A method for document image enhancement to improve template-based classification

Jiyun Li
Donghua University
2999 North Renmin Road
Songjiang, Shanghai, China
+86 21 67792293
jyli@dhu.edu.cn

Zhijie Mei
Donghua University
2999 North Renmin Road
Songjiang, Shanghai, China
+86 18801622699
2181817@mail.dhu.edu.cn

Tingting Zhang
Mid Sweden University
Holmgatan 10
Sundsvall, Sweden
+46 10 1428878
tingting.zhang@miun.se

ABSTRACT
Document classification is one of the significant procedure in paper document recognition. This article proposed a method for document image enhancement to improve the performance of classification in the convolutional neural network. An enhanced document image was generated by extracting the table frame, text region, and shape of the raw document. The template-based classification experiment on 414 customs documents and more than one thousand generated images showed the enhanced image could help CNN model achieve higher accuracies compared to the original images. It could also diminish the interference of noise and unrelated features in document classification optimizing the robustness of networks. The proposed method also demonstrated the channels of the image could provide more information except for color in deep neural networks. As the similarity in the whole image classification tasks, the conclusion might provide ideas for the training of the neural networks in other fields such as street view recognition, medical image recognition, etc.

CCS Concepts
• Computing methodologies → Computer graphics → Image manipulation → Image processing

Keywords
Image enhancement, image classification, convolutional neural network, document recognition.

1. INTRODUCTION
With the growing number of companies administrating their document by the IT system, extracting data from the previous paper documents became a large issue for them. It is easy for humans to extract the information by their knowledge and experience. The process of paper document digitization, especially in documents that had diverse structures, often required a large scale of labor cost in the past.

As the rapid development of artificial intelligence, especially on machine learning, over the past years, the related technology was expected to be applied in the document recognition. Document classification was one of the typical procedures to optimize the efficiencies and accuracies in document information extraction. It allowed us to design a suitable extraction approaches for each class of documents. Therefore, the accuracy of document classification could significantly affected the performance of the information extraction system.

In this article, a method for document image enhancement was proposed to improve the performance of template-based document image classification. In the method, we assumed the template of a document could be entirely determined by the table framework, text region, and the shape. First of all, the template-related features were extracted to generate three views of the document images. Then, the three views were merged by putting the views into the individual color channel.

In the experiment, the 414 customs invoice dataset was classified into eight classes by its template. We trained the LeNet5, AlexNet and VGG16 models on the original images, enhanced-color images, enhanced-gray images independently to compare the performance. Furthermore, an extra dataset was generated from the training dataset by converting the color scale, changing the contrast and brightness and randomly adding colorful spots on the images. The result revealed that the presented method improved on both the accuracy and robustness of the various networks.

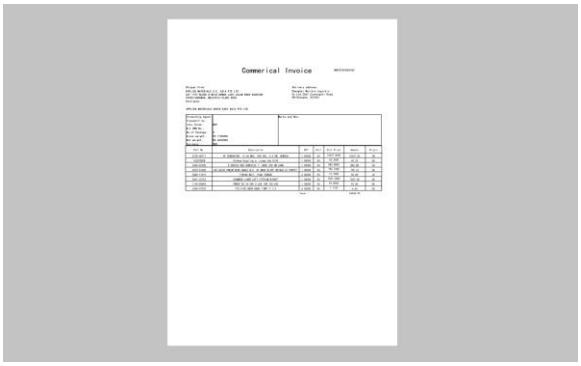

Figure 1. A typical document image.

2. PREVIOUS WORK
In recent years, the research on pattern recognition offered ideas to develop the systems for promoting automation in paper document recognition. A typical process [1] of entering paper document information to database consisted of document classification, template matching, and optical character recognition (OCR) and text structuralization. Among these procedures, classification was one of the challenges in document recognition. D. Gaceb [2] summarized low-level-image-based, physical-layout-based, logical-layout-based and textual-feature-based were four methods in document classification.

HPCCT & BDAI 2020, July 3–6, 2020, Qingdao, China.

