CHINESE IMAGE TEXT RECOGNITION ON GRAYSCALE PIXELS

Jinfeng Bai, Zhineng Chen, Bailan Feng and Bo Xu

Interactive Digital Media Technology Research Center,
Institute of Automation Chinese Academy of Sciences, Beijing 100190, China
{jinfeng.bai, zhineng.chen, bailan.feng, xubo}@ia.ac.cn

ABSTRACT

This paper presents a novel scheme for Chinese text recognition in images and videos. It's different from traditional paradigms that binarize text images, fed the binarized text to an OCR engine and get the recognized results. The proposed scheme, named grayscale based Chinese Image Text Recognition (gCITR), implements the recognition directly on grayscale pixels via the following steps: image text over-segmentation, building recognition graph, Chinese character recognition and beam search determination. The advantages of gCITR lie in: (1) it does not heavily rely on the performance of binarization, which is not robust in practical and thus severely affects the performance of OCR, (2) grayscale image retains more information of the text thus facilitates the recognition. Experimental results on text from 13 TV news videos demonstrate the effectiveness of the proposed gCITR, from which significant performance gains are observed.

***Index Terms*—** Chinese Text Recognition, Image Text, Video Text, grayscale recognition

1. INTRODUCTION

In order to quickly locate interested content in massive amounts of multimedia data, much research has been done in content based image research [1] and video story summarization [2, 3]. Textual information in image, especially the superimposed text, offers more reliable clues for it relates high level semantic summarization. While commercial optical character recognition (OCR) systems have already reached high performances for document images, the recognition of text in images or videos (e.g. superimposed text) is still a challenging problem, especially for Chinese text recognition. This is attributed to, on the one hand, the stroke number of a Chinese character with same-sized block space varies dramatically; on the other hand, a much larger category set is involved, and therefore more confusion between similar characters needs to manage.

Since the poor performance of commercial OCR is caused by complex background and low resolution when applied to image text recognition, various kinds of sophisticated binarization algorithms such as [4, 5, 6] have been developed to extract text from complex background. However, as an unsupervised process, binarization is a very challenging problem when image text suffers from several degradations such as uneven lighting, complex background, blur and so on. Furthermore, remarkable within-class variances, i.e., the same Chinese character with diverse fonts, different sizes, as well as subtle among-class differences between some category pairs, are also big and intractable problems. Therefore, instead of striving to explore a high performance binarization algorithm, we handle this challenging problem in character recognition step, i.e. recognizing the image text directly on grayscale pixels, for it's a supervised process and more capable of dealing with the problem.

Another challenge in Chinese image text recognition comes from segmentation (i.e. splitting the images of text line into pieces corresponding to single character), for a considerable part of Chinese characters inhere the structure of separation from the left and right sides. Furthermore, touching characters caused by binarization algorithm of poor performance would also bring about more improperly extracted chars that are usually impossible to recognize correctly with currently used methods. In order to settle these problems, a technique of over-segmentation is introduced, which is widely used in handwriting Chinese character string recognition [7, 8] and English text recognition [9]. It tries to find all the potential splitting points in a graphical representation of the text line and then attempt to eliminate the improper ones.

This paper proposes an over-segmentation based on a vertical histogram of the resulting image of binarization for generating potential splitting points. Based on this, a graph is built on account of geometry information of splitting points, and character recognition model is applied to every cropped image defined by two connected vertexes. Then, from start vertex step by step all the way to end vertex, a dynamical process of building and pruning the lattice is performed. At each step, every recognition result is estimated by a 3-gram language model [10] according to different histories. Along with this, a beam search by maximizing the objective function defined on the linear combination of recognition score and language s-core is performed to find the best path. The characters corresponding to arches in the best path will be the recognition results. This will not only eliminate the improper splitting points but also make use of N-best recognition results. Therefore, our method is particularly robust to complex background, low resolution or video coding artifacts. Experimental results on text from 13 TV news videos demonstrate the effectiveness of the proposed gCITR. Significant performance gains are observed compared to ABBYY FineReader [11], which is used as the baseline system at 11th [12, 13], 12th [14] the Robust Reading Competition respectively.

The rest of the paper is organized as follows. We discuss related work in Section 2. In Section 3, the proposed recognition scheme is explained in detail. Section 4 shows experimental configurations and results. The work is finally concluded in Section 5.

