LETTER

Bilateral Convolutional Activations Encoded with Fisher Vectors for Scene Character Recognition

Zhong ZHANG[†a)], ***Member***, **Hong WANG**[†], **Shuang LIU**[†], ***and*** **Tariq S. DURRANI**[††], ***Nonmembers***

SUMMARY A rich and robust representation for scene characters plays a significant role in automatically understanding the text in images. In this letter, we focus on the issue of feature representation, and propose a novel encoding method named bilateral convolutional activations encoded with Fisher vectors (BCA-FV) for scene character recognition. Concretely, we first extract convolutional activation descriptors from convolutional maps and then build a bilateral convolutional activation map (BCAM) to capture the relationship between the convolutional activation response and the spatial structure information. Finally, in order to obtain the global feature representation, the BCAM is injected into FV to encode convolutional activation descriptors. Hence, the BCA-FV can effectively integrate the prominent features and spatial structure information for character representation. We verify our method on two widely used databases (ICDAR2003 and Chars74K), and the experimental results demonstrate that our method achieves better results than the state-of-the-art methods. In addition, we further validate the proposed BCA-FV on the "Pan+ChiPhoto" database for Chinese scene character recognition, and the experimental results show the good generalization ability of the proposed BCA-FV.

key words: *bilateral convolutional activations, Fisher vectors, scene character recognition*

1. Introduction

Characters, as the basic units of texts, are of great semantic value. Many applications in computer vision and pattern recognition involve the field of scene text recognition to automatically understand the texts in images. Conventional optical character recognition (OCR) based methods [1], [2] feed the binary image into OCR engine and perform well on the scanned documents. However, the scene texts differ from the traditional scanned ones due to heavy occlusion, blur and complex background, which is hard to binarize. Over the past decades, many approaches [3]–[6] are proposed to recognize scene texts. Although these methods have been reported good performance by combining language prior [7], [8], the scene character recognition is the primary determinant to the scene text recognition. Thus, in this letter, we focus on the scene character recognition.

The challenges of accurately recognizing scene characters lie in arbitrary fonts, noises, deformations, complex background and so on. Therefore, a powerful and effective feature representation strategy is indispensable for scene character recognition. Considering the significance of feature representation, Gao *et al.* [9] build spatial embedded dictionary under the framework of the bag-of-words (BoW) model to obtain the final features. Newell *et al.* [10] extract the multiscale histogram of oriented gradient (HOG) features to recognize the characters in natural scenes. These methods achieve considerable progress in the task of scene character recognition, but they come with the problem of insufficient discrimination ability of features.

To generate more discriminative representations, Perronnin *et al.* [11], [12] utilize Gaussian mixture model (GMM) to learn codebooks and obtain Fisher vectors (FV) by taking the derivative of GMM parameters. The FV is an enriched representation and its superiority manifests that encoding the high level information is more effective than the number of occurrences of visual words. Nowadays, several researchers resort to convolutional neural network (CNN) features. Wang *et al.* [13] regard the output of the last fully-connected layer of CNN as the final image representations. Jaderberg *et al.* [14] also report impressive accuracy by using the fully-connected layer based features. In [15], [16], the convolutional activations based features replace the full-connected ones to boost the classification accuracy.

In this letter, we propose a novel representation method named bilateral convolutional activations encoded with Fisher vectors (BCA-FV) to recognize the characters in natural scenes. For character images, convolutional activations describe the particular image regions and the layout of convolutional activations corresponds to stroke structures. Hence, we first extract convolutional activation features from the convolutional layers of CNN to retain rich stroke structure information. Then, we build a bilateral convolutional activation map (BCAM) to reflect the relationship

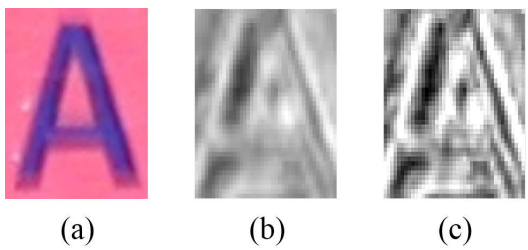

Fig. 1 Visualization of the convolutional activation map. (a) an image sample from the ICDAR2003 database, (b) the convolutional summing map (CSM), and (c) the bilateral convolutional activation map.

