IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 10, 2018, accepted June 8, 2018, date of publication June 19, 2018, date of current version July 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2848930

Consecutive Convolutional Activations for Scene Character Recognition

ZHONG ZHANG[1], (Member, IEEE), HONG WANG[1], SHUANG LIU[1], (Member, IEEE), AND BAIHUA XIAO[2]

[1]Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
[2]The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Corresponding author: Zhong Zhang (zhong.zhang8848@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61501327 and Grant 61711530240, in part by the Natural Science Foundation of Tianjin under Grant 17JCZDJC30600 and Grant 15JCQNJC01700, in part by the Fund of Tianjin Normal University under Grant 135202RC1703, in part by the Open Projects Program of National Laboratory of Pattern Recognition under Grant 201700001 and Grant 201800002, in part by the China Scholarship Council under Grant 201708120039 and Grant 201708120040, and in part by the Tianjin Higher Education Creative Team Funds Program.

ABSTRACT Driven by the rapid growth of communication technologies and the wide applications of intelligent mobile terminals, the scene character recognition has become a significant yet very challenging task in people's lives. In this paper, we design a novel feature representation scheme termed consecutive convolutional activations (CCA) for character recognition in natural scenes. The proposed CCA could integrate both the low-level and the high-level patterns into the global decision by learning character representations from several successive convolutional layers. Concretely, one shallow convolutional layer is first selected for extracting the convolutional activation features, and then, the next consecutive deep convolutional layers are utilized to learn weight matrices for these convolutional activation features. Finally, the Fisher vectors are employed to encode the CCA features so as to obtain the image-level representations. Extensive experiments are conducted on two English scene character databases (ICDAR2003 and Chars74K) and one Chinese scene character database (''Pan+ChiPhoto''), and the experimental data indicate that the proposed method achieves a superior performance than the previous algorithms.

INDEX TERMS Consecutive convolutional activations, convolutional neural network, scene character recognition.

I. INTRODUCTION

Characters, as the basic medium for image communication, are ubiquitous in images and provide valuable semantic cues for various applications like automatic geocoding [1], product search [2], robot navigation [3], and image and video indexing [4]–[6]. Scene characters, as the term suggests, are the characters extracted from scene images, and they are easily disturbed by a variety of factors, such as non-uniform illumination, complex background, font distortion, blur, various fonts, etc. Hence, accurately recognizing scene characters is a particularly challenging task. In the past of decades, scene character recognition has been a hot research focus and many interesting scene character recognition algorithms [7]–[12] have been proposed.

The design of a scene character recognition system mainly involves two parts. Firstly, the feature representation devotes to learning discriminative feature vectors for character images using different kinds of descriptors, feature coding methods and pooling methods. Secondly, discriminative feature vectors are transmitted into a classifier to obtain recognition results. Early optical character recognition (OCR) based methods [13], [14] first performed binarization for the input image and then the binarized image was delivered to the OCR engine. The OCR based methods have achieved great success in the task of scanned document recognition, however, for the scene characters, these methods are inapplicable. Followed by the OCR based methods, the object recognition based methods are introduced to recognize scene characters. De Campos *et al.* [8] assessed the character classification performance by using various features (shape contexts (SC) [15], geometric blur (GB) [16], scale invariant feature transform (SIFT) [17], spin image [18], maximum response of filters (MR8) [19] and patch descriptor (PCH) [20]), and classifiers (nearest neighbor (NN)

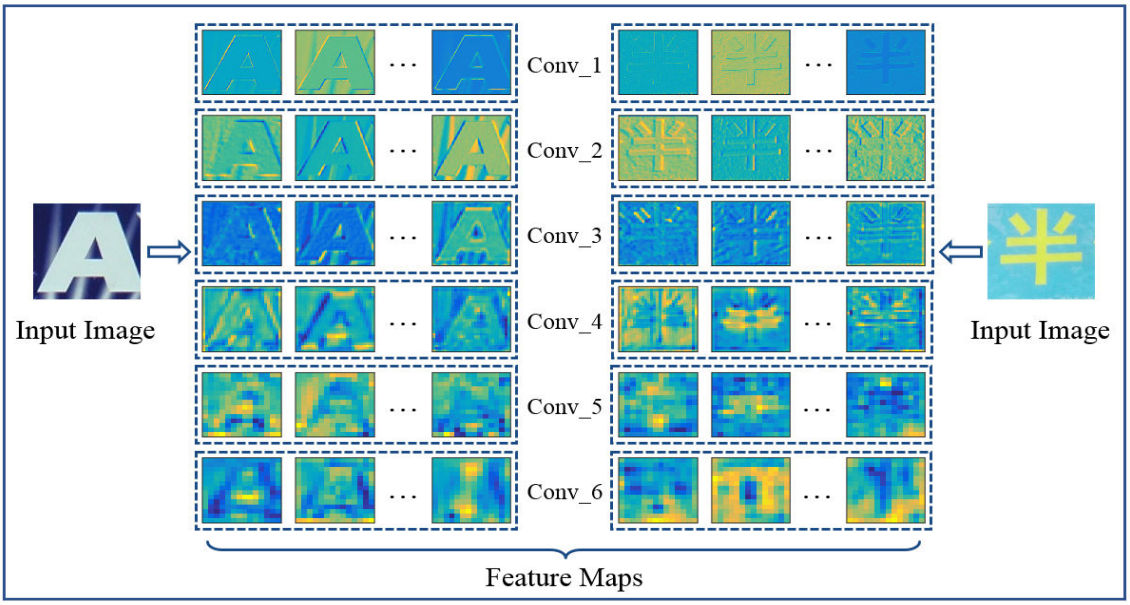

FIGURE 1. Visualization of the feature maps in different convolutional layers.

and support vector machine (SVM)). When the SC and GB in conjunction with the NN classifier, better classification performance was achieved. An elegant probabilistic graphical model [21] was built to bring together the individual and similar character appearance as well the lexicons and language statistics in the character recognition process. In [22], histograms of oriented gradients (HOG) features were extracted to describe the characters in the wild. Furthermore, several variants of HOG [10], [23] were proposed to represent character images.

