were set as suggested in [27]. The parameters for the Sauvola method were set as suggested in [2]. For [8,11,19,40], we used parameters suggested by the authors. For the proposed method, we set k to 0.4 as in [27] 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.

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.

To be able to measure the segmentation accuracy, we have performed a pixel-level annotation for ICDAR 2003 dataset.[7] This dataset contains 249 images in the training set and 251 images in the test set. The images are in full color and vary in size from 307 × 93 to 1,280 × 960 pixels.

4.3 Binarization as a preprocessing step for text localization

In the first set of experiments, we applied the methods listed above to 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.

While the methods were applied to uncropped images, for the analysis we used only the interior parts of the ground truth word bounding boxes. We cropped these bounding boxes from the results of image binarization methods and evaluated their performance across the metrics described in Sect. 4.1.

The results of evaluation are summarized in Table 1. Qualitative results are shown in Fig. 6.

The most popular methods for document image binarization like Otsu [26], Kittler and Illingworth [16], Sauvola and Pietikinen [29] show significantly degraded performance on natural scenes. In the cases when color and illumination variations are high, global thresholding methods [16,26] 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. For example, 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. [17] 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

[1] http://www.nuance.com/.

[2] http://code.google.com/p/tesseract-ocr.

[3] http://graphics.cs.msu.ru/en/research/projects/msr/text.

[4] http://liris.cnrs.fr/christian.wolf/software/binarize/index.html.

[5] http://www.comp.nus.edu.sg/~subolan/.

[6] https://sites.google.com/site/roboticssaurav/strokewidthnokia.

[7] http://graphics.cs.msu.ru/en/research/projects/msr/text.

Table 1 Comparison of the binarization methods across a number of accuracy measures on the ICDAR 2003 dataset

Method	Prec.	Rec.	*F*-sc.	PSNR	Backgr.	Whole	Fract.	Mult.	F. and M.	Mixed	OCR1 acc.	OCR2 acc.
Otsu [26]	0.79	0.85	78	8.85	1.79	0.43	0.33	0.02	0.01	0.07	47.1	34.4
Kittler and Illingworth [16]	0.70	**0.89**	72	7.36	0.93	0.32	**0.25**	0.03	**0.01**	0.01	35.1	24.3
Niblack [24]	0.90	0.80	**84**	10.05	23.57	0.60	1.48	0.02	0.02	0.04	56.0	35.8
Sauvola and Pietikinen [29]	0.90	0.66	73	9.62	4.05	0.47	0.84	0.02	0.01	**0.02**	53.8	40.1
NL Niblack [40]	0.93	0.73	79	10.34	4.05	0.47	0.84	0.02	0.01	0.02	59.3	45.4
Howe [13]	0.81	0.66	71	9.01	**0.61**	0.46	0.32	0.01	0.01	0.03	53.2	41.2
Gatos et al. [11]	0.90	0.68	75	9.80	0.88	0.50	0.56	0.02	0.01	0.03	56.2	44.8
Ezaki [8]	0.85	0.82	82	9.61	2.57	0.43	0.43	0.03	0.02	0.05	47.6	30.9
Minneto et al. [19]	0.87	0.79	82	9.41	2.90	0.50	0.42	0.02	0.02	0.05	47.3	22.8
Epstein et al. [7]	0.81	0.85	82	9.40	1.24	0.44	0.42	0.01	0.03	0.12	47.6	29.8
Wolf and Doermann [33]	0.88	0.66	72	9.59	4.17	0.48	0.78	0.02	0.01	0.02	53.4	40.4
Lu et al. [17]	0.87	0.66	73	8.80	1.92	0.43	0.63	**0.01**	0.01	0.04	52.2	40.7
Proposed	**0.91**	0.78	82	**10.44**	2.22	**0.64**	0.33	0.02	0.01	0.03	**63.5**	**48.6**

Number of segments is divided by the number of ground truth characters in dataset. See the text for more details
Bold numbers means the best result among others in each category

method and its widely used nonlinear modification achieve high OCR accuracy. While the method of Howe [13] uses Laplacian-based unary terms similar to our method, it shows significantly lower accuracy in the case of natural images with complex backgrounds. Our methods show highest OCR accuracy, which we believe is due to better choice of unary and pairwise terms inside the global optimization compared to Howe's 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 the OCR accuracy. For example, Niblack method, which has the highest *F*-score, is only fourth in terms of OCR accuracy. And vice versa, the nonlinear Niblack method which has a mediocre pixel-level results shows very high recognition accuracy. A consequence of this observation is that structured output machine learning of binarization techniques based on the 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 nontext 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. Comparsion of the words binarization is presented on Fig. 4.

4.4 Binarization for recognition of localized text

Some end-to-end text understanding pipelines first perform text detection, then apply image binarization to localized text regions, and pass the result to the OCR module. In this scenario, the approximate scale of the text can be inferred from the size of detected text region. Some binarization techniques (e.g., local binarization methods such as Niblack and Sauvola) benefit significantly from the knowledge of the text size. Their performance may increase if the parameters are selected according to the size of detected text.

