# Chapter 81
# The First Robust Mongolian Text Reading Dataset CSIMU-MTR

**Yunxue Shao, Guanglai Gao, Linbo Zhang, and Zhong Zhang**

**Abstract** Text extraction from various text containers like document, born digital images, real scenes and videos has been a continuous interest in this field for more than a decade. Although a lot of work has been done on printed Mongolian document image analysis, there has little work on Mongolian text extraction from complex images. For the design and evaluation of Mongolian text extraction algorithms and systems, the availability of large-scale dataset is important. This paper first introduces a dataset named CSIMU-MTR which is built by the College of Computer Science of Inner Mongolia University. And then presents benchmark results using two state-of-the-art methods in text detection on this new dataset. The reported results serve as a baseline for evaluating the further works.

**Keywords** Mongolian text extraction dataset • Scene text detection • Maximally stable extremal regions

## 81.1 Introduction

Past research has shown that a considerable amount of text on Web pages is presented in image form (17 %), while an important fraction of this text (76 %) is not to be found anywhere else in the Web page [1]. The use of images as text carriers stems from a number of reasons, for example in order to make the Web pages more beautiful (e.g. titles, headings etc.), to attract attention (e.g. advertisements), to hide information (e.g. images in spam emails used to

Y. Shao (✉) • G. Gao
College of Computer Science, Inner Mongolia University, Inner Mongolia, PR China
e-mail: csshyx@imu.edu.cn; csggl@imu.edu.cn

L. Zhang
China Academy of Transportation Sciences (CATS), Beijing, China
e-mail: zhanglinbo@163.com

Z. Zhang
College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin, China
e-mail: zhong.zhang8848@gmail.com



J. Mu et al. (eds.), *The Proceedings of the Third International Conference on Communications, Signal Processing, and Systems*, Lecture Notes in Electrical Engineering 322, DOI 10.1007/978-3-319-08991-1_81

avoid text-based filtering). Automatically extracting text from images would provide the enabling technology for a number of applications such as improved indexing and retrieval of Web content, enhanced content accessibility, content filtering etc. At the same time, efficient and fast comprehension of text in our environment is an important aspect of scene understanding for a variety of application areas, e.g. for automatic and assisted navigation of robots and humans respectively [2, 3].

In Asia, classical Mongolian is used by more than five millions of people. Mongolian text can be seen everywhere on the street in Inner Mongolia. Although there has been a lot of work on Latin text, Arabic text or Chinese text extraction from complex images [4–8] and born-digital images [9–12], and a lot of work on printed Mongolian document image analysis [13–16]. There has little work on Mongolian text extraction from complex images or born-digital images. For the design and evaluation of Mongolian text extraction algorithms and systems, the availability of large-scale dataset is important. This paper introduces a dataset named CSIMU-MTR built by the College of Computer Science of Inner Mongolia University. This dataset contains more than 500 complex real-scene images captured by high-resolution camera and born-digital images downloaded from Web pages. The dataset is publicly available for academic research.

The rest of the paper is organized as follows. Section 81.2 describes the CSIMU-MTR dataset. Section 81.3 presents the experimental results of two text detection methods. Finally, the last section gives conclusions.

## 81.2 Dataset

The dataset was built up similarly to ICDAR2011 robust reading datasets. Overall, we collected a set of 560 images which includes real-scene images and born-digital images. The dataset was split into a training set of 400 images and a test set of 160 images randomly. Real-scene images were captured with digital camera using auto focus and natural lighting. This kind of images containing text in a variety of colors and fonts on many different backgrounds and in various orientations, which pose considerable challenges to text detection, such as blurred or out of focus frames, low-contrast, over-exposure, uneven lighting, complex backgrounds, and lens distortion. Born-digital were collected by downloading from search engines. Born-digital images are usually low-resolution (for fast transmitting or displaying), non-uniform color and often suffer from compression artefacts and severe anti-aliasing.