2. RELATED WORK

For image text recognition, the two editions of the Robust Reading Competition at 11th [12, 13], 12th [14] International conference on Document Analysis and Recognition respectively, show active in-

The work was supported by National Nature Science Foundation of China(grant No.61202326)

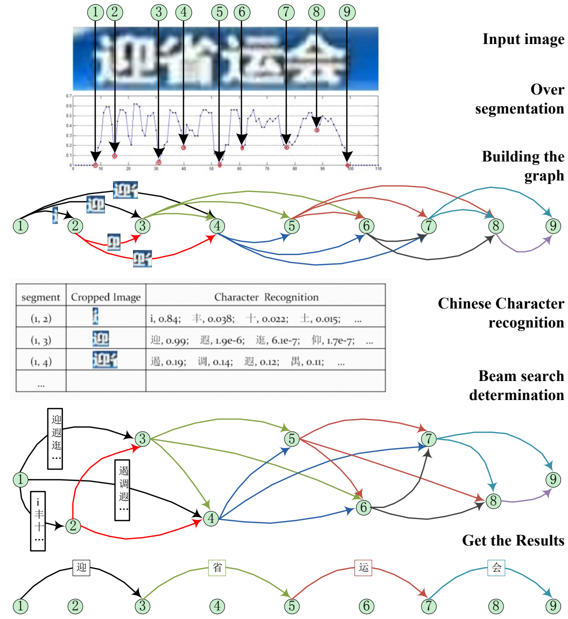
978-1-4799-2893-4/14/$31.00 ©2014 IEEE

terest in this research area. Variety kinds of algorithms have been proposed to settle the problem. Roughly, they can be classified into two categories: methods based on binarization and methods based on holistic approach. The methods (like Mishra [4], Boris [6], etc.) based on binarization dedicate to developing a high performance algorithm to binarize text image, fed the resulting image to commercial OCR engine and get the results. The methods (like Zohra [9], Yokobayashi [15], etc) based on holistic approach use an over-segmentation and beam search architecture to get recognition results, since touching characters in English words are quite common. The latter methods have shown promising performance in image text recognition in recent years. Yokobayashi et al [15] proposed two systems for character recognition in natural scene images which modified the recognition step based on an improved version of the global affine transformation (GAT) correlation technique for grayscale character images. Zohra et al [9] proposed an image Text Recognition Graph depicting the architecture, using a convolutional neural network for over-segmentation and character recognition.

In Chinese image text recognition, only the former methods [16, 17] have been reported so far as we know. An important reason is that one may not be able to find so much labeled data to train a character recognition model since Chinese character has so large categories. Take offline Chinese handwriting character recognition in 12th [18] International conference on Document Analysis and Recognition for example, the total categories and samples are 3755 and about 1.1 million respectively. Therefore, methods which could overcome this challenge would be a preferable choice.

3. CHINESE IMAGE TEXT RECOGNITION

The proposed recognition scheme for Chinese image text includes four main steps: image text over-segmentation, building recognition graph, Chinese character recognition and beam search determination. It's worth to mention that beam search determination selects the optimal recognized path based on outputs from both the recognition model and language model. Along with this, a dynamical process of building and pruning the lattice is performed and the result is obtained by maximizing objective function defined on lattice. The flowchart is shown in Fig. 1.



segment	Cropped Image	Character Recognition
(1, 2)		i, 0.84; 丰, 0.038; 十, 0.022; 土, 0.015; ...
(1, 3)		迎, 0.99; 遐, 1.9e-6; 逛, 6.1e-7; 仰, 1.7e-7; ...
(1, 4)		遏, 0.19; 调, 0.14; 遐, 0.12; 禺, 0.11; ...
...		

Fig. 1. The Flowchart of proposed recognition scheme

3.1. Image Text Over-segmentation

In order to do over-segmentation, a roughly extracted binary character image is needed. Therefore, a double-edge [19, 20] feature map is extracted from text image, which portrays the distinguishing trait of the characters. Paper [20] develops a fast and unified method of searching double-edge features occurring with a certain distance.

