Effective Candidate Component Extraction for Text Localization in Born-Digital Images by Combining Text Contours and Stroke Interior Regions

Kai Chen, Fei Yin, Cheng-Lin Liu
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation of Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing 100190, P.R. China
Email: {kchen, fyin, liucl}@nlpr.ia.ac.cn

***Abstract*—Extracting candidate text connected components (CCs) is critical for CC-based text localization. Based on the observation that text strokes in born-digital images mostly have complete contours and the text pixels have high contrast with the adjacent non-text pixels, we propose a method to extract candidate text CCs by combining text contours and stroke interior regions. After segmenting the image into non-smooth and smooth regions based on local contrast, text contour pixels in non-smooth regions are detached from adjacent non-text pixels by local binarization. Then, obvious non-text contours can be removed according to the spatial relationship of text and non-text contours. While smooth regions include stroke interior regions and non-text smooth regions, some non-text smooth regions can be easily removed because they are not surrounded by candidate text contours. At last, candidate text contours and stroke interior regions are combined to generate candidate text CCs. The CCs undergo CC filtering, text line grouping and line classification to give the text localization result. Experimental results on the born-digital dataset of ICDAR2013 robust reading competition demonstrate the efficiency and superiority of the proposed method.**

***Index Terms*—Candidate component extraction, text contours, stroke interior regions, text localization.**

I. Introduction

The text elements embedded in born-digital images, prevalent on the Web, carry salient semantic information such as advertisements and security-related information. Therefore, extracting text information from born-digital images enhances the semantic relevance of web content for indexing and retrieval. Usually, a text information extraction system consists of three steps: text localization, text segmentation and text recognition. Text localization is critical to the overall system performance and is suffering from variable image background, text color and layout.

Many methods have been proposed for text localization in images, and they roughly fall into two categories: connected component (CC)-based and texture-based.

CC-based methods have achieved state-of-the-art results on several public datasets [1]. Extracting candidate text CCs is the most critical step for those methods. On one hand, as many text CCs as possible should be recalled. On the other hand, fewer non-text CCs should be generated so that text CCs are more easily identified and grouped into text lines. Researchers frequently use color, stroke width transform (SWT) and maximally stable extremal regions (MSERs) to cluster pixels into CCs. Color clustering based methods [2, 3] adopt strategies such as k-means to segment an image in the hope that pixels belonging to the same text CCs are segmented into the same sub-image. As expected, deciding cluster number is the main difficulty. SWT based methods [4, 5] rely on the existence of two edge pixels with roughly opposite gradient directions in seeking strokes, and merge strokes with about the same stroke width into CCs. Those methods are sensitive to the defects of the edge images. MSERs based methods assume text CCs correspond to MSERs. Finding MSERs can be seen as a process to find local binarization results that are stable over a range of thresholds. To reduce the missing of text CCs, MSERs based methods [6–9] generate tremendous non-text CCs, including many ambiguous ones. Besides, as [10] pointed out, some text elements correspond to extremal regions (ERs) instead of MSERs.

Texture based methods are based on the observation that text regions in images have distinct textural properties in contrast to non-text regions. Some textual based methods [11, 12] slide a sub-window in multi-scales through all locations of the image using a trained classifier to decide whether the sub-window contains text or not. The exhaustive search makes the computation costly. There are also other methods [13–15] which generate text region proposals first and decide whether each region proposal contains text or not. Designing effective features for discriminating textual regions from non-textual regions is the main difficulty of textual based methods.

The proposed method is CC-based. Unlike that many methods have been proposed for scene text detection, few works have been published specifically for born-digital images. Because born-digital images present different characteristics from scene images, it is not necessarily true that methods developed for scene images are appropriate for born-digital images. Text strokes in born-digital images usually have complete contours and pixels on the contours have high contrast compared with the adjacent non-text pixels. This is often not true for text in scene images due to non-ideal camera-capturing environment.

978-1-5090-1792-8/16 $31.00 © 2016 IEEE
DOI 10.1109/DAS.2016.30

Based on this observation, we first detach interior region pixels from contour pixels, then detach text contours from non-text contours. As a result, a large percentage of non-text contours and non-text interior regions are removed based on their spatial relationship. Experimental results show that the proposed method recalls most text CCs but generates a relatively small number of candidates compared with MSERs based and SWT based methods.

