IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 37, No. 6, December, 2020, pp. 1103-1110
Journal homepage: http://iieta.org/journals/ts

A Hybrid Classifier for Handwriting Recognition on Multi-domain Financial Bills Based on DCNN and SVM

Shuaiwen Wang[1,2*], Bei Yuan[1,2], Di Wu[3]

[1] Harbin Bank Postdoctoral Research Station, Harbin 150010, China
[2] Harbin Institute of Technology, Harbin 150001, China
[3] College of Electronics and Communication Engineering, Tongji University, Shanghai 200082, China

Corresponding Author Email: wangshuaiwen1@hrbb.com.cn

https://doi.org/10.18280/ts.370623

Received: 15 June 2020
Accepted: 28 September 2020

Keywords:
financial bill, handwriting recognition, deep convolutional neural network (DCNN), support vector machine (SVM)

ABSTRACT

With the rapid growth of the global economy, the automatic recognition of financial bills becomes the primary way to reduce the burden of the traditional manual approach for bill recognition and classification. However, most automatic recognition methods cannot effectively recognize the handwritten characters on financial bills, especially when the bills come from different financial companies. To solve the problem, this paper fully explores the bill system in banks and the operations of bill number recognition, and then develops a hybrid classifier based on deep convolutional neural network (DCNN) and support vector machine (SVM), with the aim to recognize the handwritten numbers on financial bills in different domains. The DCNN with different channels was adopted to effectively mine the local handwritten numbers on financial bills from varied sources. Then, the extracted information was fed to the SVM to realize accurate classification of numbers. Our method makes full use of the distribution difference between information in different fields, and adapts to different fields based on the parameter sharing mechanism. Experimental results show that our method can recognize the handwritten numbers on financial bills more accurately (>3%) than benchmark methods.

1. INTRODUCTION

The application of artificial intelligence in finance blurs the boundaries between digital economy and traditional economy. Modern techniques like blockchain and cryptocurrency provide novel solutions to financial economy, and unveil a brand-new digital future with infinite possibilities.

With the development of the economy and the proliferation of financial services, the automatic recognition of financial bills has become an important task of banks and other financial institutions in the pursuit of informatization and electronification. Traditionally, the bank system handles various bills manually. The manual processing is slow and prone to errors.

Thanks to advanced electronic hardware and image recognition algorithms, banks today can read the bill information quickly, and management the entered information, ensuring the smoothness of business transactions in the financial system [1]. However, it remains a technical difficulty to recognize the key contents on the bills quickly and accurately. The common solution is to scan the bill with a high-speed scanning device, import the scan image into computer, and pinpoint and extract the area of important information, especially the amount, from that image [2]. But the figures in the amount, which is usually handwritten, is very difficult to segment. After all, each person has his/her own writing style, the handwriting often involves lots of ligatures, and the amount contains multiple figures [3].

To overcome the difficulty, many scholars have investigated handwriting recognition, and provided tools with good recognition accuracy, such as neuro-fuzzy system (NFS), support vector machine (SVM), and deep learning classifiers [4, 5]. Despite these efforts, the handwritten numbers on financial bills from various sources are still an open challenge. New techniques and methods need to be developed to further improve the recognition accuracy, operating time, and calculation complexity.

In computer vision, convolutional neural network (CNN) is widely used in recognition tasks [6]. The CNN is a feedforward network invariant to displacement and distortion, and therefore good at recognizing numbers. Deep convolutional neural network (DCNN) contains more layers and a more complex hierarchical structure than CNN [7]. This network needs to be trained with more data, and utilized with deep learning algorithms, which reduce the workload of manual operations. The feature learning ability of DCNN increases with the number of hidden layers. The application principle of DCNN in machine learning is the same as that of traditional CNN: convolution layers and max-pooling layers are arranged alternatively, followed by several fully-connected layers. The DCNNs with such a structure include AlexNet [8], ZFNet [9], VGGNet [10], GoogLeNet [11], and ResNet [12].

This paper presents a hybrid recognition framework for handwritten numbers on financial bills, which couples DCNN with SVM. The images containing handwritten numbers were selected from the bill number dataset of multiple financial companies, and the image features were extracted with the CNN. Then, the extracted numbers were recognized by the SVM. By contrast, most previous studies are based on manually extracted features. The manual feature extraction is a tedious process requiring human expertise, failing to optimize efficiency and recognition accuracy at the same time. The handling of irrelevant features might increase the computing load, and reduce the recognition performance. In this paper, the features are directly retrieved from the original images by non-manual method, which eliminates the need for collecting prior knowledge or details on feature design. In addition, the CNN ensures that the topology information in the input is invariant to basic transforms like rotation and translation, and outshines the conventional multilayer perceptron (MLP) model in the handling of complex problems. That is why the CNN was adopted to extract and segment the key points of handwritten numbers from different sources.

The main contributions of this paper are as follows:

(1) Considering the existing recognition methods for handwritten numbers on multi-domain financial bills, this paper puts forward a hybrid CNN-SVM framework, and improves the domain adaptability based on the multi-domain model sharing mechanism. The CNN was adopted to effectively extract the features from handwritten number images. These features were applied to SVM recognition, which helps to improve recognition accuracy, shorten operating time, and reduce computing complexity.

(2) The traditional method of manual feature extraction is inefficient and inaccurate in recognition. To solve the defects, this paper proposes a key point extraction and segmentation method for handwritten numbers on financial bills from various domains. The proposed method effectively simplifies the recognition of overlapping, broken, and joined numbers, despite the variation in personal writing style.

(3) The simulation results show that our method is 3% more accurate than the state-of-the-art models, and capable of accurately recognizing the financial bills from different domains. In addition, our method also boasts high stability and low complexity.