Let arbitrary pixel belonging to text image be represented by p. In one direction across p, the intensity of p being estimated to belong to a double-edge with thickness W can be calculated as follow:

$$DE^d(x) = f(x) - \min_{i\in[1,W-1]} \{\max(f(x-i), f(x+W-i))\} \tag{1}$$

Therefore, the intensity of p being estimated to belong to a double-edge with thickness W in an image can be approximated as follow:

$$DE(p) = \max\left\{0, \max_{d\in[0,3]} \{DE^d(x)\}\right\} \tag{2}$$

Here, $d = 0, 1, 2, 3$ refer to four directions $\{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$, which approximate most possible edge directions around a pixel. After this process, characters, which have trait of double-edge, stand out and backgrounds are suppressed. A global threshold T obtained by Otsu [21] could separate stroke pixels from background. In experiment, the maximum thickness W is set 10 pixels.

Character segmentation is one of the key steps in character string recognition, for improperly extracted characters not only result in wrong recognition itself but also may influence the global best path searching. In order to solve this problem, over-segmentation technique is introduced to try to find all potential splitting points of the text line and then attempt to eliminate the improper ones. Since Chinese character usually has approximate square structure and touching characters mostly caused by poor performance of binarization algorithm, methods based on vertical histogram (i.e. a vertical projection) [22] of the resulting image of binarization step are very suitable considering both efficiency and performance. Vertical histogram is a function counts the number of white pixels in each column and allows for easy detection of empty columns and simple touching characters. Since splitting points mostly occur on the local minima, classical segmentation algorithm simply chooses all local minima below the specified threshold as splitting points. This may result in under or over segmentation when applied to different datasets. On account of this, an algorithm 1 is developed based on vertical histogram to do over-segmentation, in which W and H denote the width and height of the image text line respectively.

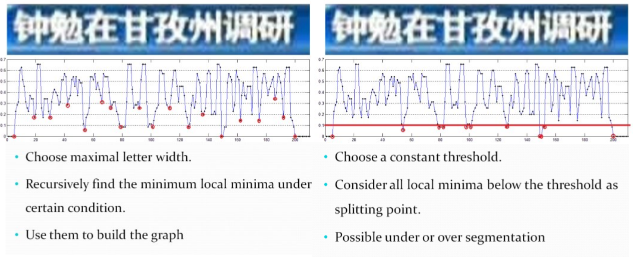
Fig. 2 shows the differences between classical segmentation [9] and over-segmentation intuitively. Obviously, there are much more splitting points found in over-segmentation in order to ensure all true splitting points included. And we can also see that in classical segmentation, under or over segmentation is inevitable. The true splitting positions will be determined by the overall process as detailed in the following paragraphs.

3.2. Graph Built On Geometry Information

This step takes the over-segmentation results as input and a graph is built by considering every splitting point as a vertex. To build the connections between vertices, we choose to apply some constraints

Algorithm 1 Over segmentation based on vertical histogram.

```
Main:
    SplitArray = []
    minStep = $max(4, H/10)$
    maxWdth = $0.7 \times H$
    oversegmentation (0, W)
    sort(SplitArray)
Function: oversegmentation (a, b)
    Return if b-a $\leq$ maxWdth - 2$\times$minStep
    Find smallest local minima $t$ falling in $[a, b]$
    If not found, $t = min(a, b)$
    Push SplitArray, t
    oversegmentation(a+minStep, t-minStep)
    oversegmentation(t+minStep, b-minStep)
```



(a) Over segmentation

(b) Classical segmentation

Fig. 2. Over-segmentation versus classical segmentation

in order to optimize the computation without losing efficiency, since the length-width ratio of a character varies in a certain range. These constraints are:

- A vertex i is connected to a vertex j $(j > i)$ if the length-width ratio of the cropped image between them is within a chosen interval. This means that we have to evaluate the minimum and the maximum width of character given the text line height.
- A vertex i cannot be connected to a vertex j $(j > i)$ if there is a former vertex k $(k \geq i)$ of j and values of vertical histogram in the interval between them are all zero. In other words, spaces between the characters will not be considered.

3.3. Chinese Character Recognition

Different from English image character recognition task, which already has many public labeled dataset such as [12, 13, 14], there is not any available labeled datasets in Chinese image character recognition. An important reason is that one may not be able to find so much data to train a character recognition model since Chinese character has so large categories. On account of this, we automatically generate the training datasets and use them to train a recognition model. Like in the paper [23] a scheme is developed to generate the noise labeled data using existing dataset to test model capacity, we use standard machine-born Chinese characters as the label dataset and noise them with proposed scheme. The detail is as follow.

