Pattern Recognition 72 (2017) 1–14

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Fisher vector for scene character recognition: A comprehensive evaluation

Cunzhao Shi*, Yanna Wang, Fuxi Jia, Kun He, Chunheng Wang, Baihua Xiao

The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, No.95 Zhongguancun East Road, Beijing 100190, PR China

ARTICLE INFO

Article history:
Received 19 October 2016
Revised 22 May 2017
Accepted 16 June 2017
Available online 23 June 2017

Keywords:
Character representation
Character recognition
Fisher vector (FV)
Gaussian Mixture Models (GMM)
Bag of visual words (BOW)

ABSTRACT

Fisher vector (FV), which could be seen as a bag of visual words (BOW) that encodes not only word counts but also higher-order statistics, works well with linear classifiers and has shown promising performance for image categorization. For character recognition, although standard BOW has been applied, the results are still not satisfactory. In this paper, we apply Fisher vector derived from Gaussian Mixture Models (GMM) based visual vocabularies on character recognition and integrate spatial information as well. We give a comprehensive evaluation of Fisher vector with linear classifier on a series of challenging English and digits character recognition datasets, including both the handwritten and scene character recognition ones. Moreover, we also collect two Chinese scene character recognition datasets to evaluate the suitability of Fisher vector to represent Chinese characters. Through extensive experiments we make three contributions: (1) we demonstrate that FV with linear classifier could outperform most of the state-of-the-art methods for character recognition, even the CNN based ones and the superiority is more obvious when training samples are insufficient to train the networks; (2) we show that additional spatial information is very useful for character representation, especially for Chinese ones, which have more complex structures; and (3) the results also imply the potential of FV to represent new unseen categories, which is quite inspiring since it is quite difficult to collect enough training samples for large-category Chinese scene characters.

1. Introduction

Characters are the fundamental elements of text, which could provide high-level semantic information. With the widespread use of smart mobile devices with digital cameras, automatically understanding the text in images is in urgent need for a growing number of vision applications, such as image understanding, automatic sign recognition, navigation, automatic geocoding [1,2], etc. Although conventional Optical Character Recognition (OCR) has been considered as a solved problem and has been successfully applied in many fields, camera-captured text images are confronted with more challenges than conventional scanned ones, such as different fonts, various deformations, complex background and so on. Over the past decades, many algorithms are proposed to deal with scene text recognition [3–12]. Although the final text recognition result could be improved by using the language prior [5,6,10] or some lexicon based postprocessing [11,13], we argue that character recognition accuracy is the primary determinant and thus of crucial importance to the final text recognition.

Scene character recognition still has many obstacles due to the large intra-class variability caused by unpredictable image collecting condition (lighting, resolution, blurring, etc.) and changeable background. To deal with all the problems, a compact and robust representation is the key for the final satisfactory performance. Most recently published methods consider scene characters as a special category of objects. They take advantages of feature extraction and representation methods that perform well in object detection or recognition tasks and apply these methods on character recognition [3,11,14,15]. The Bag-of-Visual-Words (BOW) framework is one of the most widely used image representation methods and was first introduced by De Campos et al. [14] for character recognition. They benchmarked the performance of various features to assess the feasibility of posing the problem as an object recognition task and achieved better results compared to conventional OCR methods. However, the recognition accuracy is still far from satisfactory.

The Fisher kernel is a powerful framework which combines the strengths of generative and discriminative approaches to pattern

* Corresponding author.
E-mail address: cunzhao.shi@ia.ac.cn (C. Shi).

http://dx.doi.org/10.1016/j.patcog.2017.06.022
0031-3203/

classification [16]. The idea is to characterize a signal with a gradient vector derived from a probability density function (pdf) which models the generation process of the signal [17]. When we use GMM to model the distribution of low-level features, FV extents the BOW in a way that the representation is not limited to the number of occurrences of each visual word but also encodes additional information about the distribution of the descriptors [18]. Perronnin et al. [18] proved the suitability of FV with only SIFT features and costless linear classifiers for large-scale image classification tasks.

