Scene Character Recognition Using Coupled Spatial Learning

Zhong ZHANG†‡, Member, Hong WANG†, Shuang LIU†, and Liang ZHENG††, Nonmembers

SUMMARY Feature representation, as a key component of scene character recognition, has been widely studied and a number of effective methods have been proposed. In this letter, we propose a novel method named coupled spatial learning (CSL) for scene character representation. Different from the existing methods, the proposed CSL method simultaneously discovers the spatial context in both the dictionary learning and coding stages. Concretely, we propose to build the spatial dictionary by preserving the corresponding positions of the codewords. Correspondingly, we introduce the spatial coding strategy which utilizes the spatiality regularization to consider the relationship among features in the Euclidean space. Based on the spatial dictionary and spatial coding, the spatial context can be effectively integrated in the visual representations. We verify our method on two widely used databases (ICDAR2003 andChars74k), and the experimental results demonstrate that our method achieves competitive results compared with the state-of-the-art methods. In addition, we further validate the proposed CSL method on the Caltech-101 database for image classification task, and the experimental results show the good generalization ability of the proposed CSL.

key words: coupled spatial learning, feature representation, scene character recognition

1. Introduction

Scene text recognition implies important applications in image retrieval, intelligent transportation, human-computer interaction, etc. A typical scene-text-extraction system consists of two major components: text detection and text recognition. Scene text detection has achieved impressive accuracy while scene text recognition especially scene character recognition still needs further exploration. Generally speaking, scene character recognition methods are divided into two major categories: the optical character recognition (OCR) based methods and the object recognition based methods.

Due to the limitations of OCR, such as scene text block binarization, many researchers focus on the object recognition based methods. For example, Zhang et al. [1] extract the histograms of sparse codes (HSC) features for scene character representation. Weinman et al. [2] consider the appearance and similarity in the stage of scene character recognition. Shi et al. [5] extend [4] to discriminative multi-scale stroke detector-based representation (DMSDR). Tian et al. [6] learn the spatial information by using the co-occurrence of histogram of oriented gradients (Co-HOG) which captures spatial distribution of neighboring orientation pairs. Gao et al. [4] propose the stroke bank to consider the spatial context in the stage of scene character representation. Although these methods demonstrate promising results, the spatial context among local regions is largely ignored, which plays a profound role in scene character representation. As shown in Fig. 1, patches in different characters may exhibit similar appearances. For example, the local appearance in the lower region of ‘B’ is the same as in the upper region of ‘P’. Hence, confusion will occur if spatial context is not considered.

In order to overcome the above-mentioned limitation, Gao et al. [4] propose the stroke bank to consider the spatial context in the stage of scene character representation. Shi et al. [5] extend [4] to discriminative multi-scale stroke detector-based representation (DMSDR). Tian et al. [6] learn the spatial information by using the co-occurrence of histogram of oriented gradients (Co-HOG) which captures spatial distribution of neighboring orientation pairs. Gao et al. [7] embed the spatial context in the stage of dictionary learning for scene character representation. However, these methods only consider spatial context in either the dictionary learning or coding stages, so spatial constraints may be less effectively integrated.

In recent years, deep learning method has shown promising performance in the field of pattern recognition. Wang et al. [11] utilize convolutional neural networks (CNN) to recognize scene characters and achieve impressive accuracy when using both the original training images and those synthetic ones. Jaderberg et al. [12] develop a CNN classifier that can be used for both text detection and recognition. Although deep learning method has achieved significant performance, the effectiveness depends on the far more training images. Besides, the network configuration often requires extensive fine tuning for the optimal classification performance.

In this letter, we propose a novel coupled spatial learn-
coupled spatial learning (CSL) method for scene character representation. In CSL, we learn the spatial context in the dictionary learning and coding stages simultaneously. The overall pipeline of the proposed CSL method is shown in Fig. 2. In the dictionary learning stage, the HOG features are first extracted from each image and then the local features in the same position are clustered into a sub-dictionary. Finally, all the sub-dictionaries together with their positions form the spatial dictionary. The spatial dictionary preserve the corresponding positions of the codewords in each sub-dictionary so as to contain the spatial context. In the coding stage, we propose the spatiality regularization to consider the spatial relationship among features in the Euclidean space. Meanwhile, we employ the locality regularization to reconstruct feature using several codewords. Based on the above stages, the proposed CSL method can not only effectively learn spatial context for character representation, but also alleviate the quantization error.

2. Coupled Spatial Learning Method

2.1 Building Spatial Dictionary

The spatial context of features is critical for scene character representation. However, traditional dictionary learning methods directly cluster the features, which loses the spatial cues. To overcome the drawback, we incorporate the spatial context in the dictionary learning stage. The learning procedure is listed as follows:

1. Initialization. We normalize scene character images into the same size $H \times W$ (64x32), where $H$ and $W$ are the height and width of an image respectively. Meanwhile, we transform the color images into gray ones.

