A method of synthesizing handwritten Chinese images for data augmentation

Xi Shen
École Nationale des Ponts et Chaussées
Paris, France
shenxiluc@gmail.com

Ronaldo Messina
A2iA SAS
Paris, France
Email: rm@a2ia.com

***Abstract*—The performance of printed document recognition has been significantly improved by generating synthetic images to augment the training data, particularly by providing more variability in the linguistic contents. Handwriting recognition benefits less from this data augmentation and the only variability that is usually added is via artificially generated combinations of skew, slant and noise. Generating handwritten text is complex due to variations in form, scale and spatial placement of the characters, and can be further complicated by the cursive aspects of the script.**

We propose a novel strategy, in the particular case of Chinese characters, to generate synthetic lines of text, given samples of the isolated characters. The well-known CASIA database is used to train MDLSTM-RNN models and also in the creation of synthetic line images. On an independent set of document images, a model trained only on synthetic images achieved a small relative reduction of 4.4% in the character error rate with respect to a baseline model trained exclusively on real images, while training on a combination of real and synthetic images resulted in a appreciable reduction of 10.4%.

***Keywords*-Handwritten Chinese recognition; Synthetic image; Data augmentation;**

I. Introduction

It has been a tendency to use larger and larger training datasets for neural networks. More data improves the estimation of model parameters, but in most cases, collecting or manually creating data is costly and time consuming, especially for document recognition where the annotation of the images is complex.

Certain techniques of image degradation have been proposed to augment the amount of training data, and experimental results prove that these techniques provide a good model of the degradation observed in machine-printed documents in the course of printing, photocopying, FAXing, and scanning. Baird [1], [2] introduces some techniques for degradation, such as random scaling factors, resolution, skew, etc. Kanungo [3] proposes to model the perturbation in the optical scanning and digitization process, as a result, global (perspective and non-linear illumination) and local effects (speckle, jitter, etc.) are generated. Loce [4] models the perturbation due to mechanical disturbances in high-end photocopiers. Machine-printed documents can be created in a large volume with a linguistic content and the above degradation techniques, and the accuracy of the printed document recognition benefits from those synthetic images.

However, there are few techniques concerning the generation of handwritten documents due to variations in form, scale and spatial placement of the characters, specially the cursiveness of the text lines. Graves [5] applied Recurrent Neural Networks (RNNs) to generate realistic-looking[1] sequences by a combined neural network architecture that uses predictions of the $x; y$ coordinates and the end-of-stroke markers one point at time (learned from on-line data) with a sequence of the characters to synthesize.

The Deep Recurrent Attentive Writer (DRAW) [6] composed of two RNNs : one is in charge of compressing real images, the other manages to reconstitute images after receiving codes, demonstrates its ability to improve the state-of-the-art result by generating highly realistic images from MNIST [7]. Besides, other generative models such as DBM [8], DBN [9], NADE [10], EoNADE [10], [11], DLGM [12], [13], DARN [14], prove a good performance to varying degrees.

Unlike the above models based on Neural Network, this work introduces a simple method to generate synthetic lines of text in Chinese by using the images of isolated characters in the CASIA database [15], [16] (on-line and off-line data). We train Multi-Dimensional Long Short Term Memory – MDLSTM-RNN – networks to assess the recognition performance on real images of text.

The article is organized as follows. First, we briefly describe the well-known CASIA database and explain the process to generate synthetic images. We then present our experimental results. Finally, we exploit some possible directions to generalize our method.

II. Generation processes with CASIA database

A. Database

The CASIA database [15], [16] is an on-line and off-line Chinese handwriting database. It contains samples of isolated characters (DB1.0 – 1.2) and handwritten lines (DB2.0 – 2.2) which were produced by 1,020 writers using an Anoto pen on paper. More specifically, each writer of DB1.0-1.1 (720 writers in total) put on paper about 4 000 most frequent characters (including the Latin alphabet, punctuation, and

[1] As judged by a native Chinese.

some symbols), while the DB1.2 set contains about 3 000 less common characters.