For character recognition, although standard BOW has been applied, the performance is still unsatisfactory. As FV could encode higher-order statistics, a better representation could be acquired with much less visual words compared to standard BOW. In this paper, we apply FV derived from Gaussian Mixture Models (GMM) based visual vocabularies on character recognition. Considering the unique structure information of man-made characters, we also integrate the spatial information into the representation. We give a comprehensive evaluation of FV with linear classifier on a series of challenging character recognition datasets, including two handwritten datasets–CVL Single Digit dataset (CVLSD) [19] and MNIST dataset [20], and four scene character recognition datasets–the Street View House Number (SVHN) dataset [1], ICDAR 2003 character recognition dataset (ICDAR03-CH) [21], Chars74k dataset [14] and ISI-Bengali-Character [22]. Moreover, we also collect two Chinese scene character recognition datasets—the Chinese Plate Character recognition dataset (CPC) and the much larger Chinese Street View Text dataset (CSVT) to evaluate the suitability of Fisher vector to represent Chinese characters. Through extensive experiments and analysis our contributions are three-folds: (1) we demonstrate that FV with linear classifier could outperform most of the state-of-the-art methods for character recognition, even the CNN based ones with the same amount of training samples and the superiority is more obvious when training samples are limited; (2) we show that additional spatial information is very useful for character representation, especially for Chinese ones, which have more complex structures; and (3) the results also imply the potential of FV to represent new unseen categories, which is quite inspiring since it is very difficult to collect enough training samples for Chinese scene characters which have tens of thousands of categories.

The rest of the paper is organized as follows. Section 2 briefly reviews the related work of scene character recognition. Section 3 introduces the formulation of FV and how we use it to represent characters. Experimental results and analysis are given in Section 4 and conclusions are drawn in Section 5.

2. Related work

In this section, we will review recently published work on recognition of scene characters, which share the same hurdles as generic object recognition—large intra-class variations and complex background. Since FV could be seen as extension of the BOW representation, we will also give some brief introduction of BOW and review the related work of character recognition using BOW framework. We list the overview of recently proposed methods in Table 1 and details are given below.

2.1. Handcrafted and learning based representation for characters

A good feature representation is essential to the final recognition result and researchers have tried various features for scene character recognition, including the handcrafted and unsupervised learning ones. Wang and Belongie [3] proposed to use Histograms of Oriented Gradients (HOG) [29] in conjunction with an NN classifier and reported better performance than conventional OCR. Newell and Griffin [23] presented two extensions of HOG descriptor to include features at multiple scales and their method achieved promising performance on two datasets, Chars74k [14] and ICDAR03-CH [21]. Tian et al. [24] used co-occurrence of histogram of oriented gradients to recognize scene characters and reported better results than HOG. Coates et al. [25] took an unsupervised approach to learn features from unlabeled data and the character recognition results on the ICDAR03-CH [21] are quite promising. Netzer et al. [1] adopted unsupervised feature learning to recognize digits in natural scenes and experimental results demonstrated the major advantages of learned representations over hand crafted ones.

Deep learning has achieved significant performance gain in image classification [30], speech recognition [31] and character recognition [26,32–35], especially in the presence of large amount of training data. For character recognition, Wang et al. [13] used convolutional neural networks (CNN) to recognize English and digits characters in natural scene images and achieved satisfactory performance when using the original training set as well as those synthetic ones. Alessandro et al. [8] adopted deep fully connected neural networks with five hidden layers to recognize scene characters, based on which they built an effective text recognition system. Wan et al. [26] used dropconnect to regularize neural networks and achieved satisfactory performance on SVHN dataset. Lee et al. [32] further proposed the deeply-supervised nets (DSN) and achieved extraordinary performance on MNIST and SVHN datasets. Although deep learning has shown promising performance for character recognition, their effectiveness depends on the large amount of training samples. For scene characters, it is very difficult to collect so many training samples, especially for Chinese characters, which have tens of thousands categories.

2.2. BOW based representation for characters

One of the most popular approaches to image classification is to describe images with bag-of-visual-words (BOW) histograms [18]. In a nutshell, for the BOW representation of an image, local descriptors are extracted from the image and each descriptor is assigned to its closest visual word in a “visual vocabulary”—a codebook obtained offline by clustering a large set of descriptors with k-means. Any classifier can then be used for the categorization of this histogram representation. Based on this framework, De Campos et al. [14] benchmarked the performance of various features to assess the feasibility of posing the character recognition problem as an object recognition task and showed that Geometric Blur [36] and Shape Context [37] in conjunction with Nearest Neighbor (NN) classifier, performed better than other methods. Yi et al. [15] compared different sampling methods, feature descriptors, dictionary sizes, coding and pooling schemes for character recognition and showed that Global HOG achieved better performance than local feature representations. Recently, Gao et al. [27] and Shi et al. [28] proposed to learn spatiality embedded dictionary (SED) to represent characters. Different from conventional codeword learning methods, which used k-means to cluster the descriptors regardless of their positions, SED learned position related codewords by associating each codeword with a response region. Experimental results demonstrate its effectiveness for character recognition. However, SED only used histograms of word counts to represent characters and the k-means based dictionary learning might not be as good as GMM, which is not a problem for FV derived from GMM based visual vocabularies as introduced in this paper.