In the next section, we give the details of the proposed method. Section III presents experimental results and Section IV concludes the paper.

II. Method

Our system consists of three stages as shown in Fig. 1. Text contour pixels have higher local contrast than stroke interior region pixels, and we detect them separately and combine them to generate candidate text CCs. First the image is segmented to non-smooth and smooth regions based on local contrast thresholding. Text contour pixels and non-text contour pixels in non-smooth regions are detached using local binarization. Fortunately, stroke interior regions correspond to smooth regions directly and are not attached to non-text pixels. Based on the observation that text contours are surrounded by non-text contours and stroke interior regions are surrounded by text contours, obvious non-text contours and non-text smooth regions are removed. The remaining contours and smooth regions are merged to generate candidate text CCs. This is what we do in the first stage and the main contribution of this paper. Then the CCs undergo CC filtering, line grouping and line classification to give the text localization result in the following two stages.

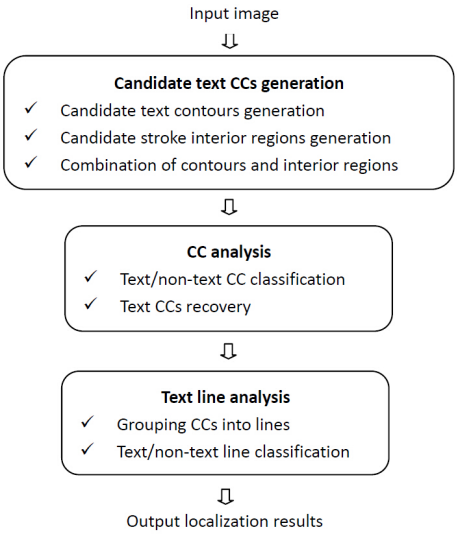


Fig. 1. Block diagram of the proposed algorithm.

A. Candidate text CCs generation

Text strokes in born-digital images mostly have complete contours. We generate candidate text contours and stroke interior regions separately, which are combined to generate candidate text CCs. As illustrated in Fig. 2, we segment an image (Fig. 2(a)) into non-smooth (Fig. 2(b)) and smooth regions (Fig. 2(c)) based on local contrast thresholding, with pixels of large local contrast constituting non-smooth regions, and pixels of small local contrast constituting smooth regions. The threshold is selected to guarantee that text contour pixels are segmented into non-smooth regions. Then we detach text contour pixels from the adjacent background pixels by local binarization (Fig. 2(d)(e)). Because text contours should be surrounded by non-text contours, obvious non-text contours are removed (Fig. 2(f)(g)). Furthermore, smooth regions which are not surrounded by candidate text contours are removed (Fig. 2(h)). At last, stroke interior regions (Fig. 2(h)) are merged with corresponding candidate text contours (Fig. 2(f)(g)) to generate candidate text CCs (Fig. 2(i)(j)).

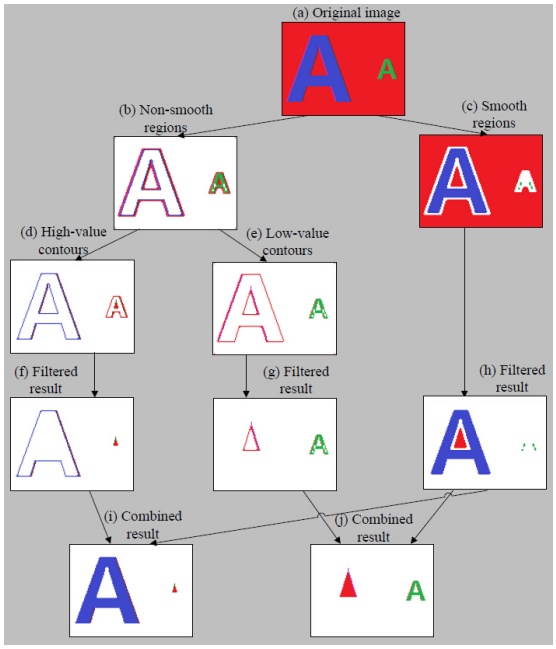


Fig. 2. An illustration for candidate text CCs generation. White pixels are used to isolate different components. (a) An input image. (b) Non-smooth regions. (c) Smooth regions. (d) High-value contours. (e) Low-value contours. (f)(g)(h) Filtered results of (d), (e) and (c), respectively. (i) Combined result of (f) and (h). (j) Combined result of (g) and (h).

