J. Vis. Commun. Image R. 25 (2014) 1387–1398

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Beyond visual word ambiguity: Weighted local feature encoding with governing region

Chunjie Zhang [a], Xian Xiao [b,*], Junbiao Pang [c], Chao Liang [d], Yifan Zhang [e], Qingming Huang [a,f]

[a] School of Computer and Control Engineering, University of Chinese Academy of Sciences, 100049 Beijing, China
[b] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[c] College of Computer Science and Technology, Beijing University of Technology, 100124 Beijing, China
[d] National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, 430072 Wuhan, China
[e] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P.O. Box 2728, Beijing, China
[f] Key Lab of Intell. Info. Process, Institute of Computing Technology, Chinese Academy of Sciences, 100190 Beijing, China

ARTICLE INFO

Article history:
Received 16 October 2013
Accepted 23 May 2014
Available online 12 June 2014

Keywords:
Visual word ambiguity
Governing region
Weighted encoding
Image classification
Sparse
Bag-of-visual words
Object categorization
Locality constraint

ABSTRACT

Typically, *k*-means clustering or sparse coding is used for codebook generation in the bag-of-visual words (BoW) model. Local features are then encoded by calculating their similarities with visual words. However, some useful information is lost during this process. To make use of this information, in this paper, we propose a novel image representation method by going one step beyond visual word ambiguity and consider the governing regions of visual words. For each visual application, the weights of local features are determined by the corresponding visual application classifiers. Each weighted local feature is then encoded not only by considering its similarities with visual words, but also by visual words' governing regions. Besides, locality constraint is also imposed for efficient encoding. A weighted feature sign search algorithm is proposed to solve the problem. We conduct image classification experiments on several public datasets to demonstrate the effectiveness of the proposed method.

1. Introduction

The bag-of-visual words (BoW) model has been very popular for image applications in recent years. It is inspired by the bag-of-words model for text analysis. Basically, the BoW model can be divided into four steps. First, it selects image regions and extracts local features to represent these regions. Second, a codebook is generated either by *k*-means clustering [1] or sparse coding [2]. Third, each local feature is then encoded accordingly to form image representation. Finally, classifiers are trained to predict the classes of images.

Of the four modules, the generation of codebook and local feature encoding modules are very important for the final image representation. A codebook is a collection of patterns which are used for local feature encoding. Usually, *k*-means clustering is used and local features are quantized by nearest neighbor assignment. However, as pointed out by many researchers [3–10], this hard assignment strategy causes sever information loss, especially when local features lie on the boundary of visual words. To reduce the information loss during local feature encoding process, a lot of works [3–10] have been done by softly assigning local features, such as kernel codebook [3] and sparse coding [2].

Although proven effective, there are some problems with previous codebook generation and local feature encoding strategies. First, each local feature is treated equally for codebook generation and image representation. The generation of codebook and encoding of local features should be task dependent for efficient image representation. Second, the extracted local features are often too many to be used for codebook generation due to the computational power and memory usage required. To solve this problem, researchers choose to randomly select local features for codebook generation. The cluster centers are viewed as visual words. However, during the new local feature encoding process, only the visual words are used. This strategy discards the useful information of local features during the codebook generation process [8]. In fact, what we learned during the codebook generation is a partition of local feature space, not merely cluster centers. We believe this information should also be made use of for efficient codebook learning and local feature encoding. Third, many previous practices only try to reduce the information loss during local feature encoding without considering the codebook generation process. If we can

[*] Corresponding author.
E-mail addresses: cjzhang@jdl.ac.cn (C. Zhang), xaioxain@gmail.com (X. Xiao), junbiao_pang@bjut.edu.cn (J. Pang), cliang@whu.edu.cn (C. Liang), yfzhang@nlpr.ia.ac.cn (Y. Zhang), qmhuang@jdl.ac.cn (Q. Huang).

http://dx.doi.org/10.1016/j.jvcir.2014.05.010
1047-3203/© 2014 Elsevier Inc. All rights reserved.

generate the codebook and encode local features by jointly reducing the information loss, we will be able to get more representative codebook and encoding parameters simultaneously.

To preserve the spatial information of local features, spatial pyramid matching (SPM) is introduced by Lazebnik et al. [11] and is widely used by researchers. This method partitions an image into increasingly finer sub-regions of different scales with $2^l \times 2^l,\ l = 0, 1, 2$. The histogram representation of each sub-region is then concatenated to form the final image representation. Many works [12–16] have demonstrated the effectiveness of this method for visual applications, such as image classification, image/video retrieval. In cases only the first scale $l = 0$ is used, SPM will reduce to BoW. Researchers have empirically found that k-means clustering based codebook generation method should be combined with SPM and non-linear kernels for efficient image classification. However, the computational cost is high. To speed up the computation and improve the performance, Yang et al. [2] proposed to use sparse coding with linear SVM classifier instead. Inspired by this, a lot of works were proposed [4–6] which further improve the performance. However, local features are still treated equally. Besides, the partition of local feature space information during the codebook generation process is still not considered by these methods.

