SCS: Style and Content Supervision Network for Character Recognition with Unseen Font Style

Wei Tang[1,2,3(✉)], Yiwen Jiang[1,2,3], Neng Gao[3], Ji Xiang[3], Yijun Su[1,2,3], and Xiang Li[1,2,3]

[1] State Key Laboratory of Information Security, Chinese Academy of Sciences, Beijing, China
{tangwei,jiangyiwen,suyijun,lixiang9015}@iie.ac.cn
[2] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[3] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{gaoneng,xiangji}@iie.ac.cn

Abstract. There is a significant style overfitting problem in traditional content supervision models of character recognition: insufficient generalization ability to recognize the characters with unseen font styles. To overcome this problem, in this paper we propose a novel framework named Style and Content Supervision (SCS) network, which integrates style and content supervision to resist style overfitting. Different from traditional models only supervised by content labels, SCS simultaneously leverages the style and content supervision to separate the task-specific features of style and content, and then mixes the style-specific and content-specific features using bilinear model to capture the hidden correlation between them. Experimental results prove that the proposed model is able to achieve the state-of-the-art performance on several widely used real world character sets, and it obtains relatively strong robustness when the size of training set is shrinking.

Keywords: Character recognition · Convolutional neural networks · Style overfitting · Style supervision

1 Introduction

In our daily life, character is the most important information carrier. There are more than tens of thousands of different characters with variable font styles, and most of them can be well recognized by most people. But in the field of artificial intelligence, character recognition is considered as an extremely difficult task due to the very large number of categories, complicated structures, similarity between characters and the variability of font styles. Because of its unique technical challenges and great social needs, there are intensive research in this field and a rapid increase of successful techniques, especially the Convolutional

T. Gedeon et al. (Eds.): ICONIP 2019, CCIS 1143, pp. 20–31, 2019.
https://doi.org/10.1007/978-3-030-36802-9_3

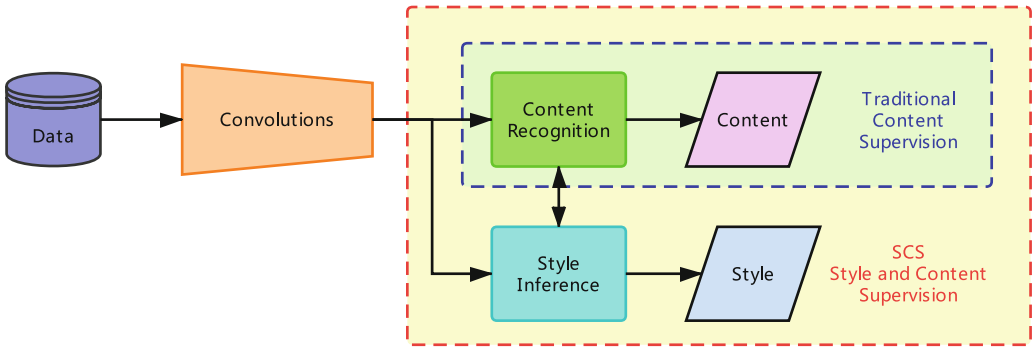

Fig. 1. Different to previous content supervision methods, SCS leverages both style and content supervision simultaneously. SCS takes into account the correlation between style and content in purpose of reducing style overfitting.

Neural Network (CNN), which has played an important role to improve recent Computer Vision studies [4,6,9,14,15].

As CNNs have achieved the great success in general object recognition, face recognition and other image recognition tasks [3,7,11,14,15,22], CNN-based methods also break the bottleneck of character recognition and achieve excellent performance even better than human on several popular large-scale benchmarks such as MNIST and ICDAR [1,2,18,21,24]. But there is a significant style overfitting problem still challenging character recognition: insufficient generalization ability to recognize characters with unseen font styles. For example, a well trained CNN model supervised by content label in a traditional way always perform poorly when test on the characters having such font style that the trained model has never seen [16].

Another limitation of traditional content supervision method is: the generalization ability of the trained model will decline dramatically when the scale of training set is shrinking [16]. Note that people learning new concepts can often generalize successfully from just a single example, yet machine learning algorithms typically require tens or hundreds of examples to perform with similar accuracy, and people can also use learned concepts in richer ways than neural-based algorithms [8]. Unlike human-level character recognition aimed to enhance generalization ability by using limited training set, in traditional applications of character recognition, the training set needs to contain as many font styles as possible to enhance the generalization ability of the trained model. But this naive method of expanding the training set is absolutely costly [6]. In addition, it is contrary to the intention of building a human-level learning model to extract rich concepts from limited data [8].