1) Smooth/non-smooth regions segmentation: Text contour pixels have higher contrast with the adjacent pixels than stroke interior region pixels, and the gradient magnitude of a pixel can be used to measure the local contrast. Therefore, we compute the gradient magnitudes of all pixels and split them into contour and non-contour pixels by a threshold T_1. As a result, pixels with magnitudes smaller and larger than T_1 constitute non-smooth (Fig. 2(b)) and smooth regions (Fig. 2(c)), respectively.

For each pixel, we calculate its gradient magnitudes in RGB channels separately using the Sobel operator, and use the largest among the three as the final magnitude.

T_1 is selected to guarantee that text contour pixels are classified into the high-contrast (non-smooth) regions. Because we test our method on the born-digital test set of ICDAR2013 robust reading competition, we select T_1 based on the statistics of the training set, which contains 410 images. Below is the histogram (Fig. 3) which shows the distribution of stroke interior pixels and text contour pixels with respect to gradient magnitudes. The bin size is set as four. Based on Fig. 3, we set

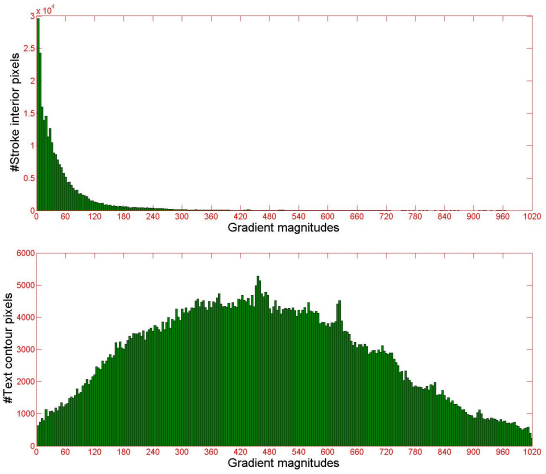


Fig. 3. Stroke interior pixels and text contour pixels distribution with respect to gradient magnitudes of the born-digital training set of ICDAR2013 Competition.

T_1 heuristically as 60, and there are 15659 (2.17 percent) out of 722169 text contour pixels whose magnitudes are smaller than T_1. To miss less text contour pixels, the 8-connected neighboring pixels of high contrast pixels are also segmented into non-smooth regions. As a result, text contour pixels with magnitudes smaller than T_1 may also be segmented into non-smooth regions. For text contour pixels in Fig. 3, 721978 pixels are segmented into non-smooth regions, and only 0.03 percent are segmented into smooth regions.

2) Local binarization of non-smooth regions: Stroke interior regions correspond to smooth regions directly. However, text contour pixels are still connected to adjacent non-text pixels, and we split them based on local binarization. For a non-smooth pixel P_c, we check a 5 * 5 window centered at P_c to decide the threshold for P_c. Because non-smooth regions have included low contrast pixels as the neighbors of high contrast pixels, a 3 * 3 window may not provide enough information. Besides, a 5 * 5 window size is more robust to noisy pixels. We adopt Otsu's binarization algorithm to decide the threshold T_2. P_c is put into high-value contours (Fig. 2(d)) if its intensity is higher than T_2, otherwise it is put into low-value contours (Fig. 2(e)). We process all pixels in non-smooth regions the same as P_c.

3) Non-text contours filtering: As we can see from Fig. 2, the blue text contour "**A**" in Fig. 2(d) is surrounded by the red non-text contour "**A**" in Fig. 2(e), and the green text contour "**A**" in Fig. 2(e) is surrounded by the red non-text contour "**A**" in Fig. 2(d). Actually, it is a general case that text contours in one contour image are surrounded by non-text contours in the other contour image. Although text interior contours (e.g. the blue triangle belong to "**A**" in Fig. 2(d)) are not surrounded, there should not be a problem after removing them. For example, we can still recognize the blue text CC "**A**" in Fig. 2(i) even though interior text contour pixels are missing, and the proposed method can still detect it. Therefore, we remove contours in one image if they are not surrounded by contours in the other image. For example, Fig. 2(f) and 2(g) are the filtered results of Fig. 2(d) and 2(e), respectively.

As illustrated in Fig. 4, for a contour C in one image, we track C's adjacent white pixels and count how many black pixels (N_b) and white pixels (N_w) there are in corresponding positions of the other image. We measure C's "surroundedness" using $N_b/(N_b + N_w)$.