Firstly, 4120 character categories are selected as recognition vocabulary, which include 3755 Chinese characters of GB2312-1, 245 extensional high frequency Chinese characters, 94 printable Ascii symbols and 26 punctuation symbols according to frequency. Secondly, plenty of patches are randomly cropped from a set of grayscale images to be used as the background, which are randomly extracted from news video and downloaded from the internet. Patches which have low pixel variance (i.e. contained little texture) were ignored. Lastly, we choose combination of 11 fonts, 5 different sizes and 3 degrees of boldness to generate the standard data and each of them is noised using 3 different variance backgrounds. In total, 495 (case number of each category) × 4120 (number of all categories) samples are generated.

After generating training data, a multilayer perception model is trained on the feature of 512d of 8 directions gradient histogram. The architecture of recognizer consists of a 512d input layer, two 2048d hidden sigmoid layers and a 4120d softmax activation output layer. Initial weights are drawn from a uniform random distribution in the range [-0.05; 0.05]. Then every cropped image corresponding to each arch in the graph built before is processed by the recognizer system and N-best results associated with their probabilities are reserved. The results of this step are given as inputs to the graph weights calculator.

3.4. Beam search determination

After processed by recognition step, each arch in the graph represents N-$best$ results corresponding to cropped image; therefore the arch is replaced by N arches having same start and end vertex but different characters associated with their probabilities. Then, from start vertex 0 step by step all the way to end vertex 9 (take the Fig. 1 for example), a dynamical process of building and pruning the lattice is performed. Along with this, a beam search is performed to find the best path. In each step from vertex i to $i + 1$, every arch that arrived to vertex $i + 1$ is processed by a character based 3-gram language model. According to the different histories (only two former character considered in our case) of the character corresponding to arch, the arch is again replaced by many arches representing each unique different one. In each unique arch, a recognition probability $Rprob$ and a 3-gram language probability $Lscore$ are recorded. The total score of the arch is calculated as follow:

$$edgescore_{i,j} = \alpha \log(Rprob) + (1 - \alpha) \log(Lprob) \quad (3)$$

Here, α is the proportion of language score and recognition score and equals 0.9 in our experiment. Since 3-gram language model is applied to every different character arrived to vertex $i+1$, quite a lot of arches are added to the vertex $i+1$. Therefore, a pruning strategy is adopted to improve the efficiency and keep recognition accuracy meanwhile. Only the top M (M is set to 50 in our case) arches that arrived vertex $i + 1$ are reserved and others are all pruned. The total score from start vertex to vertex $i + 1$ via k-th path is defined as follow:

$$totalscore_k^{i+1} = \left(\sum_k edgescore\right) / edge(k) \quad (4)$$

where $\sum_k edgescore$ represents the sum of all $edgescore$ in k-th path and $edge(k)$ represents the total arch number in the k-th path. When process arrives the end vertex, the maximum $totalscore$ path will be the best path and the characters corresponding to arches in the best path will be the recognition results. Fig.1 illustrates the whole recognition scheme with an example.

4. EXPERIMENTS AND RESULTS ANALYSIS

To test the performance of the proposed scheme, we have run a series of experiments on a dataset containing various image text, which are collected from 13 TV news videos respectively, denoting each

