MHDT: A Deep-Learning-based Text Detection Algorithm for Unstructured Data in Banking

Shenglan Ma
Division of Science and Technology,
Fujian Rural Credit Union
Fujian, China

Lingling Yang
Division of Innovation, China
UnionPay Co., Ltd
Fujian, China

Hao Wang
Department of Computer Science,
Norwegian University of Sci. & Tech.
Gjøvik, Norway

Hong Xiao
College of Computer,
Guangdong University of Technology
Guangzhou, China

Hong-Ning Dai
Faculty of Information Tech.
Macau University of Sci. and Tech.
Fujian, China

Shuhan Cheng&Tongsen Wang
Division of Science and Technology,
Fujian Rural Credit Union
Fujian, China

ABSTRACT
Text detection in natural scene images becomes highly demanded for unstructured data in banking. In this paper, we propose a new deep learning algorithm called *MSER, Hu-moment and Deep learning for Text detection* (MHDT) based on Maximum Stable Extremal Regions (MSER) and Hu-moment features. Firstly, we extract MSERs as candidate characters. Secondly, a character classifier is introduced with Hu-moment features to reduce the number of input for clustering. After single linkage clustering, a text classifier trained from a *Deep Brief Network* is used to delete non-text. The proposed algorithm is evaluated on the ICDAR database, and the experimental results show that the proposed algorithm yields high precision and recall rate.

CCS Concepts
- **Computer systems organization~Neural networks**

Keywords
Text detection; Unstructured data; Deep learning.

1. INTRODUCTION
With the continuous development of the Internet banking, unstructured data account for 80% of bank data. The integration of cloud computing, big data and artificial intelligence to effectively manage unstructured data becomes the general trend in the development of financial technology [1]. At present, most banks have established unstructured data platforms to store, access and process unstructured data. One of the most critical functions in an unstructured data platform is the ability to detect text from documentary materials in natural scenes such as vouchers, tickets, and reports in a banking scenario to enable rapid pattern recognition and OCR recognition [2].

However, it is a big challenge to detect text in natural scene images due to the complicated background and wide variety of text appearance [3]. The most representative methods can be divided into two categories: region-based and connected component-based (CC-based) [4]. Region-based methods, also known as the sliding-window-based methods, use a sliding window to find candidate texts and then get the real texts with a classifier. Lee et al. [5] put forward an *adaboost*-based algorithm introduced with CART. The algorithm first divides the original image into 16 different scales; then it combines the detecting results after applying adaboost classifiers at each scale. These methods are slow as the image must be processed at multiple scales [6]. CC-based methods extract connected components as candidate characters followed by grouping them into candidate text. Epshtein et al. [7] proposed a stroke based detect algorithm, which consists of four main processes: edge detection, stroke width conversion, text candidate detection, and deletion. Yi et al. [8] proposed a geometric grouping and adaboost text detection algorithm based on MSER in 2012. CC-based methods need additional checks to remove false positives [9].

This paper proposes an effective algorithm to detect text in natural scene images, called *MSER, Hu-moment and Deep learning for Text Detection* (MHDT). Firstly, Maximum stable extremal regions (MSERs) are extracted as candidate characters. Some non-character candidates are filtered out by an adaboost classifier, which is trained from some geometrical features. Secondly, single linkage cluster is used to combine characters into candidate texts. Lastly, we use a deep learning model—*Deep Belief Network* (DBN) to remove non-text elements. As the input of deep belief network need to be normalized, we adopt algorithm of median filtering to improve image quality. Our method is trained on the competition training dataset and tested on various benchmark datasets. The experimental results show the excellent performance of our algorithm.

This paper is divided into five sections. Section 2 introduces the unstructured data application scenarios in banking and related text detection algorithms. Section 3 describes the proposed MHDT algorithm in the unstructured data platform. In Section 4 we present experiment results. Section 5 concludes the paper.

2. RELATED WORKS
2.1 Unstructured Data Application Scenarios in Banking
In the existing system architecture of modern banks, structured data hold only a small fraction of the overall data. Compared with the structured data used in transaction information, process control and statistical analysis in business information systems, unstructured data has unique and continuous values that can be used in sharing, retrieval, and analysis. The management and use of unstructured data have an important impact on banking. Banks generate large amounts of vouchers, notes, statements, files, and other unstructured data every day. Unstructured electronic data constitutes 80% of the data field of financial institutions, doubling the number every four years, and the total amount and growth rate of data are enormous. Unstructured data is directly related to the

core value of the bank, including credit management, compliance management, risk management, and customer service. The bank's unstructured data involve many business areas, and there are a large number of business transferring and sharing requirements. The operation of unstructured data is often an important part of business processes, directly affecting the efficiency of business processes, and often has strong life cycle management and security requirements.