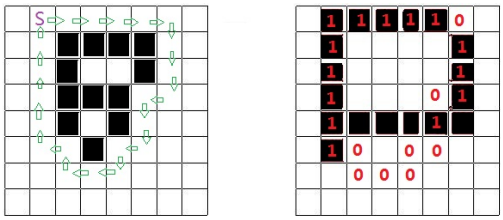


Fig. 4. The "surroundedness" of the contour in (a) by pixels in (b). Each grid represents a pixel. We track the adjacent white pixels of black pixels in (a) from the starting position **S** along the green arrow flow. Corresponding positions in (b) are marked as "1" or "0" if they are black or white pixels, respectively. In this example, N_b equals to 14 and N_w equals to 8, therefore, "surroundedness" of the contour in (a) is 0.64.

Table I depicts the distribution of text contours with respect to their "surroundedness" for the born-digital training set of ICDAR2013 Robust Reading Competition. Based on Table I, we set T_3 as 0.9, and remove a contour if its "surroundedness" is below T_3. For the born-digital training set, 147 text contours (0.897 percent of all text contours) are removed erroneously, and 26587 non-text contours (54.39 percent of all non-text contours) are removed correctly.

4) Non-text smooth regions filtering: As we can see from Fig. 2, the stroke interior regions in Fig. 2(c) are surrounded by text contours in Fig. 2(f)(g). This is not the case for all non-text smooth regions. For each smooth region R_s (Fig. 2(c)), we compute its "surroundedness" with both filtered text contour images (Fig. 2(f)(g)) and use the larger one as R_s's "surroundedness".

Table II depicts the distribution of stroke interior regions with respect to their "surroundedness" for the born-digital training set of ICDAR2013 Robust Reading Competition. Based on Table II, we set T_4 heuristically as 0.8, and remove a smooth region if its "surroundedness" is below T_4. For the born-digital training set, 31 stroke interior regions (0.824 percent of all stroke interior regions) are removed erroneously, and 11420 non-text smooth regions (45.42 percent of all non-text smooth regions) are removed correctly.

TABLE I
TEXT CONTOURS DISTRIBUTION WITH RESPECT TO "SURROUNDEDNESS" OF THE BORN-DIGITAL TRAINING SET OF ICDAR2013 COMPETITION.

"Surroundedness" Range	[0, 0.2)	[0.2, 0.3)	[0.3, 0.4)	[0.4, 0.5)	[0.5, 0.6)	[0.6, 0.7)	[0.7, 0.8)	[0.8, 0.9)	[0.9, 1.0)	1.0
# Text Contours	0	1	4	13	41	11	26	51	180	16060

TABLE II
STROKE INTERIOR REGIONS DISTRIBUTION WITH RESPECT TO "SURROUNDEDNESS" OF THE BORN-DIGITAL TRAINING SET OF ICDAR2013 COMPETITION.

"Surroundedness" Range	[0, 0.5)	[0.5, 0.6)	[0.6, 0.7)	[0.7, 0.8)	[0.8, 0.9)	[0.9, 1.0)	1.0
# Stroke Interior Regions	0	9	10	12	48	154	3529

5) Contours and interior regions merging: The remaining smooth regions (Fig. 2(h)) are merged into the one of filtered contour images (Fig. 2(f) or (g)) based on which one gives a larger "surroundedness". The candidate text CCs are in two separate images (Fig. 2(i)(j)) now.

B. CC analysis and Text line analysis

We adopt the method of our previous work [16] to select text CCs and group them into text lines.

We select text CCs by text/non-text CC classification and recovering some rejected CCs. The classifier is a linear SVM with distinguishing features for text CCs from non-text CCs, and the CCs which are labeled as non-text are removed. Then a rejected CC is recovered if it has a neighboring text CC and they meet several heuristic constraints concerning color, stroke width, distance and alignment.

Candidate text CCs are processed as follows. Two neighboring CCs are linked into a pair if they obey several heuristic constraints. Then pairs sharing a common CC are merged sequentially to construct text line candidates until no pairs can be merged. We abandon all lines consisting of less than two CCs and apply text/non-text line classification to the remaining lines. The classifier is also a linear SVM.

Finally, overlapping text lines from two different candidate images (Fig. 2(i)(j)) are merged and segmented into words as the final result.