The DB2.0 – 2.2 comprise 4 433 images. Each image is a full page of handwritten text with the ground truth values and the bounding box of the characters. For the following experiments, we just use the off-line set and randomly separate the page images into two sets: 3 840 images (corresponding to 33 334 lines) as train set and the remaining 593 images (5 727 lines) as validation set to estimate parameters of the MDLSTM-RNNs. All images were binarized with a randomly selected algorithm from either Otsu [17], Wolf-Jolion [18], or Niblack [19].

There are two test sets. The first was collected in-house and comprises 105 actual images of documents, binarized in the same fashion as the training data. The recognition of this database can be considered very difficult due to various types of documents (forms, bills, meeting notes), complicated layout and rapid writing style. An example of test image is shown in Figure 1a, and the other test set (see Figure 1b) is the one from the ICDAR 2013 Chinese Handwriting Recognition Competition (Task 4). All tests are on located lines to avoid the influence of line detection on the results.

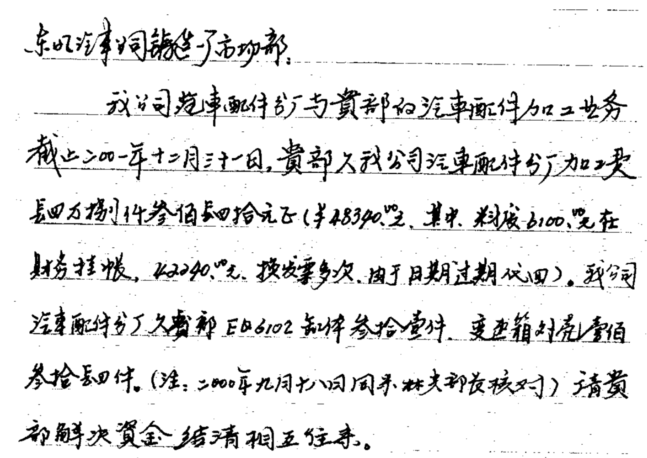
B. *Optical modeling*

A MDLSTM-RNN with the architecture shown in Figure 2 is the optical model in the experiments, following the work in [20]. This model can recognize full lines of characters without explicit segmentation into characters. Actually, an existing model is adapted on each generated training data, using RMSProp [21] and mini-batches of 16 samples. The Connectionist Temporal Classification (CTC) [22] objective function is used; it uses a "non-character" label called *blank* that is used to align the sequences of outputs of the network and the target characters which can have different lengths. Training stops when the character error rate (CER) on the validation data does not decrease within 10 epochs and the model yielding the lowest CER is retained.

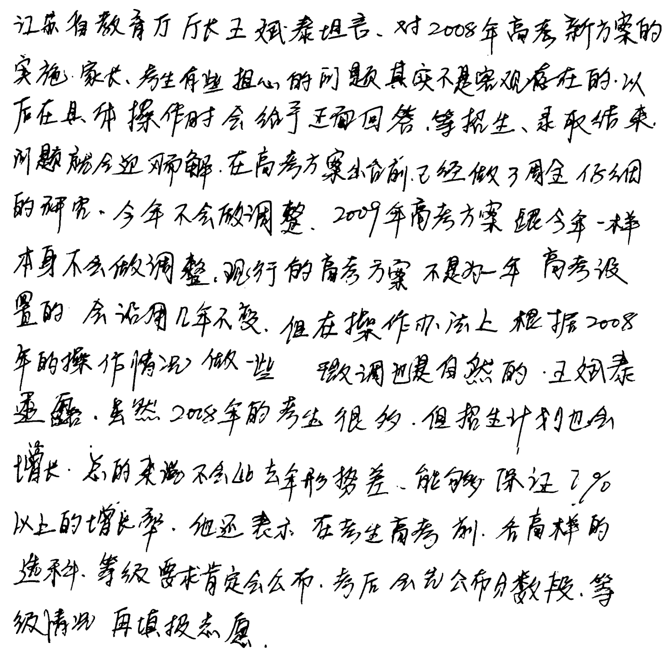
III. Generation strategies

We profit from the segmentation at character level of the CASIA database to evaluate many different strategies to generate synthetic lines of text from the isolated character snippets from the DB1.0 – DB1.2 subsets. The strategies range from simply putting the character snippets one after another to more complex processing where the individual character coordinates in annotated lines are used to create more realistic-looking images of text lines.