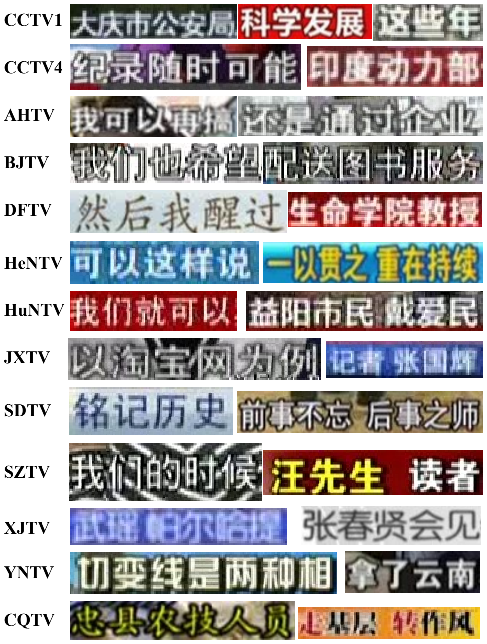


Fig. 3. Examples from different TV news

set with its original TV program name. Each of them is about 30-minute long and they are completely different from each other. Since the same text in video will last for a certain period of time, only one of them is cropped manually. And every line of the text block is fully annotated and saved separately for only the problem of image Chinese character string recognition is focused on. These text line images are of different sizes (17×85, 30×378, 49×452), different colors, different fonts, and different character pitches and suffered from different kinds of distortion. These include complex background, low resolution, similar foreground background color, uneven light and so on. Fig. 3 shows some typical type samples from each dataset intuitively. Table 1 shows recognition accuracy of each dataset and other detail parameters associated with it such as the total of text line and the total of character. There are 1131 text lines and 12700 characters in the dataset respectively.

For comparison, we test all the datasets with ABBYY FineReader 10.0 [11]. As a commercial OCR, ABBYY can recognize not only document images but also text in general images; therefore it's been used as the baseline system of the two editions of the Robust Reading Competition at 11th [12, 13], 12th [14] International conference on Document Analysis and Recognition respectively. The task of robust reading competition is locating and recognizing the English word in images. In our case, Chinese character string recognition problem is concerned and ABBYY is also suitable to be used as the baseline system since simplified Chinese recognition is supported. Recognition accuracy of both systems is showed in Fig.4. It can be found that when image text lines have clean background and high resolution, both systems could achieve a high performance, such as in DFTV dataset. But when applied to complex background and low resolution image text, our system is more robust than ABBYY. The average recognition accuracy of both system is 85.44% and 53.59% respectively. There is an enhancement of 59.4% over all the whole dataset. Since our scheme directly recognizes image in grayscale space and only generated data is used to train the recognizer, it is less dependent on dataset and more transferable over different kinds of image text. Hence, it is very robust to the complex background image text.

Table 1. Configuration of datasets

TV News	Total of Line	Total of Character	String accuracy
CCTV1	97	1026	91.91%
CCTV4	83	817	93.39%
AHTV	84	788	86.80%
BJTV	137	1359	68.43%
DFTV	204	2009	93.53%
HeNTV	126	1172	95.73%
NuNTV	111	1181	90.09%
JXTV	41	369	64.77%
SDTV	93	1022	86.69%
SZTV	44	403	76.18%
XJTV	100	904	73.86%
YNTV	59	557	84.02%
CQTV	132	1093	82.25%

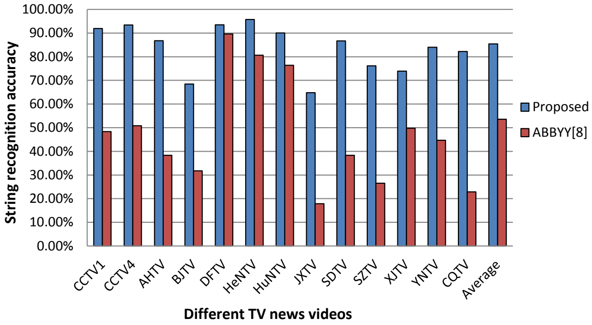


Fig. 4. Comparison of proposed algorithm and ABBYY [11]

5. CONCLUSION

In this paper, we propose a new scheme for Chinese image text recognition, directly recognizing text on grayscale image instead of binary counterpart. Therefore, it does not heavily rely on the performance of binarization thus facilitates the recognition. In order to solve the problem of touching characters and separation structure character, an over-segmentation technique is introduced to try to find all potential splitting points. And then, a dynamical process of building and pruning the lattice will not only eliminate the improper splitting points but also make use of N-best recognition results. Experimental results on text from 13 TV news videos demonstrate the effectiveness of the proposed gCITR. In future work, we intend to explore sequence featrue extraction methods by discriminating sequences as a whole [24] to improve the performance of recognition process.

6. REFERENCES

[1] Bailan Feng, Juan Cao, Xiuguo Bao, Lei Bao, Yongdong Zhang, Shouxun Lin, and Xiaochun Yun, “Graph-based multi-space semantic correlation propagation for video retrieval,” *The Visual Computer*, vol. 27, no. 1, pp. 21–34, 2011.