Since characters consist of strokes with certain structures, many researchers make attempts to incorporate the structure or stroke information into the global decision. We roughly divide these methods into part-based and stroke-based methods. As for part-based methods, the characters are first expressed as a group of parts in the training stage, and in the test stage, part structure matching is performed to obtain a category label for an image. For instance, Shi *et al*. [24], [25] made use of part-based tree structure to model each type of character where the local appearance and global structure information were captured. The part-based tree structure is utilized to characterize the co-occurrence and spatial relationship among features. However, the part-based tree structure is manually designed and the part model size is limited to a single scale, which indicates that the part-based method has its limits. The stroke-based methods incorporate the character strokes into the character recognition progress. In [26], stroke bank was established for scene character recognition. Specifically, the positive and negative training samples are first collected for training stroke detectors, and then the detectors' maximal outputs in the corresponding areas are regarded as features. The method achieved encouraging experimental results on two English character databases (ICDAR2003 [27] and Chars74K [8]). Inspired by the stroke bank, the discriminative multi-scale stroke detector based representation (DMSDR) [28] was proposed, in which both the multi-scale stroke detectors and the discriminative stroke detector selection strategy were utilized for final feature representation. To capture the co-occurrence among local strokes, the concept of spatiality embedded dictionary [28], [29] was proposed to incorporate more precise spatial contextual information. Lee *et al*. [30] proposed a mid-level feature pooling method named region-based discriminative feature pooling to integrate the low-level pixel-wise features, so that the distinctive spatial structures were able to effectively preserved for each individual character.

Recently, the convolutional neural network (CNN) has been successfully applied in speech recognition [31], brain electrical source analysis [32], [33], image classification [34], [35], object classification [36], character recognition [37]–[39] and so on. For character recognition, there are two main kinds of methods to learn visual representations from CNN model. The first one directly uses the output of fully-connected (FC) layer as the global character representation [40], [41]. The second kind of methods apply the convolutional activation features in the convolutional layer to describe character images. In [42] and [43], the final character representation was aggregated from convolutional activation features and these methods obtain significant gain. Zhang *et al*. [44] proposed a bilateral convolutional activations encoded with the Fisher vectors (BCA-FV) for scene character recognition, in which the bilateral convolutional activation map was injected into the Fisher vectors to encode convolutional activation descriptors. However, most existing approaches only utilize the activations in one convolutional layer to describe feature representations, which neglects the information provided by other layers. As shown in Figure 1, we visualize feature maps in different convolutional layers.

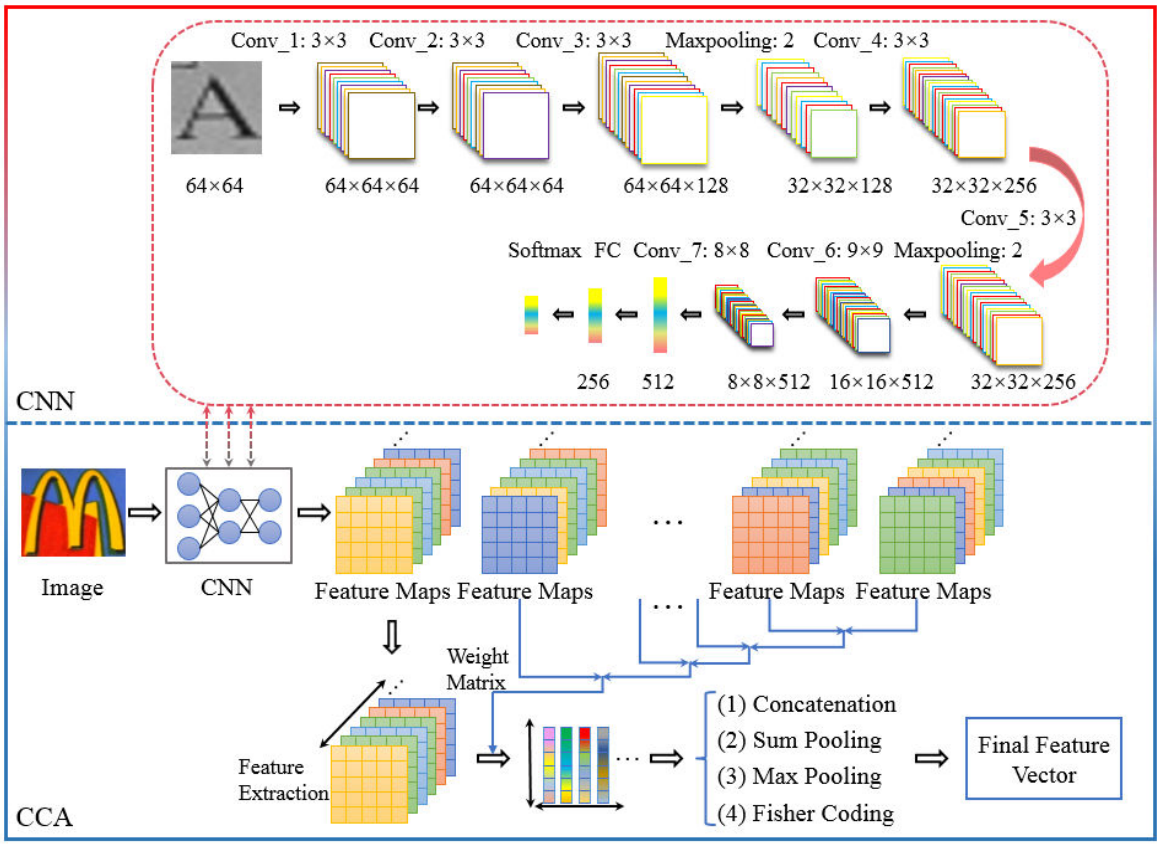

FIGURE 2. Visualization of the proposed CCA for scene character recognition.