The unstructured data of the bank comes from the following aspects: First, all the original documents of the business that need to be archived, such as the account opening application, the business approval process data and various business documents will be scanned and stored into electronic documents; Second, the service recording of the call center needs to be permanently saved; the third is the audio and video materials of some conferences; the fourth is the interactive information on the portal website. In the unstructured data application scenario, one of the most critical functions is the detection of text under the natural scene image, such as the detection of key text in the picture taken by the transaction scene picture or the voucher shooting equipment.

2.2 Text Detection in Natural Scene Images

Chen et al. proposed a maximum stable extreme value region text detection algorithm based on edge enhancement [10], which belongs to the method based on connected regions. Although MSER has many properties such as affine invariance, the MSER algorithm is very sensitive to image blur. In order to be able to detect smaller text in low-convolution images, the authors propose a method combining canny edge detection algorithm with MSER algorithm. In addition, in order to improve the efficiency of text detection, the author proposes a stroke width conversion method, which uses the stroke width image to represent the maximum stable extreme value area.

Yin et al. proposed a geometric grouping and adaboost text detection algorithm based on the maximum stable extremum region [6]. First, the algorithm uses the maximum stable extremum region as the candidate letter; second, uses the geometric properties of the letter to delete some non-letter candidates, and uses the disjoint set to combine similar candidate letters; third, calculate the characteristics of each candidate region, including vertical and horizontal rate of change, stroke width, color and other geometric features; Finally, feature training is performed by the adaboost classifier, and non-text candidates are deleted to obtain the final text detection result.

Xucheng Yin et al. proposed a robust text detection method based on MSER pruning algorithm [3]. The algorithm first extracts MSER as candidate letters, then removes non-letter candidates through pruning algorithm and finally gets the final through the adaboost classifier. text.

Deep learning is a hot topic in recent years. Its main purpose is to establish a network that can simulate human brain learning, and classify or predict by automatically learning the characteristics of objects. Hinton et al. proposed the concept of deep learning in 2006, defining a deep learning structure as a multi-layer perceptron with multiple hidden layers, by combining lower-level features to obtain a more abstract high-level representation, better describing the properties of the object or features to discover the distribution characteristics of the data [11][12].

The deep belief network (DBN) is a machine learning model that has emerged in recent years. It combines an unsupervised learning process with a supervised learning process, and has gradually achieved some achievements in the field of image processing. Literature [13] classifies images by differential deep belief network, and achieves better classification results; [14] uses SVM as the classifier of the last layer of DBN network to identify expressions in images; [15] is the application of DBN in image detection, and the vehicle in the image is detected by DBN.

3. The Proposed MHDT ALGORITHM

In this section, we present our MHDT algorithm. Note that we have developed MHDT as a SDK package, which can be called by business applications connected to the unstructured data platform.

3.1 Banking Unstructured Data Platform and MHDT Module

The overall functions of the unstructured data platform are mainly composed of service access subsystem, business processing subsystem, transmission cache subsystem, unified access subsystem, content management subsystem, configuration management subsystem and lifecycle management subsystem. The seven subsystems are shown in Fig. 1.

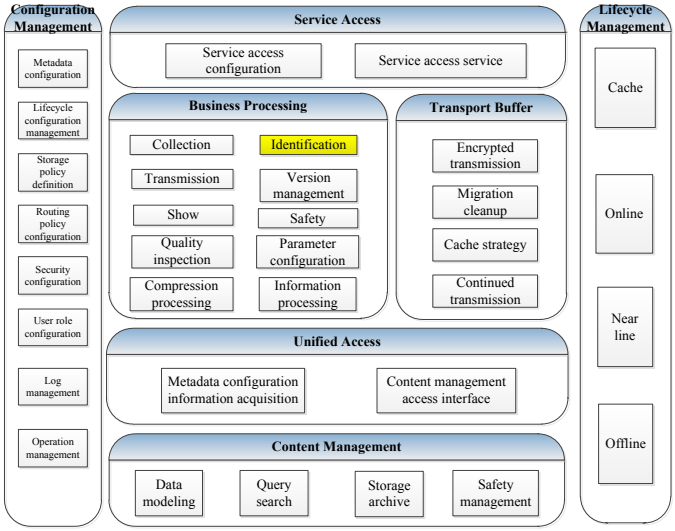

Figure 1. Unstructured Data Platform System Architecture

1. Service access subsystem

The service access subsystem provides a unified access interface for the external system, and provides URL, JAVA, and VC service access interfaces, and supports B/S and C/S mode access. By implementing the URL interface, the processing interface can be invoked to implement the application of the external system.