At the same time, text localization is not always accurate. Thereby, binarization methods have to be robust to errors in text localization, and their performance should not drop significantly in case when text is localized inaccurately. In the next series of experiments, we measure the sensitivity of binarization methods to errors in text localization.

We created a dataset of randomly shifted and scaled word bounding boxes using ICDAR 2003 ground truth annotation. The shift vector for each image was chosen at random, so as the resulting cropped image contains at least one word bounding box from the ground truth annotation. We used the following set of scaling factors: 1, 1.2, 1.5, 1.7, 2. We measured segmentation assessment metrics in the region of original word bounding box and OCR accuracy for binarized cropped word. Since generated bounding boxes sometimes contain letters from the neighboring words, recognition result of the OCR module for the generated bounding box may contain some extra symbols. Therefore, we consider that a word is correctly recognized if the resulting string contains a

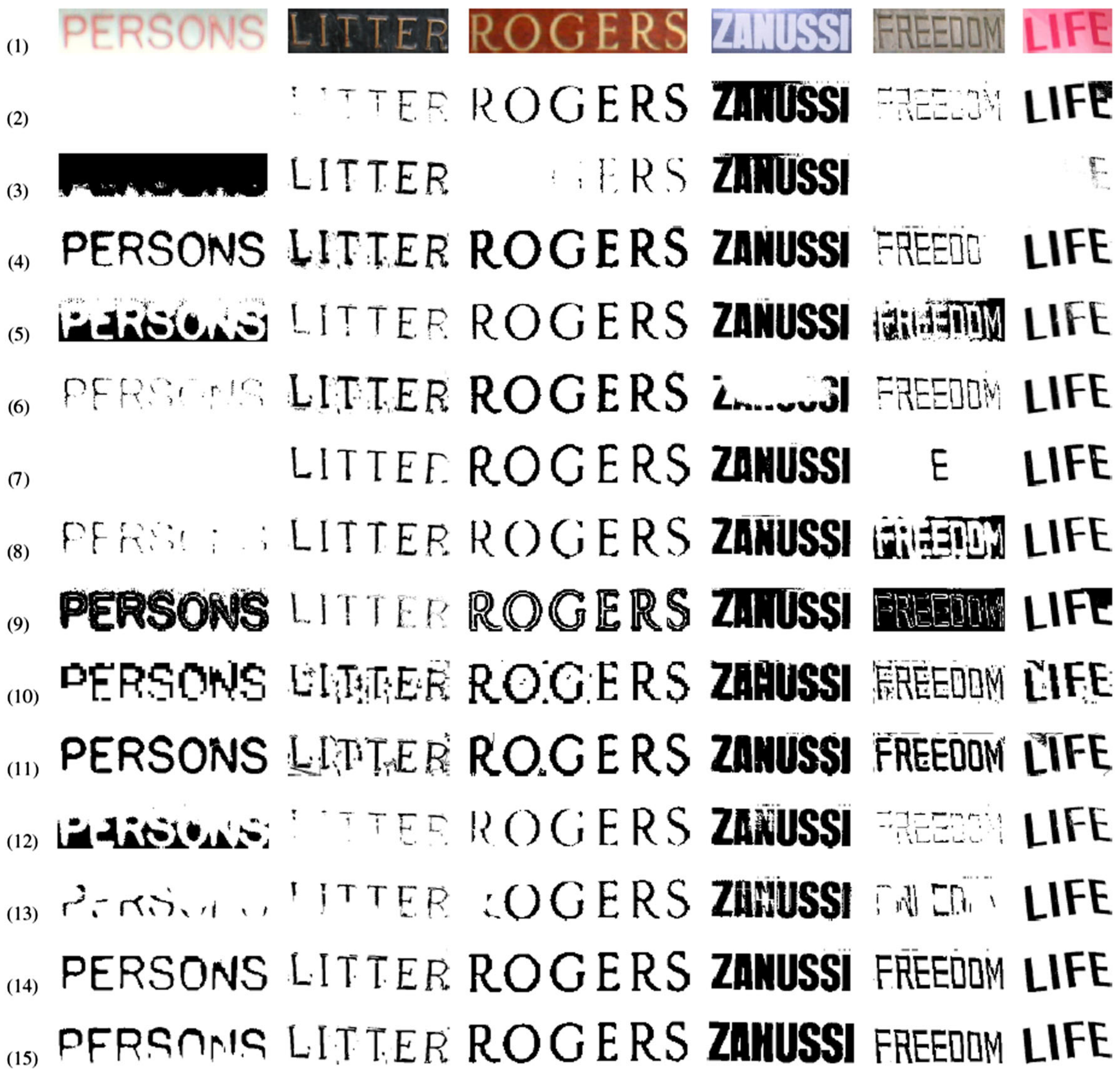


Fig. 4 Comparison of binarization methods. From *top* to *bottom line*: (*1*) original image, (*2*) Otsu, (*3*) Kittler et al., (*4*) Niblack, (*5*) Sauvola, (*6*) Nonlinear Niblack, (*7*) Howe et al., (*8*) Gatos et al., (*9*) Ezaki et al., (*10*) Minetto et al., (*11*) Epshtein, (*12*) Wolf et al., (*13*) Su et al., (*14*) Proposed (graph-cut inference), (*15*) Proposed (recursive bilateral filter inference)

word from ground truth annotation even if it contains a few more characters before or after this word.