In this paper, we propose a novel framework named Style and Content Supervision (SCS) network to separate the style-specific and content-specific features of character. Different from traditional approaches only supervised by content labels, SCS simultaneously leverages the style and content supervision and takes into account the hidden correlation of style and content. Figure 1 shows the differences between the traditional content supervised method and our simul-

taneous style and content supervised method. Leveraging style supervision, the proposed SCS could achieve better performance than content supervised models on both the limited and relatively large training set.

Another useful method for reducing overfitting is data augmentation, which produces deformed samples to increase the diversity of training set. Previous work employ affine transformation to rotate and distort the original characters [23], but it lack of the ability to control the thickness of strokes and hardly deal with the outline fonts. We find that the thickness of strokes is a key point associated with the diversity of font styles, and the outline fonts are more hard to recognize compare with ordinary fonts. According to this situation, we design a simple but efficient data augmentation approach based on morphologic operations such as erosion and dilation to change the thickness of strokes and transfer the input character into outline style.

We conduct extensive experiments on several widely used character sets: Arabic numerals, English letters and simplified Chinese characters. Experimental results demonstrate that the proposed SCS is able to achieve state-of-the-art performance on all of these character sets, and it obtain desirable robustness on both of limited training set and relatively lager training set. And the benefit of the proposed data augmentation method is also be proved in our experiments. Overall, our contributions are as follows:

1. We propose an novel end-to-end trainable framework for character recognition, called Style and Content Supervision (SCS) network, which simultaneously leverages the style and content supervision to recognize character with unseen font style.
2. We verify the effectiveness of data augmentation approach based on erosion and dilation, which alter the thickness of character strokes as well as generate outline fonts to increase the style diversity of the training set for reducing style overffiting.
3. We carry out a series of experiments to evaluate our method, and the experimental results prove that the proposed SCS is able to achieve state-of-the-art performance on multiple character sets even when the scale of training set is changing.

The rest of this paper is organized as follows. Section 2 summarizes the related works. Section 3 introduces the proposed method in details. Section 4 presents the experimental results. Finally in Sect. 5, we conclude our work and discuss the future work.

2 Related Work

In recent years, character recognition has achieved unprecedented success because of the rise of Convolutional Neural Networks. But all these successful models are training on a large scale dataset and test on a relatively small dataset without significant font style variation. Wu et al. [18] propose a character recognition model based on relaxation convolutional neural network, and

took the 1st place in ICDAR'13 Handwriting Character Recognition Competition [21]. Meier et al. [2] create a multi-column deep neural network achieving first human-competitive performance on the famous MNIST handwritten digit recognition task. And this neural based model recognize the 3755 classes of handwritten Chinese characters in ICDAR'13 with almost human performance. Zhong et al. [24] design a streamlined version of GoogLeNet for character recognition outperforming previous best result with significant gap. Chen et al. [1] propose a CNN-based character recognition framework employ random distortion [13] and multi-model voting [12]. This classifier perform even better than human on MNIST and Chinese character set. In addition to the above models, the current state-of-the-art character recognition method proposed by Xiao et al. [19] it is also challenged by the style overfitting problem.

A growing number of works are devoted to enhance the generalization ability of character recognition models to make their performance desirable in dealing with unseen font styles. Xu et al. [20] propose a artificial neural network architecture called Cooperative Block Neural Networks to address the variation in the shape of characters by considering only three different fonts. Lv [10] successfully applied the stochastic diagonal Levenberg-Marquardt method to a convolutional neural network to recognize a small set of multi-font characters that consist of the Arabic numerals and English letters without the Chinese characters. Zhong et al. [23] propose a CNN-based multi-font character recognizer using multi-pooling and affine data augmentation achieving acceptable result, but the size of the training set is more than 500% of the test set (240 font styles for training and other 40 for test) and shrinking the size of training set will significantly reduce the effect of the model. Tang et al. [16] propose a specific kernel to extract the marginal distribution of character pixels that takes account the skeleton information to enhance the generalization ability of CNNs when training on a limited data set. But all these previous works are based on a single network supervised by content label and ignore the informative style supervision.

3 Methodology

In this section, we first introduce the overall architecture of the proposed SCS. Then we present our method in detail by successively introducing the modules of data augmentation, shared convolution as well as style and content branches.