ACM ISBN 978-1-4503-7560-3/20/07…$15.00
https://doi.org/10.1145/3409501.3409531

In the real-world, most raw documents, such as notification, invoice, prescription, delivery slip, were issued based on a sort of specific templates. It is clear that involving the layout feature in the extraction process could help the systems understand the images more easily which leads to higher accuracy compared to the ways in which lots of researches only focusing on the text [2][3][4]. Hu etc. [5] introduced a feature set, interval encoding, which could encode the layout feature to a fixed-length vector and could be classified in the hidden Markov model. O. Augereau, etc. [6] combined the visual feature from Bag of Words(BoW), and the textual feature, Bag of Visual Words(BoVW), to improve classification performance.

The studies on deep learning have shown the excellent performance of the convolutional network in image classification. As early as 1998, LeCun, etc. [7] found a gradient-based algorithm that can be used to achieve high accuracy in high-dimension pattern classification with appropriate network given. They introduced a CNN network called LeNet5 to recognize the handwritten characters. The experiment result showed the error rate stabilized at 0.95% after training for ten epochs. In 2012, Alex Krizhevsky, etc. [8] proposed a deep convolutional network called AlexNet and achieved 37.5% top-1 error rate and 17.0% top-5 error rate on ImageNet dataset. VGG16 [9] is another famous convolutional network introduced by K. Simonyan and A. Zisserman in 2015. Their experiments showed an improvement could be reached by increasing the depth of the network. The convolutional neural network model could be customized for the specific domain and much effort has been made in the field of

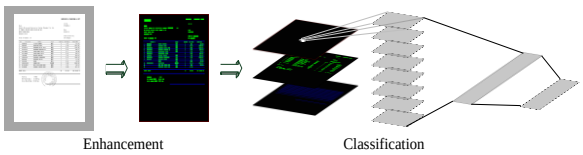

Figure 2. The process of classification with proposed enhancement.

recognition.

For document image, Sun Y [10] proposed a neural network improved on VGG16 in 2019. By adding a batch normalization layer and increasing kernel size, the model made an improvement on speed. X Chen [11] concluded his dissertation that several classic image enhancement algorithms allowed CNN to achieve higher accuracy. These facts convinced us that the performance of neural networks in document recognition could be optimized by image enhancement. Therefore, the proposed enhancement was applied to the raw images before the classification, shown in Figure 2.

3. DOCUMENT IMAGE ENHANCEMENT

In this section, the proposed method for document image enhancement was explained. We assumed that the raw document images has already been skewed to the correct angle and entirely split from the background if taken by a camera.

As Figure 3 showed, the table frame in the document was detected and separated by the morphological algorithm to generate the table-frame view. This process would be elaborated in section 3.1. Then, the text region was detected and marked by the filled rectangles to generate the text-region view. Thirdly, the frame of the original images was marked by a contour to indicate the shape of the original images. Last, all three views were merged into one image and resized to a standard resolution that could be input to the network directly. This step would be elaborated in section 3.3.

There was another deep question we tried to make clear in this article, whether it was possible to improve the performance by the experience of human beings. In detail, the features of documents were partially determined manually rather than fully recognized by the network like the process of the traditional convolutional neural network training. We hoped the proposed method could prevent interference from the unrelated information in the image by the knowledge of human beings. Empirically, distinguishing feature types helped us with the classification of images. For instance, the human could distinguish the table frame and characters on a document image even both of them are comprised of black pixels, this ability could help a lot in the process of human image classification. The procedure of distinguishing could be easily achieved by the traditional image processing algorithm or classic models. Therefore, we believed enhancing the

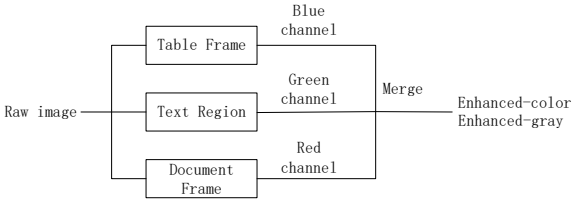

Figure 3. Schematic diagram for proposed method.

images by the traditional algorithm could improve the performance of the neural network.