[2] Bailan Feng, Peng Ding, Jiansong Chen, Jinfeng Bai, Su Xu, and Bo Xu, “Multi-modal information fusion for news story segmentation in broadcast video,” in *Acoustics, Speech and*

Signal Processing (ICASSP), 2012 IEEE International Conference on, March 2012, pp. 1417–1420.

[3] Bailan Feng, Zhineng Chen, Rong Zheng, and Bo Xu, "Style learning based story boundary detection for chinese broadcast news videos," in *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, New York, NY, USA, 2013, ICIMCS '13, pp. 42–45, ACM.

[4] A. Mishra, K. Alahari, and C.V. Jawahar, "An mrf model for binarization of natural scene text," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, sept. 2011, pp. 11 –16.

[5] Jinfeng Bai, Bailan Feng, and Bo Xu, "Binarization of natural scene text based on l1-norm pca," in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*, 2013, pp. 1–4.

[6] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2963–2970.

[7] Qiu-Feng Wang, Fei Yin, and Cheng-Lin Liu, "Handwritten chinese text recognition by integrating multiple contexts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 8, pp. 1469–1481, 2012.

[8] Liang Xu, Fei Yin, Qiu-Feng Wang, and Cheng-Lin Liu, "An over-segmentation method for single-touching chinese handwriting with learning-based filtering," *International Journal on Document Analysis and Recognition (IJDAR)*, pp. 1–14, 2013.

[9] Zohra Saidane, Christophe Garcia, and Jean Luc Dugelay, "The image text recognition graph (itrg)," in *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, Piscataway, NJ, USA, 2009, ICME'09, pp. 266–269, IEEE Press.

[10] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn, "Scalable modified Kneser-Ney language model estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 2013.

[11] ABBYY Finereader 9.0., "http://www.abbyy.com," .

[12] D. Karatzas, S.R. Mestre, J. Mas, F. Nourbakhsh, and P.P. Roy, "Icdar 2011 robust reading competition - challenge 1: Reading text in born-digital images (web and email)," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, 2011, pp. 1485–1490.

[13] A. Shahab, F. Shafait, and A. Dengel, "Icdar 2011 robust reading competition challenge 2: Reading text in scene images," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, 2011, pp. 1491–1496.

[14] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere de las Heras, "Icdar 2013 robust reading competition," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, 2013, pp. 1484–1493.

[15] Minoru Yokobayashi and T. Wakahara, "Segmentation and recognition of characters in scene images using selective binarization in color space and gat correlation," in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, 2005, pp. 167–171 Vol. 1.

[16] M.R. Lyu, Jiqiang Song, and Min Cai, "A comprehensive method for multilingual video text detection, localization, and extraction," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 2, pp. 243 – 255, feb. 2005.

[17] Tsung-Han Tsai, Yung-Chien Chen, and Chih-Lun Fang, "2dvte: A two-directional videotext extractor for rapid and elaborate design," *Pattern Recognition*, vol. 42, no. 7, pp. 1496 – 1510, 2009.

[18] Fei Yin, Qiu-Feng Wang, Xu-Yao Zhang, and Cheng-Lin Liu, "Icdar 2013 chinese handwriting recognition competition," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, 2013, pp. 1464–1470.

[19] Salim Djeziri, Fathallah Nouboud, and Rjean Plamondon, "Extraction of signatures from check background based on a filiformity criterion," *IEEE Transactions on Image Processing*, vol. 7, pp. 1425–1438, 1998.

[20] Xiangyun Ye, M. Cheriet, and C.Y. Suen, "Stroke-model-based character extraction from gray-level document images," *Image Processing, IEEE Transactions on*, vol. 10, no. 8, pp. 1152 – 1161, aug 2001.

[21] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[22] Magdalena Brodowska., "Oversegmentation methods for character segmentation in off-line cursive handwritten word recognition c an overview," *Schedae informaticae*, vol. 20, no. 285-296, pp. 23–27, 2011.

[23] Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra, and Yoshua Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proceedings of the 24th international conference on Machine learning*, New York, NY, USA, 2007, ICML '07, pp. 473–480, ACM.

[24] Bing Su and Xiaoqing Ding, "Linear sequence discriminant analysis: a model-based dimensionality reduction method for vector sequences," in *Proc. IEEE Int'l Conf. Computer Vision*. IEEE, 2013, pp. 889–896.