From Figure 1, we can see that the convolutional activations in shallow convolutional layers reflect low-level patterns, such as the structural and textural information. As the layer goes deeper which approaches to the classification layer, the convolutional activations reflect high-level patterns.

Motivated by the above observations, in this paper, we design a novel feature representation scheme named consecutive convolutional activations (CCA) for character recognition in natural scenes. The proposed CCA could integrate several successive convolutional layers into character representations. Specifically, after training a CNN model, we first extract convolutional activation features from one shallow convolutional layer to absorb the structural and textural information. Meanwhile, the convolutional activations in the higher convolutional layers reflect high-level semantic information which is particularly significant for feature representations. Hence, we condense convolutional activations in consecutive deep convolutional layers as the weights of activations from the shallow convolutional layer. Finally, we utilize the learned weights to pool the extracted convolutional activation features so as to derive discriminative and powerful character representations. The contributions of the proposed CCA lie in: 1) In order to obtain the completed features, we learn weights from the consecutive convolutional layers; 2) The proposed CCA integrates low-level and high-level patterns into the global decision.

We conduct extensive experiments on two English scene character recognition databases (ICDAR2003 [27] and Chars74K [8]) and one Chinese scene character recognition database (“Pan+ChiPhoto” [10]). Experimental data indicates that our method obtains superior performance for scene character recognition.

The paper is organized as follows. The next section gives a detailed description of the proposed CCA for scene character recognition. Section III presents experimental results and analysis of the proposed CCA. Finally, we conclude the paper in Section IV.

II. APPROACH

In this section, we elaborate the proposed CCA. Firstly, we describe the architecture of the CNN used for scene character representations. Then, we introduce the proposed CCA in detail. Finally, we interpret how to obtain the final feature representation for scene characters.

A. NETWORK ARCHITECTURE

We train a CNN for feature representations and the configuration of the CNN is shown in Figure 2. The input of the CNN is a fixed-size 64×64 image and the subtracting mean value operation is performed for the input image. As for the first layer, we utilize 64 filters with a receptive field of 3×3 to convolve the input image, generating 64 feature maps

of size 64 × 64. The 64 feature maps are then convolved by 64 filters of size 3 × 3, generating feature maps of size 64 × 64 × 64. As for the third layer, the 64 feature maps are taken as the input and convolved by 64 filters of size 3 × 3. Then, the output of the third layer is transmitted into the max pooling strategy, obtaining feature maps of size 32 × 32 × 128. As for the next two layers, 256 filters with a receptive field of 3 × 3 are used, resulting in feature maps of size 32 × 32 × 256. After that, the max pooling strategy is employed. The sequence proceeds by convolving with 512 filters and the receptive fields of 9×9 and 8×8, respectively, resulting in feature maps of size $8 \times 8 \times 512$ and $1 \times 1 \times 512$. Afterwards, one FC layer, which is a 256 dimensional vector, is followed. Finally, we utilize the softmax strategy to convert the activations into character probabilities.

The max pooling is performed within a 2 × 2 window and the stride is fixed to 2 pixels. In addition, the appropriately zero-padded is carried out so that the resulting feature map is the same size as the input one. We apply the back propagation gradient-descent algorithm [45] for updating parameters, i.e., weights and biases. The size of each mini-batch is set to 64. The gradient-descent algorithm is terminated at 90 epochs. As for the first 60 and the remaining 30 epochs, the learning rates are empirically set to 0.001 and 0.0001, respectively.

B. CONSECUTIVE CONVOLUTIONAL ACTIVATIONS

In the convolutional layer, the filters traverse the input image in a manner of sliding-window. The top-left convolutional activations are produced by the top-left image region, while the bottom-right convolutional activations are generated by the bottom-right image region. Hence, the rich spatial information is embedded into the feature maps during the convolution process. The feature maps in one convolutional layer can be regarded as a set of N-dimensional convolutional activation features extracted from $W \times H$ positions. Each convolutional activation in a feature map depicts an image part and high activation values indicate salient parts. Hence, we extract convolutional activation features for character representations.

Typically, the shallow convolutional layers capture low-level patterns, such as specific structures and textures. These low-level patterns are essential for feature representations. While the higher layers, which approach to the category label, encode high-level patterns. Hence, for obtaining powerful and discriminative feature representations, we extract convolutional activation features from feature maps in one shallow convolutional layer and utilize the next consecutive deeper convolutional layers to learn weights for these features.