Manuscript received October 30, 2017.
Manuscript revised January 21, 2018.
Manuscript publicized February 2, 2018.

[†]The authors are with Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin, China.

[††]The author is with Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow Scotland, UK.

a) E-mail: zhong.zhang8848@gmail.com

DOI: 10.1587/transinf.2017EDL8238

between the spatial structure and the activation response. As visualized in Fig. 1 (c), the BCAM keeps the highly active responses of the convolutional map and removes the less important ones. Finally, the BCAM is embedded into the FV encoding strategy to derive a powerful character representation. In fact, the highly active responses indicate the salient parts, and therefore the BCAM can be regarded as the weight map of FV, which highlights the useful information for classification.

2. The Proposed Method

2.1 Convolutional Activations

In the convolutional layer, the filter traverses the input image in a sliding-window manner to generate a convolutional map. Generally, the top-left (bottom-right) activation in a convolutional map is generated by the top-left (bottom-right) part of the input image. Therefore, the obtained convolutional map involves not only the responses of activations, but also the stroke structure information of characters. In order to mine the useful information as much as possible, we extract convolutional activation descriptors from the convolutional maps.

The convolutional maps can be viewed as a tensor of size $W \times H \times N$, which contains N convolutional maps with width W and height H. The tensor can be treated as a map consisting of $(W \times H)$ N-dimensional convolutional activation descriptors. We select one convolutional layer from the pre-trained CNN network as described in [14] for the extraction of convolutional activation descriptors. The convolutional activation descriptors extracted from different positions of the convolutional maps represent diverse parts and preserve the stroke structure information of characters.

2.2 Bilateral Convolutional Activations Map

As we known, each activation response in a convolutional map describes a local part of the input image and the high responses indicate the salient parts. To discover the important feature and spatial stroke information, we propose the bilateral convolutional activations map (BCAM) to build the relationship between the spatial structure and the activation response. Specifically, we first add all the convolutional maps of one convolutional layer to capture the completed spatial response information of characters. Let C_i denote the i-th response of the convolutional summing map (CSM) and it is formulated as:

$$C_i = \sum_{n=1}^{N} c_i^n, \tag{1}$$

where c_i^n denotes the i-th activation response of the n-th convolutional map and N is the number of the convolutional maps. The CSM is shown in Fig. 1 (b), from which we can see that those high activated positions mainly distribute in the character area. This illustrates that the value of C_i can reflect the importance of the local features. However, some high responses may be distracters, such as noise, outlier, etc. To overcome this limitation, the BCAM is proposed based on the CSM. The j-th activation response of the BCAM is formulated as:

$$O_j = \sum_i B_{ij} C_i, \tag{2}$$

where i and j represent the activation response indexes, and B_{ij} is designed as:

$$B_{ij} = exp(\frac{-|C_i - C_j|}{\sigma_1}) + \alpha exp(\frac{-|L_i - L_j|}{\sigma_2}), \tag{3}$$

where L_i and L_j are the corresponding coordinates of C_i and C_j, respectively, and α is the parameter to regulate the influence of spatial similarity. The first term of Eq. (3) indicates the importance based on the activation response differences, and the second term indicates the importance based on the spatial distance between the activation responses. Both of them are controlled by σ_1 and σ_2, respectively.

Generally, the activation responses of the neighbors in a convolutional map are similar. If the activation responses of the neighbors in the CSM are similar, the B_{ij} is large to highlight the salient parts. While if the activation responses of the neighbors in the CSM differ greatly, there may be the noise or outlier. In this situation, B_{ij} is small so as to restrain the distracters. We visualize the BCAM in Fig. 1 (c), from which we can see that the BCAM boosts the salient visual content and suppresses the interference components.