2. Feature extraction. We extract HOG features from the training set. Specifically, the HOG feature in the position $P_i$ ($i = 1, 2, \cdots, m$) is extracted for each training image, where $m$ is the number of positions as shown in Fig. 2(a). As a result, we obtain $t_{rnum}$ HOG features for each position $P_i$ in the training set, where $t_{rnum}$ is the number of training images.

3. Feature clustering. We employ $K$-means clustering algorithm on the $t_{rnum}$ HOG features in the position $P_i$, and then obtain the $i$-th sub-dictionary $C_i \in \mathbb{R}^{N \times k}$ ($i = 1, 2, \cdots, m$). Here $k$ is the number of codewords in the sub-dictionary $C_i$, and $N$ is the dimensionality of the HOG feature. Thus, for all the $m$ positions, we obtain the spatial dictionary which contains $m$ sub-dictionaries.

It is noteworthy that each sub-dictionary corresponds to a position information. From Fig. 2(b), we can see that the spatial dictionary consists of two parts, i.e., the collection of sub-dictionaries $C$ and its position $P$. Here, $C = (C_1, C_2, \cdots, C_m)$ and $P = (P_1, P_2, \cdots, P_m)$. Therefore, the spatial dictionary can be described as: $D = \{C, P\} = \{(C_1, P_1), (C_2, P_2), \cdots, (C_m, P_m)\}$. The spatial dictionary not only contains the spatial context for each codeword, but is also beneficial to integrate spatial context for the coding stage.

2.2 Learning Spatial Coding

The traditional coding methods, such as locality-constrained linear coding [10], sparse coding [8], and Fisher vector [9], only consider the relationship among features in the feature space, but neglect the spatial cues of features. In order to effectively integrate the spatial context based on the spatial dictionary, we propose the spatiality regularization which explicitly considers the spatial relationship among features in the Euclidean space. Specifically, the distance in the Euclidean space between the positions of the feature and codewords is incorporated by the spatiality regularization in the coding stage. The smaller the distance between the positions of feature and codewords is, the greater the influence of the codeword on the coding coefficients obtains. Hence, we prefer that the coding coefficient is inversely proportional to the distance between the positions of feature and codewords. According to the above considerations, the spatiality regularization is defined as:

$$
\|d_{jkl} \circ a_j\|_2^2,
$$

(1)

where $\|\cdot\|_2^2$ represents the $l_2$ norm and $\circ$ is the element-wise multiplication in two matrices. Let $f = (f_j, l_j)$ denote the feature extracted from a patch, where $f_j \in \mathbb{R}^{N \times k}$ represents the HOG feature vector and $l_j$ is the corresponding position. $a_j$ is the coding vector of $f_j$, and it follows the principle of $1^T a_j = 1(\forall j)$ which means the sum of all elements in $a_j$ is equal to 1. $d_{jkl}$ is the distance between the position of feature and codewords and it is designed as:

$$
d_{jkl} = \exp\left(\frac{\text{dist}(l_j, P_k)}{\sigma_E}\right),
$$

(2)

where $\sigma_E$ is a positive number to adjust the speed of weight decay for $d_{jkl}$. Here, $\text{dist}(l_j, P)$ is the Euclidean distance between $l_j$ and $P$:

$$
\text{dist}(l_j, P) = \left[\text{dist}(l_j, P_1), \cdots, \text{dist}(l_j, P_1), \text{dist}(l_j, P_2), \cdots, \text{dist}(l_j, P_m), \cdots, \text{dist}(l_j, P_m)\right]^T,
$$

(3)

where $\text{dist}(l_j, P_i)$ is the Euclidean distance between $l_j$ and $P_i$. Note that since there are $m$ positions ($m$ sub-dictionaries)}
and $k$ codewords with the same position for each sub-dictionary, $\text{dist}(l_j, P)$, $d_{j,k}$ and $a_j \in \mathbb{R}^{m \times 1}$ are all vectors containing $m \times k$ elements. Equation (1) can be viewed as a penalty term, which gives a high penalty if the distance between $l_j$ and $P$ is large. That is to say, it gives a penalty according to the Euclidean distance between the positions of $l_j$ and $P$. So the term is called the spatiality regularization. The objective function of the proposed spatial coding is formulated as:

$$\arg \min_A \sum_j \|f_j - C a_j\|^2 + \alpha \|d_{j,k} \odot a_j\|^2 + \beta \|d_{j,f} \odot a_j\|^2$$

$$\text{s.t. } A^T a_j = 1, \forall j,$$

where $\alpha$ and $\beta$ are the regularization parameters, and $A = \{a_1, a_2, \ldots, a_j, \ldots\}$ is the set of coding vectors for all the features. Here, $d_{j,f}$ is the distance between the feature and codewords in the feature space:

$$d_{j,f} = \exp(-\frac{\text{dist}(f_j, C)}{\sigma_F}),$$

(5)

where $\sigma_F$ is used to regulate the weight decay speed for $d_{j,f}$. The $\text{dist}(f_j, C)$ is defined as:

$$\text{dist}(f_j, C) = [\text{dist}(f_j, C_1), \text{dist}(f_j, C_2), \ldots, \text{dist}(f_j, C_m)]^T,$$