2. Business processing subsystem

Business processing subsystem provides complete, comprehensive image processing capabilities, supporting service-oriented, image-oriented controls, and image-oriented SDKs. Service-oriented provides image access unified URL interface, external system calls URL interface to realize image access application. Image-oriented control is integrated inside the control image processing SDK, and the specific implementation details are shielded externally. The external system can directly call the image control interface. The image processing API is provided for the image processing SDK, and the external system can call the image processing SDK for deep customization development.

The text detection algorithm studied in this paper is provided as a package for the SDK, as shown in the yellow box in Fig. 1.

3. Transport buffer subsystem

Transport buffer subsystem provides a secure and reliable (encrypted transmission, breakpoint resume, etc.) transmission mechanism, does not require database support, supports distributed deployment, and meets the performance requirements of business systems for unstructured data platforms.

4. Unified access subsystem

Unified access subsystem implements synchronization of user accounts of each system through unified user authentication to support unified authentication and authority authentication control requirements of the application system.

5. Content management subsystem

Content management subsystem will plan the content management space and the corresponding storage devices for the business variety according to the size of the business system data file storage.

6. Configuration management subsystem

Configuration management subsystem supports the browser-based unified configuration management tool to manage the unstructured data platform. The configuration management center can manage the metadata model of the data storage.

7. Lifecycle management subsystem

Lifecycle management subsystem is responsible for the control of the complete process of unstructured data from generation to retrieval, migration, storage, archiving, and long-term preservation management, offline, and destruction.

3.2 The Proposed Text Detection Algorithm

The structure of the MHDT, as well as the sample result of each stage is presented in Fig.2.

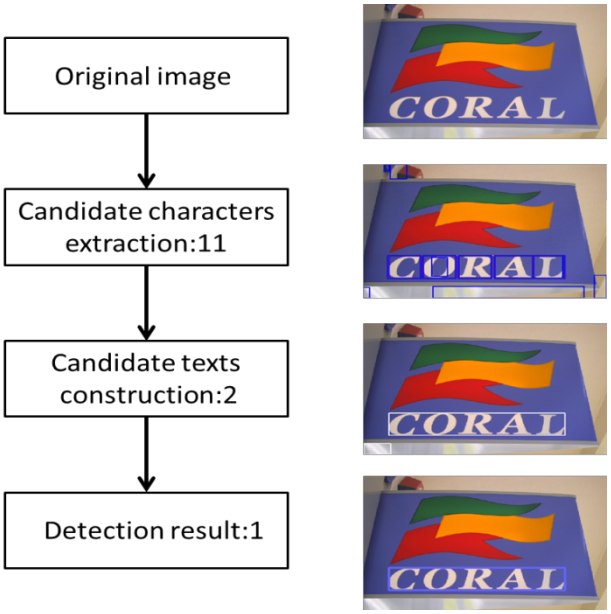

Figure 2. Flowchart of the MHDT and results after each step of the sample

1. Character candidate extraction

Maximum stable extremal region, proposed by Matas [16], is a method to find correspondences between images of different viewpoints. There is a significant difference in brightness and color between backgrounds and texts in natural scene images. MSER has been reported as one of the best region detector [17].

2. Character classifier with Hu-moment

The MSERs extracted are repeated regions of the image, making the input number for clustering too large. The extracted MSER is shown in Fig. 3.

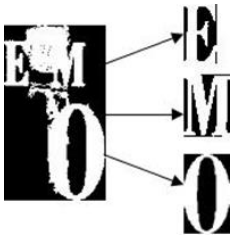

Figure 3. MSERs detected by MSER

As can be seen in Fig. 3, the regions “e”, m”, “o” are sub regions of the left region. We design a character classifier trained from geometrical features as well as Hu-moment. Moments play an important role in describing geometrical characters of images. Hu-moment contains seven invariant moments, which are proved to be rotation, translation and zoom invariant [18]. Moreover, Hu-moments are eigenvalues of contours, suitable for contours extracted as MSERS in our method. Therefore, Hu-moment is a feasible feature.

