Neurocomputing 172 (2016) 271–280

Contents lists available at ScienceDirect

Neurocomputing

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

Object recognition based on the Region of Interest and optimal Bag of Words model

Weisheng Li [a,*], Peng Dong [a], Bin Xiao [a], Lifang Zhou [b]

[a] Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, PR China
[b] College of Software, Chongqing University of Posts and Telecommunications, Chongqing 400065, PR China

ARTICLE INFO

Article history:
Received 30 October 2013
Received in revised form
6 December 2014
Accepted 19 January 2015
Available online 9 May 2015

Keywords:
Object recognition
Region of Interest
Bag of Words model
k-means++ cluster
Gaussian mixture models

ABSTRACT

Bag of Words (BoW) model has been widely used in conventional object recognition tasks. Different from the existing methods, this paper proposed a method for object recognition based on Region of Interest (ROI) and Optimal Bag of Words model. It includes the following steps: (1) ROI extraction in combination with the Shi–Tomasi corner and Itti saliency map; (2) The SIFT feature descriptors are detected and described on images of interest; (3) A visual codebook is generated through utilizing the Gaussian mixture models, which relies on the clustering results of k-means++; (4) The similarities between each visual word and corresponding local feature are computed by posterior pseudo probabilities discriminative to construct a visual word soft histogram for image representation; (5) The Support vector machine (SVM) is used to perform image classification and recognition. The experiments are performed on the MSRC 21-class database. The results show that the proposed method can be more accurately recognize images.

1. Introduction

The problem of finding and analyzing object [1] in the images or video sequences is a fundamental task in computer vision and pattern recognition. Its intention is to recognize unlabeled object in the images or video sequences. As an intermediate step to bridge the semantic gap between an image and its content meaning (object), object recognition technique plays an important role in the applications of intelligent video-surveillance [2], image retrieval [3], vehicle assistant driving [4], and so on. In recent years, the cutting-edge research on object recognition focuses on how to effectively classify and identify the large-scale object categories. Though extensive research works have been devoted to object recognition, it still remains challenging, due to tremendous variations in object appearance caused by viewpoint, illumination, background clutter, partial occlusion, geometric distortion, intra-class variation, and so on. All these factors make the object recognition be a great challenge in the analysis, modeling, learning, and recognition.

Over the last decade, in order to find an efficient method of object recognition, researchers in the field of computer vision succeed in introducing Bag of Words (BoW), which has ever been applied to the fields of document processing to object recognition field. Among various object recognition methods, the methods based on BoW have been paid much attention due to its low computation cost, robustness against illumination variation, partial occlusion and clutter background. The idea of BoW is inspired by the success of textual words in natural language processing, where documents are represented by a bag of textual words from a vocabulary dictionary. Analogously, an image can be treated as a document to detect and describe its features. Generally speaking, the first step extracts a large number of features so that codebook can be constructed; the second step uses a visual word histogram to represent image; the last step designs the classifier for the recognition task. Indeed, BoW method compares favorably to others because of its simpleness and effectiveness.

Although BoW method has achieved appealing results, various drawbacks of the method still exist. Moreover, some of them are innate, such as never distinguishing the foreground from the background of image, ignoring the relationship among features, rarely crossing the “semantic gap”, etc. Traditional BoW model has commonly adopted used the scale-invariant feature transform (SIFT), k-means clustering algorithm and Support vector machine (SVM). Recently, great strides have been made in BoW framework to improve the discriminative and descriptive power by many ways, which have been reported on papers [5–13]. Zhang et al. [5] compared some common feature detectors, descriptors, and support

