Pattern Recognition 108 (2020) 107559

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

SCUT-HCCDoc: A new benchmark dataset of handwritten Chinese text in unconstrained camera-captured documents[☆]

Hesuo Zhang, Lingyu Liang[*], Lianwen Jin

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

ARTICLE INFO

Article history:
Received 2 November 2019
Revised 24 March 2020
Accepted 20 July 2020
Available online 28 July 2020

Keywords:
Document analysis and recognition
Handwritten Chinese text recognition
Handwritten Chinese text detection
Benchmark dataset

ABSTRACT

In this paper, we introduce a large-scale dataset, called SCUT-HCCDoc, to address challenging detection and recognition problems of handwritten Chinese text (HCT) in the camera-captured documents. Despite extensive studies of optical character recognition (OCR) and offline handwriting recognition for document images, text detection and recognition in the camera-captured documents remains an unsolved problem that is worth for extensive study and investigation. With recent advances in deep learning, researchers have proposed useful architectures for feature learning, detection, and recognition for the scene text. However, the performance of deep learning methods highly depends on the amount and diversity of training data. Previous OCR and offline HCT datasets were built under specific constraints, and most of the recent scene text datasets are for non-handwritten text. Hence, there is a lack of a comprehensive scene handwritten text benchmark. This study focuses on scenes with handwritten Chinese text. We introduce the SCUT-HCCDoc database for HCT detection, recognition and spotting. SCUT-HCCDoc contains 12,253 camera-captured document images with 116,629 text lines and 1,155,801 characters. The diversity of SCUT-HCCDoc can be described at three levels: (1) **image-level diversity**: image appearance and geometric variances caused by camera-captured settings (such as perspective, background, and resolution) and different applications (such as note-taking, test papers, and homework); (2) **text-level diversity**: variances of text line length, rotation, etc.; (3) **character-level diversity**: variances of character categories (up to 6109 classes with additional English letters, and digits), character size, individual writing style, etc. Three kinds of baseline experiments were conducted, where we used several popular text detection methods for text line detection, CTC-based/attention-based methods for text line recognition, and combine text detectors with CTC-based recognizer to achieve end-to-end text spotting. The results indicate the diversity of SCUT-HCCDoc and the challenges of HCT understanding in document images. The dataset is available at https://github.com/HCIILAB/SCUT-HCCDoc_Dataset_Release.

1. Introduction

Text detection and recognition in images are important problems for pattern recognition, which aim to locate the regions of text on an image by using bounding boxes, and transform the cropped text instance image into linguistic symbols. Optical character recognition (OCR) and offline handwriting recognition for document images have been thoroughly researched over decades.

Recently, active research has started in scene text detection and recognition in natural images [1], which may lead to useful applications, such as real-time translation, robot navigation, traffic sign recognition for automatic driving, and image-based information retrieval. Despite years of studies, scene text detection and recognition is far from solved. The challenges lie in the diversity of text in natural environment, the variability of unpredictable natural scene backgrounds, and uncontrolled imaging conditions.

Deep neural networks were successful for pattern recognition and image understanding [2–4]. Many architecture of feature learning, detection, and recognition were introduced and modified for handwritten or scene text understanding [5]. Yet, the performance of deep learning methods highly depends on the size and diversity of the dataset (such as ImageNet dataset [6], Microsoft COCO dataset [7], and VOC Pascal dataset [8] for object recognition and scene understanding). Hence, benchmark datasets are indispensable for scene text understanding.

[☆] This research is supported in part by National Nature Science Foundation of China (Nos. 61936003, 61872151), Natural Science Foundation of Guangdong Province (No.2017A030312006, No.2019A1515011045), the National Key Research and Development Program of China (No. 2016YFB1001405), and Fundamental Research Funds for the Central Universities (x2dxD2190570, 2019MS023).

[*] Corresponding author.

E-mail addresses: eehesuo.zhang@mail.scut.edu.cn (H. Zhang), lianglysky@gmail.com (L. Liang), lianwen.jin@gmail.com (L. Jin).

https://doi.org/10.1016/j.patcog.2020.107559
0031-3203/© 2020 Elsevier Ltd. All rights reserved.

Table 1
Detailed comparison of representative datasets for scene (handwritten) text detection and recognition, where the top 17 datasets are for non-handwritten text, and the bottom 10 datasets are for handwritten text.

Dataset			Det. Task	Rec. Task	Size	Language
Non-Handwritten Char./Text	Natural Images	IC2013 [9]	√	√	462 images	EN
		IC2015 [10]	√	√	1500 images	EN
		IC2017-MLT [11]	√	×	18k images	Multi-lingual
		IC2019-MLT [12]	√	√	20k images	Multi-lingual
		IC2019-LSVT [13]	√	√	50k images	CN
		IC2019-ART [14]	√	√	10,166 images	EN,CN
		IC2019-ReCTS [15]	√	√	25k images	CN
		CTW [16]	√	√	32,285 images	CN
		RCTW [17]	√	√	12,263 images	CN
		CASIA-10k [18]	√	√	10k images	CN
		Total-Text [19]	√	√	1555 images	EN,CN
		SVT [20,21]	√	√	350 images	EN
		SCUT-CTW1500 [4]	√	×	1500 images	EN
		MSRA-TD500 [22]	√	×	500 images	EN,CH
		III5K [23]	×	√	5000 cropped words	EN
		SVT-Perspective [24]	×	√	238 images	EN
		SVHN [25]	×	√	about 630k digits	Digit
Dataset			Writer Num.	Class Num.	Size	Online/Offline
Handwritten Char./Text (Chinese)	Constrained Images	HCL2000 [26]	1000	3755	3.755 M isolated char.	Offline
		HIT-OR3C [27]	142	6825	909 K char.	Online
		CASIA-OLHWDB 1.0–1.2 [28]	1020	7185	3.7 M char.	Online
		CASIA-HWDB 1.0–1.2 [28]	1020	7185	3.7 M char.	Offline
		SCUT-COUCH 2009 [29]	190	6763	3.6 M char.	Online
		HIT-MW [30]	780	3041	8.6 K text lines, 186K char. in text	Offline
		CASIA-OLHWDB 2.0–2.2 [28]	1019	2655	52 K text lines, 1.3M char. in text	Online
		CASIA-HWDB 2.0–2.2 [28]	1019	2703	52 K text lines, 1.3M char. in text	Offline
	Unconstrained Images	SCUT-EPT [31]	2986	4250	50 K text lines, 1.27M char. in text	Offline
		SCUT-HCCDoc	Countless	6109	116 K text lines, 1.15M char. in text	Offline

Table 1 shows a list of datasets collected for text detection and recognition. Before the deep learning era, most datasets were collected for constrained or specific applications, which have been surveyed in Doermann et al. [1]. With recent research progress in scene text detection and recognition, more datasets were constructed using natural images. However, most of these datasets are for non-handwritten text understanding, such as text lines in books, historical documents, bills, packages, street billboards, and shops. The datasets for scene handwritten text is insufficient.

Because handwritten (HW) texts are prevalent in daily life and significantly differ in their variety from printed text, it is essential to collect HW datasets. Table 1 illustrates some typical HW datasets; it indicates that many HW datasets are collected using constrained images, which may fail to capture the variances in natural scenes. Recently, Zhu. et al. [31] presented a new dataset and benchmark for scene Chinese text understanding, but it focuses on text line in examination papers and only considers the recognition task. Therefore, it would be important to construct a comprehensive scene handwriting text benchmark dataset.

In this study, we address the challenges of scene handwritten Chinese text (HCT) understanding in natural settings. We propose a new dataset, called SCUT-HCCDoc, for both HCT detection and recognition. SCUT-HCCDoc contains 12,253 camera-captured document images with 116,629 text lines and 1,155,801 characters. The diversity of SCUT-HCCDoc can be described at three levels: (1) **image-level diversity**: image appearances and geometric variances caused by camera-captured settings (such as perspective, background, and resolution) and different applications (such as note taking, test papers, and homework); (2) **text-level diversity**: variances of text line length, rotation, etc.; (3) **character-level diversity**: variances of character categories (up to 6109 classes with additional English letters, and digits), character size, individual writing style, etc. We obtain the baseline results for three kinds of HCT recognition problems, including several state-of-the-art text detection methods for text line detection, CTC-based/attention-based methods for text line recognition, and the combination of detectors with CTC-based recognizer for end-to-end text spotting.

In this study, the main contributions can be summarized as follows:

(1) **Dataset construction:** We propose a new large-scale SCUT-HCCDoc dataset containing 12,253 camera-captured document images with 116,629 text lines and 1,155,801 characters. The dataset can used for text detection, recognition or end-to-end text spotting.
(2) **Benchmark analysis:** We statistically analyzed the samples and annotations of SCUT-HCCDoc. Moreover, we systematically analyzed the sample diversity of image level, text level, and character level.
(3) **Detection evaluation:** We use different state-of-the-art methods for baseline evaluation of text line detection. We analyze the results for the complete SCUT-HCCDoc dataset and its subsets with different variances.
(4) **Recognition evaluation:** We use both CTC-based and attention-based methods for baseline evaluation of text line recognition. The comparison with the related dataset indicates the diversity and challenges of the SCUT-HCCDoc dataset for HCT recognition.
(5) **End-to-end text spotting evaluation:** We use detectors combining with CTC-based recognizer to achieve end-to-end text spotting. The results of text spotters with different detectors are presented.

2. Related work

Many handwritten text datasets, which can be used for offline or online recognition, have been released. Offline handwriting datasets include CEDAR English words and characters [32], RIMES French paragraph dataset [33], IFN/ENIT database of Arabic words [34], and Chinese dataset CASIA-HWDB1.0–1.2/2.0–2.2 [28]. Online handwriting datasets include IAM English sentence database [35], LMCA of Arabic words, characters and digits [36], Japanese text datasets Kondate [37], and Chinese dataset SCUT-COUTH2009[29].