Note that we select the t-th convolutional layer as the shallow layer where the size of feature maps is $W_t \times H_t \times N_t$. As shown in Figure 3, we directly concatenate activation values in the position p of all feature maps in the t-th convolutional layer to form a feature vector, and denote it as the convolutional activation feature $x_t(p)$ which is a N_t dimensional feature vector. Then, we utilize feature maps

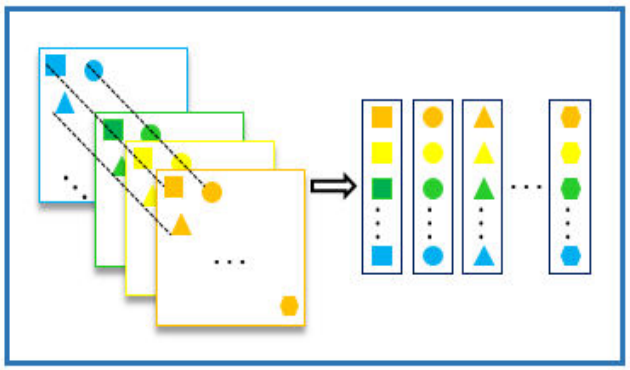

FIGURE 3. The process of extracting convolutional activation features.

in the $\{(t+1), (t+2), \cdots, (t+l), \cdots, (t+L-1)\}$-th convolutional layers to learn weight matrices. L is the total number of the selected feature maps. One feature map in a convolutional layer corresponds to one weight matrix. The j-th weight matrix in the $(t+L-2)$-th convolutional layer is formulated as:

$$S^j_{(t+L-2)} = \sum_i M^j_{(t+L-2)} \odot S^i_{(t+L-1)}, \tag{1}$$

where $\odot$ is the element-wise multiplication in two matrices, $M^j_{(t+L-2)}$ is the j-th feature map in the $(t+L-2)$-th convolutional layer and $S^i_{(t+L-1)}$ is the i-th weight matrix at the $(t+L-1)$-th convolutional layer. Note that, in Equation 1, the prerequisite is that the two matrices must be dimensional consensus. Hence, we employ the bilinear interpolation algorithm for filling the smaller matrix. In Equation 1, with a recursive method, the final weight matrices can be derived from initial weight matrices which follows the principle of:

$$S^j_{(t+L-1)} = M^j_{(t+L-1)}, \tag{2}$$

where $M^j_{(t+L-1)}$ is the j-th feature map in the $(t+L-1)$-th convolutional layer. According to Equation 1 and Equation 2, the $\{(t+L-2), \cdots, (t+l), \cdots, (t+2), (t+1)\}$-th weight matrices can be sequentially obtained. The weight matrices in the $(t+1)$-th convolutional layer, which contain the information from the $(t+1)$-th to the $(t+L-1)$-th convolutional layers, are the final weight matrices. Assume the size of feature maps in the $(t+1)$-th convolutional layer is $W_{(t+1)} \times H_{(t+1)} \times N_{(t+1)}$, the final weight matrices are of the same size as the feature maps in the $(t+1)$-th convolutional layer, i.e., $W_{(t+1)} \times H_{(t+1)} \times N_{(t+1)}$. The final weight matrices are used for encoding the extracted convolutional activation features.

Let f_j be the j-th CCA feature and it can be obtained using the following equation:

$$f_j = \sum_p x_t(p) S^j_{(t+1)}(p), \tag{3}$$

where $S^j_{(t+1)}(p)$ is the weight value at position p in the j-th weight matrix of the final weight matrices.

The dimensionality of f_j is N_t. As a result, each character image can be expressed by the CCA feature set F:

$$F = \{f_1, f_2, \cdots, f_j, \cdots, f_{N_{(t+1)}}\}, \tag{4}$$

where $N_{(t+1)}$ is the number of feature maps (weighted matrices) in the $(t+1)$-th convolutional layer.

C. IMAGE-LEVEL REPRESENTATION

We employ four approaches to these CCA features so as to obtain powerful and discriminative image-level representation. The detail is illustrated as follows:

(1) Concat+CCA: all the CCA features are directly concatenated into an image-level feature vector $F_{Concat+CCA}$:

$$F_{Concat+CCA} = (f_1, f_2, \cdots, f_j, \cdots, f_{N_{(t+1)}}). \tag{5}$$

In addition, the principal component analysis (PCA) is employed for dimensionality reduction.

(2) SP+CCA: we apply the sum pooling for all the CCA features and the final image-level representation F_{SP+CCA} is:

$$F_{SP+CCA} = \sum_j f_j. \tag{6}$$

(3) MP+CCA: we employ the max pooling for all the CCA features and the final image-level representation can be defined as:

$$F_{MP+CCA} = max\{f_1, f_2, \cdots, f_j, \cdots, f_{N_{(t+1)}}\}. \tag{7}$$

(4) FV+CCA: we utilize the Fisher Vectors [46], [47] to learn the high-order statistic information into the image-level representation. The N_t dimensional CCA feature derivatives with respect to the statistic parameters of k-th Gaussian mixture model (GMM) are denoted as:

$$g_{\mu_k} = \frac{1}{N_{(t+1)}\sqrt{w_k}} \sum_{j=1}^{N_{(t+1)}} \phi_j(k)\left(\frac{f_j - \mu_k}{\sigma_k}\right), \tag{8}$$

$$g_{\sigma_k} = \frac{1}{N_{(t+1)}\sqrt{w_k}} \sum_{j=1}^{N_{(t+1)}} \phi_j(k)\left[\frac{(f_j - \mu_k)^2}{\sigma_k^2} - 1\right], \tag{9}$$

where w_k, μ_k, σ_k are the weight, mean vector and diagonal variance vector of the k-th Gaussian component, respectively. $\phi_j(k)$ represents the soft assignment weight of f_j to the k-th Gaussian component. We concatenate g_{μ_k} and g_{σ_k} for all the K Gaussian components to generate a $2N_tK$ dimensional final image-level representation:

$$F_{FV+CCA} = (g_{\mu_1}, g_{\sigma_1}, g_{\mu_2}, g_{\sigma_2}, \cdots, g_{\mu_k}, g_{\sigma_k}, \cdots, g_{\mu_K}, g_{\sigma_K}). \tag{10}$$

III. EXPERIMENTAL RESULTS

In this section, we assess the effectiveness of the proposed method for scene character recognition. In Section III-A, we first introduce the databases and experimental setup, and we then analyze the effect of different convolutional layers in the feature representation process in Section III-B. In Section III-C, we compare the performance of proposed method with other representative methods. In Section III-D, we investigate the influence of different encoding and pooling approaches for the proposed method.