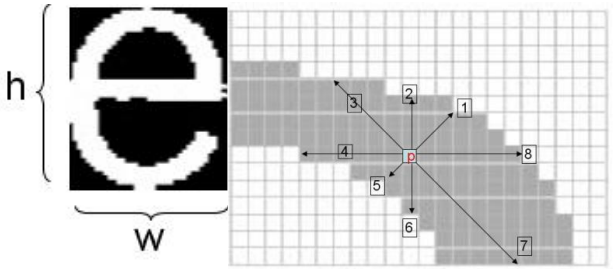

Figure 4. Candidate characters (left) and stroke width (right)

Other geometrical features are as follows. As can be seen in Fig. 4 (left), the height of a candidate character is the height of external matrix for a MSER. Second is the width. The third feature is aspect ratio, namely h divides w. The forth one equals to the number of pixels in white divides h and w. Stroke width of a pixel is computed by averaging widths of eight directions, as is shown in Fig. 4 (right) and the calculation function is shown as follow:

$$swp(p) = \min\{(1+5),(2+6),(3+7),(4+8)\}$$

And stroke width of a candidate character is the average of all white pixels.

3. Single linkage clustering for text candidates

Considering the bottom-up feature of single linkage clustering and inspired by Yin’s work [3] , we obtain text candidates by single linkage clustering. Each candidate letter is treated as a separate class, and the clustering algorithm combines

the two closest classes by calculating the distance between the classes. Fig. 5 shows example results of clustering.

Figure 5. Result of clustering

4. Filtering and Normalizing

Image noise filtering is an effective means to remove image noise without losing image details. As the input of deep learning has to be normalized, there must be some noise after normalizing. So in our experiment, the quality of images is improved by median filtering, which is proved to be effective by PSNR [19], the most widely used evaluation parameter. In consideration of speed and precision, the normalized size is 25 multiplied by 50.

5. Deep belief network for non-text deletion

Deep belief network, composed of several layers of RBM, is of high precision if there are enough training data [20]. The precision of DBN is mostly decided by its structure, such as the number of hidden layers, hidden units and so on. In this paper, we choose the commonly used three layers. When it comes to the number of hidden units, we follow the recipe proposed in [21]. The procedure mainly includes four parts: up-going1, down-going, up-going2 and update. Each layer of RBM network is trained separately and unsupervised to ensure that the feature information is retained as much as possible when the feature vector is mapped to different feature spaces. The last layer of the BP network is set up to receive the output feature vector of the RBM as its input feature vector, and the entity relationship classifier is trained supervised.

Fig.6 shows the training procedure of deep belief network. There are 2000 training images with 1250 pixels each. The number of hidden units in first hidden layer is about equal to 200. As the estimate number is small, the following two layers should use more units. In our experiment, they are 150 and 100. As to the procedure for learning weight and biases, we use the Contrastive Divergence criterion (CD) [20]. Inspired by the recipe for dividing the training set into mini-batches in [21], we set 10 batches for each layer.

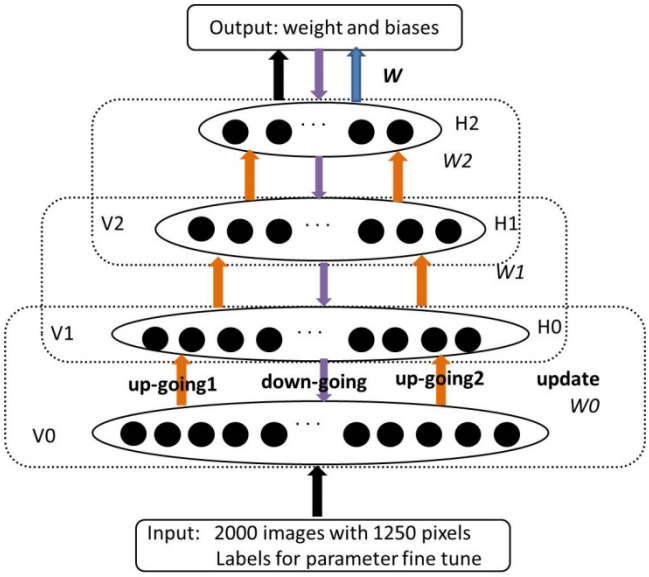

Figure 6. Deep learning procedure

4. EXPERIMENTS

In this section, first we present experiments on main components (character classify and deep belief network) of the proposed method. Then we compare our approach with several state-of-the-art methods on ICDAR 2011 database [22].