In this idea, the assumption that templates could be determined by the feature of the table frame, text region, and shape of the document is believed to be the experience held by humans. We tried to 'guide' the neural network via the proposed image enhanced method.

There were two ways to prove the idea. Firstly, we tried to actively overlook the unrelated information. For example, the characters were replaced with the filled rectangles as the exact characters are not the feature of images but the regions they lied on. Secondly, in almost CNN experiment, the values of RGB were input into the network as the three channels. Rather than ordinary images, the color was barely useful in template-based classification as almost all elements on the documents were in a single or similar color. We tried to map the different classes of features to different image channels. This procedure helped the network distinguish the different types of features. The networks were expected to be trained combined with the senses of human beings by the above approaches.

3.1 Table Frame

Table frame was to be considered as one of the features of images in this article. Tables in most documents comprised of several horizontal and vertical straight lines. The goal of this step was extracting the lines from the raw document to a single-channel image.

Due to the various condition of optical equipment, the raw documents had a wide range of equality. An adaptive binarization algorithm was applied to the document images at first. The algorithm would adjust the threshold with the Gaussian weight sum of the neighborhood area in order to adapt to the different brightness and contrast around.

The next step was removing all characters by erosion and dilation. Figure 4 was the image after bitwise-or operation of horizontal,

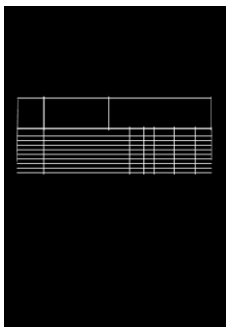

Figure 4. An example of table-frame view.

vertical erosion and dilation showing the table frame has been extracted. Straight lines detection was applied to the image after table extraction. The detected lines were redrawn on a black image to make the table frame clearer.

3.2 Text Region

This step aimed to mark the text region by the bright filled rectangles to generate the text region view. This view provided the information about the position and size of each word on the documents. This article was not stressed on the characters detection since many researches [12][13][14][15] on characters or word detection had been conducted these years. Either approach could be used as long as the text region could be detected. Figure 5 showed an example of text-region view.

Figure 5. An example of text –region view.

3.3 Document frame

The shape was another important information in template-based classification. The document-frame views displayed the contour of the whole documents that could indicate the shape. The lines of contour must be thick enough to affect the network. Figure 6 showed an example of document-frame view.

Figure 6. An example of document-frame view.

3.4 Image merging

The goal of image merging was combining the generated three single-channel original-size views to one three-channel resized images which could be input to the neural network.

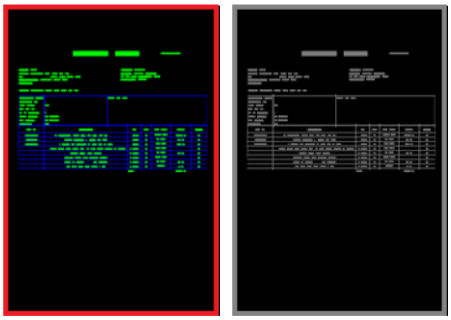

Figure 7. Examples of enhanced-color images and enhanced-gray images.

The three views showed above would be merged into a new image showed in Figure 7, in which the table-frames view as the blue channel, text-region view as the green channel and document-frame view as the red channel. As the aspect ratio was significant in classification, the new image cannot be resized directly. The merging processing must ensure all enhanced images in the same size without deformation. The merged images were converted to gray-scale images as a comparison. The color-scale images were named as enhanced-color image and the grayscale images as enhanced-gray images.

4. EXPERIMENT

In this paper, we conducted experiments on the classic CNN networks LeNet5, AlexNet, and VGG16. Early stopping point is set during training that ends the training when the test loss no longer decreased over the latest five epochs. The performance of our enhanced method was evaluated by comparing the accuracies of the same model independently trained on three groups of images, original images, enhanced-color images, and enhanced-gray images.

