J. Vis. Commun. Image R. 40 (2016) 24–33

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

A novel method for image classification based on bag of visual words ☆

Ronggui Wang, Kai Ding, Juan Yang *, Lixia Xue

School of Computer and Information, Hefei University of Technology, 230009 Hefei, China

ARTICLE INFO

Article history:
Received 27 January 2016
Revised 5 May 2016
Accepted 30 May 2016
Available online 4 June 2016

Keywords:
BOW
Saliency
Topological structure
Delaunay triangulation
Image classification

ABSTRACT

Bag of words (BOW) model is widely applied in image classification. The traditional BOW neglects the spatial information and object shape information, which fails to distinguish between image features absolutely. In this paper, we combine the salient region with visual words topological structure. It not only can produce more representative visual words, but also can avoid the disturbance of complex background efficiently. Firstly, the salient regions are extracted and the BOW model is built on salient regions. Secondly, in order to describe the characteristics of the image more accurately and to resist the influence of background information, the visual words topological structure and Delaunay triangulation method are employed, which is able to integrate into the global and local information. The performance of the proposed algorithm is tested on several datasets, and compared with other models. The experiment results seem to demonstrate that the proposed method provide a higher classification accuracy.

1. Introduction

With the rapid development of Internet and multimedia technology, the number of digital images, videos and other multimedia information is exploding. In order to organize, manage and retrieve images accurately and efficiently, computers are required to understand the content of images accurately. Image classification is an important approach to solve the problem of image understanding which also plays an important role in promoting the development of multimedia retrieval technology. In addition, as the basis of image processing, image features representation is the key study content in this field, as its performance affects the results of image classification and object recognition directly.

The bag of visual words model proposed by Sivic and Zisserman [1] is widely used in the classification of objects and scene, and it has showed excellent performance. The model is originally inspired by related concepts and ideas in the field of text analysis which has been used in text mining, it does not need to analyze the composition of specific objectives in the image except the overall statistics of the image. The low-level image features are quantized as visual words to express the image content through the distribution of visual words.

In recent years, many state of the art technologies of image feature representation based on the bag of words model are proposed. Lazebnik et al. [2] proposed spatial pyramid matching method (SPM), which works by repeatedly subdividing an image and computing histograms of image features over the resulting subregions. The color correlograms have been introduced by Huang et al. in [3]. The idea was to represent the distribution of color pairs as a function of the spatial distance between these colors in the image. Savarese et al. extend this representation to the distribution of pairs of visual words in the image plane. Jiang et al. [4] introduced image contextual information to form image features by distinct statistical word pairs and visual word group. In [5], they consider only pairs of visual words located in a local neighborhood whose radius is related to the scale of the associated SIFT keypoint and use the supervised TF-IDF information for selection of the best visual phrases. In addition, the descriptive visual words and descriptive visual phrases are proposed as the visual correspondences to text words and phrases by Zhang et al. [6], where each visual word co-occurring with the visual word center within certain distance composes a DVP. Liu et al. [7] considered semantic co-occurrence probability of visual words and introduced Markov random field theory to improve the accuracy of visual semantic word. Li et al. [8] proposed a new algorithm, named CBOW, modeling the conceptual context and neighboring context of local patches based on the BOW. Sivic et al. [9] select the 100 most frequent visual words and each visual word is paired with its 5 nearest neighbor in the image space. Deng et al. [10] proposed a graph assignment method with maximal mutual information (GAMI) regularization. GAMI takes the power of manifold structure to better reveal the relationship of massive number of local features by nonlinear graph metric. Yang et al. [11] used an extension of the SPM

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author at: School of Computer and Information, Hefei University of Technology, Hefei, Anhui 230009, China.
E-mail address: yangjuan6985@163.com (J. Yang).

http://dx.doi.org/10.1016/j.jvcir.2016.05.022
1047-3203/

method, by generalizing vector quantization to sparse coding followed by multi-scale spatial max pooling, and propose a linear SPM kernel based on SIFT sparse codes. Although the traditional BOW model achieved good results in image classification, the method assumes that visual words are independent. Therefore, there are limitations exist in image representation. The bag of words model used statistical methods which neglected spatial information between visual words, so we can improve the performance in image classification by adding spatial information to represent more discriminative features.

The further study found that the human vision often focus on a salient region when the object is identified. The disturbances of complex information on background can be avoided by extracting salient regions of the image. In this paper, the visual saliency and visual words similarity topological constraint [12–14,17] based on BOW model are combined for image classification. Firstly, the global constraints are applied to the position of words on salient regions. Secondly, the Delaunay triangulation network is constructed to further constrain the local position information of words. We propose global topological constraint and local position constraint to improve traditional BOW, the proposed method can also improve the efficiency of image classification.