Reading text in images consisted of two steps. The first step is to identify text regions and mark their location with axis-aligned rectangular bounding boxes. The second step is to recognize cropped word images of scene text. Accordingly, the ground truth is prepared in two phases. In the first phase, we prepared text location ground truth. The bounding boxes are tight so they touch most of the boundary pixels of a word. In the second phase, we will prepare word recognition ground truth.


**Fig. 81.1** An example image in the dataset and its corresponding ground truth file

At the current stage, the first phase is finished and the dataset is open for the study of text localization. Each image in the dataset corresponds to a ground truth TXT file which contains the set of bounding rectangles. Each bounding rectangle is determined by the top-left point and the bottom-right point. Each line in the TXT file corresponds to: x-axis value and y-axis value of the top-left point, x-axis value and y-axis value of the bottom-right point. An example image and its corresponding ground truth file is shown in Fig. 81.1.

## 81.3 Benchmark Results

A text localization system generally consists of two major components: candidate text region extraction and text region filtering. In candidate text region extraction step, according to the features utilized, text localization methods can be categorized into region-based and texture-based. The problem with traditional texture-based methods is their computational complexity in the texture classification stage, which accounts for most of the processing time. In particular, texture-based filtering methods require an exhaustive scan of the input image to detect and localize text regions. This makes the convolution operation computationally expensive. Region-based methods use the properties of the color or gray scale in a text region or their differences with the corresponding properties of the background. These methods can be further divided into two sub-approaches: connected component (CC)-based and edge-based. These two approaches work in a bottom-up fashion: by identifying sub-structures, such as CCs or edges, and then merging these sub-structures to mark bounding boxes for text. Due to their relatively simple implementation and effectiveness, region-based methods are widely used. In this paper, we present benchmark results using two region-based methods on this new dataset. One is based on

the Canny edges and the other one is based on the Maximally Stable Extremal Regions (MSER). In recent years, MSER is usually used for extracting character like regions and ICDAR2013 robust reading competition results [17] demonstrate that MSER based methods perform better.

The flowchart of edge based text detection method used in our experiments is shown in Fig. 81.2. In the pre-processing stage, the input image is first resized into different scales. And at each image scale, the resized image is smoothed by 5 × 5 gaussian mask and Canny edge image is computed on the pre-processed image. Before edge merging, the edge image is first smoothed into a gray image and then been binarized for eliminating small edges and connecting adjacent edges. In the edge merging step, some heuristic knowledge, such as closeness, alignment, and comparable height is used and the candidate text regions are generated in this step. For each candidate region, a SVM classifier with RBF kernel is used to determinate whether or not this region is a Mongolian text like region, this string level classifier is denoted as StringSVM in this paper. StringSVM was trained on a set of 2,000 Mongolian text regions and 5,000 non-Mongolian text regions obtained by manually selected from candidated text regions extracted by the edge merging step. Bag of visual words (BoVW) model is used for feature extraction on each region. Finally, some refinement processes, such as region reduction or expansion are done on the text like regions. Figure 81.3 illustrates an example of the results in each step.

In the BoVW model, the extracted feature is a vector of occurrence counts of a vocabulary of local image features. To achieve this, it usually includes three steps:

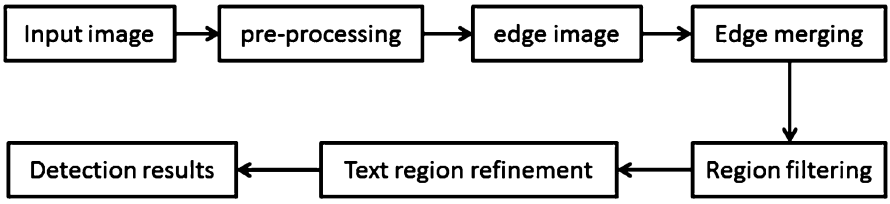


**Fig. 81.2** The flowchart of edge based text detection method

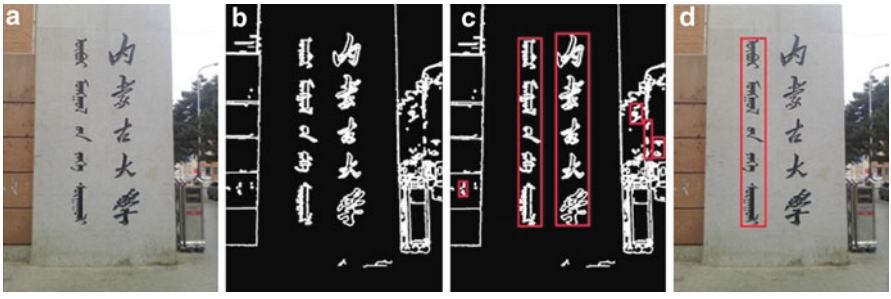


**Fig. 81.3** An example of the results in each step of edge based method. (**a**) the input image (**b**) the smoothed Canny edge image (**c**) edge merging result (**d**) text detection result

feature detection, feature description and codebook generation. In this paper, dense sampling with different sliding window size is used in the feature detection step. In the feature description step, three HOG [18] feature vectors are extracted on each sampled image patch at three scales. Three codebooks are generated at each scale accordingly by K-means clustering method. Finally, each HOG feature is projected into its corresponding codebook. Join these histograms together results in the final feature vector. The feature extraction method is illustrated in Fig. 81.4.

The flowchart of MSER based text detection method used in our experiments is shown in Fig. 81.5. The pre-processing and the text region refinement methods used in this method are the same as used in edge based method. In the MSER detection stage, the Vlfeat open source library [19] was used to detect MSER regions. Then the MSER regions are merged into a candidate text regions using the same heuristic knowledge as used in edge based method. Finally, the candidate text regions are classified by StringSVM. An example of the results in each step is illustrated in Fig. 81.6.

For the evaluation of text localization results, the framework proposed by Wolf and Jolion [20] is used in our experiments. The key principle of the scheme is that evaluation is done at the object level over the whole collection, taking into account the quality of each match between detected and ground truth text boxes. Matches are first determined based on area overlapping. Then different weights for one-to-one, one-to-many and many-to-one matches are used when pooling together the results.