For Chinese handwriting recognition, many large-scale datasets have been built for offline/online character or text line recognition, as listed in Table 1. For character recognition, the representative datasets since 2000 include HCL2000 [26], HIT-OR3C [27], CASIA-OLHWDB1.0–1.2 [28], CASIA-HWDB1.0–1.2 [28] and SCUT-COUTH2009 [29]. HCL2000 [26] is built by constraining the handwritten Chinese characters within the preprinted boxes, and it only covers 3755 categories, which is much less than the total number of Chinese characters. HIT-OR3C [27] is a database for character recognition that consists of 5 subsets: "GB1", "GB2", "Letter", "Digit" and "Document". The first 4 corpora contain 6825 categories with 832,650 samples produced by 122 persons; the document corpus contains 2442 categories with 77,168 samples produced by 20 persons. CASIA-OLHWDB1.0–1.2 and CASIA-HWDB1.0–1.2 [28] are two series of systematically-built large-scale datasets for offline and online character recognition, respectively. SCUT-COUCH2009 database [29] is for unconstrained online character recognition; it contains 11 subsets of isolated characters and 3.6 million samples contributed by more than 190 writers.

For text line recognition, the representative datasets include HIT-MW [30], CASIA-OLHWDB2.0–2.2 [28], CASIA-HWDB2.0–2.2 [28], and SCUT-EPT [31]. HIT-HW [30] contains 853 samples, 186,444 characters, and a lexicon with 3041 entries. In CASIA-OLHWDB2.0–2.2 and CASIA-HWDB2.0–2.2 [28], the samples were produced on papers using Anoto pen. SCUT-EPT [31] is a collection of examination papers; it contains variances of character erasure, text line supplement, character/phrase switching, noisy background, nonuniform word size, and unbalanced text length.

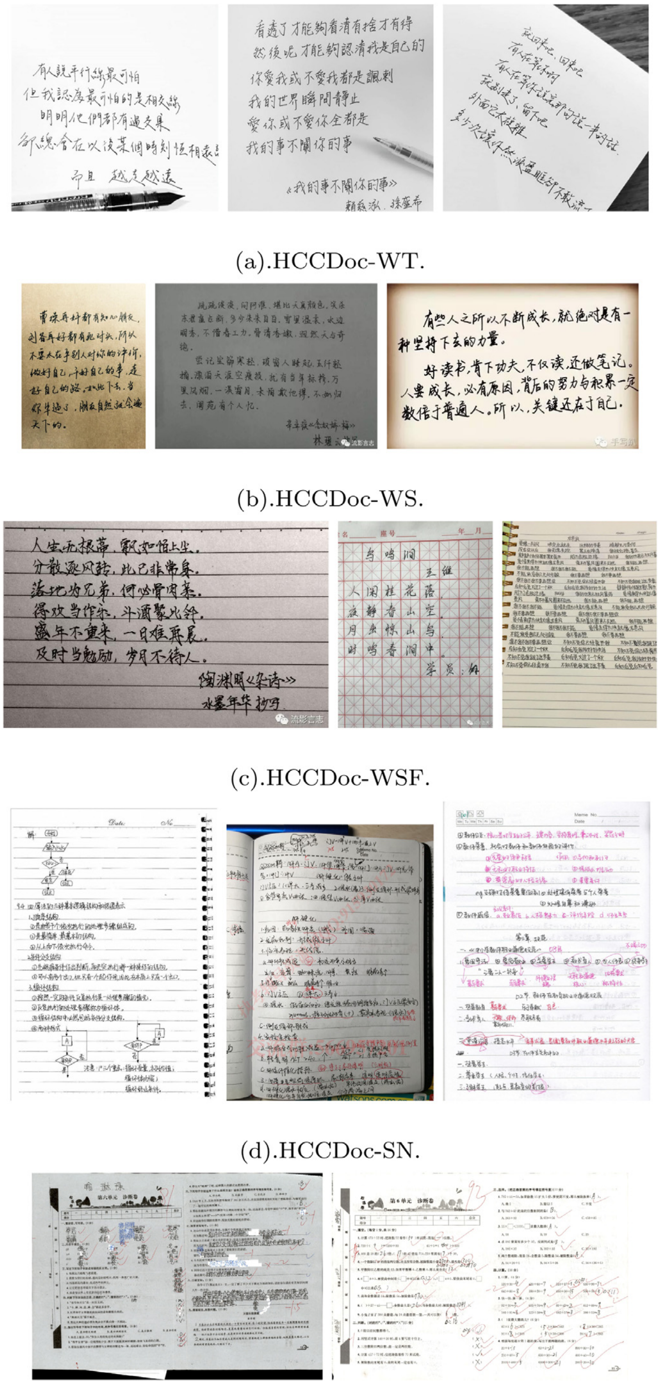
Table 1 indicate a lack of large-scale handwritten text database, especially for handwritten Chinese text in different natural images. Hence, the proposed SCUT-HCCDoc dataset aims to provide diverse text lines in images of various natural scenes, with systematical labels for both HCT detection and recognition.

3. Construction of SCUT-HCCDoc database

3.1. Images collection

All the images of SCUT-HCCDoc were obtained by Internet search. The sources included Instagram,[1] WeChat's official account, Baidu Images,[2] Sina Weibo,[3] and other image sharing websites. We collected more than 100 thousand candidate images. Finally 12,253 images within these candidates were manually selected based on the following principles (shown in Fig. 1):

- We mainly considered the images with horizontal text. Slanted or vertical text was also occasionally included, so that the distribution of text in various directions would be consistent with the reality. More details on the distribution will be given in the following section.
- We excluded images with calligraphy text, too complex mathematical formulas, or any other "fancy style" text possibly generated by image processing software.
- Because of the variety of text topics in web image, we excluded the images with text not suitable for external dissemination (such as text related to personal privacy, sexual content, or violence).


(a).HCCDoc-WT.

(b).HCCDoc-WS.

(c).HCCDoc-WSF.

(d).HCCDoc-SN.

(e).HCCDoc-EP.

Fig. 1. Samples of five subsets of SCUT-HCCDoc. More details of the subset are shown in Table 2 and discussed in Section 4.1.

3.2. Image annotation

The annotation of SCUT-HCCDoc is in text-level, where all text line instances in each image were located using quadrangular bounding box with four points around a text line, and all the text in the bounding box (including simplified Chinese text, traditional Chinese text, English text, digits or punctuation) was labeled; as shown in Fig. 2.

We ignore incomplete characters, printed text, or illegible handwriting that can hardly be recognized with the eyes. Additionally,

[1] https://www.instagram.com/.
[2] https://image.baidu.com/.
[3] https://weibo.com/.

Table 2
Sample distribution of five subsets of SCUT-HCCDoc.

Subset of SCUT-HCCDoc	Image Num.	Char. Num.	Text Num.	Classes Num.	Class Num. of L1/L2	Description Num.
HCCDoc-WT (15.05%)	4705	173,970	22,847	3354	1871/258	Short text, traditional Chinese characters
HCCDoc-WS (29.74%)	3657	343,762	32,920	4091	3171/697	White background, diversity of text length
HCCDoc-WSF (26.61%)	3060	307,573	30,565	3757	2972/568	formatted background, diversity of text length
HCCDoc-EP (18.15%)	300	120,753	15,569	2431	2234/111	Large scale, high text density, printed text inference
HCCDoc-SN (10.45%)	531	209,743	14,728	3386	2913/350	Large scale, high text density, long text
ALL	12,253	1,155,801	116,629	6109	3556/1,154	–

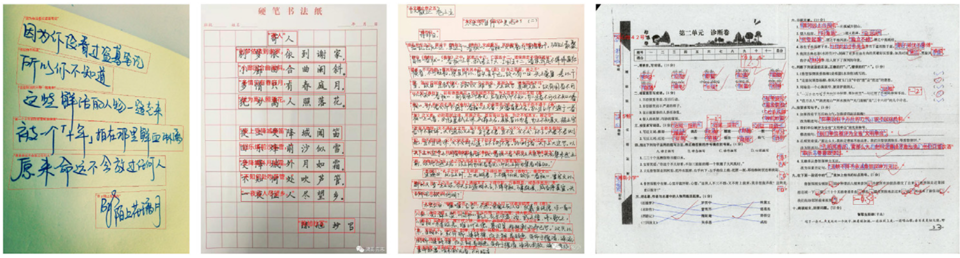


Fig. 2. Samples of annotation of the SCUT-HCCDoc dataset.

if individual characters are unrecognized in a text line, they are annotated with "##". Because of the uncertainty about the number of unrecognized characters, it is unknown how many characters "##" represents. Additionally, a few illegible mathematical formulas are annotated with "##".

3.3. Dataset splitting

SCUT-HCCDoc is randomly split into the training and test sets with a ratio of 4:1. After the splitting, the training set contains 9801 images with 93,411 text instances and 925,733 characters. The test set contains 2452 images with 23,218 text instances and 230,068 characters. The division is random, with equal characteristics and distributions of the training and test sets. Therefore, we directly made a benchmark analysis of the complete dataset in Section 4.

4. Benchmark analysis of SCUT-HCCDoc

We perform a benchmark analysis for samples and annotation of SCUT-HCCDoc for image-level diversity, text-level diversity, and character-level diversity.

4.1. Image-level diversity

For image-level diversity, we consider image appearances and geometric variances caused by camera-capture settings (such as perspective, background, and resolution) and application scenes (such as noting, test papers, and homework).