2.3 Encoding with BCAM

In order to combine the local and global information, we propose bilateral convolutional activations encoded with Fisher vectors (BCA-FV) for character representation. Specifically, the BCAM is embedded into FV to encode the convolutional activation descriptors. Assuming that the dimensionality of convolutional activation descriptor is N, the N-dimensional derivatives with respect to the mean vector μ_k and diagonal variance vector σ_k of the k-th GMM are denoted as:

$$f_{\mu_k} = \frac{1}{M\sqrt{w_k}} \sum_{j=1}^{M} O_j \phi_j(k) (\frac{x_j - \mu_k}{\sigma_k}), \tag{4}$$

$$f_{\sigma_k} = \frac{1}{M\sqrt{w_k}} \sum_{j=1}^{M} O_j \phi_j(k) [\frac{(x_j - \mu_k)^2}{\sigma_k^2} - 1], \tag{5}$$

where w_k denotes the weight of the k-th Gaussian component, $\phi_j(k)$ is the soft assignment weight of convolutional activation descriptor x_j to the k-th Gaussian component, and M is the total number of convolutional activation descriptors in an image. We concatenate f_{μ_k} and f_{σ_k} for all the K Gaussian components to generate the feature vector for the MCA-FV which is a $2NK$-dimensional vector. The BCAM can be regarded as a weight map of FV. By injecting the BCAM into FV encoding, the proposed BCA-FV can automatically

select the useful descriptors for characters, leading to a more powerful feature representation.

3. Experiments

3.1 Databases and Implementation Details

We first evaluate the proposed method on two typical scene character recognition databases: ICDAR2003 [17] and Chars74K [18]. Both of them contain 52 classes English letters, i.e., lower English letters a-z, upper English letters A-Z, and 10 classes Arabic numbers 0-9. The ICDAR2003 database contains 6,185 training and 5,430 test images, and these images undergo extensive variances such as nonuniform illumination, distortions and complex backgrounds. The Chars74K database collected from various natural scenes has totally 12,503 images. The samples in this database vary in color, size, font, background, etc. When performing Chars74K evaluation, we randomly select 30 images for each class, in which 15 images are used for training and the remaining are used for testing as described in [18], [19]. We also conduct experiments on the Chinese scene character database "Pan+ChiPhoto" [20] and adopt the same experimental setup as described in [20].

In the experiments, we utilize the pre-trained CNN network in [14] for the extraction of convolutional activation descriptors. We employ 64 convolutional maps of the second convolutional layer (*conv*_2), and therefore the dimensionality of the convolutional activation descriptor is 64. The character images are normalized into 24×24. The parameters σ_1 and σ_2 are empirically set to 0.05 and 6, respectively, and α is set to 1.5. The parameters σ_1, σ_2 and α are chosen on the training set of ICDAR2003 database, and we directly utilize the same values on the other databases.

3.2 Evaluation of Vocabulary Size of BCA-FV

The vocabulary size K is an importance parameter because it determines the dimensionality of the final feature vector and effects the classification results of scene characters.

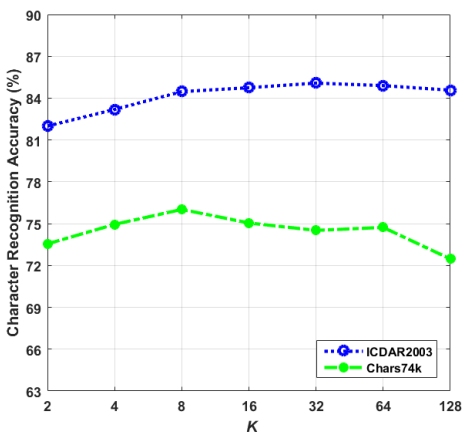

Fig. 2 Performance of the proposed BCA-FV method under different size of vocabulary on the ICDAR2003 and Chars74K databases.