The rest of this paper is organized as follows: Section 2 reviews the relevant research in this field; Section 3 extracts and segments the key points from financial bills in different domains; Section 4 verifies our method through simulation; Section 5 summarizes the findings and discusses the future research.

2. LITERATURE REVIEW

In 1995, the SVM was applied to the optical character recognition (OCR) of handwritten numbers for the first time [13]. Since then, SVM classifiers have become the default choice for various supervised classification problems, namely, character recognition [14], face detection [15], and target recognition [16]. For example, Shareef and Altayar [17] studied the feature extraction of handwritten numbers with the SVM.

In recent years, the CNN has achieved good results in handwritten number recognition. For example, King et al. [18] recognized more than 98% of handwritten numbers in MNIST database with the CNN. Memon et al. [19] merged multiple CNNs into an integrated network, which achieved a high accuracy of 98.73%. Jiang and Learned-Miller [20] expanded the early 7-layer CNN into a 35-layer network, and increased the recognition accuracy to 99.70%. Torres et al. [21] developed a three-layer deep belief network for handwritten number recognition, based on the greedy algorithm. Using the CNN, Echegaray et al. [22] recognized the feature map on the bending directions of outdoor handwritten Chinese characters. Romanuke [23] designed a feature extractor for MNIST database based on the structure of the CNN.

The impressive performance of the CNN in Dayyeh's research [24] fully demonstrates the effectiveness of the network in feature extraction. Therefore, this paper aims to create a deep learning method for handwritten number recognition, drawing on the deep feature extraction ability of the CNN, and the structural risk minimization ability of the SVM. Both abilities have been proved in various fields [25, 26].

The fusion of CNN and SVM is very useful in handwritten number recognition. For example, Katib and Medhi [27] integrated CNN with SVM to recognize the numbers in MNIST database. Phinyomark et al. [28] combined hidden Markov model (HMM) with CNN into a hybrid model to recognize the house numbers from street view images. Wan et al. [29] proved that the CNN-SVM hybrid model greatly improves the recognition accuracy of handwritten numbers.

3. EXTRACTION AND SEGMENTATION OF MULTI-DOMAIN BILLS

3.1 Extraction and segmentation of bill images

As mentioned in the Introduction, our handwritings often contain lots of ligatures. This leaves many redundant strokes between characters, making it difficult to segment bill characters. Therefore, this paper puts forward a novel way to extract and segment characters:

Step 1. For a particular type of bill, the amount and the identity (ID) number should be filled into fixed areas with fixed sizes. Based on prior knowledge, two relatively large areas were extracted corresponding to the two information.

Step 2. The extracted image areas were binarized.

Step 3. The grid size was determined based on the physical information of the bill. Then, horizontal and vertical projections were implemented to find the boundaries of the grids for the amount area and the ID number area.

Step 4. The maximum area of each character was estimated in turn. During handwriting, the maximum area of each character is slightly larger than the original grid. Within the maximum area, a vertical projection value was calculated. The width and value of the vertical projection determine whether the grid is empty or filled with a character. If it is empty, the grid was not processed.

In general, the first grid in a character area is filled with the currency symbol ¥. For any other grid in that area, the grid was defined as the point with the largest vertical projection value, and the middle black block was found on the corresponding vertical projection line. Then, the black block was treated as the connected area of the character at the starting point, using the template A in Figure 1.

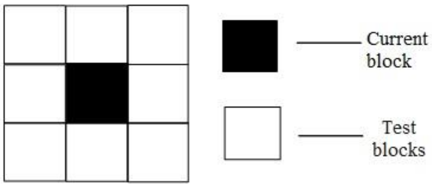


Figure 1. Template A

In Figure 1, the middle black block is the current color block, which is a part of the character. Whether the current point has any adjacent black block could be judged by the gray dots around. The adjacent black blocks were found and recorded. Then, each adjacent black block was taken as the center to find its adjacent black blocks. The search was terminated when no more adjacent black block could be found. In this way, the maximum area of the estimated character was determined in the connected area, and the image on the corresponding connected area was obtained, solving the difficulty in recognizing overlapping characters.

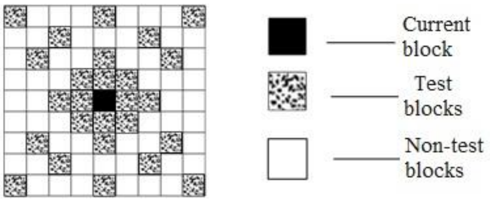
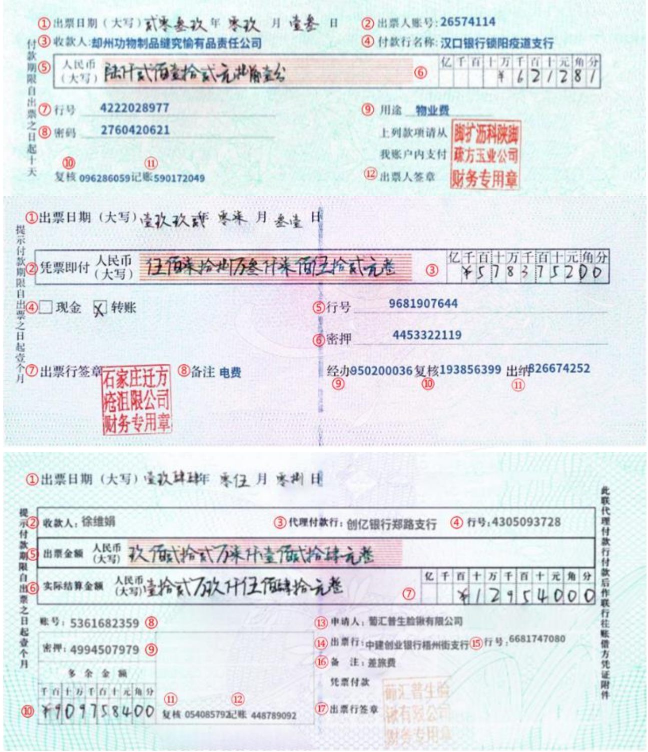


Figure 2. Template B

However, template A could not easily segment broken characters. Thus, template B in Figure 2 was introduced to discover the black blocks that are not adjacent to the current block, but belong to the same character as the latter. This template can effectively segment broken characters.