3.1 Overall Architecture

An overview of our framework is illustrated in Fig. 2. The backbone of the style inference branch is a full connection network with a softmax classifier supervised by style label, which has a extra mixer to receive the feedback from content branch. And the content recognition branch obtain the same architecture with the style inference branch, which is supervised by content label and also has a mixer to receive the feedback from style branch. We adopt a convolutional network as the main framework of shared convolution module. This shared

convolution module is widely used to extract the shared feature of input image. In addition, we use data augmentation to enhance the style diversity of the training data.

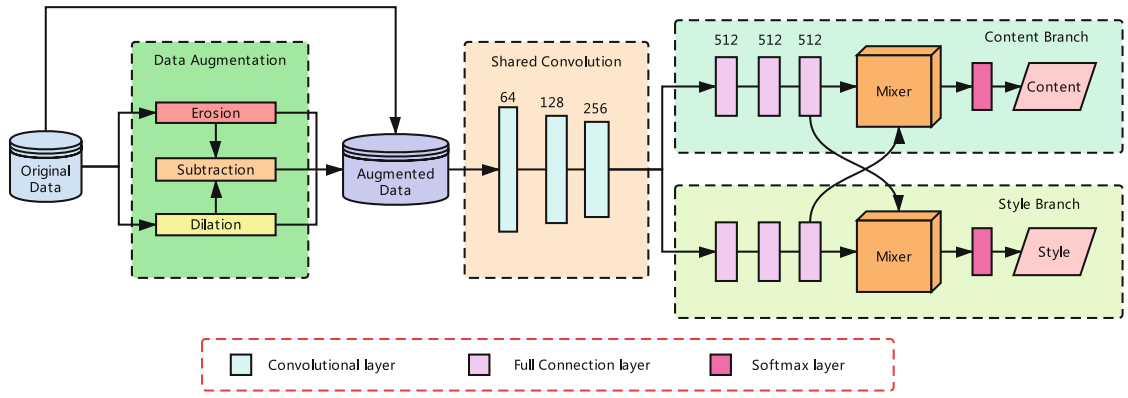

Fig. 2. Overall Architecture of the proposed SCS network. From left to right, first the original training data is augmented to enhance the style diversity; then the augmented data is pass through the shared convolutional layers for extracting texture features; next the shared features is fed to separated supervision branches to learn the hidden representations for different tasks; finally the mixed representations are fed to softmax classifiers to make task-specific predictions. This end-to-end learning approach takes into account the hidden correlations of style and content, and it is supervised by these two factors simultaneously.

3.2 Data Augmentation

It is proved that the diversity of training data is significant for enhancing the generalization ability of deep model [8]. Previous work employ affine transformation to rotate and distort the original characters [23], but it is difficult to control the thickness of character strokes and can not generate outline fonts. However, such style-related factors are typically the important features of character diversity.

According to this situation, we propose an effective data augmentation method in purpose of altering the thickness of character strokes and producing outline fonts, which could considerably increase the style diversity of training data. In our approach, the morphologic operations of dilation and erosion are respectively used to increase and reduce the thickness of character strokes, and then the outline character is produced by the subtraction of the dilated character and the eroded one. This approach do not require any training process and could be a complement to traditional affine transformation. Figure 3 presents some generated samples produced by the proposed data augmentation method.

3.3 Shared Convolution

Convolutional Neural Network (CNN) is widely used to extract the texture feature of image [6,9]. We leverage the strong feature learning power of CNN to

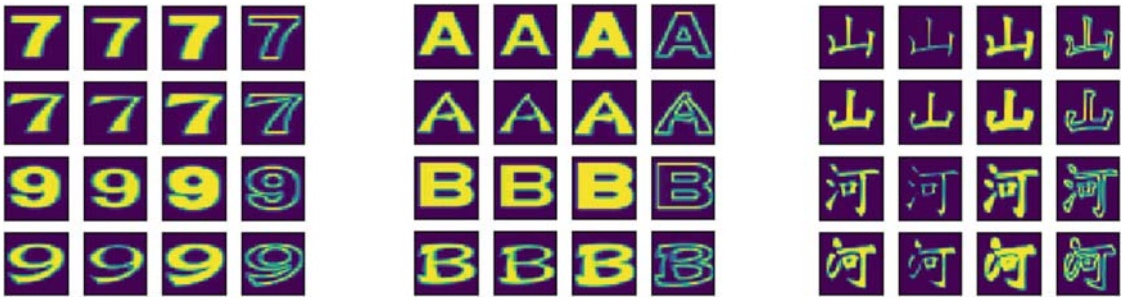

Fig. 3. Samples produced by the proposed data augmentation method. In this figure, the left, center and left parts respectively present several results of Arabic numerals, English letters and Chinese characters. And for each character set, the 1-th, 2-th, 3-th and 4-th columns respectively list the original, the dilated, the eroded and the outline fonts.

capture the shared texture features of input characters, and then feed these shared features into separate learning branches to obtain content-specific feature and style-specific feature.