Hence, we investigate the influence of the vocabulary size K of BCA-FV for scene character recognition. Figure 2 shows character recognition accuracy on the ICDAR2003 and Chars74K databases when K = 2, 4, 8, 16, 32, 64, 128. As can be seen, the character recognition accuracy improves with the increasing size of vocabulary in a range, but when K comes to a certain point, the performance drops a little or fluctuates. For the ICDAR2003 database, the best performance can be obtained when K is equal to 32. While on the Chars74K database, the highest accuracy is obtained when K is equal to 8. For the "Pan+ChiPhoto" database, K is empirically set to 32. With a small size of vocabulary, the proposed BCA-FV method can achieve high scene character recognition accuracy.

3.3 Comparison with Other Methods

In Table 1, we compare the proposed BCA-FV method with other state-of-the-art methods. From Table 1, we can see that our method is superior to other published methods including HOG based and CNN (using fully-connected layer features) based methods. Compared with HOG+SVM [19] and Co-HoG [20] which encode the spatial information by considering the co-occurrence of orientation pairs, our method achieves superior performance. Compared with SED [9], DSEDR [19], DMSDR [19] and Stoke Bank [21], which utilize the HOG as the local descriptors to capture stroke structure information, the proposed BCA-FV method outperforms them by more than 3% (8%), 2% (4%), 3% (9%) and 5% (10%) on the ICDAR2003 (Chars74K) database, respectively. Compared with CNN+softmax [22], FV+SVM [22] and MCA-FV [23], the superiorities of our method lie in: (1) the convolutional activation descriptors extracted from the second convolutional layer possess stronger discriminative ability; (2) the proposed BCA-FV method can automatically select the useful descriptors with prominent feature and stroke structure information for characters. The CSM-FV method directly inject the convolutional summing map (CSM) into FV for encoding convolutional activation descriptors, and achieves the accuracies of 82.98% and 73.12% on the ICDAR2003 and Chars74K databases, respectively. Noticeably, the performance of the

Table 1 Recognition accuracies (%) of different methods on the ICDAR2003, Chars74K and "Pan+ChiPhoto" databases.

Algorithm	ICDAR	Chars74K	"Pan+ChiPhoto"
HOG+SVM [19]	77.00	62.00	59.20
Co-HoG [20]	80.50	-	64.30
Stoke Bank [21]	79.80	65.90	-
SED [9]	82.00	67.10	-
DSEDR [19]	82.60	71.80	-
DMSDR [19]	81.70	66.10	-
CNN+softmax [22]	81.57	73.52	53.40
FV+SVM [22]	84.40	74.80	-
MCA-FV [23]	83.40	-	76.70
CNN [14]	86.80	-	61.5
CSM-FV	82.98	73.12	75.34
BCA-FV	**85.08**	**76.02**	**77.30**

BCA-FV gains higher accuracy than the CSM-FV, because our method can automatically select the useful descriptors and remove the less importance ones, leading to a more powerful representation. The CNN [14] achieves 86.80% on the ICDAR2003 database for character recognition which partly attributes to the large amount of additional training data, i.e., 107k. With limited training data, i.e., 6k, the proposed BCA-FV achieves the accuary of 85.08%.

The proposed BCA-FV achieves the best result on the Chinese scene character database, i.e., "Pan+ChiPhoto" database, and the results demonstrate the good generalization ability of the proposed method. The proposed BCA-FV outperforms CNN [14] by more than 15% on the Pan + Chiphoto database. Since the training samples are limited, the advantage of CNN could not present and it obtains lower accuracy.

4. Conclusion

In this letter, we have proposed the BCA-FV, a novel feature encoding method for recognizing characters in natural scenes, to automatically select the useful descriptors in the encoding process. The proposed BCA-FV method builds the BCAM for each image to reflect the relationship between the spatial structure and the activation response, and embeds the BCAM into FV to encode the convolutional activation descriptors for scene character representation. The proposed BCA-FV method has been validated on three well-known databases, i.e., the ICDAR2003, Chars74K and "Pan+ChiPhoto" databases, and the experimental results outperform the other previous methods in scene character recognition.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant No. 61501327 and No. 61711530240, Natural Science Foundation of Tianjin under Grant No. 17JCZDJC30600 and No. 15JCQNJC01700, the Open Projects Program of National Laboratory of Pattern Recognition under Grant No. 201700001 and No. 201800002, the China Scholarship Council No. 201708120039 and No. 201708120040, and the NSFC-Royal Society grant.