Step 5. The joined characters were segmented. The characters were divided into the minimum vertical projection values at the horizontal position, based on the vertical projection of the common part of the characters. Figures 3 and 4 provide the example financial bills, and the binarized and segmented images on the amount area, respectively.



①出票日期（大写）贰零叁玖 年 零玖 月 壹叁 日 ②出票人账号:26574114
③收款人:却州功物制品缝究愉有品责任公司 ④付款行名称:汉口银行锁阳疫道支行
⑤人民币（大写）陆什贰佰壹拾贰元捌角壹分 ⑥ 亿 千 百 十 万 千 百 十 元 角 分 ¥ 6 2 1 2 8 1
⑦行号 4222028977 ⑨用途 物业费
⑧密码 2760420621 上列款项请从 我账户内支付
⑩ ⑪ 复核 096286059 记账 590172049 ⑫出票人签章
脚扩沥科陕脚 疎方玉业公司 财务专用章
付款期限自出票之日起十天

①出票日期（大写）壹玖玖贰年 零柒 月 叁壹 日
②凭票即付 人民币（大写）伍佰柒拾捌万叁仟柒佰伍拾贰元整 ③ 亿 千 百 十 万 千 百 十 元 角 分 ¥ 5 7 8 3 7 5 2 0 0
④□现金 ☒转账 ⑤行号 9681907644
⑥密押 4453322119
⑦出票行签章 ⑧备注 电费 经办950200036 复核193856399 出纳826674252 ⑨ ⑩ ⑪
石家庄迁方疮沮限公司 财务专用章
提示付款期限自出票之日起壹个月

①出票日期（大写）壹玖肆肆年 零伍 月 零捌 日
②收款人：徐维娟 ③代理付款行：创亿银行郑路支行 ④行号：4305093728
⑤出票金额 人民币（大写）玖佰贰拾贰万柒仟壹佰贰拾肆元整
⑥实际结算金额 人民币（大写）壹拾贰万玖仟伍佰肆拾元整 ⑦ 亿 千 百 十 万 千 百 十 元 角 分 ¥ 1 2 9 5 4 0 0 0
账号：5361682359 ⑧ ⑬申请人：蓟汇普生脍做有限公司
密押：4994507979 ⑨ ⑭出票行：中建创业银行梧州街支行 ⑮行号：6681747080
多余金额 ⑯备 注：差旅费
千 百 十 万 千 百 十 元 角 分 凭票付款
⑩ ¥ 1 0 9 7 5 8 4 0 0 ⑪ 复核 054085792 ⑫ 记账 448789092 ⑰出票行签章
蓟汇普生脍 诚有限公司 财务专用章
提示付款期限自出票之日起壹个月
此联代理付款行付款后作联行往账借方凭证附件

Figure 3. Original images on financial bills

Note: ①Date of issuance (handwritten in the format of Y/M/D); ② Number of issuer account (printed); ③ Payee (printed); ④ Password (printed); ⑤ RMB (handwritten); ⑥ Reviewer (printed); ⑦ Line number (printed); ⑧ Name of payment bank (printed); ⑨ Purpose (printed); ⑩ Bookkeeper (printed) ⑪ Signature of the issuer (seal)

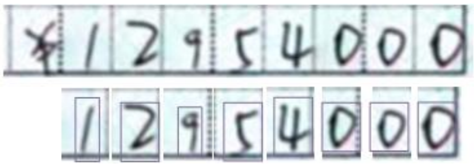


Figure 4. Binarized and segmented images on amount area

3.2 Feature extraction

Feature preprocessing of data helps to improve the learning of neural network, enhancing its ability to differentiate between classes of data. Normally, feature preprocessing reduces the computing complexity by extracting features through transforms (e.g., Z-transform), or dimensionality reduction, namely, principal component analysis (PCA) and T-distributed Stochastic Neighbor Embedding (tSNE), to convert high-dimensional data to low-dimensional feature space, while retaining the key information [30].

Since the numbers on financial bills are handwritten, this paper extracts features based on both statistical features and character structure. Specifically, each selected bill image was covered with k grids. The value of k determines the effect of the CNN on the bill image. The larger the k value, the greater the number of model parameters, and the more the extracted features. After k was determined, the features were constructed as per k value and character features. Let (f_l) be the feature of the number of pixels belonging to k as a proportion of the entire character. Then, we have:

$$f_{i,1} = \frac{ni}{N} \tag{1}$$

where, the subscript is the specific number of k; N is the number of character features in the grid; f_l is the number of pixels in $f_{i,1} = \frac{ni}{N}$ grids. Hence, the numbers in the bill image were divided into individual key points by the grids. Then, the eigenvector of each grid was extracted for model learning. Here, two features are adopted to represent the fitting vector *i*, *ni*:

$$f_{i,2} = \frac{2b_i}{1+b_i^2} \quad f_{i,3} = \frac{1-b_i^2}{1+b_i^2} \tag{2}$$

where, $f_{i,2} = \frac{2b_i}{1+b_i^2}$ is the eigenvector of the *bi*-th $f_{i,3} = \frac{1-b_i^2}{1+b_i^2}$ grid. The angle between the two features was set to θ=0, π/5, 3π/10, 2π/5, 3π/5, 7π/10, 4π/5, 9π/10. The corresponding bi values were 0, 0.3249, 0.7265, 1.3764, 3.0777, -1.3764, -0.7265, and -0.3249.