The same text in the lines of the DB2.0 – DB2.2 subsets is used to create new images. After determining the best strategy, we plan to synthesize images with arbitrary text. The synthetic images have no background, which is the same as in the original CASIA data; rendering a realistic background is not in the scope of this work, but it is an interesting subject to provide data that is more similar to actual scanned images of documents. In the following sections we described the methods we evaluated.



(a) A sample of actual images



(b) A sample of ICDAR competition

Figure 1: Sample of test sets

1) LineSeparate – LS: The first strategy involves making each line separately. More precisely, for each line we randomly choose a writer from the isolated characters database. To place the characters in the line image, we suppose that the median line is horizontal with random small vertical displacements for each character and the spacing between them is also randomly chosen. In the case of missing characters (certain writers have an incomplete set), other writers will be chosen to complete the line without any normalization. We choose to do this in order to keep using

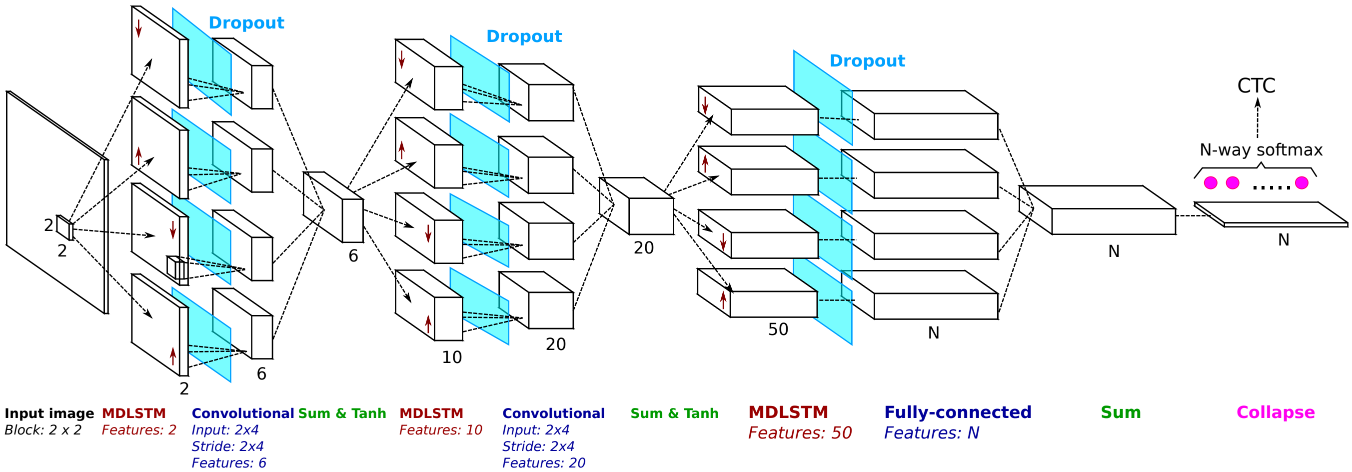


Figure 2: Architecture of the MDLSTM-RNN used in our experiments, N corresponds to the number of predicted labels, 2 667 in our experiments.

the same quantity of training samples, even if the result may look unnatural. However, the number of samples with missing characters is quite reduced (1 324) compared to the whole database (33 334), so we decided to neglect the effects of mixing data for two writers and just used the character snippets as they are. Samples of this strategy are shown in Figure 3.