To solve the problems mentioned above, in this paper, we propose a novel image representation method by going one step beyond visual word ambiguity and consider the governing regions of visual words generated by weighted local features. Each local feature is weighted according to specific visual application task by using the output of pre-trained traditional classifiers [1]. Each weighted local feature is then encoded not only by considering its similarities with visual words, but also by visual words' governing regions. A visual word's governing region is defined as the area covered by local features that are encoded with this visual word. We first find the local features nearby used during the codebook generation process, each weighted local feature is encoded by considering its similarities between visual words and the coding parameters of these nearby neighbors. This governing region based encoding scheme is consistently used both during the codebook generation and local feature encoding processes. We propose a weighted feature sign search algorithm to solve this governing region based weighted sparse coding problem. Experiments on several public datasets demonstrate the effectiveness of the proposed governing region based weighted sparse coding algorithm for image representation.

The rest of this paper is organized as follows. Section 2 gives the related work. In Section 3, we give the details of the proposed weighted local feature encoding by governing region method for image representation and conduct experiments on several public datasets to demonstrate its effectiveness in Section 4. Finally, we conclude in Section 5.

2. Related work

Inspired by text analysis, the bag-of-visual word model (BoW) [1] has been widely used for visual applications. One important component of the BoW model is the generation of codebook and encoding of local features. Traditional BoW model used the k-means clustering algorithm and viewed the cluster centers as visual words. Local features are then quantized to the nearest visual word. However, this scheme causes sever information loss which limits its discriminative power. To reduce the information loss during the local feature encoding process, a lot of works have been proposed [2–10]. Yang et al. [2] proposed to use sparse coding for codebook generation. Linear classifier was then trained to save computational cost. Gemert et al. [3] explored the visual word ambiguity problem and tried to quantize local features softly with codebook uncertainty. Wang et al. [4] found that locality is also very important and proposed the locality constrained linear coding algorithm which can be viewed as an extension of the sparse coding algorithm [2]. To consider the smooth property of sparse coding parameters, Gao et al. [5] proposed a Laplacian sparse coding algorithm and improved the image classification performance. The sparse coding along with max pooling strategy is suboptimal because negative coding parameters are useless for image representation. To solve this problem, non-negative sparse coding along with max pooling is proposed by Zhang et al. [6]. Liu et al. [7] conducted experiments on several public datasets and found soft assignment is very necessary for local feature encoding. Perronnin and Dance [8] proposed to use the fisher kernel for codebook construction which extended the BoW model by going beyond counting the zero order statistics and tried to encode second order statistics. Jurie and Triggs [9] proposed an acceptance-radius based clustering method to generate better codebooks than k-means clustering. To avoid the heavy computation of k-means clustering, Randomized clustering forest was proposed by Moosmann et al. [10] which speeded up the computation and improved the final image classification performance.

The traditional BoW model lacks the spatial information, inspired by the work of Grauman and Darrell [12] and Lazebnik et al. [11] proposed the spatial pyramid matching (SPM) algorithm which was widely used by researchers. Motivated by the SPM algorithm, a lot of works [13–16] have been done to combine the spatial information of local features. Chen et al. [13] proposed a hierarchical matching method with side information and used it for image classification. A weighting scheme was used to select discriminative visual words. Yao et al. [14] combined randomization and discrimination into a unified framework and used it for fine-grained image categorization. Zhang et al. [15] proposed to represent images with components and used a bilinear model for object recognition. A pose pooling kernel was proposed by Zhang et al. [16] to recognize sub-category birds.

To avoid the information loss, researchers also tried to classify images by using local features directly [17–19]. Boiman et al. [17] proposed to measure the similarities of local features and used it for image classification with encouraging results. To speed up the computation, Timofte and Gool [18] proposed to iteratively search for the nearest neighbors and used it for image classification and dimensionality reduction. A local naive Bayes nearest neighbor method [19] was also proposed by McCann and Lower. The use of neighbor information speeds up the computation and also helps to improve classification performance. In fact, the use of local neighbor information has been widely used for visual applications, such as locality constrained linear coding [4], Laplacian sparse coding [5] and kernel codebook [3].

The use of sparsity [20] has also been very popular in recent years. Sparse coding [21] is used by Yang et al. [2] for image classification. Sparse representation is used for robust face recognition by Wright et al. [22]. Yang et al. [23] combined fisher discrimination information with sparse coding to learn the dictionary for face, digit and gender classification. Structured sparse representation is also proposed by Elhamifar and Vidal [24] for robust face classification.

Researchers [13,25–33] have also tried to treat local features differently which has shown its effectiveness for various image classification tasks. Chen et al. [13] combined object confidence map and visual saliency map for side information which achieved good image classification performance. Fernando et al. [25] tried to fuse discriminative feature with a regression model to boost image classification performance. Cinbis et al. [26] viewed local features as non-iid sampled and used the fisher kernel to classify images. Sharma et al. [27] also explored discriminative saliency with spatial information. Moosmann et al. [28] tried to learn the saliency