FIGURE 4. Some samples from the ICDAR2003, Chars74K, and "Pan+ChiPhoto" databases. (a) ICDAR2003. (b) Chars74k. (c)"Pan+ChiPhoto".

A. DATABASES AND EXPERIMENTAL SETUP

ICDAR2003 Database: ICDAR2003 database [27] is an English scene character database. It consists of 6,185 character images for training and 5,430 character images for test. These images are distributed in 62 classes, i.e., A-Z, a-z and 0-9. The character images are subjected to various factors, such as blur, complex background, font variants, distortions, illumination and so on. A few samples taken from this database are shown in Figure 4 (a).

Chars74K Database: Chars74K database [8] is an English scene character database. It includes 12,503 character images of 62 classes, i.e., A-Z, a-z and 0-9. As in [8] and [28],

TABLE 1. Evaluation results (%) of different convolutional layers. The Conv_x denote the x-th convolutional layer. The three accuracy numbers are the evaluation results on the ICDAR2003, Chars74K and "Pan+ChiPhoto" databases, respectively.

Shallow Layers	Deeper Layers				
Conv_1	Conv_2	Conv_{2,3}	Conv_{2,3,4}	Conv_{2,3,4,5}	Conv_{2,3,4,5,6}
	(78.05, 67.10, 68.46)	(79.01, 67.53, 69.98)	(82.98, 71.83, 75.80)	(84.00, 74.19, 77.23)	(84.99, 75.27, 78.01)
Conv_2	Conv_3	Conv_{3,4}	Conv_{3,4,5}	Conv_{3,4,5,6}	-
	(81.03, 68.82, 73.05)	(81.84, 69.89, 74.52)	(85.30, 75.81, 78.75)	(**85.82**, **76.34**, **79.53**)	-
Conv_3	Conv_4	Conv_{4,5}	Conv_{4,5,6}	-	-
	(80.66, 67.20, 71.17)	(81.00, 68.60, 72.84)	(82.54, 73.33, 76.13)	-	-
Conv_4	Conv_5	Conv_{5,6}	-	-	-
	(77.13, 64.52, 65.74)	(77.81, 65.05, 67.95)	-	-	-

we randomly choose 30 character samples from each category, among which 15 samples are taken as training images and the remaining ones are regarded as test images. The Chars74K database is more challenging than the ICDAR2003 database where the most of character images are cropped from various natural scenes, such as products from stores, advertisement signs and so on. Figure 4 (b) shows some character samples from the Chars74K database.

"Pan+ChiPhoto" Database: "Pan+ChiPhoto" database [10] is a Chinese scene character database. It has totally 10,658 Chinese character images of 1,443 classes. In our implementation, we apply the same database settings as the works of [10] and [43]. These images in the "Pan+ChiPhoto" database are mainly captured from outdoors of Beijing and Shanghai, China, which involve various natural scenes like shop sign boards, road signs, banners, etc. A few of these Chinese character samples are shown in Figure 4 (c).

In the experiment, all images are normalized to 64×64. For the concatenated image-level feature vector, we perform PCA on the $F_{Concat+CCA}$ to reduce the dimensionality to 1,200 dimensions. For the FV, the number of Gaussian components K is ultimately set to 4, 2 and 8 on the ICDAR2003, Chars74K and "Pan+ChiPhoto" databases, respectively. In addition, we utilize L_2 normalization for the final image-level feature vectors. Note that, as for the Concat+CCA, SP+CCA, MP+CCA and FV+CCA, the FV+CCA is optimal for the task of scene character classification. Hence, in Section III-B and Scetion III-C, we only list the evaluation results of the FV+CCA.

B. EFFECT OF USING DIFFERENT CONVOLUTIONAL LAYERS

In a CNN, the shallow convolutional layers encode specific structures and textures of characters while the deeper convolutional layers usually encode high-level semantic information. In order to obtain powerful and discriminative feature representations, we first select feature maps in one shallow convolutional layer for extracting convolutional activation features, and then utilize the next consecutive deeper convolutional layers for learning weights.

The proposed method is evaluated from the viewpoint of using different convolutional layers. For convolutional activation feature extraction, the convolutional layer index varies from 1 to 4, as listed in the columns of Table 1. For weight matrix learning, the number of the next consecutive deeper convolutional layers varies from 1 to 4, as listed in the rows of Table 1. From Table 1, on the three databases, when extracting convolutional activation features from the feature maps in the shallow layer *Conv*_2 and learning weights from the feature maps in the next for deeper layers, i.e., *Conv*_{3, 4, 5, 6}, the proposed method achieves the highest accuracy. Furthermore, by analyzing the experimental results, we conclude the following three points:

First, when only learning weights from the shallow convolutional layers, for example, choosing the *Conv*_{2, 3} or *Conv*_{2, 3, 4}, the experimental results are unsatisfactory. Intrinsically, the shallow convolutional layers can only reflect the low-level patterns, and therefore the high-level patterns are ignored.