To verify its effectiveness, our feature extraction method was applied to 2,000 financial bills. The 28×28 original single-character images were processed by 16×7 grids, producing 16×3=48 eigenvalues. Then, the 60k features of the 2,000 bills were reduced to 3k.

4. HYBRID CLASSIFIER

This section describes the process of recognizing handwritten numbers in financial bills with the DCNN-SVM hybrid classifier, which combines the merits of SVM and neural network in classification.

The CNN, composed of multiple fully-connected layers, implements a supervised learning mechanism, and works similarly to humans. The network can learn local invariant features excellently, and extract the most discriminative information from the original number images. In the CNN, the output of each layer serves as the input of the next layer. Effective sub-regions could be calculated from the original number image, using the receptive fields of the CNN.

The SVM separates different types of data elements with a hyperplane, thereby representing multidimensional datasets. The SVM classifier could minimize the generalization error of unknown data, and perform well on multivariate classification. Its performance is not greatly affected by the noises in the data [31].

4.1 Single-channel DCNN

The proposed single-channel DCNN has two network streams: S_{cl} and S_{com} (Figure 5). The former helps to identify the region of the object; the latter ensures the discovery of all these regions.

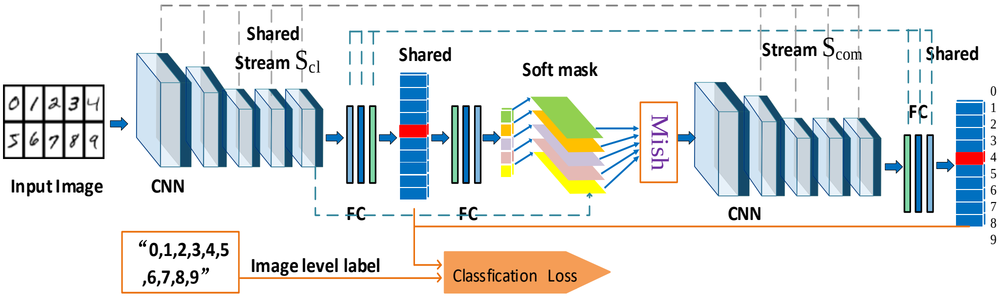


Figure 5. Single-channel DCNN

For the first network flow, the CNN contains classic optimization techniques like pooling and dropout. The network is followed by three fully-connected layers, which were connected to the loss function. The parameters of the three layers were shared with the following neural networks.

The soft mask [32] was adopted to configure the importance of each parameter. This function reflects the regions on the input image that support network prediction, and monitors the attention during network training for interesting tasks. Thus, the network prediction will focus on the region of interest. This is realized by training the soft mask in an end-to-end manner. In addition, the soft mask was optimized by the latest Mish activation function, before being imported to the CNN of the second network flow.

Besides, the idea of ResNet was referred to prevent vanishing gradients. The parameters of the CNN of the first network flow were shared with that of the second network flow, e.g., S_{cl} in Figure 5. The sharing makes the soft mask more complete, accurate, and suitable for segmentation. For a given image I, the activation function of unit k on level l in S_{cl} is denoted as $f_{l,k}$. Then, the gradient of the classification probability for a class c from the ground truth label was calculated relative to the activation map $f_{l,k}$. The calculated gradients were diverted back to the global average pooling layer to obtain the importance weight of each node $\omega_{l,k}^{c}$:

$$\omega_{l,k}^{c} = GAP\left(\frac{\partial s^{c}}{\partial f_{l,k}}\right) \tag{3}$$

where, GAP(·) is the global average pooling. After the end of parameter update, backpropagation was performed to obtain $\omega_{l,k}^{c}$, which indicates the importance of activation map $f_{l,k}$ in support of the prediction of class c. Taking weight matrix ω^{c}as the kernel, two-dimensional (2D) convolution was applied on the activation map matrix f_l to integrate all activation maps. After that, Mish operation [33] was conducted to obtain the soft mask A^{C} that supports online training:

$$A^{C} = Mish(conv(f_l, \omega^{c})) \tag{4}$$

where, f_l is the last convolution. The trainable A^{C} was used to generate the soft mask for the original input image, revealing the spatial details of the image features in advanced semantics. Then, the regions I^{*c} in class c that interest the network can be obtained by:

$$I^{*c} = I - (T(A^{c}) \odot I) \tag{5}$$

where, $\odot$ is the element-wise multiplication; $T(A^{c})$ is a masking function based on threshold operation. To make it differentiable, $T(A^{c})$ value was approximated by sigmoid function:

$$T(A^{c}) = \frac{1}{1+exp(-\omega(A^{c}-\sigma))} \tag{6}$$

where, σ is a threshold matrix whose elements are all equal to σ; ω is the scale parameter ensuring that $T(A^{c})_{i,j}$ is approximately 1, when $(A^{c})_{i,j}$ is greater than σ or 0.