All experiments were all performed on Dell G3 series laptop with Intel Core i5 8300H, NVIDIA GeForce 1050Ti and Windows 10. Google open source OCR engine Tesseract [16] was used to layout analysis and catch the text region. We applied OpenCV to extract the table frame and reconstructed the enhanced images.

4.1 Dataset

The dataset included 414 customs document images and was classified eight classes by its template. The resolution of the raw scanned image was approximately 2400*3500 and would be adjusted to exact 421*297 after the enhancement. Apart from one of the labels that meant the template was previously unseen and couldn`t be classified, the samples with the same label were issued based on the same templates despite minor differences in structure, font, and content.

In the real-world, the qualities of images might be various under different circumstances. That was another common issue caused recognition failure. In order to assess the introduced method as realistically as possible, an extra dataset was provided. The extra dataset included 1242 images and all of them were generated from

the raw images by three approaches, converting the document images from color scale to grayscale, changing the contrast and brightness of document images and drawing random spots on the images. The methods could simulate the document scanning process to assess the robustness of the trained models.

4.2 Neural Network

The number of output layer was eight in accordance with the classes of the dataset. All three networks were simplified by decreasing the number of both the filters in convolutional layers and units in fully-connected layers. The simplification made the differences of accuracies between groups more distinct. The structures of the simplified LeNet5 were shown in Figure 8.

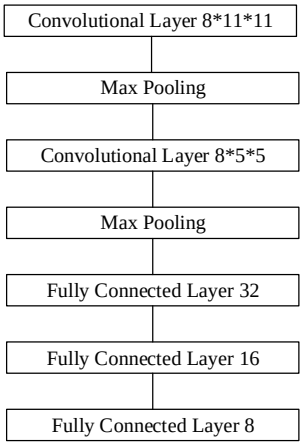

Figure 8. Simplified LeNet5.

4.3 Result

Table 1 showed the test accuracies and epochs at the early stopping point in training on LeNet5, AlexNet, and VGG16. For LeNet5, the training stopped at around twelve epochs. The networks on both enhanced images achieved close accuracies, about 70%, higher than the accuracy of original images, 60%. For AlexNet, the three groups of training stopped with the same time consumed, about thirteen epochs. The two proposed enhanced images had higher accuracies throughout the training. The ultimate accuracies were approximately 10 percent higher than the original images. For VGG16, all three networks stopped at nearly twenty epochs achieving the accuracies of 83.75%, 85.9%, and 86.25% respectively. The result suggested that the enhanced images could help the model reach higher accuracies without more time cost.

Table 1. Epochs and accuracies at the early stopping point

	LeNet5	AlexNet	VGG16
Origin	12/62.50%	13/77.50%	19/83.75%
Enhanced-color	13/70.51%	14/87.18%	18/85.90%
Enhanced-gray	12/71.25%	14/86.25%	20/86.25%

The result of the evaluation was listed in Table 2 showing the accuracies of evaluation on the generated dataset. The accuracies of original images were consistent in test dataset and generated dataset over all three networks. The enhanced-color images performed better with the accuracy rising from 85.9% on the test dataset to 90.54% on the generated dataset in VGG16. The enhanced-gray images enjoyed better accuracies in LeNet5 improving the accuracies from 71.25% to 78.13%.

Overall, the enhanced images got higher accuracies on all three networks. By comparing the accuracies between the two datasets, the enhanced images could weaken the influence of noise added into images maintaining the higher accuracies. The enhanced-color performed better on the deeper networks compared to the enhanced-gray images.

Table 2. The accuracies of evaluation on the generated dataset

	LeNet5	AlexNet	VGG16
Origin	64.98%	77.88%	83.33%
Enhanced-color	70.51%	87.42%	90.54%
Enhanced-gray	78.13%	88.38%	85.74%

5. CONCLUSION

In this article, a method for document image enhancement was proposed. The proposed method was proved to improve the performance of document classification in CNN. Firstly, the three types of image features which were table frame, text region and shape of documents, were extracted and generate three views of images. Then, the three views were merged by putting them into individual color channel and resized the merged image to a standard size. The resized images were input to three classic CNN models to evaluate the proposed method.