According to different application scenes, SCUT-HCCDoc can be roughly divided into five subsets:

- **HCCDoc-WT**: images of traditional Chinese characters;
- **HCCDoc-WS**: images of simplified Chinese characters without a formatted background;
- **HCCDoc-WSF**: images of simplified Chinese characters with the formatted background;
- **HCCDoc-SN**: images of student notes;
- **HCCDoc-EP**: images of examination papers.

A sample distribution of these five subsets is listed in Table 2, which illustrates the sample diversity at the image-level for different application scenes. In Fig. 3, we show the average box number (ABN) and the average character number (ACN) of five subsets. Hence, the ABN and ACN of five subsets are unbalanced. The ABN and ACN of HCCDoc-EP and HCCDoc-SN are much larger than the others. In addition, HCCDoc-SN tends to have longer text than HCCDoc-EP. Most of the traditional Chinese characters are in HCCDoc-WT.

Furthermore, we also analyze variances in image aspects. Fig. 4 illustrates the aspect ratio of images in SCUT-HCCDoc, where we observe that the ratios of most images are mainly between 0.6 and 1.8.

4.2. Text-level diversity

For text-level diversity, we consider variances of text line length, rotation, etc.

We analyze and visualize the distribution of text-level diversity. Fig. 5(a–c) shows the aspect ratio of cropped images of text line instance, the number of images containing specific text line instance, and the number of cropped images of text line instance for a specific height, which illustrates the variances of text line instance. Fig. 5(d) illustrates the percentage of text line boxes' angle, which is roughly measured by the upper boundary of the quadrilateral of text line instance. It can be observed that most of the text lines in SCUT-HCCDoc are horizontal text or slightly slanted text.

4.3. Character-level diversity

For character-level diversity, we consider variance in character categories (up to 6109 classes with additional English letters, and digits), character size, individual writing style, etc. We analyze and visualize the distribution of character-level diversity, as

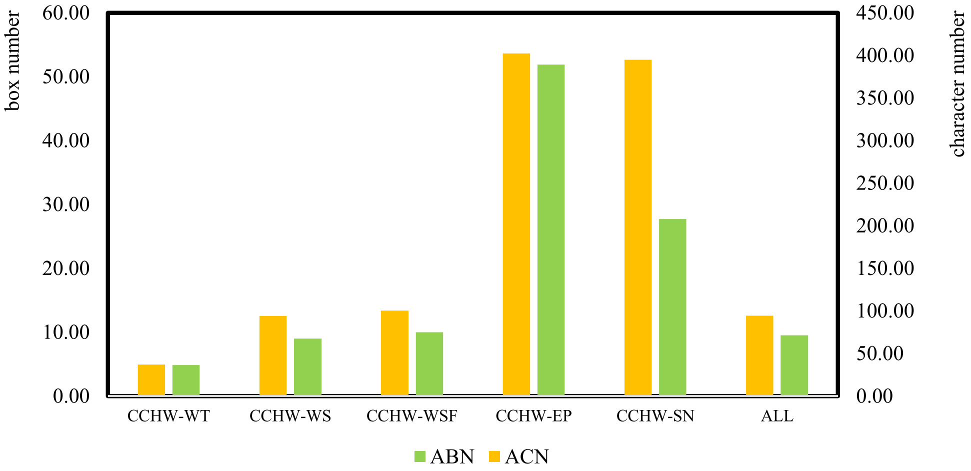


Fig. 3. Comparison of the five subsets of SCUT-HCCDoc in terms of the character number and text line box number. ABN is average box number; ACN is average character number.

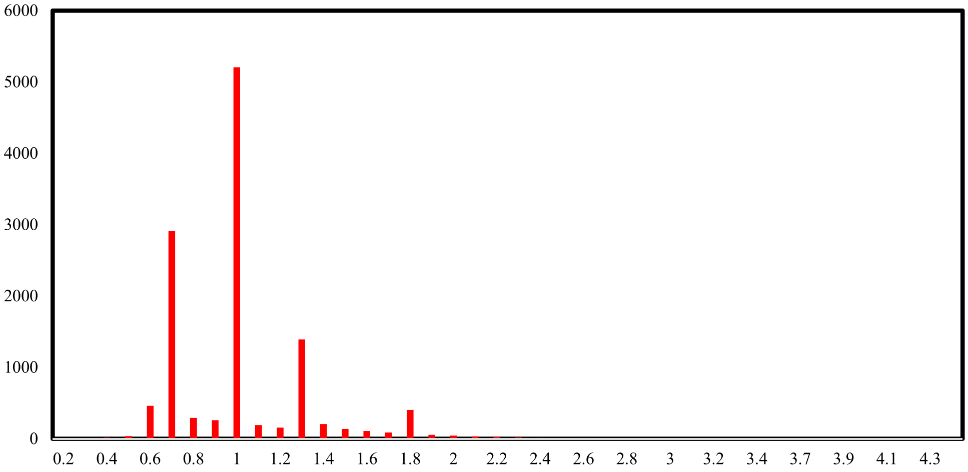


Fig. 4. Distribution of image aspect ratio.

Table 3
Number of character categories contained in occurrence frequency intervals.

Character Num. Interval	(10,000,)	(5000, 10,000]	(1000, 5000]	(500, 1000]	(100, 500]	(10, 100]	(, 10]
Category Num.	10 (< 0.2%)	7 (< 0.2%)	224 (3.67%)	242 (3.96%)	1056 (17.29%)	1912 (31.30%)	2658 (43.05%)

Table 4
Distribution of character category of training and test set of each subset.

Subset	ALL	HCCDoc-WT	HCCDoc-WS	HCCDoc-WSF	HCCDoc-EP	HCCDoc-SN
Class Num. of training set	5925	3201	3951	3726	2318	3206
Class Num. of test set	4437	2192	2864	2618	1741	2238
Class Num. of OOV	184	153	140	162	113	180

shown in Fig. 6. Fig. 6(a) shows the number of text line containing specific characters. Fig. 6(b) illustrates the top 50 most frequently observed character categories and the number of character instances in each category. Obviously, all of the text contains commonly used Chinese characters. Table 3 shows the number of characters in different character number intervals: 74% of the character categories appear no more than 100 times, and 43% appear even less than 10 times. That is, the number of frequently observed characters is larger, but the number of least used characters is much smaller. The imbalance of character frequency is consistent with real life and it becomes a challenge of HCTR. Moreover, we analyze the distribution of the character category of training and test set of each subset, respectively, as shown in Table 4. The out-of-vocabulary(OOV) characters refer to the characters that exist in the test set but not in the training set.

5. Text detection evaluation

For the text detection task on SCUT-HCCDoc, at least three challenges exits. First, some samples contain printed text or non-text interferences. The text detection algorithm must segment the written text and ignore the printed text or other interferences. Second, SCUT-HCCDoc contains the slanted text or vertical text. Sometimes, it is difficult to determine whether a text is written verti-

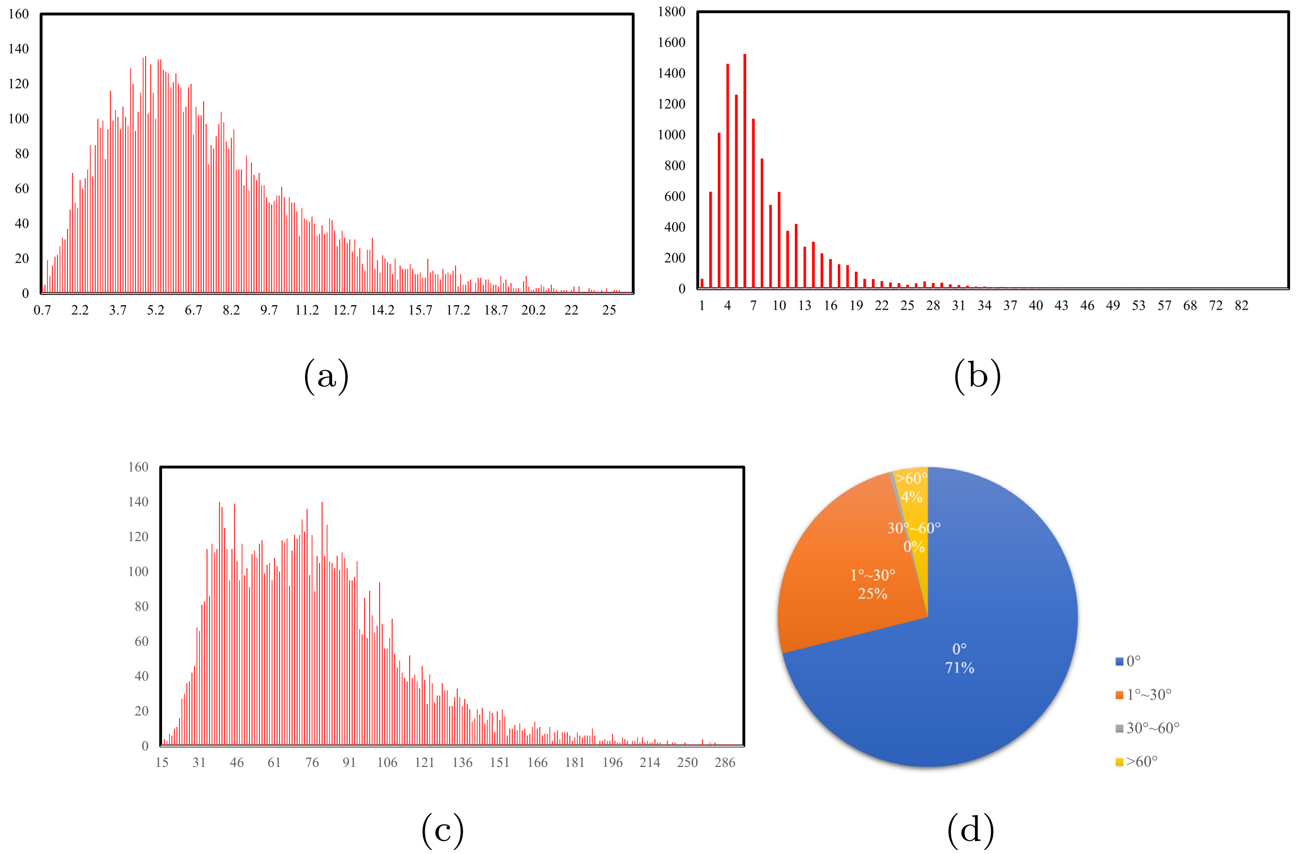


Fig. 5. Analysis and visualization for text-level diversity. (a) The aspect ratio of cropped images of the text line instance. (b) Number of images containing a specific text line instance. (c) Number of cropped images of the text line instance for a specific height. (d) Percentage of text line box angles.

cally or horizontally without additional semantic information. Examples will be given in the following experimental results. Third, the imbalance of text length and character size or resolution may also be a challenge.