Then, the class prediction probability was obtained with I^{*c} as the input to S_{com}. Our goal is to guide the network to focus on the regions of interest. Therefore, the I^{*c} operation should cover as few features in the target class as possible, that is, the areas other than the high response area in the soft mask region must not contain single pixels that trigger the recognition of class c targets. From the angle of loss function, this is equivalent to minimizing the prediction probability for I^{*c} to belong to class c. To realize this goal, the loss function was designed as:

$$L_{\mathrm{am}} = \frac{1}{n}\sum_{c} s^{c}(I^{*c}) \tag{7}$$

where, $s^{c}(I^{*c})$ is the prediction probability for I^{*c} to belong to class c; n is the number of ground truth labels of image I. The final global loss function L_{self} was established by adding up the classification losses L_{cl} and L_{am}:

$$L_{\mathrm{self}} = L_{\mathrm{cl}} + \alpha L_{\mathrm{am}} \tag{8}$$

where, L_{cl} is the classification loss of multiple labels and classes (in this paper, this loss is characterized by multi-label soft margin loss); α is the weighting parameter (for all experiments, α=1). Under the guidance of L_{self}, the network learning expands the region of interest that benefit target classification, so that soft mask could adapt to the task of semantic segmentation.

4.2 CNN-SVM

In the above subsection, the network is enabled to evaluate the importance of each parameter. On this basis, this subsection makes it possible for the network to adapt to the task of interest by controlling the extra supervised learning with a small soft mask.

Following the thought of imposing additional supervision on the attention map, the SVM was extended to seamlessly integrate additional supervision into the weakly supervised learning framework. Then, the weakly supervised semantic segmentation was improved under the self-guided gain framework. Under the test data and training data of different distributions, the CNN-SVM was applied to guide the network learning, such that the network could robustly handle the bias in the dataset, and acquire strong generalization ability.

Drawing on the advantages of the CNN in subsection 4.1, the authors set up mixed channels and adopted integrated learning to recognize handwritten numbers of different types, writing rules, or angles. The multi-channel CNN-SVM hybrid classifier is illustrated in Figure 6.

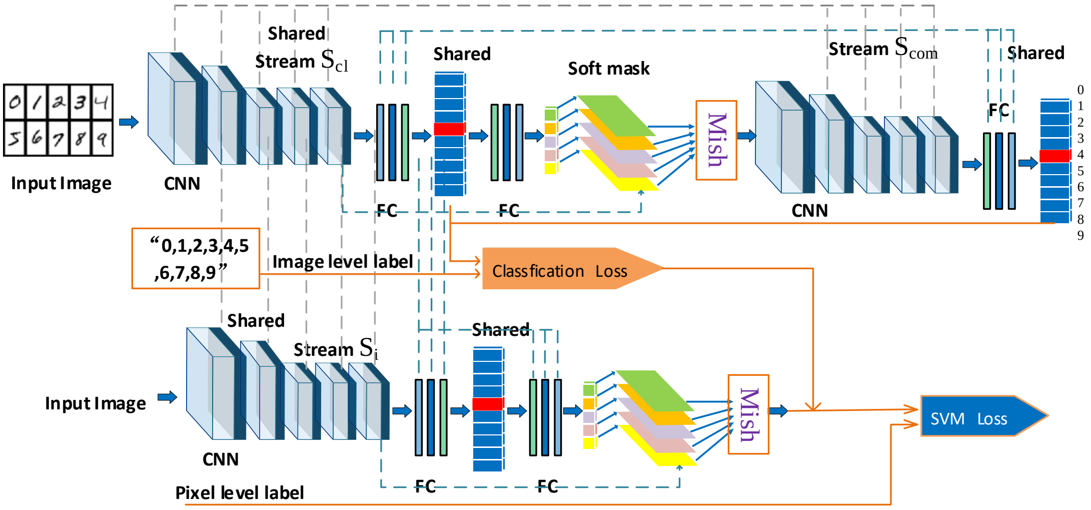


Figure 6. Multi-channel CNN-SVM hybrid classifier

As shown in Figure 6, the proposed model is deep and wide. Considering the difficulty in recognizing the overlapped, broken, and joined numbers in various financial bills with diverse writing styles, the model parameters in different channels were shared to enhance the adaptability of the entire classifier. The features were effectively extracted by the CNN from each handwritten number image, and imported to SVM for number recognition at a high accuracy, in a short time, and with a low computing complexity. Apart from L_{cl} and L_{am}, another objective function was designed: the SVM loss under given external supervision L_e:

$$L_e = \frac{1}{n}\sum_c (A^c - H^c)^2 \tag{9}$$

where, H^c is the additional supervision, such as the pixel-level segmentation mask in Figure 4. Since it takes an excessively long time to generate pixel-level segmentation maps, this paper intends to find a classification method using a few data under external supervision. The model framework in Figure 6 could serve this purpose. As shown in Figure 6, the two network flows share all the parameters with the SVM. The inputs to the SVM include image-level labels and pixel-level segmentation masks. Using the SVM, the classifier performance was improved with a few pixel-level labels. The final loss function L_{ext} of our method was defined as:

$$L_{\text{ext}} = L_{\text{cl}} + \alpha L_{\text{am}} + \omega L_e \tag{10}$$

where, L_{cl} and L_{am} are of the same meanings as in Subsection 4.1; ω is the weighting parameter depending on the desire for extra supervision (for all experiments, ω=10).

Our model can be easily modified to suit other tasks. Once the activation map $f_{l,k}$ corresponding to the final output of the network is obtained, the L_e can be used to guide the network to learn the key regions in the task of interest. The multi-channel design ensures that the network could robustly handle the bias in the dataset, and acquire strong generalization. In this case, the extra supervision of the SVM plays an auxiliary role to the CNN.

5. EXPERIMENTS

5.1 Experimental setup

Our method was compared with benchmark methods like AlexNet, ResNet, and CNN. The neural networks were modeled under PyTorch, and deployed on a server operating on CentOS-7, using two Tesla V100 graphics cards (32G), and a central processing unit (CPU) of 2.20GHz.