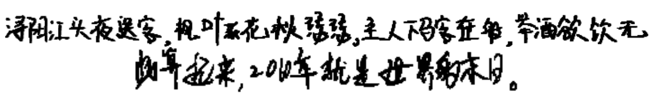
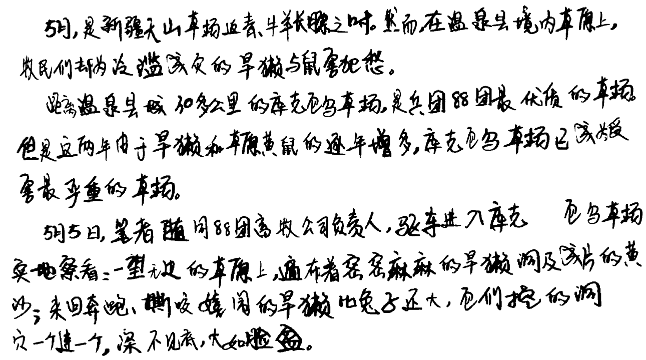


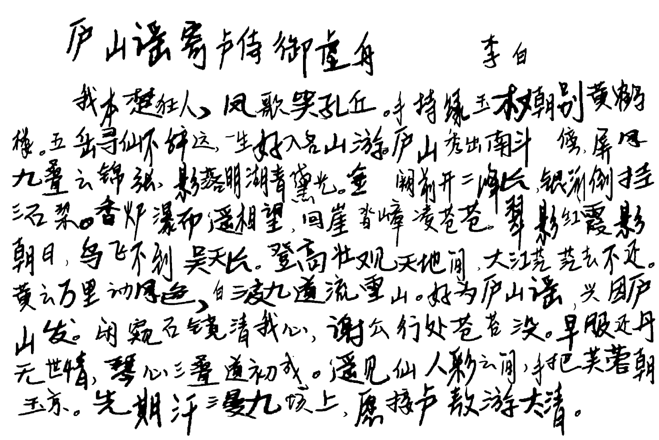
Figure 3: Sample of LS

2) CompletePage – CP: Striving to create more realistic images, we start to use the segmentation information provided in DB2.0 – 2.2. Instead of generating one line, this strategy makes a full document each time. For each image in DB2.0 – 2.2, a random writer from the DB1.x is chosen. The character size and position are exactly the same as the original image to keep the placement of the characters and their spacing realistic. In the case of missing characters, we will crop the characters from the original images. This strategy generates more credible samples (see Figure 4a) than the previous one, as well as some samples that are not realistic (see Figure 4b).

3) CompletePageWriter – CPW: We realized that the bad samples of CompletePage are probably caused by the incorrect adaptation of characters to the size in the source image. This can be distinguished into two cases: in the first the characters are excessively deformed (see Figure 5a), the scale factor of width and height have considerable differences between the source image and the character snippet; the second case is that the scale factor is significantly large, so that we have a visual variation of "boldness", illustrated in Figure 5b. This "boldness" effect can be statistically



(a) Realistic Sample



(b) Unrealistic Sample

Figure 4: Samples of CP

quantified by a parameter named **s**troke. Ryu [23] and Tseng [24] proposed methods to evaluate this parameter.

Aiming at reducing the effect of the first case, we analyze the global ratio between height and width for the character

(a) Blue characters present excessive deformation.

(b) Red characters have a large scale factor.

Figure 5: Examples of incorrect adaptation of character size.

snippets of each writer, as shown in Figure 6. This ratio is calculated for all characters in the source image, then the distances between the ratios in the image and all writers are stored. From the writers having small distance, we randomly choose one to synthesize the corresponding image.

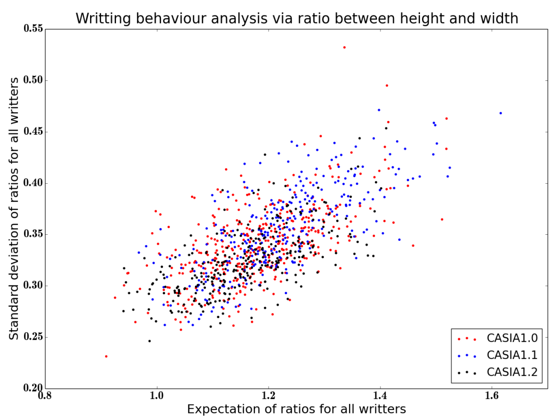


Figure 6: Analysis of writers: each point represents a writer, and the X-axis and the Y-axis are related to the mean and the standard deviation values of ratios, respectively

To enable have more writing style, the number of images written by a single person are limited to 100.