(6)

where $\text{dist}(f_j, C_i)$ is a vector with $k$ elements which represents the distance between $f_j$ and $k$ codewords in the sub-dictionary $C_i$ ($i = 1, 2, \ldots, m$). Hence, there are $m \times k$ elements in the vector $\text{dist}(f_j, C) \in \mathbb{R}^{m \times 1}$. $\|d_{j,f} \odot a_j\|^2$ can be treated as the locality regularization which considers the relationship among features in the feature space. The non-zero elements in $a_j$ are used to reconstruct features to alleviate the reconstruction error. Note that when $\beta = 0$, the coding strategy of CSL will reduce to LLC, and thus the dictionary does not contain any spatial information. Besides, it is worth mentioning that Eq. (4) has an analytical solution that can be derived from Eq. (7):

$$\alpha_j = (A_j + \text{diag}(d_{j,k}) + \beta \text{diag}(d_{j,f})) \backslash 1,$$

(7)

where $A_j = (C^T - 1 f_j^T)(C^T - 1 f_j^T)^T$ is the data covariance matrix. We further normalize $\alpha_j$ using $a_j = \alpha_j/1^T \alpha_j$. Taking advantage of the analytical solution rather than complex iterative optimization can reduce the computational cost.

After learning all the coding for each image, we employ max pooling strategy [10], [13] to obtain the final feature vector, and these features are classified by a SVM with linear kernel (see Fig. 2 (d) and (e)).

3. Experiments

3.1 ICDAR2003 Database


On this database, we first study the influence of the regularization parameters $\alpha$ and $\beta$ in Eq. (4). We evaluate $\alpha$ as it varies from 10 to 40 with a step of 10 and evaluate $\beta$ as it varies from 40 to 70 with a step of 10. From Fig. 3, the experimental results indicate that when $\alpha = 20$ and $\beta = 50$ we achieve the best results. Table 1 compares our method with the other competing methods, from which we can see that the proposed CSL outperforms LLC [10] by more than 4%, and the superiority of our method lies in simultaneously considering the spatiality and the locality regularization terms. Compared with DMSDR [5] and SED [7] which only consider to learn the spatial context in one of learning stages, our method achieves superior performance. Compared with the proposed CSL method, DSED [5] utilizes a WOSVM (weight of SVM) strategy which trains a one-vs-all classifier for each category of characters and selects the discriminative codewords corresponding to the large absolute weights. For fair comparison, we combine our approach with the WOSVM strategy to choose discriminative codewords and the CSL+WOSVM obtains higher accuracy than DSED on the ICDAR2003 database. In [12], it achieves 86.8% on 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 CSL+WOSVM method achieves the accuracy of 84.8%.

3.2 Chars74k Database

The Chars74k [16] database contains 62 character classes (0-9, a-z, A-Z). For each class, 30 images are randomly selected, among which 15 are used for training and the re-
maining are used for testing as [5], [16]. With the same parameters discussed on the ICDAR2003 database, the CSL and CSL+WOSVM achieves the recognition accuracies of 68.2% and 72.5% respectively. Table 1 compares our result with the state-of-the-art methods, where the results clearly show that our approach is superior to other published methods. Compared with DMSDR [5], DSED [5], SED [7] and LLC [10], the superior results manifest that the spatial context incorporated in the learned feature vector is effective in improving the recognition accuracy.

3.3 Caltech-101 Database

In this section, the proposed CSL method is further validated on the Caltech-101 database [17] which contains 9144 images in 101 classes (animals, flowers and ect.), to demonstrate the generalization ability. In this experiment, we partition the whole database into 15, 20, 25 and 30 training images and no more than 50 testing images per class. We use average accuracy over 102 classes (containing a background class). From Table 2, we can see that the proposed CSL method outperforms the LLC method by more than 2 percent.

4. Conclusion

In this letter, we propose a novel feature representation method named coupled spatial learning (CSL) method for scene character recognition. The proposed CSL method explicitly considers the spatial context in the dictionary learning and coding stages simultaneously, which can effectively discover the spatial context. The proposed CSL method has been validated on three challenging databases (ICDAR2003, Chars74k and Caltech-101), and the experimental results outperform the other previous methods.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant No. 61501327, No. 6161101323 and No. 61401309, Natural Science Foundation of Tianjin under Grant No. 17JCYBJC30600, and No. 15JCQNJC01700, the Open Projects Program of National Laboratory of Pattern Recognition under Grant No. 201700001, and Doctoral Fund of Tianjin Normal University under Grant No. 5RL134 and No. 52XB1405.

References