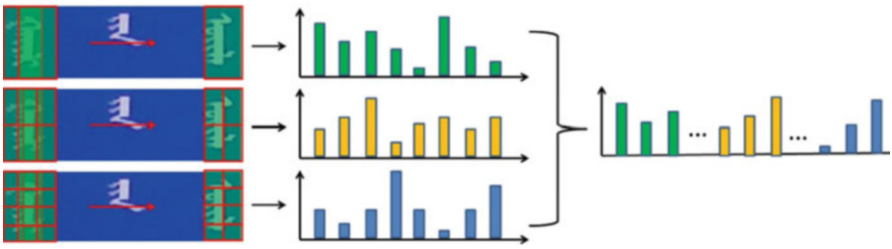

**Fig. 81.4** Feature extraction method used in text region classification

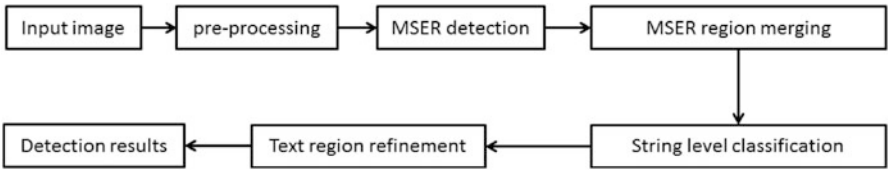


**Fig. 81.5** The flowchart of MSER based text detection method

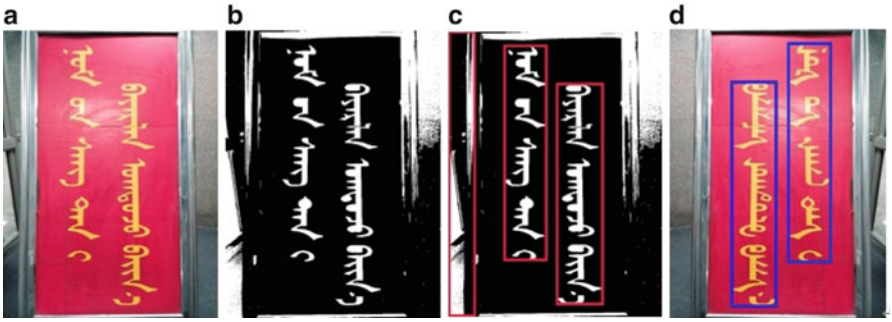


**Fig. 81.6** An example of the results in each step of MSER based method. (**a**) the input image (**b**) MSER regions (**c**) region merging result (**d**) text detection result

**Table 81.1** Text localization results

| Method | Recall (%) | Precision (%) | F-score (%) |
|---|---|---|---|
| Edge based | 61.13 | 72.36 | 66.27 |
| MSER based | 63.56 | 74.25 | 68.49 |

Results of this two method on this new dataset are shown in Table 81.1. MSER based method performs a little better than edge based method. Form these results we can see that this problem is still a challenge and it opens a large room for research and improvement.

An crucial problem of the proposed edge based method is that if text edges connect with background edges, the edge merging step would fail to give an accurate candidate text region. The problem of the proposed MSER based method is that if the gray value or color in the same text differs, the MSER detection step would fail to detect an accurate MSER region. Some examples of detection failures are depicted in Fig. 81.7. Most of them are caused by nonuniform color, lower resolution, uneven illumination, low-contrast or over-exposure which cannot generate complete MSERs or Canny edges for component merging. Some others are caused by non-regular text such as hand-written fonts and art fonts, since they are difficult to be correctly predicted by text classifier.

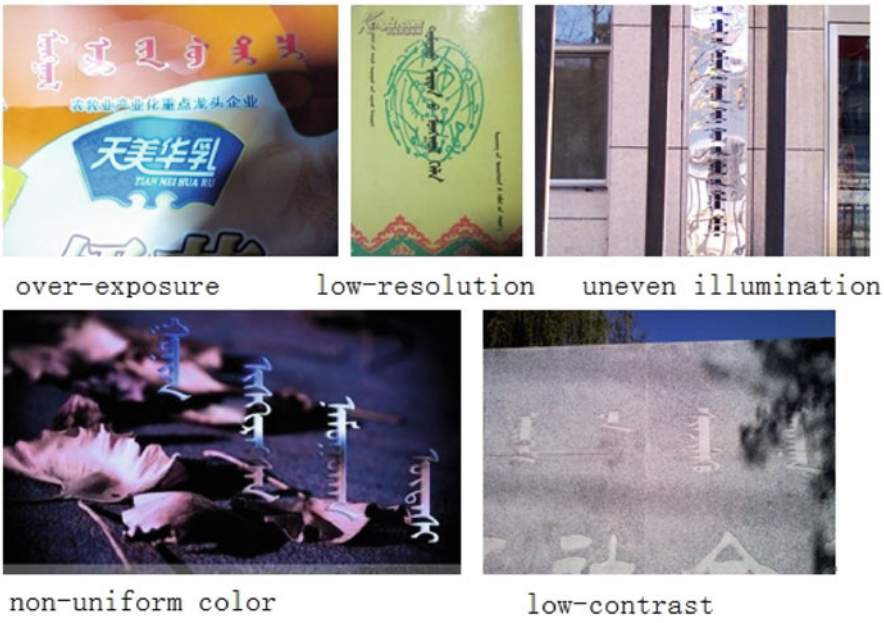


**Fig. 81.7** Detection failures of MSER based or edge based method

**Conclusion**

In this paper, the first robust Mongolian text reading dataset named CSIMU-MTR which is built by the College of Computer Science of Inner Mongolia University is introduced. This dataset poses considerable challenges to text detection such as blurred image, low-contrast, over-exposure, uneven lighting, complex backgrounds, non-uniform color. Benchmark results are presented using two state-of-the-art methods in text detection on this new dataset. The reported results serve as a baseline for evaluating the further works.