The experiment showed that the proposed enhancement method could optimize the accuracy of all three classic CNN models. It also could diminish the interference of noise added in the generated dataset improving the overall accuracies under various circumstances, such as the conditions of scanning device, qualities of raw documents. The enhanced-color images could achieve better performance on the deeper network whereas the enhanced-gray images were better on shallower network. The reason why the generated dataset had higher accuracies than the test dataset is that they were generated from the training dataset which means they still kept the most features of training images.

For the first way mentioned in section 3, the features other than table frame, text region and shape of the document were defined as template-unrelated features. All of them were actively overlooked in the process of the proposed method. However, the accuracies of the model became higher even some information in the images was lost. This facts suggested the experience of humans helped the network learn the feature of images more precisely. For the second way, the experiments on LeNet-5(five layers), AlexNet(eight layers) and VGG16(sixteen layers) showed that splitting the features by putting the same type of features into individual image channels could improve the accuracies in the deeper networks. However, no evidence had been seen that it could offer an improvement on network with fewer layers.

To sum up, the accuracy of the network could be optimized by the proposed image enhancement. The result of the experiment showed the experience of humans might be effective in the neural network learning. As the features of images in other image classification tasks can be easily split by traditional algorithms, the proposed method might offers ideas on the fields such as street view recognition, medical image recognition, etc.

6. REFERENCES

[1] Benjamin Seidler, Markus Ebbecke, and Michael Gillmann. 2010. SmartFIX statistics: towards systematic document

analysis performance evaluation and optimization. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (DAS '10). Association for Computing Machinery, New York, NY, USA, 333–340. DOI:https://doi.org/10.1145/1815330.1815373.

[2] Gaceb, D., Eglin, V., & Lebourgeois, F. (2014). Classification of business documents for real-time application. Journal Of Real-Time Image Processing, 9(2), 329-345.

[3] Larry M. Manevitz and Malik Yousef. 2002. One-class svms for document classification. J. Mach. Learn. Res. 2 (March 2002), 139–154.

[4] Palm, R., Winther, O., & Laws, F. (2017). CloudScan - A configuration-free invoice analysis system using recurrent neural networks.

[5] Jianying Hu, R., Kashi, & Wilfong. (1999). Document image layout comparison and classification. Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318), 285-288.

[6] Augereau, O., Journet, N., Vialard, A., & Domenger, J. (2014). Improving Classification of an Industrial Document Image Database by Combining Visual and Textual Features. 2014 11th IAPR International Workshop on Document Analysis Systems, 314-318.

[7] Lecun, Y., Bottou, Bengio, & Haffner. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 6 (May 2017), 84–90. DOI:https://doi.org/10.1145/3065386

[9] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition.

[10] Sun, Y., Zhang, J., Meng, Y., Yang, J., & Gui, G. (2019). Smart Phone-Based Intelligent Invoice Classification Method Using Deep Learning. IEEE Access, 7(99), 118046-118054.

[11] Chen X. Image enhancement effect on the performance of convolutional neural networks. Blekinge Institute of Technology 2019.

[12] Liao, M., Shi, B., Bai, X., Wang, X., & Liu, W. TextBoxes: A Fast Text Detector with a Single Deep Neural Network. Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[13] Liao M, Shi B, Bai X. Textboxes++: A single-shot oriented scene text detector. IEEE transactions on image processing, 2018, 27(8): 3676-3690.

[14] Chen K, Seuret M, Wei H, et al. Ground truth model, tool, and dataset for layout analysis of historical documents.Document Recognition and Retrieval XXII. International Society for Optics and Photonics, 2015, 9402: 940204.

[15] Chen K, Liu C L, Seuret M, et al. Page segmentation for historical document images based on superpixel classification with unsupervised feature learning. 2016 12th IAPR Workshop on Document Analysis Systems (DAS). IEEE, 2016: 299-304.

[16] Smith R. An overview of the Tesseract OCR engine. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). IEEE, 2007, 2: 629-633.