Second, when choosing one deeper layer, such as the *Conv*_4, for convolutional activation feature extraction, the important structural and textural information are abandoned, which is not conducive to generate powerful and discriminative feature vectors.

Third, it is inadequate of learning weights from only one or two deeper layers, for instance, choosing the *Conv*_3, *Conv*_4 or *Conv*_{3, 4}. It is because the other important high-level sematic cues are neglected.

C. COMPARISON TO THE STATE-OF-THE-ART METHODS

In Table 2, we compare the FV+CCA with the handcrafted feature, the part, the stroke and the CNN based methods on two English scene character databases (ICDAR2003 and Chars74K). The recognition results are in Table 2 and the following several conclusions can be drawn. First, the proposed method achieves the highest accuracies of 85.82% and 76.34% on the ICDAR2003 and Chars74K databases, respectively. Second, the proposed method obviously outperforms the handcraft features, i.e., HOG+SVM [28], Co-HOG [10] and ConvCoHOG [10]. These handcraft features utilize the gradient information

TABLE 2. Recognition accuracies (%) of different methods on the ICDAR2003 and Chars74K databases.

Algorithm	ICDAR2003	Chars74K
HOG+SVM [28]	77.00	62.00
Co-HOG [10]	80.50	-
ConvCoHOG [10]	81.70	-
TSM [24]	77.90	-
SED [28]	82.00	67.10
DSEDR [28]	82.60	71.80
FV+SVM [48]	84.40	74.80
CNN+softmax [48]	81.50	74.40
MCA-FV [43]	83.40	-
FV+CCA	**85.82**	**76.34**

of one signal pixel or the co-occurrence of neighboring pixel pairs, while the proposed method absorbs the structural, textural and the high-level sematic information with the help of the elaborate deep learning network. Third, compared with TSM, SED, DSEDR and FV+SVM which directly capture part, stroke or high-order statistic information from original character images, the advantages of the proposed method are: 1) the proposed CCA extracts features from convolutional layers learned by CNN; (2) the proposed CCA contains low-level and high-level patterns by extracting features from one shallow convolutional layer and learning weights from the consecutive convolutional layers. Fourth, the proposed method obtains better results than CNN+softmax and MCA-FV. The CNN+softmax and MCA-FV utilize the convolutional activations in one convolutional layer to describe feature representations, which neglects the information provided by other layers. While different from them, the proposed method not only captures the important textural and structural information by extracting convolutional activation features from one shallow convolutional layer, but also explores high-level information by learning weights from the consecutive convolutional layers.

Besides the two English scene character databases, we also evaluate the FV+CCA on one Chinese scene character database, i.e., "Pan+ChiPhoto" database. Table 3 lists the recognition accuracies on the "Pan+ChiPhoto". It shows that the proposed method outperforms other approaches and could correctly identify 79.53% of the test samples. The results indicate the good generalization ability and the effectiveness of the FV+CCA on this challenging Chinese scene character database.

D. INFLUENCE OF DIFFERENT ENCODING AND POOLING STRATEGIES UPON CCA

We evaluate four different encoding and pooling approaches, i.e., concatenation (Concat), sum pooling (SP), max pooling (MP) and Fisher vectors (FV) for the proposed CCA. We conduct experiments on the ICDAR2003, Chars74K and "Pan+ChiPhoto" databases to demonstrate which aggregation approach is the best for feature representations.

TABLE 3. Recognition accuracies (%) of different methods on the "Pan+ChiPhoto" database.

Algorithm	"Pan+ChiPhoto"
HOG+SVM [28]	59.20
Co-HOG [10]	65.40
ConvCoHOG [10]	71.20
CNN+softmax [48]	53.40
MCA-FV [43]	76.70
CNN [41]	61.50
FV+CCA	**79.53**

TABLE 4. Recognition results(%) of different aggregation methods upon CCA on the ICDAR2003, Chars74K, and "Pan+ChiPhoto" databases.

Databases	Concat+CCA	SP+CCA	MP+CCA	FV+CCA
ICDAR2003	82.91	83.17	84.73	**85.82**
Chars74K	73.33	74.09	75.27	**76.34**
"Pan+ChiPhoto"	76.19	77.38	78.63	**79.53**

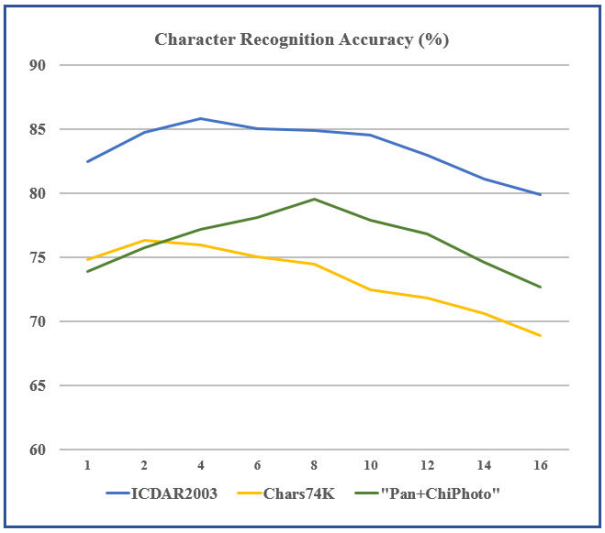

FIGURE 5. Performance of the proposed method under different K.

The experimental results of different aggregation methods upon CCA are reported in Table 4 and show that the FV+CCA is optimal for scene character recognition.