4) CompletePageWriterNoise – CPWN: We introduce certain types of noise to images created by CPW, including horizontal or vertical lines (simulating line crops in tables, as in Figure 7), noisy crop, resize and JPEG noise. The noisy crop reduces the sensibility to the positions of bounding boxes and the resize concerns lines with different sizes of characters. About 20% of the images are noisy.

IV. Experimental results

A baseline model is trained on the binarized CASIA images. To make the comparison on the in-house test set fair, we created a baseline that was artificially noised, denoted “BaselineNoise”. We also created a “Mix” model that was trained on an equal amount of real and synthetic images from the CPWN strategy (50%-50%, keeping the same number of lines as in the other experiments).

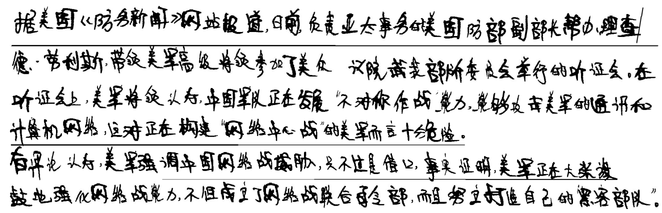


Figure 7: A sample of image using *C*PWN .

Experimental results on the in-house test set are shown in Table I. Even if this data cannot be made publicly available, we believe it is interesting to show the performance on document data, which is noisier and in some aspects, harder to recognize than those in the CASIA database.

Table I: Test result on document images.

Model/Strategy	%CER
Baseline	47.34
BaselineNoise	44.59
LS	69.28
CP	60.09
CPW	56.35
CPWN	45.30
Mix	42.95

The results with the simpler strategies LS and CP are much worse than the baseline, as it could be expected. Visually, the synthesized images are quite different from real images, and we could not hope the network to generalize properly.

Without the noise, the CPW model shows a relative degradation of 20% respect to the baseline model, but it shows an improvement comparing to a random choice of writer. Adding noise ameliorates so that the model CPWN obtains a relative reduction of 4.4% compared to the baseline. And the model trained on “Mix” data, which introduces more variabilities shows a relative 10% reduction to the baseline and presents slightly better results than the baseline trained on noisy data.

On the publicly data from the ICDAR 2013 competition, we obtained the results presented in Table II. The results are presented for both gray-scale and binary images; we performed those tests to assess the influence of the type of image to the results.

Table II: Test result on the ICDAR 2013 set.

Model/Strategy	%CER	
	Gray-scale	Binary
Baseline	23.21	22.73
BaselineNoise	24.58	24.05
CPW	27.11	27.74
CPWN	29.12	28.38
Mix	24.04	23.79

Adding noise helped improve the results on the in-house data, while degraded on the ICDAR 2013 test set (which was collected in the same setup as the CASIA database). The most parts of the binary test results are slightly better than the gray-scale results simply because all the models are trained on binary images. In term of the binary test, the CPW deteriorates relatively 22% in comparison with the Baseline due to the unnatural samples, after all, 27.74% of CER still proves that CPW generates a great number of good samples. And we are not surprised to obtain 23.79% for the Mix which has an improvement compared to CPW and still above the result of Baseline. Since there are still quite great differences between the LS, CP and CPW on the first test, we did not evaluate them on ICDAR 2013 set.

V. Analysis

The strategy of generating full pages instead of single line created more realistic images by keeping the placement of the characters in the lines of text. The relative placement of the characters in neighbor lines is also maintained. We believe those are important characteristics of real images that are not easily replicated artificially.

The improvement of the CPW strategy compared to CP shows that the scale factors of width and height between the characters should not be very different. The fact of enforcing a single writer to synthesize a page helps to keep the same writing style of documents, which is more natural, but we might propose a strategy to choose one writer for each character, which might be more precise to reduce the deformation caused by varying height/width ratio.

Finally, the modeling of noise can bring a remarkable amelioration for recognizing actual images when the initial training data was collected in a laboratory-controlled environment.

VI. Conclusion

This article proposes different strategies to obtain synthetic handwritten Chinese documents by using an existing segmented database at character level, in this case we used the CASIA database. The strategy we called CPWN gave the best results on an in-house database comprising images from documents.