This shared convolution network consist of a series of Convolution-BatchNorm-LeakyReLU down-sampling blocks that yield the texture features of input images. Here in our model, we totaly stack 3 down-sampling blocks, and in some advanced versions it could be stacked more blocks if necessary. The output channels of convolutional layers are 2, 4, 8 times of 32 respectively. The first convolution layer is with 5×5 kernel and stride 1 and the rest are with 3×3 kernel and stride 1, and all ReLUs are leaky with slope 0.1.

3.4 Style and Content Branches

We design separated learning branches to capture style-specific feature and content-specific feature. In each branch we adopt stacked full connection layers to learn such task-specific features, and then we use the mixers based on bilinear model to communicate each branches for taking into account the correlation between style and content. Finally we employ softmax classifiers to infer the style and recognize the content.

Full connection network is widely used in hidden feature learning on account of their excellent learning ability and desirable scalability [6]. In each branches, we stack 3 full connection layers, and each layers contain 512 cells with dropout rate 0.1. Note that it could be stacked more full connection layers for enhancing the learning ability, but adding more full connection layer could lead to a sharp increase in the number of trainable parameters, which bring high costs.

We combine the style feature and content feature in the mixer, which is a bilinear model. Bilinear model is a two-factor model with the mathematical property of separability: their outputs are linear in either factor when the others held constant, which has been demonstrated that the influences of two factors

can be efficiently separated and combined in a flexible representation [17]. The combination function can be formulated as:

$$M_s = F_s W_s F_c$$

$$M_c = F_c W_c F_s$$

where W_s and W_c is the trainable parameters of bilinear models, F_s and F_c denote the feature representations of style and content, M_s and M_c is the mixed feature representations of style and content.

With the mixed feature obtained by the previous bilinear model, we obtain the prediction probability for the style y_s and content y_c of a given character C by softmax classifiers:

$$p(y_s|C) = softmax(V_s M_s)$$

$$p(y_c|C) = softmax(V_c M_c)$$

where V_s and V_c is the trainable parameters, which is responsible for converting the mixed feature representation for each task into predictions through linear transformation.

The goal of traditional character recognition is to infer the contents of characters from their images. It typically minimize the sum of the negative log-likelihoods. In addition to content supervision, here we take into account style supervision. Combined with content loss and style loss, the full loss function is:

$$Loss = -\sum_i (\lambda logp(y_c^i|C^i) + (1-\lambda) logp(y_s^i|C^i))$$

where the hyper-parameter λ controls the trade-off between content loss and style loss. Considering that content recognition is more important in our experiments, we set λ to be 0.9.

4 Experiments

We conduct extensive experiments to evaluate the effectiveness of our approach. In this section, we present the details of our experiments and analyze the experimental results.

4.1 Data

We build several dataset to evaluate our method, including Arabic numerals (10 content classes), upper-case and lower-case English letters (52 content classes), and simplified Chinese characters (3755 content classes). All of these character sets are extracted from True Type font (TTF) files, which jointly launched by Apple Inc. and Microsoft Corp. as a standard font format file supporting their operating systems. We extract all the character images from 91 TTF files with widely varying font styles. Each character with a certain style and content is presented by a 32×32 PNG image. Figure 4 presents the samples of data. Note that the reason why we abandon the popular benchmarks such as MNIST and ICDAR is that they lack of explicit style labels.

Fig. 4. Samples of data. From left part to right part in this figure, it respectively present several characters of Arabic numerals, English letters and Chinese characters. In each part, there are 91 character images obtain same content but different styles.

4.2 Baseline Models

We compare our models with several baselines, all these models are design for character recognition with unseen styles. The description of these baselines and our model are listed below:

CMP. Convolutional Multiple Pooling network is a CNN-based model adopting multi-pooling and affine transformation. It could achieve desirable performance on a relatively large training set [23].