5.1. Mask R-CNN for HCT detection

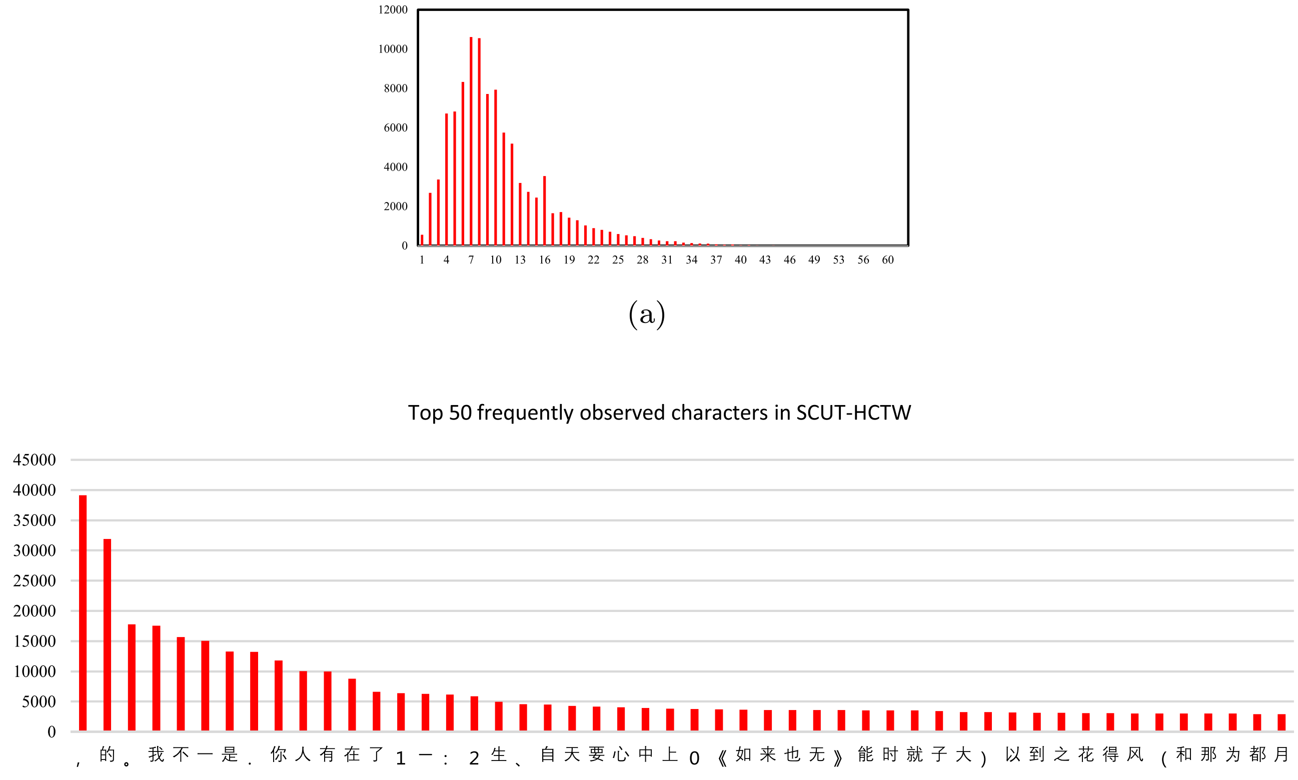
If we regard the separate text as a special instance, we can get inspiration from the instance segmentation task. Instance segmentation task attempts to determine the location of each occurrence of a given object within the image; hence it has much in common with text line detection [38]. We formulate the HCT detection of SCUT-HCCDoc using an instance segmentation problem and make a baseline evaluate using the Mask R-CNN [39]. As a state-of-the-art algorithm in instance segmentation, Mask R-CNN has a good performance in end-to-end text detection and recognition in many recent works. Unlike common instance segmentation datasets, SCUT-HCCDoc has no pixel-level text/non-text annotations, so we treat the pixels in the polygon as text, and the pixels outside as non-text, then we get an instance of the text area. Given an image, the output of Mask R-CNN is the mask of text line instance, which then leads to bounding boxes of text. The backbone of Mask R-CNN we used is based on ResNet50 [40] with FPN [41]. With SGD as an optimizer, the initial learning rate is 0.01. The model was pretrained on ImageNet and was subsequently trained for 180,000 iterations on SCUT-HCCDoc, with 2 images fed at one iteration. We only use the training set of SCUT-HCCDoc for training, and the SCUT-HCCDoc test set for testing.

As previously mentioned, SCUT-HCCDoc is divided into five subsets according to the image and text features. To further illustrate the different characteristics of each subset, we evaluated each subset separately. For text detection, a precise text bounding box is significant for the following recognizer. Therefore, we evaluated the detection results with different IOU constraints. Moreover, we evaluate SCUT-HCCDoc by re-implementing several state-of-the-art scene text detection methods, including PSENet [42], TextField [43], SBD [44] and DBNet [45].

5.2. Results and analysis

To evaluate the text detection, we use the popular evaluation criteria: Recall, Precision and F-score. The results are shown in Table 5 and Table 6. We can observe that Mask R-CNN obtained relatively good performance on the text detection task of SCUT-HCCDoc. However, when comparing results among different subsets, there are distinctions worth noting. For example, HCCDoc-SN and HCCDoc-EP have inferior results than the others because of the complex layout, dense text, and printed text interference. There are also slight differences when comparing the performance of various popular scene text detection methods on SCUT-HCCDoc. In Table 6, we can observe that it is still a challenging problem for the detection system to locate the text accurately (under higher IOU constrain). Take the detection results of Mask R-CNN as an example, we have analyzed the different detection errors for each subset.

For HCCDoc-WT/-WS, the detection errors are mainly caused by the confusion of two adjacent texts in one line, the background interference similar to text and the arbitrary-shaped text. For HCCDoc-WSF, the text with low resolution, the confusion of vertical or horizontal text on squared paper also caused the errors. For HCCDoc-EP, the dense text, the interference of printed text, the corrected text and the special handwritten text such as the score marking of examination paper, cause errors. For HCCDoc-SN, the



(a)

(b)

Fig. 6. Analysis and visualization for character-level diversity. (a) Number of text lines containing specific characters. (b) Number of character instances for the 50 most frequently observed character categories in the SCUT-HCCDoc.

Table 5
Results of text detection using different methods in different subsets.

Subset	All	HCCDoc-WT	HCCDoc-WS	HCCDoc-WSF	HCCDoc-EP	HCCDoc-SN
Precision(%)						
Mask R-CNN [39]	92.59	97.47	93.26	91.35	88.64	89.92
DBNet [45]	95.18	98.76	95.65	95.94	87.30	93.08
SBD [44]	**97.82**	**99.02**	**98.83**	**96.62**	**96.71**	**96.82**
PSENet [42]	87.91	92.17	89.4	86.84	85.59	81.61
TextField [43]	89.35	95.79	91.98	90.01	78.13	80.27
Recall(%)						
Mask R-CNN	**94.23**	97.56	**96.54**	**95.93**	**85.22**	**89.39**
DBNet	88.55	97.21	94.32	93.11	56.76	85.66
SBD	87.87	96.68	92.48	92.65	62.58	79.40
PSENet	86.86	93.71	93.47	92.23	51.03	86.64
TextField	89.15	**97.82**	95.24	94.16	59.45	82.12
F-score(%)						
Mask R-CNN	**93.40**	97.51	94.87	93.58	**86.90**	**89.65**
DBNet	91.75	**97.98**	94.98	94.50	68.79	89.22
SBD	92.58	97.84	**95.55**	**94.60**	75.99	87.25
PSENet	87.38	92.93	91.39	89.45	63.94	84.05
TextField	89.25	96.79	93.49	92.04	67.52	81.18

Table 6
Results of text detection using different methods in different IOU constraints.