In the experiments from this section, we compared only the methods designed specifically for text binarization in localized text regions and excluded the methods designed for generating connected components for text detection. Namely, we did not include the methods by Ezaki [8], Minnetto et al. [19], Epshtein et al. [7], nonlinear Niblack [40] and Lu [17] in this comparison.

We used the adaptive window size estimated from the height of a word bounding box. The ratio between the window size and bounding box height was validated on the training part of ICDAR 2003 database so as to maximize the OCR accuracy.

The results are shown in Fig. 5. The segmentation assessment metrics show strong correlation with OCR accuracy. With increased size of bounding box, the number of well-segmented characters does not change noticeably, but the number of background and mixed segments increases, which causes degraded OCR accuracy.

In case of precisely localized text, the methods based on global thresholding (e.g., Otsu) may provide good results. With inaccurate text localization, global thresholding becomes more problematic, which significantly decreased the accuracy of such methods.

Information about text size is crucial for the performance of methods based on local thresholding. Appropriate choice of window size allows them to segment the letters accurately

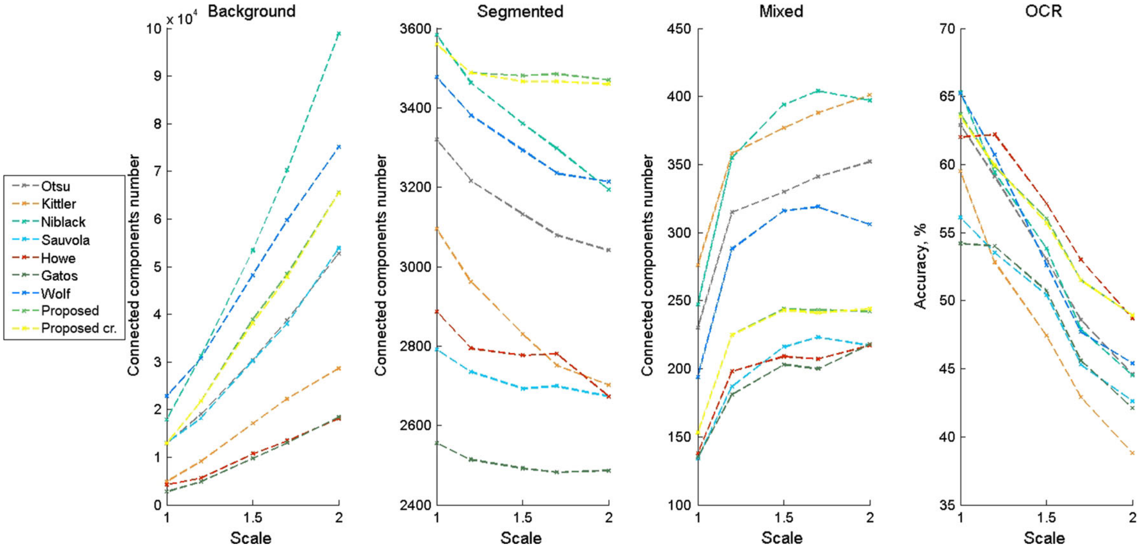


Fig. 5 Performance evaluation of binarization methods for processing of localized text regions. The plots of morphological metrics versus the size of the localized text region are shown. Larger size of localized region corresponds to more inaccurate text localization

and suppress noise which leads to good OCR accuracy. However, while popular binarization techniques like Niblack perform well in case of precise text localization, their performance drops when localization errors occur and the choice of window size for local thresholding becomes less effective.

On average, the highest OCR accuracy is achieved with the use of binarization methods based on nonlocal aggregation of image evidence such as our method and Howe's method. These methods have also demonstrated robustness to errors in text localization.

Howe method however has shown mediocre results for the binarization of whole images. This significant difference in performance can be explained as follows. In localized text regions, large values of Laplacian usually correspond to the boundary of text with rare exceptions. The asymmetric bias on background label used in this method can handle a reasonable amount of outlier pixels, and thus, it produces fewer noisy segments and as a consequence achieves higher OCR accuracy. At the same time, these properties of the Howe method lead to degraded performance in case of large images where the distribution of Laplacian is more complex compared to cropped images.