The rest of the paper is organized as follows: in Section 2, the original BOW and a brief introduction to the concepts of saliency detection are presented, respectively. In Section 3, the improved BOW algorithm is proposed based on global topological constraint and local position constraint, besides that, some kernel functions are further modified. The experimental results are demonstrated in Section 4. Finally, Section 5 concludes the paper.

2. Related work

2.1. Image classification method based BOW

Bag of words model was originally applied to classify and identify documents in the field of text processing. The BOW model has been widely used to tackle various applications due to its simplicity and effectiveness. The basic principle is that the document is regarded as a set of disorderly keywords, and then applying to image classification by frequency statistics for each keyword which appears in a single document. The process of image classification based on BOW includes the following three steps: feature extraction and description, construct visual dictionary and training classifier.

The implement of image classification based on BOW can be described as follow:

Step 1: Feature extraction and description. Given an image, the main task is to extract global and local features to describe the image. The SIFT descriptors are applied to achieve this process in the most of traditional methods, each feature is characterized by 128-dimension vector.

Step 2: Construct visual dictionary. The k-means clustering algorithm is applied to cluster a large number of feature points obtained in Step 1. The cluster center is defined as visual word, all the words are combined as visual dictionary and the size of dictionary is the number of words. In traditional methods, images are commonly represented as the histograms of visual words.

Step 3: Training classifier. SVM is one of the most popular classifier and its implementation is simple in various classifiers. Besides, the main goal of Support Vector Machine is to find a decision plane for separating between a set of objects which have different class memberships. SVM is initially applied to two-class classification problems and now it has been gradually used to solve multi-class high-dimension classification problems. It can be described as the following optimization problem:

$$\min_{\omega,\xi}\left\{\frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i\right\}$$

Subject to the constraints:

$$y_i(\omega \cdot x_i - b) \geqslant 1 - \xi_i, \quad \xi_i \geqslant 0, \quad i = 1, \ldots, n.$$

where ω is the vector of coefficients, C is the capacity constant, b denotes a constant, ξ_i represents parameters for handling nonseparable data. The index i labels the N training cases and $y \in \pm 1$ represents the class labels.

In this paper, the improvement of BOW model with spatial information is introduced. We introduce salient region combined with the spatial information of visual words to improve the traditional bag of words model, using the SVM classifier to optimize image classification.

2.2. Saliency detection

In this section, we introduce a contrast analysis method [15], region contrast, so as to integrate spatial relationships into region-level contrast computation. Firstly, we segment the image into regions, then compute color contrast at the region level, and finally define saliency for each region as the weighted sum of the region's contrasts to all other regions in the image. The weights are set according to the spatial distances with farther regions being assigned smaller weights.

We utilized a regional contrast based salient object detection algorithm, which simultaneously evaluates global contrast differences and spatial weighted coherence scores. Humans pay more attention to image regions with high contrast to their surroundings.

We first segment the image into regions using a graph-based image segmentation method. Then we build the color histogram for each region. For a region r_k, we compute its saliency value by measuring its color contrast to all other regions in the images,

$$S(r_k) = \sum_{r_k \neq r_i} w(r_i) D_r(r_k, r_i) \tag{1}$$

where $w(r_i)$ is the weight of region r_i and $D_r(\cdot,\cdot)$ is the color distance metric between the two regions. We weight the distances by the number of pixels in r_i as $w(r_i)$ to emphasize color contrast to bigger regions. The color distance between two regions r_1 and r_2 is,

$$D_r(r_1, r_2) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} f(c_{1,i}) f(c_{2,j}) D(c_{1,i}, c_{2,j}) \tag{2}$$

where $f(c_{k,i})$ is the probability of the i-th color $c_{k,i}$ among all n_k colors in the kth region r_k, $k = \{1, 2\}$. Note that we use the probability of a color in the probability density function of the region as the weight for this color to further emphasize the color differences between dominant colors.

We further incorporate spatial information by introducing a spatial weighting term in (1) to increase the effects of closer regions and decrease the effective of farther regions. Specifically, for any region r_k, the spatially weighted region contrast based saliency is:

$$S(r_k) = w_s(r_k) \sum_{r_k \neq r_i} e^{\frac{D_s(r_k, r_i)}{-\sigma_s^2}} w(r_i) D_r(r_k, r_i) \tag{3}$$

where $D_s(r_k, r_i)$ is the spatial distance between regions r_k, r_i, σ_s controls the strength of spatial distance weighting, $w(r_i)$ is the weight