As for FV, the number of Gaussian components K is a crucial parameter as it determines the dimensionality of the final feature vector. We study the impact of K on all three databases. Figure 5 shows the recognition results when $K = 1, 2, 4, 6, 8, 10, 12, 14, 16$. From Figure 5, we can see that larger K towards to superior performance, however, the recognition result starts to drop when K coming to a fixed value. As for the ICDAR2003, Chars74K and "Pan+ChiPhoto" databases, K is ultimately set to 4, 2 and 8, respectively, where the proposed FV+CCA achieves the highest accuracy.

IV. CONCLUSION

In this paper, we have presented a novel feature representation approach for scene character recognition. The proposed

method (1) can mine the low-level patterns, such as structures, textures and so on, by extracting features from one shallow convolutional layer, and (2) can capture the high-level sematic information by learning weights from the consecutive convolutional layers. The proposed method has been validated on the ICDAR2003, Chars74k and ''Pan+ChiPhoto'' databases, and the experimental results outperform other typical methods for character recognition in natural scenes.

REFERENCES

[1] P. Sermanet, S. Chintala, and Y. LeCun, ''Convolutional neural networks applied to house numbers digit classification,'' in *Proc. Int. Conf. Pattern Recognit.*, Tsukuba, Japan, Nov. 2012, pp. 3288–3291.

[2] J. He *et al.*, ''Mobile product search with bag of hash bits and boundary reranking,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 3005–3012.

[3] G. N. Desouza and A. C. Kak, ''Vision for mobile robot navigation: A survey,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 237–267, Feb. 2002.

[4] H. Yang and C. Meinel, ''Content based lecture video retrieval using speech and video text information,'' *IEEE Trans. Learn. Technol.*, vol. 7, no. 2, pp. 142–154, Apr. 2014.

[5] M. Tzelepi and A. Tefas, ''Deep convolutional image retrieval: A general framework,'' *Signal Process., Image Commun.*, vol. 63, pp. 30–43, Apr. 2018.

[6] B. Yang, X. Shang, and S. Pang, ''Isometric hashing for image retrieval,'' *Signal Process., Image Commun.*, vol. 59, pp. 117–130, Nov. 2017.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ''Gradient-based learning applied to document recognition,'' *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[8] T. E. de Campos, B. R. Babu, and M. Varma, ''Character recognition in natural images,'' in *Proc. Int. Joint Conf. Comput. Vision Imag. Comput. Graph. Theory Appl.*, Lisboa, Portugal, Feb. 2009, pp. 273–280.

[9] A. Coates *et al.*, ''Text detection and character recognition in scene images with unsupervised feature learning,'' in *Proc. Int. Conf. Document Anal. Recognit.*, Beijing, China, Sep. 2011, pp. 440–445.

[10] S. Tian *et al.*, ''Multilingual scene character recognition with co-occurrence of histogram of oriented gradients,'' *Pattern Recognit.*, vol. 51, pp. 125–134, Mar. 2016.

[11] Z. Zhang, H. Wang, S. Liu, and L. Zheng, ''Scene character recognition using coupled spatial learning,'' *IEICE Trans. Inf. Syst.*, vol. 100, no. 7, pp. 1546–1549, 2017.

[12] X. Qu, W. Wang, K. Lu, and J. Zhou, ''In-air handwritten Chinese character recognition with locality-sensitive sparse representation toward optimized prototype classifier,'' *Pattern Recognit.*, vol. 78, pp. 267–276, Jun. 2018.

[13] N. Arica and F. T. Yarman-Vural, ''Optical character recognition for cursive handwriting,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 6, pp. 801–813, Jun. 2002.

[14] C.-L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, ''Handwritten digit recognition: Benchmarking of state-of-the-art techniques,'' *Pattern Recognit.*, vol. 36, no. 10, pp. 2271–2285, Oct. 2003.

[15] S. Belongie, J. Malik, and J. Puzicha, ''Shape matching and object recognition using shape contexts,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[16] A. C. Berg, T. L. Berg, and J. Malik, ''Shape matching and object recognition using low distortion correspondences,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2005, pp. 26–33.

[17] D. G. Lowe, ''Object recognition from local scale-invariant features,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2. Sep. 1999, pp. 1150–1157.

[18] S. Lazebnik, C. Schmid, and J. Ponce, ''A sparse texture representation using local affine regions,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1265–1278, Aug. 2005.

[19] M. Varma and A. Zisserman, ''Classifying images of materials: Achieving viewpoint and illumination independence,'' in *Proc. Eur. Conf. Comput. Vis.*, Copenhagen, Denmark, vol. 2, May 2002, pp. 255–271.

[20] M. Varma and A. Zisserman, ''Texture classification: Are filter banks necessaryn,'' *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2003, pp. 691–698.

[21] J. J. Weinman, E. Learned-Miller, and A. R. Hanson, ''Scene text recognition using similarity and a lexicon with sparse belief propagation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1733–1746, Oct. 2009.

[22] K. Wang and S. Belongie, ''Word spotting in the wild,'' in *Proc. Eur. Conf. Comput. Vis.*, Crete, Greece, Sep. 2010, pp. 591–604.

[23] A. J. Newell and L. D. Griffin, ''Multiscale histogram of oriented gradient descriptors for robust character recognition,'' in *Proc. Int. Conf. Document Anal. Recognit.*, Beijing, China, Sep. 2011, pp. 1085–1089.

[24] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhang, ''Scene text recognition using part-based tree-structured character detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 2961–2968.

[25] C.-Z. Shi, C.-H. Wang, B.-H. Xiao, S. Gao, and J.-H. Hu, ''Scene text recognition using structure-guided character detection and linguistic knowledge,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 7, pp. 1235–1250, Jul. 2014.