Interestingly, our method shows very similar results when applied to whole images and when applied to cropped text regions. To demonstrate this, we added the results for crops from the original large images binarized with our proposed method ("Proposed cr.") to Fig. 5. One can observe that the OCR accuracy for such crops is very close to the OCR accuracy for binarization results of the cropped regions. Therefore, with our method, there is no need to perform re-binarization of the localized regions, since the binarization of the whole image can be used for both text detection and further text recognition.

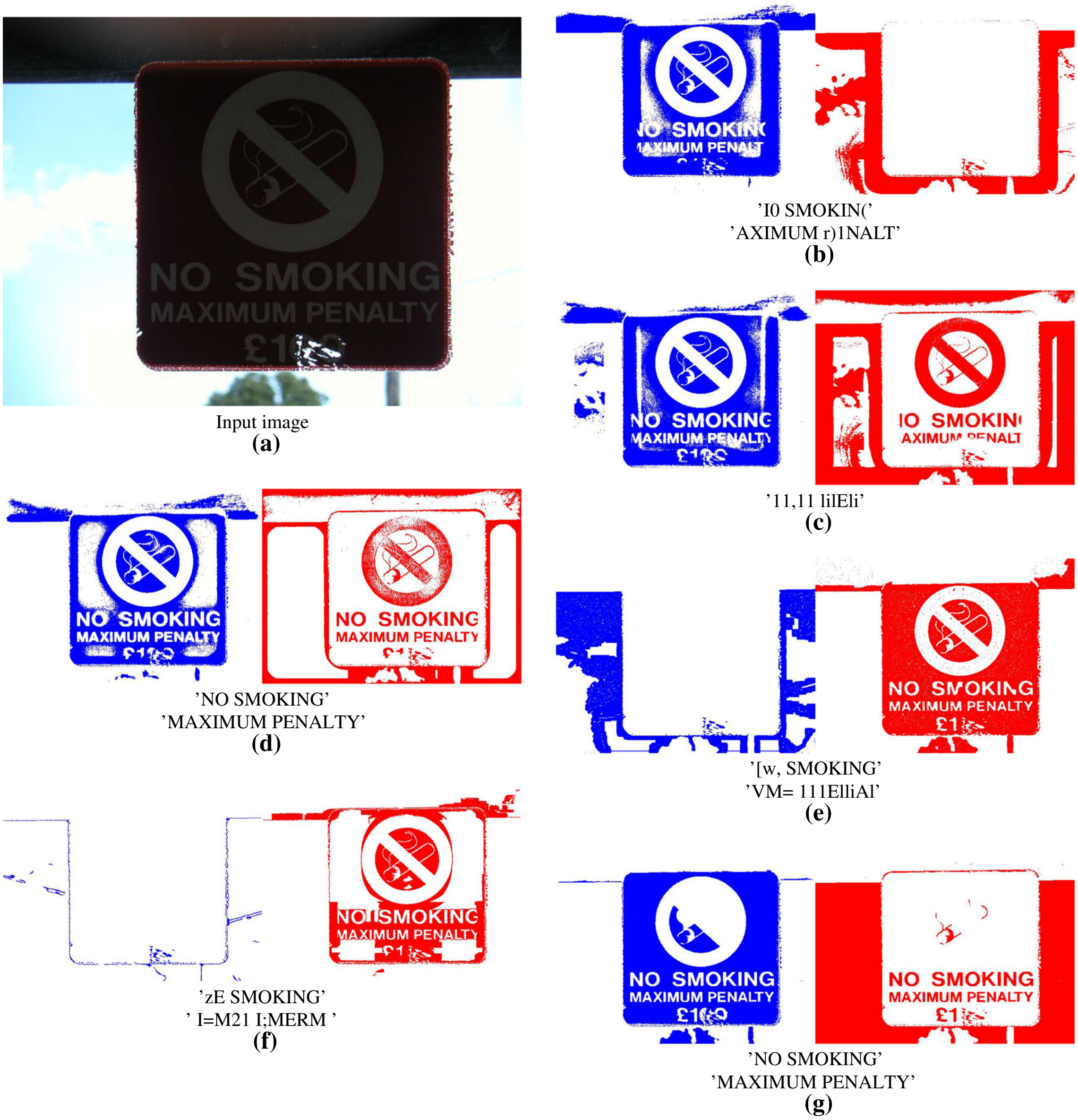
Finally, we performed evaluation on cropped word binarization and recognition on test dataset from ICDAR 2013 Robust Reading Competition Challenge 2 Task 3 [14]. We applied our binarization method to the test images and then passed the result to OCR engine. The recognition accuracy of this framework was 60.3 %. Compared to top submissions participating in this challenge that use binarization and recognition by OCR engine, such as *NESP* (64.2 %), *MAPS* (62.7 %) and *PLT*(62.4 %) [14], our method has close performance although it was not designed for cropped word binarization as these methods (Fig. 6).

4.5 End-to-end text understanding

4.5.1 Implementation details

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/noncharacter 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 the *Matlab Image Processing Toolbox*[8] (area, width, height, aspect ratio, length ratio,

[8] http://www.mathworks.com/products/image/.