5.2 Dataset

The experiments were conducted on an open dataset (https://mp.weixin.qq.com/s/3d9sQlRKpr7TqO_7iytqwg). The original data were divided randomly into 12,000 numbers.

The first 10,000 were adopted for training, and the latter 2,000 for testing.

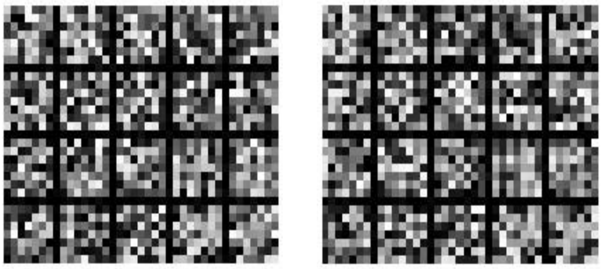
To prevent the forgetting phenomenon and domain transfer, this paper adopts a batch training method that mixes the financial bills from different domains. Each time, the eigenvectors of ten number samples from 0 to 9 were read, and imported to the network; then, the difference between the actual output and the expected value was taken as the objective function for weight adjustment.

5.3 Results

Table 1. Recognition accuracy of each method

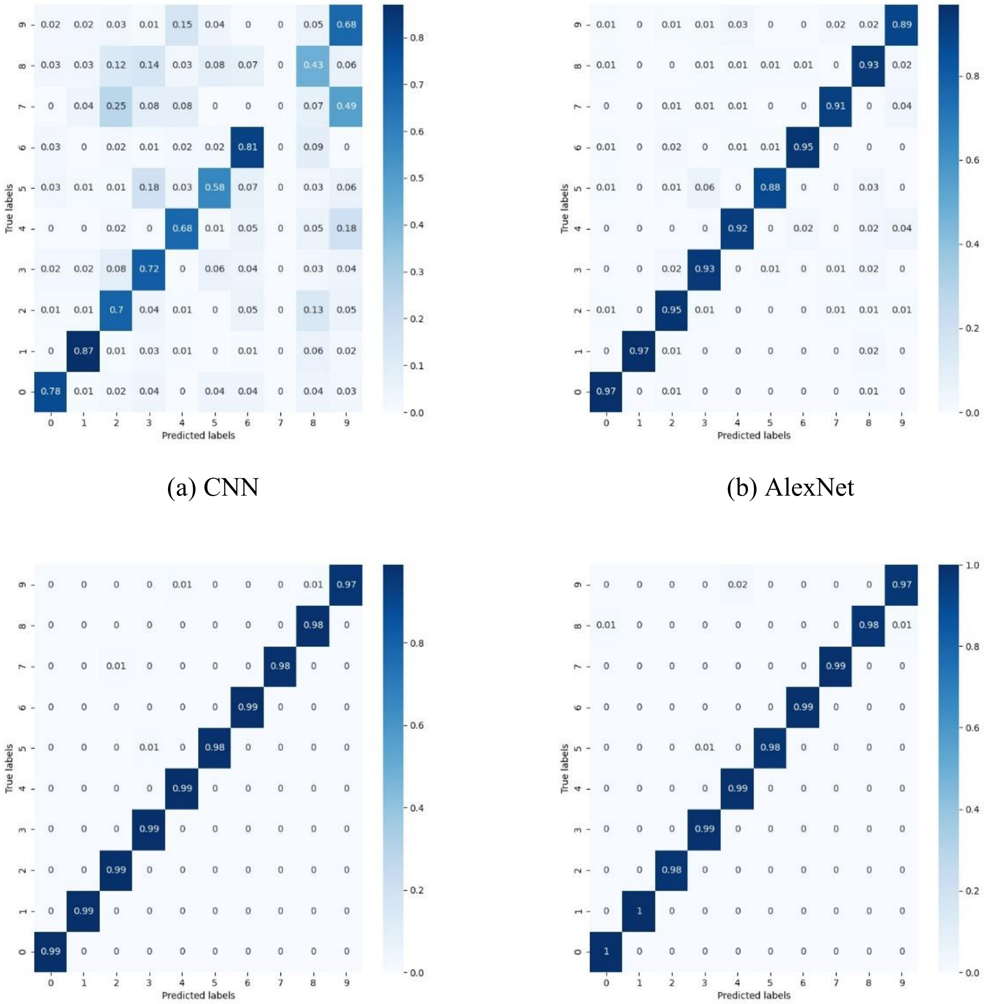
Method	Training set	Test set
CNN	96.23%	94.56%
AlexNet	97.45%	96.58%
ResNet	97.56%	96.23%
DCNN	99.21%	97.73%

Table 1 compares the recognition accuracies of the four method. It can be seen that our method surpassed the benchmark methods by 3% in the recognition accuracy of handwritten numbers on various financial bills. Even the best benchmark method was outshined by our DCNN, because the comparative methods cannot grasp the details on the bills from various domains. Meanwhile, our model could learn the pixel-level features of the original images, and preprocesses the features of handwritten numbers through segmentation and extraction. To further clarify the ability of our method in grasping pixel-level features, the convolutional kernels in two channels of the first network flow were visualized (channel 1 from Figure 7(a), and channel 2 from Figure 7 (b). As shown in Figure 7, no obvious patterns of similar local structure were observed, which is no surprise. After all, our method aims to effectively learn images from different domains, rather than the visually attractive features commonly pursued by other models. The superiority of our method stems from the grasping of image differences from varied domains.