Method	Recall(%)				Precision(%)				F-score(%)			
IOU	0.5	0.6	0.7	0.8	0.5	0.6	0.7	0.8	0.5	0.6	0.7	0.8
Mask R-CNN [39]	**94.23**	**92.61**	**88.57**	**72.11**	92.59	91.00	87.03	70.86	**93.40**	**91.80**	87.79	**71.48**
DBNet [45]	88.55	85.88	74.80	44.61	95.18	92.32	80.41	47.96	91.75	88.98	77.50	46.23
SBD [44]	87.87	86.99	83.56	66.92	**97.82**	**96.85**	**93.03**	**74.50**	92.58	91.65	**88.04**	70.50
PSENet [42]	86.86	84.20	78.87	58.28	87.91	85.22	79.82	58.99	87.38	84.71	79.34	58.63
TextField [43]	89.15	86.77	82.30	65.26	89.35	86.97	82.49	65.41	89.25	86.87	82.39	65.33

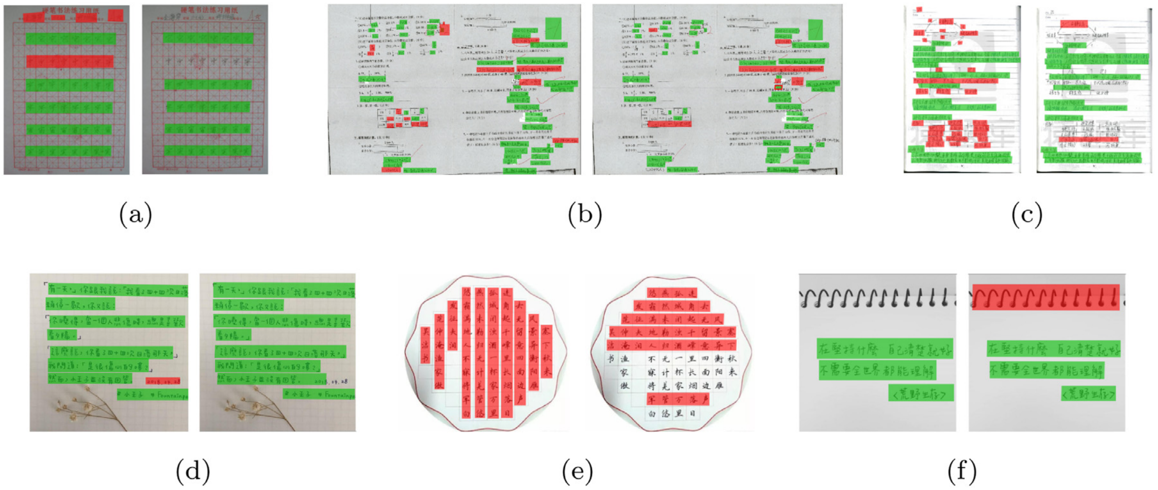

(a) (b) (c) (d) (e) (f)

Fig. 7. Some typical results of Mask R-CNN for HCT detection with different document images. (zoom in for better visualization) (a) shows the errors with undetected text, caused by illegible handwriting and the squared paper background. In (b-d), many texts are not detected because of printed text interference and small text scale. In (e), the vertical texts are detected as the horizontal because of the imbalance of arbitrary orientations in the training set, as shown in Fig. 5. In (f), there is a redundant detection instance because of the background interference which looks very similar to text.

main factors is the uncertainty and complexity of document layout, the non-texture handwriting and small text.

In Fig. 7, there are some typical detection results. The annotated images are one the left; the detection results are on the right. The green and red regions represent true positives and false positives. Visualizations are captured from the ICDAR official online evaluation system.[4]

6. Text recognition evaluation

Handwritten text recognition has been studied for more than 40 years [46,47]. Research problems includes online and offline handwriting segmentation and recognition [2,48–50]. However, offline HCT recognition (HCTR) remains a challenging problem due to the large character set, the diversity of writing style, the unconstrained language domain, and the difficulty of HCT detection. HCTR has two research directions: over-segmentation methods [2,51–53] and segmentation-free methods [50,54,55]. In recent years, the HCTR systems based on segmentation-free methods are more popular because they provide training using only the sequence-level labels as supervision and avoid the character-segmentation errors. The most popular segmentation-free systems can be roughly divided into the CTC-based [56–58] and the attention-based [59]. In the following experiments, we obtain two baseline results based on these two methods.

6.1. CTC-based recognizer for HCT recognition

As a popular choice for text recognition [56–58,60], CTC can perform seq-to-seq transcription without explicit detection information or prior alignment between the input image and its text label sequence. Improving the CRNN [61] structure, we construct our own framework with a customized CNN, multilayered residual LSTM, and transcription layer. The network architecture of our handwriting recognition system is shown in Fig. 8. First, feature vectors are extracted from the feature maps produced by the deep convolution network. The structure is: 8C3 - MP2 - 32C3 -MP2 - 128C3 - MP2 - 256C3 * 4 - MP2 - 512C3 - 512C3*1 - 512C2*1 - BN, where xCy represents a convolutional layer with kernel size of y × y and output channels of x, MPx denotes a maximum pooling layer with kernel size of x. Next, three layers of LSTMs with shortcut connections [40], are built on the top of the convolutional layers, as the recurrent layers. The LSTMs have a strong capability of capturing contextual information within a sequence. With the "shortcut connection", the network has a better ability to fuse the contextual information learned from the LSTM network and the character recognition information learned from the CNN. Finally, the per-frame predictions made by LSTM are converted into label sequences by CTC transcription. The class number of characters is expanded to 8615, which is the sum of character categories of SCUT-HCCDoc and CASIA [28].

6.1.1. Preprocessing

In the HCTR task, to reduce variation in the handwritten texts and preserve information that is relevant for recognition, several preprocessing steps are usually applied first. During the training, for the CASIA data [28], we first randomly cut a text line with a width of less than 576 from the original image. Next the cropped curved text is preprocessed as horizontal text by regressing the center line of the text. For the images with horizontal text, we determine two baselines (one above and one below), according to the following principle: the character pixels between the two baselines are 80% of the total pixels, where the character pixels refer to the pixels in bounding boxes for any characters. Then, we keep the area between the two baselines with a height of 64, and resize the whole images with the same ratio. Finally, by cropping or padding, images with height 126 and width 576 are obtained. Besides, we use isolated characters from CASIA-HWDB1.0–1.2 [28] to synthesize the semantic-free data that contains 750,000 text line images with a normalized size of 576 × 126.

For SCUT-HCCDoc, we consider the effect of image binarization preprocessing. In general, image binarization can make the text more prominent. Without the interference of the background, it is easier for the network to capture text features for learning. However, the camera-captured handwritten document images have various background interferences; hence, the binarization processing may have a variety of different effects on images. As shown in Fig. 9, the processed images of (a) and (b) are more recognizable because their backgrounds are clean. For the images (c) and (d), processed images are more difficult to read than the original. We can hardly recognize the processed images of (e) and (f) because of their non-white background and low resolution. Therefore, we experiment with the original images and binarization-processed

[4] http://rrc.cvc.uab.es/?ch=4&com=evaluation&task=1.

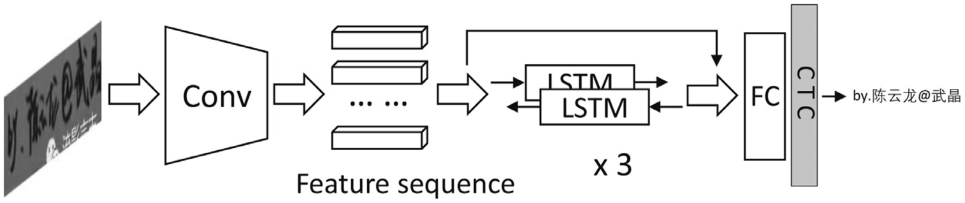
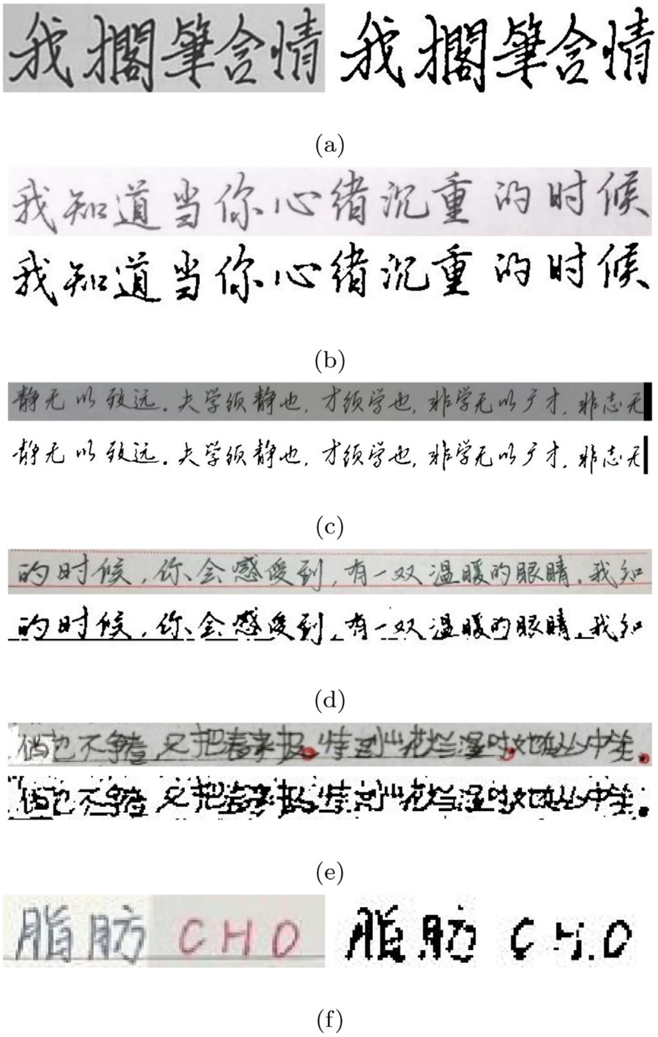


Fig. 8. Network architecture of our CTC-based recognizer.


(a)

(b)

(c)

(d)

(e)

(f)

Fig. 9. Some cropped images sample after adaptive binarization processing.

images of SCUT-HCCDoc. When training, all the images of SCUT-HCCDoc are resized to 576 × 126.