Input image
(a)

’I0 SMOKIN(’
’AXIMUM r)1NALT’
(b)

’11,11 lilEli’
(c)

’NO SMOKING’
’MAXIMUM PENALTY’
(d)

’[w, SMOKING’
’VM= 111ElliAl’
(e)

’zE SMOKING’
’ I=M21 I;MERM ’
(f)

’NO SMOKING’
’MAXIMUM PENALTY’
(g)

Fig. 6 An example of the image binarization (*red* segments correspond to light text regions on dark background, *blue* segments correspond to dark text regions on light background): **a** original image, **b** Sauvola, **c** Niblack, **d** Nonlinear Niblack, **e** Howe, **f** Lu et al., **g** the proposed approach (color figure online)

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 (the ratios of differences between their top and bottom vertical coordinates to the minimum of their heights are not exceed 0.75, their intersection on horizontal axis should be lower than 0.75 of the minimum of their widths, and their horizontal distance should be lower and two maximums of their widths, (2) they had labels of the same type (“dark text” or “light text”), and (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 cannot 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 candidate that passed the filters, we computed the average probabilistic classifier output for the segments that

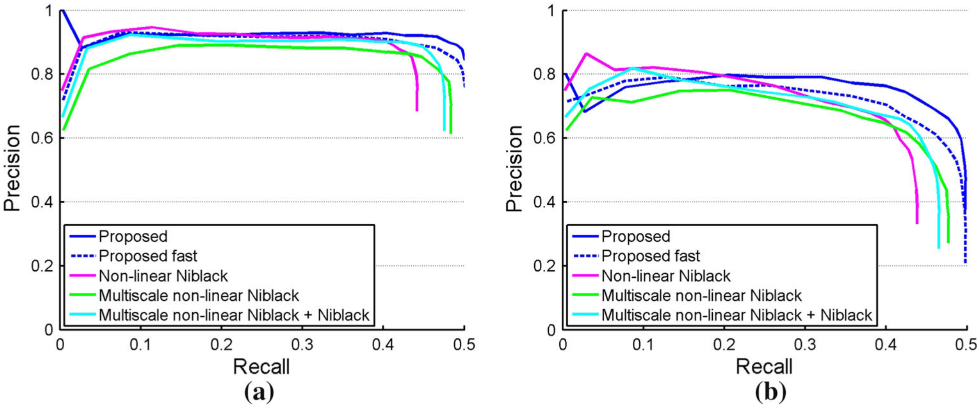


Fig. 7 Precision–recall curve for text localization and recognition on ICDAR 2003 test set. Left plot (**a**) end-to-end text recognition without lexicon, right plot (**b**) with fixed lexicon composed form all words in the ICDAR 2003 dataset. Note that multiscale nonlinear Niblack shows higher recall than its single-scale version; however, the precision of multiscale Niblack is significantly lower. Our method shows higher recall compared to single-scale nonlinear Niblack and higher precision compared to multiscale nonlinear Niblack

Table 2 The end-to-end text understanding accuracy on the ICDAR 2003 dataset

Method	Precision	Recall	*F*-measure
Wang [31] (no lexicon)	0.54	0.30	0.38
Neumann and Matas (no lexicon) [21]	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
Multiscale NL Niblack + Niblack (no lexicon)	0.64	0.43	0.51
Proposed (no lexicon)	**0.66**	**0.48**	**0.55**
Proposed fast (no lexicon)	0.61	0.45	0.52
NL Niblack (no lexicon) α-num	0.61	0.47	0.53
Multiscale NL Niblack (no lexicon) α-num	0.56	0.50	0.52
Multiscale NL Niblack + Niblack (no lexicon) α-num	0.58	0.50	0.54
Proposed (no lexicon) α-num	**0.66**	0.52	**0.58**
Proposed fast (no lexicon) α-num	0.56	**0.53**	0.54
Wang [31] (fixed lexicon)	0.45	**0.54**	0.51
NL Niblack (fixed lexicon)	0.85	0.44	0.58
Multiscale NL Niblack (fixed lexicon)	0.81	0.47	0.59
Multiscale NL Niblack + Niblack (fixed lexicon)	0.83	0.47	0.60
Proposed (fixed lexicon)	**0.88**	0.50	**0.63**
Proposed fast (fixed lexicon)	0.84	0.48	0.61
Wang [32] α-num (fixed lexicon)	–	–	**0.67**
NL Niblack (fixed lexicon) α-num	0.77	0.51	0.61
Multiscale NL Niblack (fixed lexicon) α-num	0.72	0.54	0.62
Multiscale NL Niblack + Niblack (fixed lexicon) α-num	0.77	0.53	0.63
Proposed (fixed lexicon) α-num	**0.80**	0.57	**0.67**
Proposed fast (fixed lexicon) α-num	0.76	0.56	0.64

The ability to correctly localize and recognize words is evaluated. The fixed lexicon comprises all words that occur in the ICDAR 2003 dataset Bold numbers means the best result among others in each category

constitute this word (a sigmoid transform [9] is used to map the outputs of boosted classifier to probabilities). By varying the threshold on the average probability, we generated the recall–precision curve.