CSK. Convolutional Skeleton Kernel network is a CNN-based model, which use skeleton kernel to capture the skeletons of characters. It perform well on a limited training set [16].

SCS. Style Content Supervised network is our proposed model. It leverages the style and content supervision simultaneously as well as takes into account the hidden correlations between style and content.

4.3 Implementation Details

We design two experiments for test the baselines and our model training on relatively large dataset and limited dataset respectively. In first case, we randomly choose 70 styles for training and use other 21 styles for test. In second case, the number of training styles: the number of test styles is set to be constantly $1:2$ and the number of training styles is dynamically changing to verify the robustness of the tested models. In order to get a more convincing conclusion, we conduct each experiment 10 times to calculate the average accuracy and shuffle the datasets after each evaluation.

When training, we use random values drawn from the Gaussian distribution with 0 mean and 0.01 standard deviation to initialize all trainable weights, and all bias are initialized at 0. Adam [5] is used as the optimization algorithm and the mini-batch size is 128. The learning rate is set to be $1e^{-4}$. After each epoch, we shuffle the training data to make different mini-batches. Furthermore, we adopt 2×2 morphology kernel for the proposed data augmentation method, where all the morphology operations are provided by OpenCV.

4.4 Performance Analysis

Large Training Set. Table 1 presents the experimental results of the baselines and our model training on relatively large dataset. Based on these results, we have the following findings:

Table 1. Results of training on relatively large dataset.

Model name	Data augmentation	Character set		
		Arabic numerals	English letters	Chinese characters
CMP	AF[a]	96.63%	92.32%	84.89%
CMP	ED[b]	96.94%	92.85%	85.33%
CMP	AF+ED[c]	97.21%	93.09%	85.58%
CSK	AF	96.41%	92.15%	84.97%
CSK	ED	96.72%	92.65%	85.39%
CSK	AF+ED	97.04%	92.98%	85.62%
SCS	AF	97.12%	93.16%	87.59%
SCS	ED	97.46%	93.54%	87.68%
SCS	AF+ED	**97.69%**	**93.82%**	**88.06%**

[a]Using traditional data augmentation based on affine transformation.
[b]Using proposed data augmentation based on erosion and dilation.
[c]Using two types of data augmentation methods simultaneously.

1. As we emphasized in this paper, the style supervision is important to character recognition. Leveraging this informative style supervision, SCS obtain the ability of style-aware to find that which styles of the input characters are more likely to be. This style inference is a helpful signal for content recognition because the content features of characters often have a rich correlation with their style features. Giving the credit to style supervision, our SCS models outperform the baselines with a significant gap, especially on the hard task of Chinese character recognition.
2. Altering the thickness of character strokes is a helpful approach to reduce style overfitting. The proposed data augmentation method based on erosion and dilation essentially enhance the style diversity of training set. Experimental results proves that the proposed data augmentation method based on erosion and dilation is better than traditional affine transformation on character recognition, and it is also a beneficial complement to affine transformation.

Limited Training Set. Figure 5 shows the results of the baselines and our SCS model training on limited dataset. In this experiment, we abandon data augmentation to evaluate the authentic generalization abilities of all tested models.

When training on limited dataset, our SCS model also outperform all the baselines. Even when the size of training set is changing, SCS maintains the superior position. This evidence demonstrates the robustness of the style supervision in extreme adverse conditions where few styles are available for training.

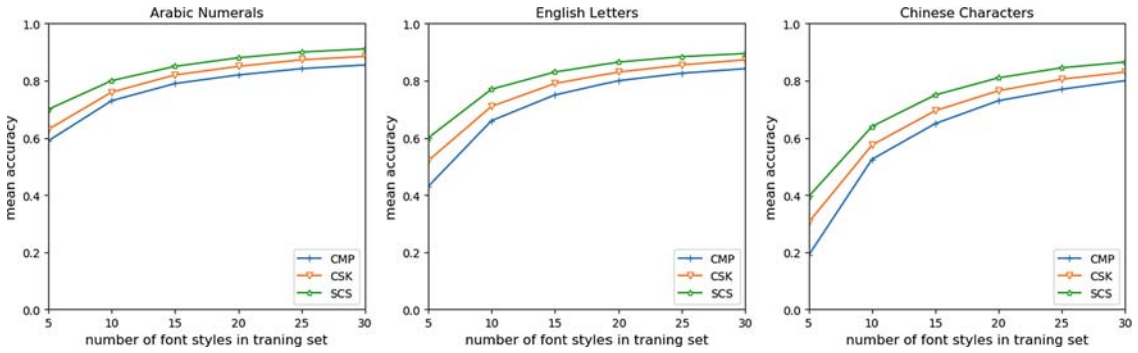

Fig. 5. Results of training on limited dataset.