6.1.2. Training

As far as we know, the text recognition task usually requires a large amount of training data, especially for the Chinese language with a large character vocabulary. Therefore, we use extra CASAI data when training the recognizer to explore if extra external training data beneficial to the recognizer. First, the model is trained in a 1:1 ratio of mixed samples of CASIA training data and our CASIA synthetic data. Next, we train it with CASIA data and SCUT-HCCDoc data at a 1:1 ratio. For the CASIA training data, we randomly shuffle the characters in a text line during the early stage of training. SCUT-HCCDoc is the first camera-captured handwritten document dataset. We note significant differences between SCUT-HCCDoc and the previous datasets. At the image level, we note background interferences and geometric variances. At the text level, we have text rotation and text length imbalance. At the character level, we note diverse character resolutions and diverse people writing styles. Examples of the differences between SCUT-HCCDoc and CASIA dataset are shown in Fig. 10.

Table 7
Text recognition results for the CTC-based recognizer on SCUT-HCCDoc.

Training set	Test set	LM	CR(%)	AR(%)	ACC(%)
CASIA	COM.	√	97.10	96.79	–
		×	92.22	91.59	–
	HCCDoc	√	28.55	25.91	–
		×	23.88	23.00	–
HCCDoc	HCCDoc	√	81.01	78.65	33.67
		×	79.61	77.70	32.13
CASIA+HCCDoc	HCCDoc	√	78.50	78.45	30.37
		×	77.67	75.59	28.65

Table 8
Text recognition results for the CTC-based recognizer on SCUT-HCCDoc (with image binarization processing).

Training set	Test set	LM	CR(%)	AR(%)	ACC(%)
CASIA	COM.	√	97.10	96.79	–
		×	92.22	91.59	–
	HCCDoc	√	56.81	46.49	–
		×	52.19	46.65	-
HCCDoc	HCCDoc	√	78.01	73.49	30.51
		×	77.05	74.87	34.44
CASIA+HCCDoc	HCCDoc	√	80.43	75.50	34.00
		×	79.91	78.44	38.90

To compare the performance improvement with the additional CASIA data, we train two baselines with the SCUT-HCCDoc training set only and with extra CASIA data(including CASIA training set and our CASIA synthesis data).

6.1.3. Result and analysis

We use the evaluation correct rate (CR), accuracy rate (AR) [62] and the accuracy of the whole sequence (ACC) to evaluate the performance of the text recognition model. They are given by

$$CR = (N - D_e - S_e)/N,$$
$$AR = (N - D_e - S_e - I_e)/N,$$
$$ACC = A_c/A,$$

where N is the total number of characters in the ground truth, D_e, S_e and I_e represent deletion, substitution, and insertion errors, respectively. A_c represents the number of text lines correctly predicted, and A represents the number of all the text lines in test set.

The experiment results of our proposed CTC-based recognition system are shown in Tables 7 and 8. From Fig. 9, we know that the binarization processing of images significantly affects the legibility of text. In Table 7, it is noteworthy that the proposed recognition system can achieve a state-of-the-art result (CR 92.22%/97.10%

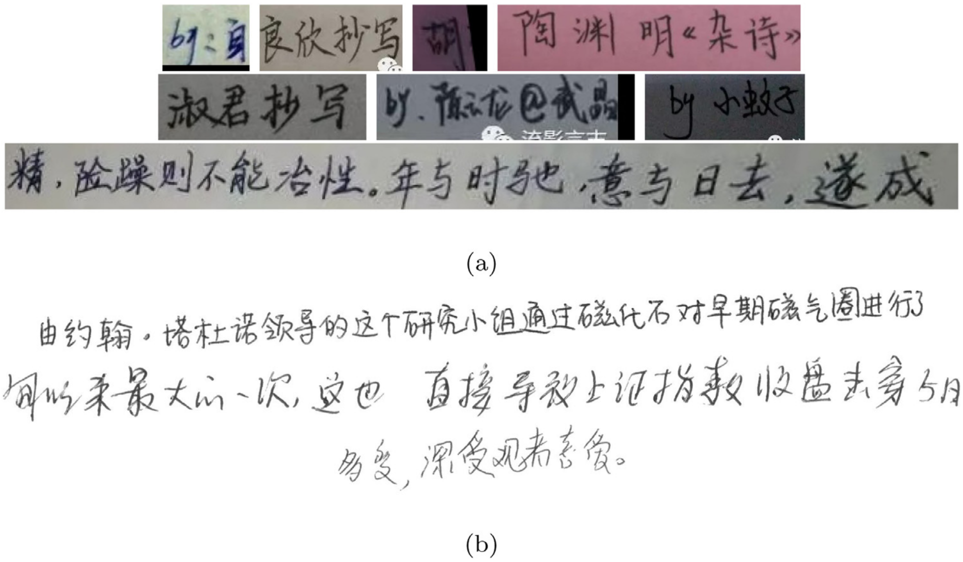


Fig. 10. Same cropped images sample of SCUT-HCCDoc (a) and CASIA (b). (a) Images with interference elements, complex background and great variation in the text. (b) Carefully scanned images with clean backgrounds.

and AR 91.59%/96.79% without/with language model) in the popular ICDAR2013 competition set (COM.) [62]. However, when applied to SCUT-HCCDoc, the system performance is much worse, which also proves that SCUT-HCCDoc is relatively challenging. Moreover, we can observe that when using the language model to decode, network performance usually has a relatively large improvement, for example, a 5% improvement of CR or AR when testing on the ICDAR2013 competition set. However, the improvement is limited for SCUT-HCCDoc. This is because the semantic information of text in SCUT-HCCDoc is varies: student notes, medical notes, poetry or lyrics. However, the applied language model does not cover such semantic information.

Besides, the cross-test shows that the network trained with CASIA data can only achieve the CR 28.55% (56.81%) and AR 25.91% (46.49%) on SCUT-HCCDoc (with binarization processing). Besides, when training with the CASIA training data together directly (without image binarization), there is no improvement (79.61% → 77.67%) for the system. It proves that there are great differences between the image features of CASIA and SCUT-HCCDoc. In Table 8, we present the results of SCUT-HCCDoc with binarization processing. Comparing it with Table 7, we can observe that the performances are worse for training with SCUT-HCCDoc data only. However, when adding the CASIA data to train, the network achieve a superior performance (74.87% → 79.91%). This is because the binarized images of SCUT-HCCDoc will have more common features with the white-background images of CASIA. Thanks to this, the network captures the image and text features better.

In Fig. 11, we show some recognition results for the original and binarization processed images, where green color indicates deletion error and red color indicates substitution and insertion error. The "pred1" labels are the prediction results of the original images, and "pred2" are results of the binarization processed. We can observe that the binarization processing makes the images easier to recognize for some samples, such as (e) and (f). However, for the others, such as (a), (b), and (g), it is not the case. For (a), there is red stroke interference. From the original images, it is easier to recognize for its distinctive color. However, the color information is missing in the binarization processed image, and the extra strokes will cause interference for the recognizer. For (b) and (g), because of low image resolution, the processed text is even harder to read.

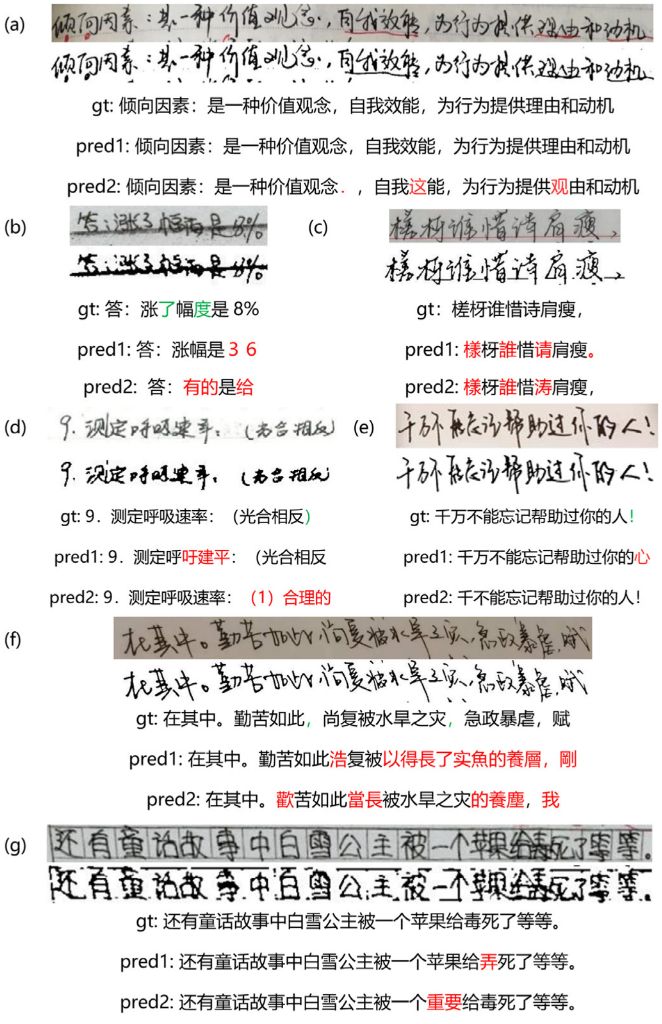

(a)
gt: 倾向因素：是一种价值观念，自我效能，为行为提供理由和动机
pred1: 倾向因素：是一种价值观念，自我效能，为行为提供理由和动机
pred2: 倾向因素：是一种价值观念. ，自我这能，为行为提供观由和动机
(b)
gt: 答：涨了幅度是 8%
pred1: 答：涨幅是 3 6
pred2: 答：有的是给
(c)
gt：槎枒谁惜诗肩瘦，
pred1: 樣枒誰惜请肩瘦。
pred2: 樣枒誰惜涛肩瘦，
(d)
gt: 9．测定呼吸速率：（光合相反）
pred1: 9．测定呼吁建平：（光合相反
pred2: 9．测定呼吸速率：（1）合理的
(e)
gt: 千万不能忘记帮助过你的人！
pred1: 千万不能忘记帮助过你的心
pred2: 千不能忘记帮助过你的人！
(f)
gt: 在其中。勤苦如此，尚复被水旱之灾，急政暴虐，赋
pred1: 在其中。勤苦如此浩复被以得長了实魚的養層，剛
pred2: 在其中。歡苦如此當長被水旱之灾的養塵，我
(g)
gt: 还有童话故事中白雪公主被一个苹果给毒死了等等。
pred1: 还有童话故事中白雪公主被一个苹果给弄死了等等。
pred2: 还有童话故事中白雪公主被一个重要给毒死了等等。

Fig. 11. Visualization of the recognition results. The "pred1" and "pred2" labels are the prediction results of the original and binarization processed images, respectively.