As shown in the previous experiment, the binarization of localized text region can provide more accurate results of characters segmentation for further recognition, compared to the result of whole image binarization. So, one of the

Table 3 End-to-end text understanding accuracy on the ICDAR 2011 dataset

Method	Precision	Recall	F-measure
Neumann and Matas [22]	0.37	0.37	0.36
Neumann and Matas [23]	0.45	0.45	0.45
Proposed (no lexicon)	0.66	0.46	0.54
Proposed fast (no lexicon)	0.64	0.43	0.52
Proposed (fixed lexicon) α-num	**0.89**	**0.49**	**0.64**
Proposed fast (fixed lexicon) α-num	0.88	0.48	0.62

The combination of image binarization and OCR surpasses the results of the recently presented system [22] considerably
Bold numbers means the best result among others in each category

approaches for recognition of localized text regions is to perform its binarization using an appropriate method.

4.5.2 Evaluated methods

Here, we report results for three different binarization strategies: (1) single-scale nonlinear Niblack, (2) multiscale nonlinear Niblack, (3) multiscale nonlinear Niblack with Niblack method binarization of localized text regions, (4) our binarization with graph-cut-based global optimizations and the (5) fast variant of our method using recursive bilateral filter. Nonlinear Niblack method was selected because other binarization methods showed clearly inferior performance on whole image binarization, and Niblack method was selected to its top results for localized word binarization. Nonlinear Niblack has been used in several previous works (e.g., [35]) in multiscale fashion in order to achieve higher recall. In our experiments, we used three scales inside the nonlinear Niblack method with varying window size and performed nonmaxima 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 Fig. 7.

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 [21]). The results are presented in the Table 2. In the second case, we used the lexicon composed of all words from ICDAR 2003 dataset (about 1,000 words in total). The results are presented in Table 2. The results are presented for considering all ground truth words as well as alpha-numeric metric (ground truth words were considered whose length is >2 and contain only character or numeric symbols). We can see that proposed binarization method significantly outperforms NL Niblack, even with the step with additional binarization of localized text regions by Niblack method. So, finally we selected our method and performed experiments on the ICDAR 2011 dataset with the results presented in Table 3 comparing to the very recent result of Neumann and Matas [22,23] (to the best of our knowledge, this is the only published result for the end-to-end text understanding on this dataset).

The performance of text localization step (generating words candidates) on ICDAR 2013 dataset is *precision* = 0.65, *recall* = 0.54 and *f-measure* = 0.59 using methodology proposed by Wolf [34] that was used in ICDAR 2011 and ICDAR 2013 Robust Reading Competitions [14,28]

One can see that results of the fast variant of our method are close to the results of the version that uses the graph-cut inference. The running time of this modification is significantly lower (about 0.05 s compared to 0.5–2 s for graph-cut-based version on Intel i7 2.8 GHz). Example of the end-to-end recognition output of the proposed method is presented on Fig. 8.

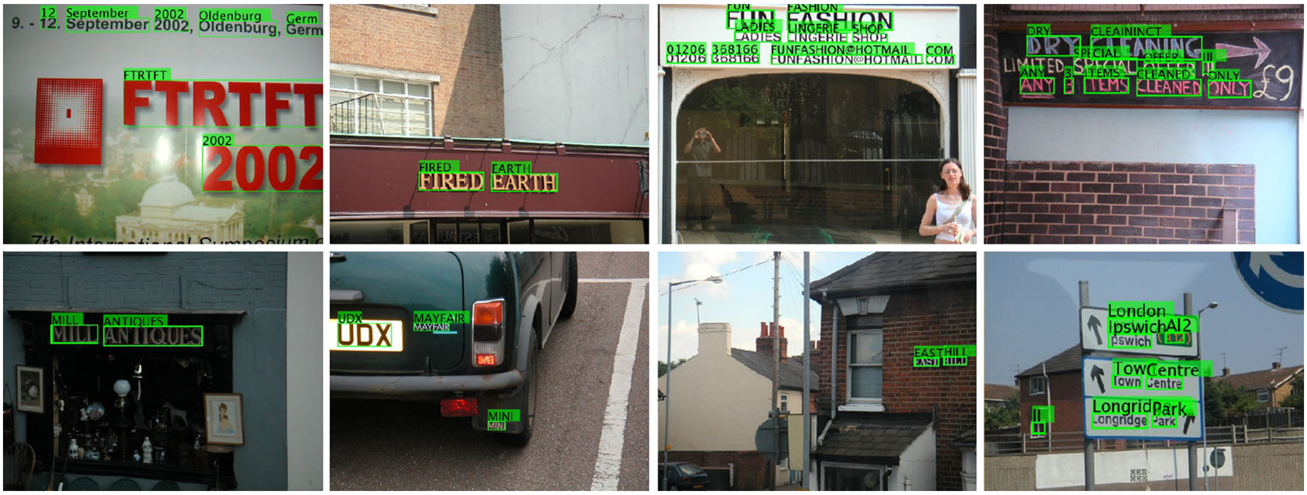


Fig. 8 Text localization and recognition results for proposed method without provided lexicon