5 Conclusion and Future Work

Our purpose is to recognize the character content when the observed characters obtain unseen font styles. It is challenged by style overfitting problem posed by the characters do not have an explicit separation between style and content. To address this issue, we propose a novel model of Style and Content Supervision (SCS) network, which integrate style and content supervision to resist style overfitting. Experimental results demonstrate that SCS achieve the state-of-the-art performance on several widely used character sets.

In the future, we would like to conduct more experiments and find a more effective way to reduce style overfitting. We also hope to verify the effectiveness of our method in a larger number of other tasks, which need more datasets and annotations. More advanced models are expected as some shortcomings are still existed in current models, such as lack of external knowledge and excessive reliance on annotated data.

Acknowledgment. We thank all reviewers for their helpful advice. This work is supported by the National Key Research and Development Program of China, and National Natural Science Foundation of China (No. U163620068).

References

1. Chen, L., Wang, S., Fan, W., Sun, J.: Beyond human recognition: a CNN-based framework for handwritten character recognition. In: IAPR Asian Conference on Pattern Recognition, pp. 695–699 (2015)
2. Dan, C., Meier, U.: Multi-column deep neural networks for offline handwritten Chinese character classification. In: International Joint Conference on Neural Networks, pp. 1–6 (2015)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. Computer Science, pp. 580–587 (2013)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 770–778 (2016)

5. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR abs/1412.6980 (2014)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Annual Conference on Neural Information Processing Systems 2012, NIPS 2012, Lake Tahoe, Nevada, United States, 3–6 December 2012, pp. 1106–1114 (2012)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
8. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015)
9. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
10. Lv, G.: Recognition of multi-fontstyle characters based on convolutional neural network. In: Fourth International Symposium on Computational Intelligence and Design, pp. 223–225 (2011)
11. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: International Conference on Neural Information Processing Systems, pp. 91–99 (2015)
12. Schmidhuber, J., Meier, U., Ciresan, D.: Multi-column deep neural networks for image classification. In: Computer Vision and Pattern Recognition, pp. 3642–3649 (2012)
13. Simard, P.Y., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks applied to visual document analysis. In: International Conference on Document Analysis and Recognition, p. 958 (2003)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
15. Szegedy, C., et al.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015, pp. 1–9 (2015)
16. Tang, W., et al.: CNN-based Chinese character recognition with skeleton feature. In: Cheng, L., Leung, A.C.S., Ozawa, S. (eds.) ICONIP 2018, Part V. LNCS, vol. 11305, pp. 461–472. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04221-9_41
17. Tenenbaum, J.B., Freeman, W.T.: Separating style and content. In: Annual Conference on Neural Information Processing Systems 1996, NIPS 1996, Denver, CO, USA, 2–5 December 1996, pp. 662–668 (1996)
18. Wu, C., Fan, W., He, Y., Sun, J., Naoi, S.: Handwritten character recognition by alternately trained relaxation convolutional neural network. In: International Conference on Frontiers in Handwriting Recognition, pp. 291–296 (2014)
19. Xiao, X.F., Jin, L., Yang, Y., Yang, W., Sun, J., Chang, T.: Building fast and compact convolutional neural networks for offline handwritten chinese character recognition. Pattern Recogn. **72**, 72–81 (2017)
20. Xu, N., Ding, X.: Printed Chinese character recognition via the cooperative block neural networks. In: IEEE International Symposium on Industrial Electronics, vol. 1, pp. 231–235 (1992)
21. Yin, F., Wang, Q.F., Zhang, X.Y., Liu, C.L.: ICDAR 2013 Chinese handwriting recognition competition (ICDAR), pp. 1464–1469 (2013)

22. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
23. Zhong, Z., Jin, L., Feng, Z.: Multi-font printed Chinese character recognition using multi-pooling convolutional neural network. In: International Conference on Document Analysis and Recognition, pp. 96–100 (2015)
24. Zhong, Z., Jin, L., Xie, Z.: High performance offline handwritten Chinese character recognition using googlenet and directional feature maps. In: International Conference on Document Analysis and Recognition, pp. 846–850 (2015)