4.1 Character Classifying

Adaboost introduced with Hu-moment and that without Hu-moment are presented for the performance of character classifying. Fig.7 show the difference between characters and non-characters in seven dimensions of Hu-moment. The experimental result shows that Hu-moment is appropriate to describe geometrical features of candidate characters.

Figure 7. Candidate characters and non-characters

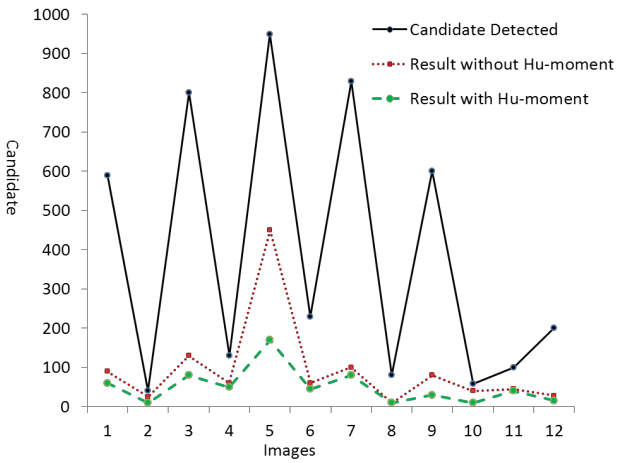

Figure 8. Line of character classifier result

Figure 9. Results without Hu-moment (Colume 1 and 3) and with Hu-moment (Column 2 and 4)

Fig. 8 and Fig. 9 is the comparison result of classifiers with and without Hu-moment, which indicates that our proposed new feature is effective in deleting non-characters. In Fig. 8, the black broken line indicates the number of candidate characters extracted through algorithm of maximum stable extremal region; the red line presents the result of classifier without Hu-moment; and the green line indicates the result of classifier introduced with Hu-moment.

4.2 Text Classifying

In this section, we present the performance comparison of two text classifying algorithms: adaboost trained with geometrical features and the auto learning deep belief network. The geometrical features of adaboost classifier are drawn on [4], with a total number of 15. Table 1 shows the comparison results on text detection precision.

Table 1. PRECISION COMPARISON OF TEXT CLASSIFIERS (DN- Detected number, Pre - Precision)

Img	Candidate	Text	Result of DBN		Result of adaboost	
			DN	Pre	DN	Prec
1	5	4	4	1	4	0.75
2	16	7	8	0.875	10	0.7
3	9	5	4	1	7	0.57
4	5	4	4	1	5	0.8
5	4	4	4	1	4	1
6	6	4	4	0.75	5	0.6
7	9	3	3	1	6	0.5
8	4	4	4	1	4	1
9	5	2	3	0.67	4	0.25
10	13	5	6	0.67	8	0.375

We can see from table III that precision of DBN is far better than that of adaboost. Therefore, DBN as a text classifier is effective.

4.3 Comparisons with Other Algorithms

The training and test dataset of MHDT are from ICDAR 2011 robust reading competition. The whole dataset consists of 485 images containing text of various fonts and colors with different backgrounds. We randomly choose 150 images.

Table 2 compares the performance of MHDT, Boris's method, a very recent MSER-based method by Yin [3] and some of the top scoring methods (Breiman [23], Chen [10], Kim [24]) from ICDAR 2011 Competition. MHDT shows a competitive f measure of 0.77, higher than all other methods.

Table 2. PRECISION COMPARISON OF TEXT CLASSIFIERS (DN- Detected number, Pre - Precision)

Method	Recall rate	Precision	F-measure
MHDT	0.68	0.89	0.77
Yin Xucheng **Error! Reference source not found.**	0.68	0.86	0.76
Epshtein [7]	0.60	0.73	0.66
Breiman [23]	0.62	0.81	0.70
Chen Huizhong[10]	0.60	0.73	0.66
Kim [24]	0.62	0.83	0.71

5. CONCLUSIONS

Text detection in banking scenes, as an important function of image processing in the unstructured platform, provides carrier support for paperless processing, business operation process optimization, and marketing transformation.