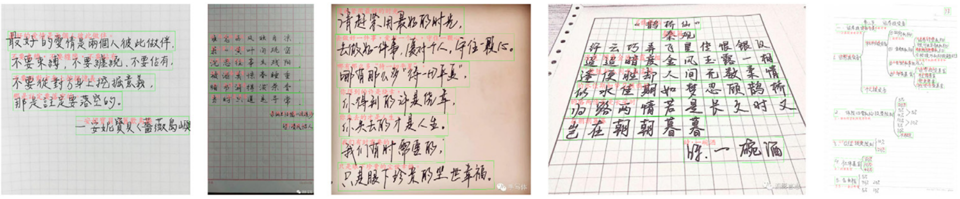


Fig. 12. Examples of results for end-to-end text spotting of SCUT-HCCDoc. Take the results of Mask R-CNN + CTC-based recognizer as an examples.

Table 9
Text recognition results for the attention-based recognizer.

Training set	Binarized	CR(%)	AR(%)
HCCDoc	√	59.79	56.06
	×	70.55	67.28
CASIA+HCCDoc	√	65.17	61.02
	×	73.70	70.81

Table 10
End-to-end text spotting results.

Method	LM	CR(%)	AR(%)	ACC(%)
Mask R-CNN [39]+	√	78.01	73.49	30.51
CTC-based recognizer	×	78.34	76.93	**36.88**
DBNet [45]+	√	**82.68**	**79.84**	29.54
CTC-based recognizer	×	81.93	79.50	26.81
SBD [44]+	√	78.55	75.59	25.84
CTC-based recognizer	×	76.85	74.28	20.87
PSENet [42]+	√	77.54	74.95	24.02
CTC-based recognizer	×	75.41	73.27	18.88
TextField [43]+	√	79.67	76.76	26.21
CTC-based recognizer	×	78.16	75.75	21.74

6.2. Attention-based recognizer for HCT recognition

The attention mechanism was first presented in Bahdanau et al. [63] to improve the performance of neural machine translation systems. But now it's flourished in many machine learning application domains including text recognition. For the attention-based recognizer, we conducted familiar experiments. The backbone of the network is the same as the CTC-based recognizer mentioned above, while the transcription method is the attention sequence-to-sequence model [64] instead. The attention-based decoder using the attention mechanism to generate output, conditioned on the input sequence produced by the previous feature extraction network. We use Adadelta optimizer with the initial learning rate of 1.0 to train. The network is trained for 50 epoches with batch size of 64. If extra CASIA training data is used, we randomly pick CASIA and HCCDoc data with a ratio of 1:1 to combine a batch used for training.

The results of our experiment are shown in Table 9. It can be observed that the attention-based recognizer has inferior performance than the CTC-based. The network trained with the attention mechanism has the best performance of CR 73.70%, worse than the CTC-based network with a CR of 81.01%. Especially, for the binarization processed images, the system performance is much more worse. As analyzed in Section 6.1.1, some binarization processed images might be more difficult to recognize, and this problem is particularly severe for attention-based systems.

Although, in recent years, attention-based methods have been a popular choice for the English text recognition problem [3,59,65,66], achieving state-of-the-art performance. For the HCTR problem, very few studies were reported to have successfully applied attention-based methods to deal with the HCTR problem. This is because missing or superfluous characters can easily cause the misalignment problem and mislead the training process for the attention module [67]. Moreover, this phenomenon becomes more severe in the HCTR problem, in which the character set is much larger, and the text length is much longer. The difference between these two tasks is analyzed in Zhu et al. [31].

7. End-to-end text spotting evaluation

In this section, we conduct an experiment for end-to-end handwritten text spotting in camera-captured document images.

Successively, the data are fed to the trained text detection system in Section 5 and text recognition system in Section 6.1, and we obtain an end-to-end text spotting result of SCUT-HCCDoc. It should be noted that only the predicted text bounding box with a matched ground truth box with IOU greater than 0.5 is sent to the recognizer. The evaluations of CR and AR are used to evaluate the performance. Thanks to the end-to-end experiment, we measured the reliability of the text bounding box generated by the detection system. In our experiment, we compared the quality of text bounding boxes generated by different text detection methods. The results are presented in Table 10. From the Table, we can observe that those text detection method get comparable performance. It is worth noting that the methods designed for text detection, such as DBNet, SBD and TextField, tend to have better performance than general object segmentation method Mask R-CNN. Some visualization results are shown in Fig. 12.

8. Conclusion

In this paper, we propose a new large-scale SCUT-HCCDoc dataset containing 12,253 camera-captured document images with 116,629 text lines and 1,155,801 characters. The dataset can be used for the construction, analysis and evaluation for scene handwritten Chinese text detection, recognition or end-to-end spotting. Both the samples and annotation of SCUT-HCCDoc were statistically analyzed. We systematically analyzed the sample diversity in SCUT-HCCDoc at the image level, text level and character level. Three kinds of baseline experiments were conducted, including several popular methods for text line detection, CTC-based/attention-based methods for text recognition and the combination of detectors with recognizer for end-to-end text spotting. The results indicate the diversity of SCUT-HCCDoc and the challenges of HCT understanding in natural camera-captured document images.

Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

References

[1] D. Doermann, K. Tombre, et al., Handbook of Document Image Processing and Recognition, Springer, 2014.
[2] Y.-C. Wu, F. Yin, C.-L. Liu, Improving handwritten chinese text recognition using neural network language models and convolutional neural network shape models, Pattern Recognit. 65 (2017) 251–264.

[3] J. Zhang, J. Du, S. Zhang, D. Liu, Y. Hu, J. Hu, S. Wei, L. Dai, Watch, attend and parse: an end-to-end neural network based approach to handwritten mathematical expression recognition, Pattern Recognit. 71 (2017) 196–206.

[4] Y. Liu, L. Jin, S. Zhang, C. Luo, S. Zhang, Curved scene text detection via transverse and longitudinal sequence connection, Pattern Recognit. 90 (2019) 337–345.

[5] S. Long, X. He, C. Yao, Scene text detection and recognition: The deep learning era, abs/1811.04256, 2018.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Proceedings of Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 248–255.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: Proceedings of European Conference on Computer Vision (ECCV), Springer, 2014, pp. 740–755.

[8] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[9] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L.G. i. Bigorda, S.R. Mestre, J. Mas, D.F. Mota, J.A. Almazán, L.P. de las Heras, Icdar 2013 robust reading competition, in: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 1484–1493.

[10] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V.R. Chandrasekhar, S. Lu, Icdar 2015 competition on robust reading, in: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 1156–1160.

[11] N. Nayef, F. Yin, I. Bizid, H. Choi, Y. Feng, D. Karatzas, Z. Luo, U. Pal, C. Rigaud, J. Chazalon, et al., Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-RRC-MLT, in: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), 1, 2017, pp. 1454–1459.

[12] N. Nayef, Y. Patel, M. Busta, P.N. Chowdhury, D. Karatzas, W. Khlif, J. Matas, U. Pal, J.-C. Burie, C.-l. Liu, et al., Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition–RRC-MLT-2019, in: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019, pp. 1582–1587.

[13] Y. Sun, Z. Ni, C.-K. Chng, Y. Liu, C. Luo, C.C. Ng, J. Han, E. Ding, J. Liu, D. Karatzas, et al., Icdar 2019 competition on large-scale street view text with partial labeling–RRC-LSVT, in: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019.

[14] C.-K. Chng, Y. Liu, Y. Sun, C.C. Ng, C. Luo, Z. Ni, C. Fang, S. Zhang, J. Han, E. Ding, et al., Icdar2019 robust reading challenge on arbitrary-shaped text (RRC-ART), in: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019.

[15] X. Liu, R. Zhang, Y. Zhou, Q. Jiang, Q. Song, N. Li, K. Zhou, L. Wang, D. Wang, M. Liao, M. Yang, X. Bai, B. Shi, et al., Icdar 2019 robust reading challenge on reading chinese text on signboard, in: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019.

[16] T. Yuan, Z. Zhu, K. Xu, C. Li, T. Mu, S. Hu, A large chinese text dataset in the wild, J. Comput. Sci. Technol. 34 (3) (2019) 509–521.

[17] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, X. Bai, Icdar2017 competition on reading chinese text in the wild (RCTW-17), in: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), 1, IEEE, 2017, pp. 1429–1434.

[18] W. He, X.-Y. Zhang, F. Yin, C.-L. Liu, Multi-oriented and multi-lingual scene text detection with direct regression, IEEE Trans. Image Process. 27 (11) (2018) 5406–5419.

[19] C.K. Ch'ng, C.S. Chan, Total-text: A comprehensive dataset for scene text detection and recognition, in: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), 01, 2017, pp. 935–942.

[20] Kai Wang, B. Babenko, S. Belongie, End-to-end scene text recognition, in: Proceedings of International Conference on Computer Vision (ICCV), 2011, pp. 1457–1464.