**Acknowledgements** This work was supported by program of higher-level talents of Inner Mongolia University.

## References

1. Antonacopoulos A, Karatzas D, Ortiz Lopez J (2001) Accessing textual information embedded in internet images. In: Proceedings of SPIE, Internet Imaging II, vol 4311, pp 198–205
2. Merino-Gracia C, Lenc K, Mirmehdi M (2012) A head-mounted device for recognizing text in natural scenes. In: CBDAR'11, LNCS, vol 7139, pp 29–41

3. Gao J, Yang J (2001) An adaptive algorithm for text detection from natural scenes. In: CVPR'01, vol 2, pp 84–89
4. Epshtein B, Ofek E, Wexler Y (2010) Detecting text in natural scenes with stroke width transform. In: 2010 I.E. conference on computer vision and pattern recognition, pp 2963–2970
5. Weinman JJ, Butler Z, Knoll D, Feild J (2014) Toward integrated scene text reading. IEEE Trans Pattern Anal Mach Intell 36(2):375–387
6. Yi C, Tian Y (2011) Text string detection from natural scenes by structure-based partition and grouping. IEEE Trans Image Process 20(9):2594–2605
7. Koo H, Kim D (2013) Scene text detection via connected component clustering and nontext filtering. IEEE Trans Image Process 22(6):2296–2305
8. Shahav A, Shafait F, Dengel A (2011) ICDAR 2011 robust reading competition challenge 1: reading text in scene images. In: 2011 international conference on document analysis and recognition, pp 1491–1496
9. Karatzas D, Mestre SR, Mas J, Nourbakhsh F, Roy PP (2011) ICDAR 2011 robust reading competition-challenge 1: reading text in born-digital images (web and email). In: 2011 international conference on document analysis and recognition (ICDAR), pp 1485–1490
10. Karatzas D, Antonacopoulos (2007) A colour text segmentation in web images based on human perception. Image Vision Comput 25(5):564–577
11. Lopresti D, Zhou J (2000) Locating and recognizing text in WWW images. Inf Retr 2:177–206
12. Perantonis SJ, Gatos B, Maragos V (2003) A novel web image processing algorithm for text area identification that helps commercial OCR engines to improve their web image recognition efficiency. In: Second international workshop on web document analysis (WDA2003), pp 61–64
13. Gao G, Li W, Hou H, Li Z (2003) Multi-agent based recognition system of printed Mongolian characters. In: Proceedings of the international conference on active media technology, pp 376–381
14. Gao G, Su X, Wei H, Gong Y (2011) Classical Mongolian words recognition in historical document. In: Proceedings of the 11th international conference on document analysis and recognition (ICDAR), pp 692–697
15. Wei H, Gao G (2014) A keyword retrieval system for historical Mongolian document images. Int J Doc Anal Recognit 17(2):33–45
16. Peng L, Liu C, Ding X et al (2010) Multi-font printed Mongolian document recognition system. Int J Doc Anal Recognit 13(2):93–106
17. Karatzas D et al (2013) ICDAR 2013 robust reading competition. In: 2013 international conference on document analysis and recognition, pp 1484–1493
18. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 886–893
19. Vedaldi A, Fulkerson B (2008) An open and portable library of computer vision algorithms. http://www.vlfeat.org/
20. Wolf C, Jolion J.-M (2006) Object count/area graphs for the evaluation of object detection and segmentation algorithms. Int J Doc Anal Recognit 8(4):280–296