(a) Channel 1 (b) Channel 2

Figure 7. Convolutional kernels in two channels of the first network flow



(a) CNN

(b) AlexNet

(c) ResNet

(d) Our method

Figure 8. Confusion matrices on the test set

The confusion matrices of CNN, AlexNet, ResNet, and our method on the test set are compared in Figure 8, where the diagonals represent correct classification. Obviously, our method has the best ability to correctly classify numbers in the confusion matrix. In the confusion matrix of CNN (Figure 8(a)), 5, 6, and 9 were mistaken as 0. In the confusion matrix of AlexNet (Figure 8(b)), 9 and 7 were confused 44 times; the two numbers had the lowest probability of correct classification. This is understandable: the two numbers are so similar that their handwritten versions are easily misidentified. In the confusion matrix of our model, 3 and 7 were both recognized without any mistake; the other numbers were recognized more accurately than the other methods. The advantage of our method comes from the optimization of data and model, which is not done in any baseline method.

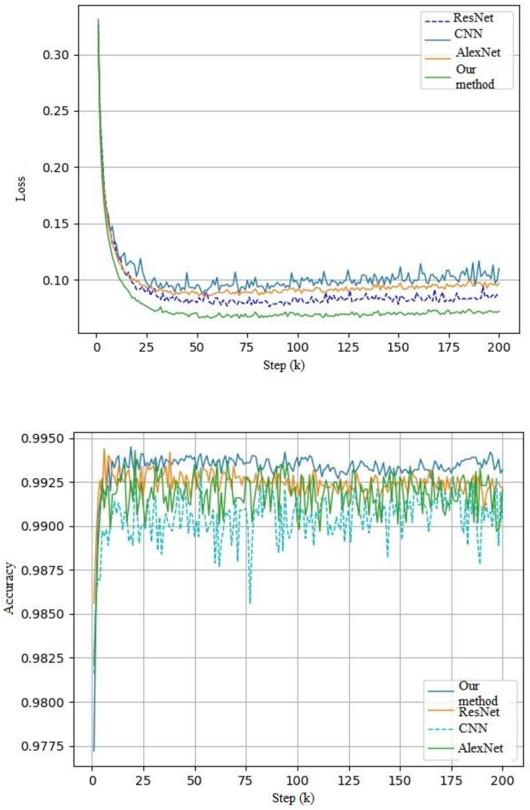


Figure 9. Training losses and accuracies of different methods

As shown in Figure 9, our method converged faster and remained more stable in training than all the baseline methods. Our method completed training after 27k steps, while the latest AlexNet needed 32k steps. In addition, our accuracy increased linearly, and stayed ahead of all methods. The slight oscillations were within the allowable range. The fast convergence and stability of our method are attributable to the following facts: the multi-channel design facilitates adaptive learning, while the parameter sharing between two network flows solves the problem of multiple domains, stabilizes model training, and prevents vanishing gradients.

6. CONCLUSIONS

To recognize the handwritten numbers in financial bills from varied domains, this paper presents a DCNN-SVM hybrid classifier, which combines the merits of SVM and neural network. Multiple fully-connected layers were arranged in the CNN, such that the network could learn local invariant features, and extract the most discriminative information from the original number images. Meanwhile, the SVM was coupled with the extracted features to make better classification of numbers. Experimental results show that our method far exceed the benchmark methods in accuracy. In future, the proposed model will be applied to recognize many other handwritten characters, e.g., those in the MNIST database, and various scenarios, e.g., the recognition of different languages. The authors will further improve the model to promote its universality and stability facing different domains or modes.

ACKNOWLEDGMENT

This paper is made possible thanks to the generous support from the Postdoctoral Special Funding Project in Heilongjiang Province (Grant No.: LBH-Z20182).

REFERENCES

[1] Cui, J., Yu, H., Chen, S., Chen, Y., Liu, H. (2019). Simultaneous estimation and segmentation from projection data in dynamic PET. Medical Physics, 46(3): 1245-1259. https://doi.org/10.1002/mp.13364

[2] Huang, Y., Sun, X., Lu, M., Xu, M. (2015). Channel-max, channel-drop and stochastic max-pooling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 9-17.

[3] Uskul, A.K., Paulmann, S., Weick, M. (2016). Social power and recognition of emotional prosody: High power is associated with lower recognition accuracy than low power. Emotion, 16(1): 11-15. https://doi.org/10.1037/emo0000110

[4] Heisig, J.P., Schaeffer, M., Giesecke, J. (2017). The costs of simplicity: Why multilevel models may benefit from accounting for cross-cluster differences in the effects of controls. American Sociological Review, 82(4): 796-827. https://doi.org/10.1177/0003122417717901

[5] Huguet, G., de la Llave, R., Sire, Y. (2010). Computation of whiskered invariant tori and their associated manifolds: new fast algorithms. arXiv preprint arXiv:1004.5231.

[6] Nguyen, S.D., Nguyen, Q.H., Choi, S.B. (2015). Hybrid clustering based fuzzy structure for vibration control–Part 1: A novel algorithm for building neuro-fuzzy system. Mechanical Systems and Signal Processing, 50-51: 510-525. https://doi.org/10.1016/j.ymssp.2014.04.021

[7] Ahlawat, S., Choudhary, A. (2020). Hybrid CNN-SVM classifier for handwritten digit recognition. Procedia Computer Science, 167: 2554-2560. https://doi.org/10.1016/j.procs.2020.03.309

[8] Hu, W., Huang, Y., Wei, L., Zhang, F., Li, H. (2015). Deep convolutional neural networks for hyperspectral image classification. Journal of Sensors, 2015: Article ID 258619. https://doi.org/10.1155/2015/258619

[9] Zhang, K., Zhang, Z., Li, Z., Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10): 1499-1503. https://doi.org/10.1109/LSP.2016.2603342

[10] Yan, Z., Zhang, H., Piramuthu, R., Jagadeesh, V., DeCoste, D., Di, W., Yu, Y. (2015). HD-CNN: hierarchical deep convolutional neural networks for large scale visual recognition. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2740-2748.