[21] K. Wang, S. Belongie, Word spotting in the wild, in: K. Daniilidis, P. Maragos, N. Paragios (Eds.), Proceedings of European Conference on Computer Vision (ECCV), Springer, 2010, pp. 591–604.

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

[23] A. Mishra, K. Alahari, C. Jawahar, Scene text recognition using higher order language priors, in: Proceedings of British Machine Vision Conference (BMVC), BMVA Press, 2012, pp. 127.1–127.11.

[24] T. Quy Phan, P. Shivakumara, S. Tian, C. Lim Tan, Recognizing text with perspective distortion in natural scenes, in: Proceedings of International Conference on Computer Vision (ICCV), 2013, pp. 569–576.

[25] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Ng, Reading digits in natural images with unsupervised feature learning, Adv. Neural Inf. Process. Syst. (NIPS) (2011).

[26] H. Zhang, J. Guo, G. Chen, C. Li, Hcl2000-a large-scale handwritten chinese character database for handwritten character recognition, in: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2009, pp. 286–290.

[27] S. Zhou, Q. Chen, X. Wang, Hit-or3c: an opening recognition corpus for chinese characters, in: Proceedings of International Workshop on Document Analysis Systems (IWDAS), ACM, 2010, pp. 223–230.

[28] C.-L. Liu, F. Yin, D.-H. Wang, Q.-F. Wang, Online and offline handwritten chinese character recognition: benchmarking on new databases, Pattern Recognit. 46 (1) (2013) 155–162.

[29] L. Jin, Y. Gao, G. Liu, Y. Li, K. Ding, Scut-couch2009-a comprehensive online unconstrained chinese handwriting database and benchmark evaluation, Int. J. Document Anal. Recognit. (IJDAR) 14 (1) (2011) 53–64.

[30] T. Su, T. Zhang, D. Guan, Hit-mw dataset for offline chinese handwritten text recognition, in: Proceedings of International Workshop on Frontiers in Handwriting Recognition (IWFHR), Citeseer, 2006.

[31] Y. Zhu, Z. Xie, L. Jin, X. Chen, Y. Huang, M. Zhang, Scut-ept: new dataset and benchmark for offline chinese text recognition in examination paper, IEEE Access 7 (2018) 370–382.

[32] J.J. Hull, A database for handwritten text recognition research, IEEE Trans. Pattern Anal. Mach. Intell. 16 (5) (1994) 550–554.

[33] E. Augustin, M. Carré, E. Grosicki, J.-M. Brodin, E. Geoffrois, F. Prêteux, Rimes evaluation campaign for handwritten mail processing, in: Proceedings of International Workshop on Frontiers in Handwriting Recognition (IWFHR), 2006, pp. 231–235.

[34] M. Pechwitz, S.S. Maddouri, V. Märgner, N. Ellouze, H. Amiri, et al., Ifn/enit-database of handwritten arabic words, in: Proceedings of CIFED'2002: colloque international francophone sur l'crit et le document (Hammamet, 21–23 octobre 2002), 2, Citeseer, 2002, pp. 129–136.

[35] U.-V. Marti, H. Bunke, The IAM-database: an english sentence database for offline handwriting recognition, Int. J. Doc. Anal. Recognit. 5 (1) (2002) 39–46.

[36] M. Kherallah, A. Elbaati, H. Abed, A. Alimi, The on/off (LMCA) dual arabic handwriting database, in: Proceedings of International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2008.

[37] T. Matsushita, M. Nakagawa, A database of on-line handwritten mixed objects named" kondate", in: Proceedings of International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2014, pp. 369–374.

[38] P. Schone, C. Hargraves, J. Morrey, R. Day, M. Jacox, Neural text line segmentation of multilingual print and handwriting with recognition-based evaluation, in: Proceedings of International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2018, pp. 265–272.

[39] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proceedings of International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 2961–2969.

[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 770–778.

[41] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2117–2125.

[42] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, S. Shao, Shape robust text detection with progressive scale expansion network, in: Proceedings of Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 9336–9345.

[43] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, X. Bai, Textfield: learning a deep direction field for irregular scene text detection, IEEE Trans. Image Process. 28 (11) (2019) 5566–5579.

[44] Y. Liu, S. Zhang, L. Jin, L. Xie, Y. Wu, Z. Wang, Omnidirectional scene text detection with sequential-free box discretization, in: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), AAAI Press, 2019.

[45] M. Liao, Z. Wan, C. Yao, K. Chen, X. Bai, Real-time scene text detection with differentiable binarization, in: Proceedings of Association for the Advancement of Artificial Intelligence (AAAI), 2020, pp. 11474–11481.

[46] F. Kimura, K. Takashina, S. Tsuruoka, Y. Miyake, Modified quadratic discriminant functions and the application to chinese character recognition, IEEE Trans. Pattern Anal. Mach. Intell. (1) (1987) 149–153.

[47] R. Dai, C. Liu, B. Xiao, Chinese character recognition: history, status and prospects, Front. Comput. Sci. China 1 (2) (2007) 126–136.

[48] W. Yang, L. Jin, D. Tao, Z. Xie, Z. Feng, Dropsample: a new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten chinese character recognition, Pattern Recognit. 58 (2016) 190–203.

[49] X. Xiao, L. Jin, Y. Yang, W. Yang, J. Sun, T. Chang, Building fast and compact convolutional neural networks for offline handwritten chinese character recognition, Pattern Recognit. 72 (2017) 72–81.

[50] Z. Xie, Z. Sun, L. Jin, H. Ni, T. Lyons, Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (8) (2017) 1903–1917.

[51] S. Wang, L. Chen, L. Xu, W. Fan, J. Sun, S. Naoi, Deep knowledge training and heterogeneous CNN for handwritten chinese text recognition, in: Proceedings of International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2016, pp. 84–89.

[52] X.-D. Zhou, D.-H. Wang, F. Tian, C.-L. Liu, M. Nakagawa, Handwritten chinese/japanese text recognition using semi-Markov conditional random fields, IEEE Trans. Pattern Anal. Mach. Intell. 35 (10) (2013) 2413–2426.

[53] Q.-F. Wang, F. Yin, C.-L. Liu, Unsupervised language model adaptation for handwritten chinese text recognition, Pattern Recognit. 47 (3) (2014) 1202–1216.

[54] Z.-R. Wang, J. Du, Writer code based adaptation of deep neural network for offline handwritten chinese text recognition, in: Proceedings of International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2016, pp. 548–553.

[55] W. Wang, J. Du, Z.-R. Wang, Parsimonious HMMS for offline handwritten chinese text recognition, in: Proceedings of International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2018, pp. 145–150.

[56] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (5) (2008) 855–868.
[57] P. Voigtlaender, P. Doetsch, H. Ney, Handwriting recognition with large multidimensional long short-term memory recurrent neural networks, in: Proceedings of International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2016, pp. 228–233.
[58] K. Dutta, P. Krishnan, M. Mathew, C. Jawahar, Improving CNN-RNN hybrid networks for handwriting recognition, in: Proceedings of International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2018, pp. 80–85.
[59] J. Sueiras, V. Ruiz, A. Sanchez, J.F. Velez, Offline continuous handwriting recognition using sequence to sequence neural networks, Neurocomputing 289 (2018) 119–128.
[60] A. Graves, J. Schmidhuber, Offline handwriting recognition with multidimensional recurrent neural networks, in: Proceedings of Advances in Neural Information Processing Systems (NIPS), 2009, pp. 545–552.
[61] B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, IEEE Trans. Pattern Anal. Mach. Intell. 39 (11) (2016) 2298–2304.
[62] F. Yin, Q.-F. Wang, X.-Y. Zhang, C.-L. Liu, Icdar 2013 chinese handwriting recognition competition, in: Proceedings of International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2013, pp. 1464–1470.
[63] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proceedings of International Conference on Learning Representations (ICLR), 2015.
[64] J.K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, Attention-based models for speech recognition, in: Proceedings of Advances in Neural Information Processing Systems (NIPS), 2015, pp. 577–585.
[65] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, X. Bai, Aster: an attentional scene text recognizer with flexible rectification, IEEE Trans. Pattern Anal. Mach. Intell. 41 (9) (2018) 2035–2048.
[66] C. Luo, L. Jin, Z. Sun, Moran: a multi-object rectified attention network for scene text recognition, Pattern Recognit. 90 (2019) 109–118.
[67] F. Bai, Z. Cheng, Y. Niu, S. Pu, S. Zhou, Edit probability for scene text recognition, in: Proceedings of Computer Vision and Pattern Recognition (CVPR), IEEE, 2018, pp. 1508–1516.

Hesuo Zhang received the B.S. degree from the school of Electronic and Information Engineering at the South China University of Technology, Guangzhou, China in 2019. He is currently pursuing the master degree in information and communication engineering at the South China University of Technology, Guangzhou, China. His current research interests include machine learning, deep learning, handwritten text segmentation and recognition.

Lingyu Liang received a B.E. and a Ph.D. degree from South China University of Technology (SCUT), in 2009 and 2014, respectively. From 2014 to 2016, he was a post-doctoral fellow with the School of Computer Science and Engineering, SCUT. From 2016 to 2017, he was an honorary post-doctoral fellow with The Chinese University of Hong Kong. He is currently an associate professor at SCUT. His research interests include image analysis and recognition, machine learning and computational photography.

Lianwen Jin received the B.S. degree from the University of Science and Technology of China, Anhui, China, and the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 1991 and 1996, respectively. He is currently a Professor with the School of Electronic and Information Engineering, South China University of Technology. He is the author of more than 100 scientific papers. Dr. Jin was a recipient of the award of New Century Excellent Talent Program of MOE in 2006 and the Guangdong Pearl River Distinguished Professor Award in 2011. His research interests include image processing, handwriting analysis and recognition, machine learning, and intelligent systems.