The model trained only on synthetic images achieved lower error compared to the one trained exclusively on real images, and a model trained on a combination of 50% real and 50% synthetic images brought a relative improvement of 10.4%; part of the gains are due to a mismatch of images, namely gray-scale for modeling and binary for evaluation. With respect to a baseline trained on binary data and also augmented by artificial noise, the improvement in character error rate is about 6% relatively. This result shows the interest of this approach for data augmentation in the training of neural network based recognizers.

With the character stroke [23], [24] parameter, we can identify and even partially correct the unrealistic results caused by large scale factors. Modifying the stroke of each scaled character snippet, so the average value in each page follows the distribution observed in real images, might be a potential approach to generate more natural-looking images.

Generating handwritten image with arbitrary text remains complex, but we plan to study ways to circumvent issues related to char

References

[1] H. S. Baird, "Document image defect models," in *Structured Document Image Analysis*. Springer, 1992, pp. 546–556.

[2] ——, "Calibration of document image defect models," in *Proc. of Second Annual Symposium on Document Analysis and Information Retrieval*, 1993, pp. 1–16.

[3] T. Kanungo, R. M. Haralick, and I. Phillips, "Global and local document degradation models," in *Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on*, 1993, pp. 730–734.

[4] R. P. Loce, W. L. Lama, and M. S. Maltz, "Modeling vibration-induced halftone banding in a xerographic laser printer," *Journal of Electronic Imaging*, vol. 4, no. 1, pp. 48–61, 1995.

[5] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[6] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra, "Draw: A recurrent neural network for image generation," *arXiv preprint arXiv:1502.04623*, 2015.

[7] Y. LeCun, C. Cortes, and C. J. Burges, "The mnist database of handwritten digits," 1998.

[8] R. Salakhutdinov and G. E. Hinton, "Deep boltzmann machines," in *International conference on artificial intelligence and statistics*, 2009, pp. 448–455.

[9] R. Salakhutdinov and I. Murray, "On the quantitative analysis of deep belief networks," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 872–879.

[10] B. Uria, I. Murray, and H. Larochelle, "A deep and tractable density estimator," *arXiv preprint arXiv:1310.1757*, 2013.

[11] T. Raiko, Y. Li, K. Cho, and Y. Bengio, "Iterative neural autoregressive distribution estimator nade-k," in *Advances in Neural Information Processing Systems*, 2014, pp. 325–333.

[12] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:1401.4082*, 2014.

[13] T. Salimans, D. P. Kingma, and M. Welling, "Markov chain monte carlo and variational inference: Bridging the gap," *arXiv preprint arXiv:1410.6460*, 2014.

[14] K. Gregor, I. Danihelka, A. Mnih, C. Blundell, and D. Wierstra, "Deep autoregressive networks," *arXiv preprint arXiv:1310.8499*, 2013.

[15] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang, "Casia online and offline chinese handwriting databases," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 37–41.

[16] D.-H. Wang, C.-L. Liu, J.-L. Yu, and X.-D. Zhou, "Casia-olhwdb1: A database of online handwritten chinese characters," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 2009, pp. 1206–1210.

[17] N. Otsu, "A Thresholding Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[18] C. Wolf, J.-M. Jolion, and F. Chassaing, "Text localization, enhancement and binarization in multimedia documents." in *ICPR (2)*. IEEE Computer Society, 2002, pp. 1037–1040.

[19] W. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, N.J.: Prentice Hall, 1986, pp. 115–116.

[20] R. Messina and J. Louradour, "Segmentation-free handwritten chinese text recognition with lstm-rnn," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015, pp. 171–175.

[21] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, vol. 4, 2012.

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[23] J. Ryu, H. I. Koo, and N. I. Cho, "Language-independent text-line extraction algorithm for handwritten documents," *Signal Processing Letters, IEEE*, vol. 21, no. 9, pp. 1115–1119, 2014.

[24] Y.-H. Tseng and H.-J. Lee, "Recognition-based handwritten chinese character segmentation using a probabilistic viterbi algorithm," *Pattern Recognition Letters*, vol. 20, no. 8, pp. 791–806, 1999.