This paper proposes an effective algorithm for text detection in natural scene images. The algorithm introduces Hu-moment as training features for character classification after extracting candidate characters with MSER. Then it uses the median filter to improve the quality of training images. After that, a deep belief network is employed to remove non-text candidates. The experimental results demonstrate the excellent performance of the proposed algorithm. Our future work will incorporate the detection for Chinese text and deploy it in more business scenarios.

6. ACKNOWLEDGMENTS

This work is partially funded by the Fujian Fumin Foundation and is supported by the National Natural Science Foundation of China under Grant (No. 61672170 and No. 61871313), the Science and Technology Planning Project of Guangdong Province (No. 2017A050501035), and Science and Technology Program of Guangzhou (No. 201807010058).

7. REFERENCES

[1] Balducci, Bitty, and Detelina Marinova. "Unstructured data in marketing." Journal of the Academy of Marketing Science (2018): 1-34.

[2] Edge, Darren, Jonathan Larson, and Christopher White. "Bringing AI to BI: Enabling Visual Analytics of Unstructured Data in a Modern Business Intelligence Platform." Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, 2018.

[3] Yin, Xu-Cheng, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao. "Robust text detection in natural scene images." IEEE transactions on pattern analysis and machine intelligence 36, 5 (2014): 970-983.

[4] Chen, Xiangrong, and Alan L. Yuille. "Detecting and reading text in natural scenes." Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Vol. 2. IEEE, 2004.

[5] Lee, Jung-Jin, et al. "Adaboost for text detection in natural scene." Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE, 2011.

[6] Yin, Xuwang, et al. "Effective text localization in natural scene images with MSER, geometry-based grouping and AdaBoost." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.

[7] Epshtein, Boris, Eyal Ofek, and Yonatan Wexler. "Detecting text in natural scenes with stroke width transform." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[8] Yi, Chucai, and Yingli Tian. "Localizing text in scene images by boundary clustering, stroke segmentation, and string fragment classification." IEEE Transactions on Image Processing 21, 9 (2012): 4256-4268.

[9] Yi, Chucai, and Yingli Tian. "Text extraction from scene images by character appearance and structure modeling." Computer Vision and Image Understanding 117, 2 (2013): 182-194.

[10] Chen, Huizhong, et al. "Robust text detection in natural images with edge-enhanced maximally stable extremal regions." Image Processing (ICIP), 2011 18th IEEE International Conference on. IEEE, 2011.

[11] Liu, Weibo, et al. "A survey of deep neural network architectures and their applications." Neurocomputing 234 (2017): 11-26.

[12] Kim, Yelin, Honglak Lee, and Emily Mower Provost. "Deep learning for robust feature generation in audiovisual emotion recognition." Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013.

[13] Zhou, Shusen, Qingcai Chen, and Xiaolong Wang. "Discriminative deep belief networks for image classification." Image Processing (ICIP), 2010 17th IEEE International Conference on. IEEE, 2010.

[14] Huang, Chenchen, et al. "A research of speech emotion recognition based on deep belief network and SVM." Mathematical Problems in Engineering 2014 (2014).

[15] Wang, Hai, Yingfeng Cai, and Long Chen. "A vehicle detection algorithm based on deep belief network." The scientific world journal 2014 (2014).

[16] Matas, Jiri, et al. "Robust wide-baseline stereo from maximally stable extremal regions." Image and vision computing 22, 10 (2004): 761-767.

[17] Mikolajczyk, Krystian, et al. "A comparison of affine region detectors." International journal of computer vision 65, 1-2 (2005): 43-72.

[18] Hu, Ming-Kuei. "Visual pattern recognition by moment invariants." IRE transactions on information theory 8, 2 (1962): 179-187.

[19] "Peak Noise to Signal Ratio". [online]. Available: http://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio

[20] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18, 7 (2006): 1527-1554.

[21] Hinton, Geoffrey E. "A practical guide to training restricted Boltzmann machines." Neural networks: Tricks of the trade. Springer, Berlin, Heidelberg, 2012. 599-619.

[22] Shahab, Asif, Faisal Shafait, and Andreas Dengel. "ICDAR 2011 robust reading competition challenge 2: Reading text in scene images." Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE, 2011.

[23] Breiman, Leo. Classification and regression trees. Routledge, 2017.

[24] Koo, Hyung Il, and Duck Hoon Kim. "Scene text detection via connected component clustering and nontext filtering." IEEE transactions on image processing 22, 6 (2013): 2296-2305.