Fig. 5 shows the stepwise results of an example image.

III. EXPERIMENTAL RESULTS

We evaluate the performance of our method on the database of the ICDAR2013 robust reading competition - Challenge 1: Reading Text in Born-Digital Images (Web and Email) [17].

The born-digital dataset comprises images extracted from web pages (news, personal, commercial, social, government, etc) and email messages (spam, newsletters, etc). There are 551 of them with a minimum size of 100 * 100 pixels, out of which 410 and 141 images are used for training and test, respectively.

A. Character detection rate

To demonstrate the effectiveness of the proposed local contrast based method, we compare it with MSER [6, 7] and SWT [4] with respect to the text candidate extraction ability. We adopt the MSER algorithm implemented in OpenCV, with the parameters delta, min-area, max-area, max-variation and min-diversity being equal to 2, 6, input image area, 0.5 and 0.33, respectively. MSER is very sensitive to parameter setting, and we set parameters as listed as the result of a tradeoff between recall and precision rate. The minimum area and maximum area of text candidates the proposed method extracts are the same. SWT are implemented following the framework of [4].

TABLE III
RANKING OF OUR METHOD AND THE METHODS SUBMITTED TO ICDAR2013 ROBUST READING COMPETITION ON BORN-DIGITAL DATASET.

Method Name	Recall (%)	Precision (%)	F-score(%)	Number of Proposals
Proposed	90.51	48.83	63.44	15994
MSER [6, 7]	83.73	28.77	42.83	48258
SWT [4]	71.79	41.54	52.62	14992

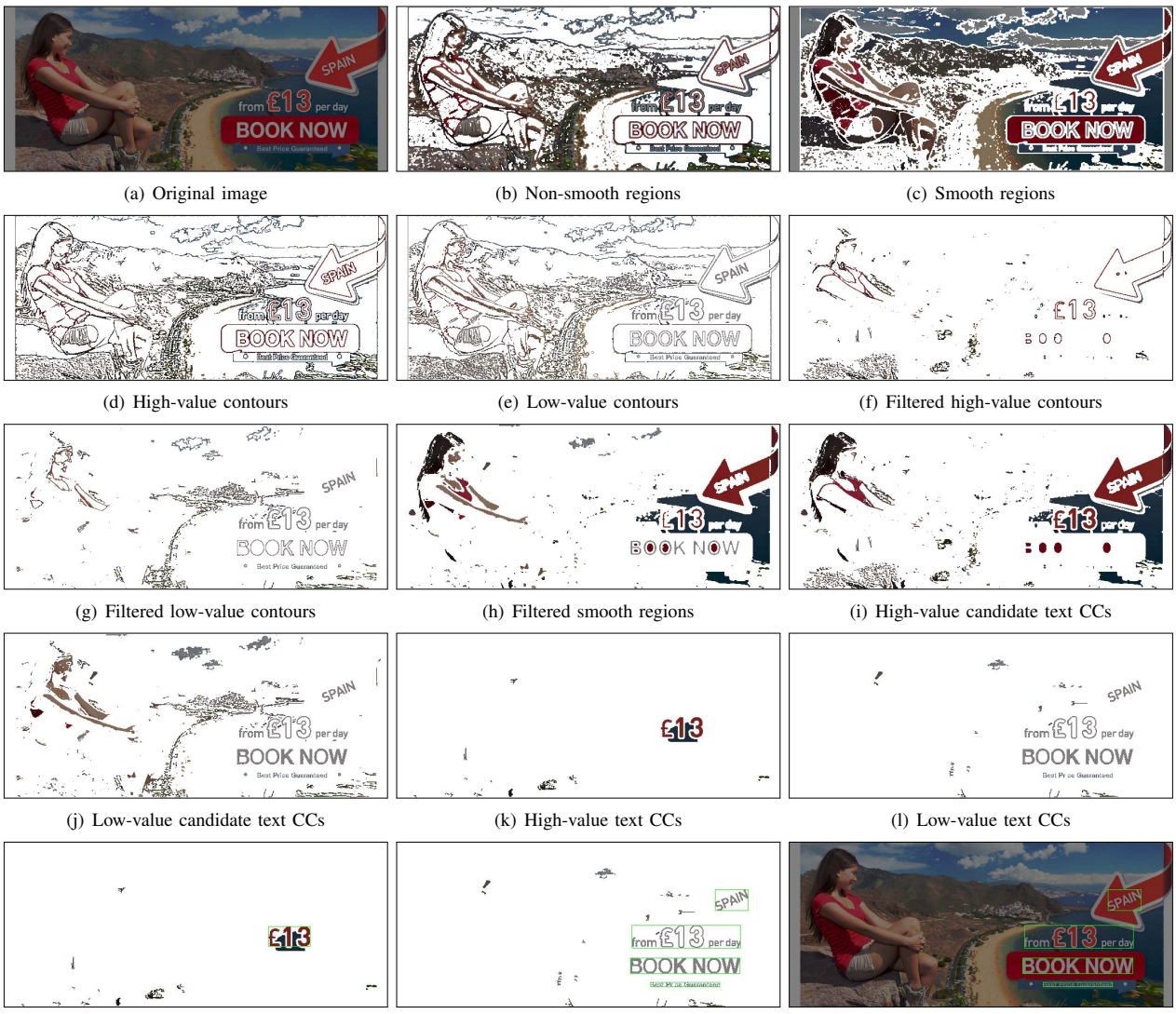
The evaluation protocol is described in the report of the competition [17]. It was proposed to evaluate word-level match over the whole test set, taking into account one-to-one, one-to-many and many-to-one matches between detected and ground truth word boxes. Here we adjust it into the character-level. That is to say, detected and ground truth boxes are labelled in the character-level.