[26] S. Gao, C. Wang, B. Xiao, C. Shi, and Z. Zhang, ''Stroke bank: A high-level representation for scene character recognition,'' in *Proc. Int. Conf. Pattern Recognit.*, Stockholm, Sweden, Aug. 2014, pp. 2909–2913.

[27] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, ''ICDAR 2003 robust reading competitions,'' in *Proc. Int. Conf. Document Anal. Recognit.*, Edinburgh, U.K., Aug. 2003, pp. 682–687.

[28] C.-Z. Shi, S. Gao, M.-T. Liu, C.-Z. Qi, C.-H. Wang, and B.-H. Xiao, ''Stroke detector and structure based models for character recognition: A comparative study,'' *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4952–4964, Dec. 2015.

[29] S. Gao, C. Wang, B. Xiao, C. Shi, W. Zhou, and Z. Zhang, ''Learning co-occurrence strokes for scene character recognition based on spatiality embedded dictionary,'' in *Proc. IEEE Int. Conf. Image Process.*, Paris, France, Oct. 2014, pp. 5956–5960.

[30] C.-Y. Lee, A. Bhardwaj, W. Di, V. Jagadeesh, and R. Piramuthu, ''Region-based discriminative feature pooling for scene text recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 4050–4057.

[31] G. E. Dahl, D. Yu, L. Deng, and A. Acero, ''Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,'' *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[32] D. Ahmedt-Aristizabal, C. Fookes, S. Dionisio, K. Nguyen, J. P. S. Cunha, and S. Sridharan, ''Automated analysis of seizure semiology and brain electrical activity in presurgery evaluation of epilepsy: A focused survey,'' *Epilepsia*, vol. 58, no. 11, pp. 1817–1831, 2017.

[33] A. Chern, Y.-H. Lai, Y.-P. Chang, Y. Tsao, R. Y. Chang, and H.-W. Chang, ''A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom,'' *IEEE Access*, vol. 5, pp. 10339–10351, 2017.

[34] Z. Wei and M. Hoai, ''Region ranking SVM for image classification,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 2987–2996.

[35] S. Khawaldeh, U. Pervaiz, A. Rafiq, and R. S. Alkhawaldeh, ''Noninvasive grading of glioma tumor using magnetic resonance imaging with convolutional neural networks,'' *Appl. Sci.*, vol. 8, no. 1, p. 27, 2018.

[36] S. T. H. Rizvi, G. Cabodi, and G. Francini, ''Optimized deep neural networks for real-time object classification on embedded GPUs,'' *Appl. Sci.*, vol. 7, no. 8, p. 826, 2017.

[37] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus, ''Regularization of neural networks using DropConnect,'' in *Proc. Int. Conf. Mach. Learn.*, Atlanta, GA, USA, Jun. 2013, pp. 1058–1066.

[38] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, ''Deeply-supervised nets,'' in *Proc. Artif. Intell. Statist.*, San Diego, CA, USA, May 2015, pp. 562–570.

[39] Z. Zhang, H. Wang, S. Liu, and B. Xiao, ''Deep contextual stroke pooling for scene character recognition,'' *IEEE Access*, vol. 6, pp. 16454–16463, 2018.

[40] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, ''End-to-end text recognition with convolutional neural networks,'' in *Proc. Int. Conf. Pattern Recognit.*, Tsukuba, Japan, Nov. 2012, pp. 3304–3308.

[41] M. Jaderberg, A. Vedaldi, and A. Zisserman, ''Deep features for text spotting,'' in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, Sep. 2014, pp. 512–528.

[42] M. Cimpoi, S. Maji, and A. Vedaldi, ''Deep filter banks for texture recognition and segmentation,'' in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3828–3836.

[43] Y. Wang, C. Shi, C. Wang, B. Xiao, and C. Qi, "Multi-order co-occurrence activations encoded with Fisher Vector for scene character recognition," *Pattern Recognit. Lett.*, vol. 97, pp. 69–76, Oct. 2017.
[44] Z. Zhang, H. Wang, S. Liu, and T. S. Durrani, "Bilateral convolutional activations encoded with Fisher vectors for scene character recognition," *IEICE Trans. Inf. Syst.*, vol. 101, no. 5, pp. 1453–1456, 2018.
[45] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. Int. Conf. Comput. Statist.*, Paris, France, Aug. 2010, pp. 177–186.
[46] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.
[47] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, Crete, Greece, Sep. 2010, pp. 143–156.
[48] C. Shi, Y. Wang, F. Jia, K. He, C. Wang, and B. Xiao, "Fisher vector for scene character recognition: A comprehensive evaluation," *Pattern Recognit.*, vol. 72, pp. 1–14, Dec. 2017.

ZHONG ZHANG (M'14) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently an Associate Professor at Tianjin Normal University, Tianjin, China. He has published about 80 papers in international journals and conferences, such as the *Pattern Recognition*, the IEEE TRANSACTIONS ON CIRCUITS SYSTEMS VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, *Signal Processing* (Elsevier), CVPR, ICPR, and ICIP.

HONG WANG is currently pursuing the master's with Tianjin Normal University, Tianjin, China. Her research interests include scene character recognition and machine learning.

SHUANG LIU (M'18) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. She is currently an Associate Professor at Tianjin Normal University, Tianjin China.

BAIHUA XIAO received the B.S. degree in electronic engineering from Northwestern Polytechnical University, Xi'an, China, and the Ph.D. degree in pattern recognition and artificial intelligence from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1995 and 2000, respectively. Since 2005, he has been a Professor at the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. His research interests include pattern recognition, computer vision, image processing, and machine learning.