References

[1] X. Chen and A.L. Yuille, "Detecting and reading text in natural scenes," International Conference on Computer Vision and Pattern Recognition (CVPR), pp.366–373, 2004.
[2] L. Neumann and J. Matas, "Real-time scene text localization and recognition," International Conference on Computer Vision and Pattern Recognition (CVPR), pp.3538–3545, 2012.
[3] A. Mishra, K. Alahari, and C.V. Jawahar, "Top-down and bottom-up cues for scene text recognition," International Conference on Computer Vision and Pattern Recognition (CVPR), pp.2687–2694, 2012.
[4] J.J. Weinman, Z. Butler, D. Knoll, and J. Feild, "Toward integrated scene text reading," IEEE Trans. Pattern Anal. Mach. Intell., vol.36, no.2, pp.375–387, 2014.
[5] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets: A learned multi-scale representation for scene text recognition," International Conference on Computer Vision and Pattern Recognition (CVPR), pp.4042–4049, 2014.
[6] B. Alessandro, C. Mark, N. Yucal, and N. Hartmut, "PhotoOCR: Reading text in uncontrolled conditions," International Conference on Computer Vision (ICCV), pp.785–792, 2013.
[7] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhang, "Scene text recognition using part-based tree-structured character detection," International Conference on Computer Vision and Pattern Recognition (CVPR), pp.2961–2968, 2013.
[8] A. Mishra, K. Alahari, and C.V. Jawahar, "Scene text recognition using higher order language priors," British Machine Vision Conference (BMVA), pp.127.1–127.11, 2012.
[9] S. Gao, C. Wang, B. Xiao, Z. Wen, and Z. Zhang, "Scene text character recognition using spatiality embedded dictionary," IEICE Trans. Inf. & Syst., vol.E97-D, no.7, pp.1942–1946, 2014.
[10] A.J. Newell and L.D. Griffin, "Multiscale histogram of oriented gradient descriptors for robust character recognition," International Conference on Document Analysis and Recognition (ICDAR), pp.1085–1089, 2011.
[11] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," International Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2007.
[12] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," European Conference on Computer Vision (ECCV), vol.6314, pp.143–156, 2010.
[13] T. Wang, D.J. Wu, A. Coates, and A.Y. Ng, "End-to-end text recognition with convolutional neural networks," International Conference on Pattern Recognition (ICPR), pp.3304–3308, 2012.
[14] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," European Conference on Computer Vision (ECCV), pp.512–528, 2014.
[15] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," International Conference on Computer Vision and Pattern Recognition (CVPR), pp.3828–3836, 2015.
[16] A.B. Yandex and V. Lempitsky, "Aggregating local deep features for image retrieval," International conference on computer vision (ICCV), pp.1269–1277, 2015.
[17] S.M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," International Conference on Document Analysis and Recognition (ICDAR), pp.682–687, 2003.
[18] T.E. de Campos, B.R. Babu, and M. Varma, "Character recognition in natural images," International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), pp.273–280, 2009.
[19] 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. on Image Process., vol.24, no.12, pp.4952–4964, 2015.
[20] S. Tian, U. Bhattacharya, S. Lu, B. Su, Q. Wang, X. Wei, Y. Lu, and C.L. Tan, "Multilingual scene character recognition with co-occurrence of histogram of oriented gradients," Pattern Recognition, vol.51, pp.125–134, 2016.
[21] S. Gao, C. Wang, B. Xiao, C. Shi, and Z. Zhang, "Stroke bank: A high-level representation for Scene Character Recognition," International Conference on Pattern Recognition (ICPR), pp.2909–2913, 2014.
[22] C. Shi, Y. Wang, F. Jia, K. He, C. Wang, and B. Xiao, "Fisher vector for scene character recognition: a comprehensive evaluation," Pattern. Recogn., vol.72, pp.1–14, 2017.
[23] 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. Recogn. Lett., vol.97, pp.69–76, 2017.