The results in Table III shows that the proposed method extracts a relatively small number of proposals but recalls more text CCs.

B. Text detection performance

The organizers of the competition maintain a website [18] where the dataset with ground truth can be downloaded after registration. The website also accepts detected results and outputs the evaluation results. The evaluation protocol is the same as described in Sect.III-A except that it is based on word-level as it was proposed. We present the results of *StradVision* (the only one method which achieves better results than our method on website [18] by ICDAR 2013 evaluation protocol), our method and the methods submitted to the 2013 competition in Table IV. Brief descriptions of the submitted methods can be found in [17]. The winner of the 2013 competition was USTB_TexStar, a method based on MSER segmentation. It also won the task of scene text localization. *StradVision* is based on extremal regions (ER) segmentation.

The results in Table IV shows that our proposed method



(a) Original image (b) Non-smooth regions (c) Smooth regions

(d) High-value contours (e) Low-value contours (f) Filtered high-value contours

(g) Filtered low-value contours (h) Filtered smooth regions (i) High-value candidate text CCs

(j) Low-value candidate text CCs (k) High-value text CCs (l) Low-value text CCs

(m) High-value text lines (n) Low-value text lines (o) Combined text lines

Fig. 5. Stepwise results of an example image.

yields superior performance compared to the competition results in ICDAR2013. After CC segmentation, we use conventional techniques for CC filtering, line grouping and word partitioning. The superior performance indicates that the proposed local contrast-based segmentation method is promising. The *StradVision* method generates even more candidate CCs than MSERs based methods and applies a powerful classifier to filter non-text CCs, which gives a direction for our future improvements.

Fig. 6 shows some examples of successful detection on images in the dataset.

Fig. 7 shows two failure cases in our experiments. For Fig. 7(a), we miss the text because our method cannot handle curved text lines. In Fig. 7(b), we identify non-text regions as text because they have text-like features.

IV. Conclusion

We proposed a new CC based method for text localization in born-digital images. The proposed method generates character candidates effectively by first detecting text contours and stroke interior regions separately and then combining them. Experimental results show that the proposed method generates a relatively small number of candidates but recalls most text CCs compared with MSERs based and SWT based methods. The CCs undergo CC filtering, line grouping and line classification to give the final result. Our method has achieved state-of-the-art performance on the born-digital dataset of ICDAR2013 Competition, convincingly demonstrating the effectiveness. We anticipate higher performance if we improve the algorithms for CC classification and line grouping in the future.

TABLE IV
RANKING OF OUR METHOD AND THE METHODS SUBMITTED TO ICDAR2013 ROBUST READING COMPETITION ON BORN-DIGITAL DATASET.