[11] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90. https://doi.org/10.1145/3065386

[12] Matthew, D., Fergus, R. (2014). Visualizing and understanding convolutional neural networks. In Proceedings of the 13th European Conference Computer Vision and Pattern Recognition, Zurich, Switzerland, pp. 6-12.

[13] Ke, H., Chen, D., Li, X., Tang, Y., Shah, T., Ranjan, R. (2018). Towards brain big data classification: Epileptic EEG identification with a lightweight VGGNet on global MIC. IEEE Access, 6: 14722-14733. https://doi.org/10.1109/ACCESS.2018.2810882

[14] Sam, S.M., Kamardin, K., Sjarif, N.N.A., Mohamed, N. (2019). Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogLeNet Inception-v1 and Inception-v3. Procedia Computer Science, 161: 475-483. https://doi.org/10.1016/j.procs.2019.11.147

[15] He, K., Zhang, X., Ren, S., Sun, J. (2016). Identity mappings in deep residual networks. In European Conference on Computer Vision, 9908: 630-645. https://doi.org/10.1007/978-3-319-46493-0_38

[16] Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132-7141.

[17] Shareef, A.Q., Altayar, S.M. (2015). OCR-ANN back-propagation based classifier. International Journal of Computer Science and Mobile Computing, 307-314.

[18] King, M.T., Grover, D.L., Kushler, C.A., Stafford-Fraser, J.Q. (2015). U.S. Patent No. 9,008,447. Washington, DC: U.S. Patent and Trademark Office.

[19] Memon, J., Sami, M., Khan, R.A., Uddin, M. (2020). Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). IEEE Access, 8: 142642-142668. https://doi.org/10.1109/ACCESS.2020.3012542

[20] Jiang, H., Learned-Miller, E. (2017). Face detection with the faster R-CNN. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 650-657. https://doi.org/10.1109/FG.2017.82

[21] Torres, G., Jaime, K., Ramos, F., Garcia, G. (2011). Brain architecture for visual object identification. In IEEE 10th International Conference on Cognitive Informatics and Cognitive Computing (ICCI-CC'11), pp. 33-40. https://doi.org/10.1109/COGINF.2011.6016119

[22] Echegaray, S., Nair, V., Kadoch, M., Leung, A., Rubin, D., Gevaert, O., Napel, S. (2016). A rapid segmentation-insensitive "digital biopsy" method for radiomic feature extraction: method and pilot study using CT images of non–small cell lung cancer. Tomography, 2(4): 283-294. https://doi.org/10.18383/j.tom.2016.00163

[23] Romanuke, V.V. (2016). Training data expansion and boosting of convolutional neural networks for reducing the MNIST dataset error rate. Наукові вісті Національного технічного університету України Київський політехнічний інститут, (6): 29-34. http://nbuv.gov.ua/UJRN/NVKPI_2016_6_6

[24] Dayyeh, B.K.A., Thosani, N., Konda, V., Wallace, M.B., Rex, D.K., Chauhan, S.S., ASGE Technology Committee. (2015). ASGE Technology Committee systematic review and meta-analysis assessing the ASGE PIVI thresholds for adopting real-time endoscopic assessment of the histology of diminutive colorectal polyps. Gastrointestinal Endoscopy, 81(3): 502-e1-502-e16. https://doi.org/10.1016/j.gie.2014.12.022

[25] Yuan, F., Zhang, L., Wan, B., Xia, X., Shi, J. (2019). Convolutional neural networks based on multi-scale additive merging layers for visual smoke recognition. Machine Vision and Applications, 30(2): 345-358. https://doi.org/10.1007/s00138-018-0990-3

[26] Giryes, R. (2016). A greedy algorithm for the analysis transform domain. Neurocomputing, 173: 278-289. https://doi.org/10.1016/j.neucom.2015.02.100

[27] Katib, I., Medhi, D. (2013). Network protection design models, a heuristic, and a study for concurrent single-link per layer failures in three-layer networks. Computer Communications, 36(6): 678-688. https://doi.org/10.1016/j.comcom.2012.09.008

[28] Phinyomark, A., Khushaba, R.N., Ibáñez-Marcelo, E., Patania, A., Scheme, E., Petri, G. (2017). Navigating features: a topologically informed chart of electromyographic features space. Journal of The Royal Society Interface, 14(137): 20170734. https://doi.org/10.1098/rsif.2017.0734

[29] Wan, W., Zhong, Y., Li, T., Chen, J. (2018). Rethinking feature distribution for loss functions in image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9117-9126.

[30] Shima, Y., Nakashima, Y., Yasuda, M. (2017). Pattern augmentation for handwritten digit classification based on combination of pre-trained CNN and SVM. In 2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT), pp. 1-6. https://doi.org/10.1109/ICIEV.2017.8338575

[31] Guo, Q., Wang, F., Lei, J., Tu, D., Li, G. (2016). Convolutional feature learning and Hybrid CNN-HMM for scene number recognition. Neurocomputing, 184: 78-90. https://doi.org/10.1016/j.neucom.2015.07.135

[32] Li, Z., Wang, S.H., Fan, R.R., Cao, G., Zhang, Y.D., Guo, T. (2019). Teeth category classification via seven - layer deep convolutional neural network with max pooling and global average pooling. International Journal of Imaging Systems and Technology, 29(4): 577-583. https://doi.org/10.1002/ima.22337

[33] Zhan, Y., Wang, J., Shi, J., Cheng, G., Yao, L., Sun, W. (2017). Distinguishing cloud and snow in satellite images via deep convolutional network. IEEE Geoscience and Remote Sensing Letters, 14(10): 1785-1789.