Table 4 Text segmentation results of ICDAR 2013 robust reading competition challenge 2: reading text in scene images

Method	Pixel-Level			Atom Level							
	Recall	Precision	*F*-score	Well Segmented	Merged	Broken	Lost	False positives	Recall	Precision	*F*-score
I2R_NUS_FAR	**74.73**	**81.70**	**78.06**	**4,027**	297	11	**1,532**	368	**68.63**	80.36	**74.03**
NSTextractor	60.71	76.28	67.61	3,719	106	10	2,033	**338**	63.38	**83.70**	72.14
USTB_FuStar	69.58	74.45	71.93	3,992	279	12	1,585	941	68.03	72.79	70.33
I2R_NUS	73.57	79.04	76.21	3,540	637	7	1,684	353	60.33	76.69	67.53
NSTsegmentator	68.41	63.95	66.10	3,989	148	25	1,706	2,819	67.98	54.14	60.28
Text Detection	64.74	76.20	70.01	3,640	314	27	1,884	1,888	62.03	57.40	59.63
OTCYMIST	46.11	58.53	51.58	2,451	276	23	3,118	4,573	41.77	31.49	35.91

NSTextractor corresponds to using proposed pipeline for text segmentation with high threshold, *NSTsegmentator* with low threshold
Bold numbers means the best result among others in each category

4.6 Text segmentation

Our pipeline for end-to-end scene text recognition can be used for text segmentation as well. We participated in text segmentation task of the ICDAR 2013 Robust Reading Competition Challenge 2: Reading Text in Scene Images. This task was introduced for the first time in this challenge. We used the same pipeline as for end-to-end text understanding experiments with proposed binarization method, but without using any off-the-shelf OCR module to filter out nontext groups of segments by their recognition results. So, we used only single threshold on average probabilistic output for each word to filter out nontext segmentations. We additionally performed nonmaximum suppression according to words scores.

We submitted two variants of our method: For the first one (called *NSTextractor*), threshold was selected to obtain large number of well-segmented characters, while keeping the number of background segments low, and for the second one (called *NSTsegmentator*), threshold is selected to maximize the number of well-segmented characters. Results of the competition presented in Table 4 [14]. For evaluation of text segmentation results, the framework by Clavelli et al. [6] was used.

Although our method uses rather straightforward steps for text extraction, it took second place in the competition. With high threshold (*NSTextractor*), our method generates the smallest amount of background segments. Low threshold allows to segment much more characters (*NSTsegmentator*), but the number of background segments also increases. So, using more sophisticated approaches for segments grouping as well as their character/noncharacter classification can improve overall system performance.

5 Discussion

We have evaluated several image binarization techniques on the ICDAR 2003 and the ICDAR 2011 benchmarks. Perhaps surprisingly, a pipeline based on image binarization and an off-the-shelf OCR achieves higher accuracy than some of the recent more sophisticated pipelines. Nonlinear Niblack and the proposed method show better performance for text recognition. Fast variant of the proposed method based on recursive bilateral filter achieves competitive accuracy compared to slower global optimization and achieves real-time performance. Proposed method shows how the mature technology of OCR combined with rather straightforward image preprocessing steps can efficiently solve hard problem of end-to-end scene text understanding without using complex pipelines. Results of the well-known ICDAR competition showed that proposed framework for extracting text candidates without using of-the-shelf OCR can achieve good results for natural scene text segmentation as well.

References

1. Adams, A., Gelfand, N., Dolson, J., Levoy, M.: Gaussian KD-trees for fast high-dimensional filtering. ACM Trans. Graph. (TOG) **28**(3), 21 (2009)
2. Badekas, E., Papamarkos, N.: Automatic evaluation of document binarization results. In: CIARP, pp. 1005–1014 (2005)
3. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: ICCV, pp. 105–112 (2001)
4. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2004)
5. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans. Pattern Anal. Mach. Intell. **26**(9), 1124–1137 (2004)
6. Clavelli, A., Karatzas, D., Lladós, J.: A framework for the assessment of text extraction algorithms on complex colour images. In: Document Analysis Systems, pp. 19–26 (2010)
7. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: CVPR (2010)
8. Ezaki, N.: Text detection from natural scene images: towards a system for visually impaired persons. In. International Conference on Pattern Recognition, pp. 683–686 (2004)

9. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Ann. Stat. **38**(2), 337–407 (2000)
10. Gatos, B., Ntirogiannis, K., Pratikakis, I.: Icdar 2009 document image binarization contest (dibco 2009). In: ICDAR, pp. 1375–1382 (2009)
11. Gatos, B., Pratikakis, I., Perantonis, S.J.: Text detection in indoor/outdoor scene images. In: CBDAR'05, pp. 127–132 (2005)
12. He, K., Sun, J., Tang, X.: Guided image filtering. In: Computer vision-ECCV 2010, pp. 1–14. Springer (2010)
13. Howe, N.: A laplacian energy for document binarization. In: ICDAR, pp. 6–10 (2011)
14. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., de las Heras, L.P.: ICDAR 2013 robust reading competition. In: 2013 International Conference on Document Analysis and Recognition (ICDAR). IEEE (2013)
15. Kimmel, R., Bruckstein, A.M.: Regularized Laplacian zero crossings as optimal edge integrators. Int. J. Comput. Vis. **53**(3), 225–243 (2003)
16. Kittler, J., Illingworth, J.: Minimum error thresholding. Pattern Recogn. **19**, 41–47 (1986)
17. Lu, S., Su, B., Tan, C.L.: Document image binarization using background estimation and stroke edges. IJDAR **13**(4), 303–314 (2010)
18. Milyaev, S., Barinova, O., Novikova, T., Lempitsky, V., Kohli, P.: Image binarization for end-to-end text understanding in natural images. In: ICDAR (2013)
19. Minetto, R., Thome, N., Cord, M., Stolfi, J., Precioso, F., Guyomard, J., Leite, N.J.: Text detection and recognition in urban scenes. In: ICCV Workshops, pp. 227–234 (2011)
20. Mishra, A., Alahari, K., Jawahar, C.V.: An mrf model for binarization of natural scene text. In: ICDAR, pp. 11–16 (2011)
21. Neumann, L., Matas, J.: Estimating hidden parameters for text localization and recognition. In: Computer Vision Winter Workshop (2011)
22. Neumann, L., Matas, J.: Real-time scene text localization and recognition. In: CVPR, pp. 3538–3545 (2012)
23. Neumann, L., Matas, J.: Scene text localization and recognition with oriented stroke detection. In: 2013 IEEE International Conference on Computer Vision (ICCV 2013), pp. 97–104 (2013)
24. Niblack, W.: An introduction to digital image processing. Strandberg Publishing, Denmark (1985)
25. Ntirogiannis, K., Gatos, B., Pratikakis, I.: An objective evaluation methodology for document image binarization techniques. In: DAS, pp. 217–224 (2008)
26. Otsu, N.: A threshold selection method from gray level histograms. IEEE Trans. Syst. Man Cybern. **9**, 62–66 (1979)
27. Pan, Y.F., Hou, X., Liu, C.L.: Text localization in natural scene images based on conditional random field. In: ICDAR, pp. 6–10 (2009)
28. Pratikakis, I., Gatos, B., Ntirogiannis, K.: ICDAR 2011 document image binarization contest (DIBCO 2011). In: ICDAR, pp. 1506–1510 (2011)
29. Sauvola, J., Pietikinen, M.: Adaptive document image binarization. Pattern Recogn. **33**, 225–236 (2000)
30. Wakahara, T., Kita, K.: Binarization of color character strings in scene images using k-means clustering and support vector machines. In: ICDAR, pp. 274–278 (2011)
31. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: IEEE International Conference on Computer Vision (ICCV). Barcelona, Spain (2011)
32. Wang, T., Wu, D.J., Coates, A., Ng, A.Y.: End-to-end text recognition with convolutional neural networks. In: 21st International Conference on Pattern Recognition (ICPR), pp. 3304–3308 (2012)
33. Wolf, C., Doermann, D.: Binarization of low quality text using a markov random field model. In: Proceedings of International Conference on Pattern Recognition, pp. 160–163 (2002)
34. Wolf, C., Jolion, J.M.: Object count/area graphs for the evaluation of object detection and segmentation algorithms. Int. J. Doc. Anal. Recogn. **8**(4), 280–296 (2006)
35. Yamazoe, T., Etoh, M., Yoshimura, T., Tsujino, K.: Hypothesis preservation approach to scene text recognition with weighted finite-state transducer. In: ICDAR, pp. 359–363 (2011)
36. Yang, Q., Tan, K.H., Ahuja, N.: Real-time o (1) bilateral filtering. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, pp. 557–564. IEEE (2009)
37. Yang, Q.: Recursive bilateral filtering. In: ECCV (1), pp. 399–413 (2012)
38. Yao, C., Bai, X., Liu, W., Ma, Y., Tu, Z.: ICDAR 2011 document image binarization contest (DIBCO 2011). In: CVPR (2012)
39. Yoon, K.J., Kweon, I.S.: Adaptive support-weight approach for correspondence search. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 650–656 (2006)
40. Zhu, K., Qi, F., Jiang, R., Xu, L., Kimaci, M., Wu, Y., Aizawa, T.: Using adaboost to detect and segment characters from natural scenes. In: Camera-Based Document Analysis and Recognition (CBDAR) (2005)