Method Name	Recall (%)	Precision (%)	F-score(%)
StradVision	85.54	**95.21**	**90.12**
Our method	**87.95**	91.14	89.51
USTB TexStar [8]	82.38	93.83	87.74
TH-TextLoc	75.85	86.82	80.96
I2R NUS FAR	71.42	84.17	77.27
Baseline	69.21	84.94	76.27
Text Detection [19, 20]	73.18	78.62	75.81
I2R NUS	67.52	85.19	75.34
BDTD CASIA	67.05	78.98	72.53
OTCYMIST [21]	74.85	67.69	71.09
Inkam	52.21	58.12	55.00

Fig. 6. Examples of successful text localization results.

Fig. 7. Two failure cases of text localization.

ACKNOWLEDGMENT

This work has been supported in part by the National Basic Research Program of China (973 Program) Grant 2012CB316302, the National Natural Science Foundation of China (NSFC) Grant 61411136002, the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant XDA06040102), and the Xinjiang Uygur Autonomous Region Science and Technology Project (Grant 201230122).

REFERENCES

[1] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura et al., ICDAR 2015 competition on robust reading, *Proc. 13th ICDAR*, 2015, pp. 1156-1160.

[2] J. Park, H. Yoon, G. Lee, Automatic segmentation of natural scene images based on chromatic and achromatic components, *Computer Vision/Computer Graphics Collaboration Techniques*, Springer Heidelberg, 2007, pp. 482-493.

[3] S. Tehsin, A. Masood, S. Kausar, Y. Javed, A caption text detection method from images/videos for efficient indexing and retrieval of multimedia data, *Int. J. Pattern Recognition and Artificial Intelligence*, 29(1), 2015.

[4] B. Epshtein, E. Ofek, Y. Wexler, Detecting text in natural scenes with stroke width transform, *Int. Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 2963-2970.

[5] W. Huang, Z. Lin, J. Yang, J. Wang, Text localization in natural images using stroke feature transform and text covariance descriptors, *Proc. 14th ICCV*, 2013, pp. 1241-1248.

[6] L. Neumann, J. Matas, A method for text localization and recognition in real-world images, *Proc. 10th ACCV*, 2010, pp. 770-783.

[7] L. Neumann, J. Matas, Real-time scene text localization and recognition, *Int. Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 3538-3545.

[8] X.-C. Yin, X. Yin, K. Huang, H.-W. Hao, Robust text detection in natural scene images, *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(5): 970-983, 2014.

[9] L. Sun, Q. Huo, W. Jia, K. Chen, A robust approach for text detection from natural scene images, *Pattern Recognition*, 48(9): 2906-2920, 2015.

[10] M.-C. Sung, B. Jun, H. Cho, D. Kim, Scene text detection with robust character candidate extraction method, *Proc. 13th ICDAR*, 2015, pp. 426-430.

[11] K.I. Kim, K. Jung, J.H. Kim, Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm, *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(12): 1631-1639, 2003.

[12] Y.-F. Pan, X. Hou, C.-L. Liu, A hybrid approach to detect and localize texts in natural scene images, *IEEE Trans. Image Process.*, 20(3): 800-813, 2011.

[13] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Reading text in the wild with convolutional neural networks, *Int. J. Comput. Vis.*, 1-20, 2014.

[14] L. Gomez, D. Karatzas, Object proposals for text extraction in the wild, *Proc. 13th ICDAR*, 2015, pp. 206-210.

[15] R. Mehta, O. Chum, J. Matas, Towars visual words to words: text detection with a general bag of words representation, *Proc. 13th ICDAR*, 2015, pp. 641-645.

[16] K. Chen, F. Yin, A. Hussain, C.-L. Liu, Efficient text localization in born-digital images by local contrast-based segmentation, *Proc. 13th ICDAR*, 2015, pp. 291-295.

[17] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, S.R. Mestre, J. Mas et al., ICDAR 2013 robust reading competition, *Proc. 12th ICDAR*, 2013, pp. 1484-1493.

[18] http://dag.cvc.uab.es/icdar2013competition/

[19] J. Fabrizio, B. Marcotegui, M. Cord, Text detection in street level images, *Pattern Anal. Appl.*, 16(4): 519-533, 2013.

[20] J. Fabrizio, B. Marcotegui, M. Cord, Text segmentation in natural scenes using toggle-mapping, *Proc. 16th ICIP*, 2009, pp. 2373-2376.

[21] D. Kumar, A.G. Ramakrishnan, OTCYMIST: Otsu-Canny minimal spanning tree for born-digital images, *Proc. 10th ICDAR*, 2012, pp. 389-393.