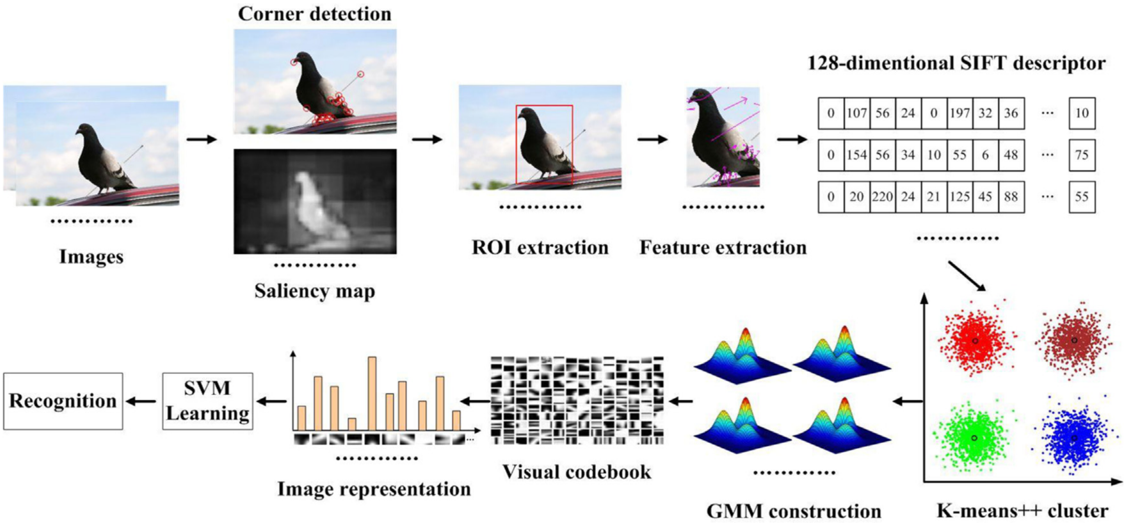
* Corresponding author. Tel.: +86 23 62471342.
E-mail addresses: liws@cqupt.edu.cn (W. Li), clasdong@qq.com (P. Dong), xiaobin@cqupt.edu.cn (B. Xiao), zhoulf@cqupt.edu.cn (L. Zhou).

http://dx.doi.org/10.1016/j.neucom.2015.01.083
0925-2312/© 2015 Elsevier B.V. All rights reserved.

vector machine kernel functions. Zhu et al. [6] involved a computationally expensive training phase to generate an effective descriptor of object recognition in natural scenes. Deselaers et al. [7] introduced principal component analysis to reduce SIFT descriptors dimensions, and employed unsupervised training of Gaussian mixture model (GMM) to create visual vocabulary. Wu et al. [8] demonstrated that Histogram Intersection Kernel could also be used in an unsupervised manner to significantly improve the generation of visual codebooks. Based on SIFT descriptors, Lienhart et al. [9] used probabilistic Latent Semantic Analysis (PLSA) for images recognition. Wang et al. [10] proposed a visual word soft-histogram to represent image by using statistical modeling and discriminative learning of visual words. Jiang et al. [11] investigated different representation choices, such as vocabulary size, weighting scheme, stop word removal, feature selection, and kernel selection in SVM, and then provided some very practical insights in algorithm design. Hong et al. [12] proposed an MIL method with discriminative feature mapping and feature selection, which is able to explore both the positive and negative concept correlations. It also can select the effective features from a large and diverse set of low-level features for each concept under MIL settings. Wang et al. [13] proposed a scheme that is able to automatically turn a movie clip to comics. The scheme mainly contains three components: script-face mapping, descriptive picture extraction, and cartoonization. These methods just pay attention to the whole image while ignore the local target object in an image. In practice, the goal of object recognition is only a portion of the whole image, which should be taken into account. Otherwise, detection approach will be relatively time-consuming, and susceptible to interference of complex background. As human is inclined to pay more attention to regions of interest (ROI), visual attention model is a feasible method to find ROI and measure the degree of visual interest to the ROI of the image. Harris et al. [14] proposed a robust interest points detection method, which could reduce the effect of image rotation, translation, illumination variation, etc. Tomasi et al. [15] improved the performance of Harris corner detector and proposed the Shi–Tomasi corner detector method. Itti et al. [16,17] computed color, intensity and orientation feature maps to obtain visual saliency of an image. In the method, ROI are consistent with the human visual system and real object. The Region of Interest (ROI) extraction is widely employed in image retrieval and compression.

In conclusion, most of existing methods apply a scale-invariant feature transform (SIFT) descriptor in the image pixel domain and a k-means cluster in the visual codebook to create the BoW. The foreground and background of an image cannot be well distinguished by these methods. In addition, the image representation only matches each local feature vector to the only visual word, which ignores the similarity between them. In this paper, a novel approach for object recognition based on Region of Interest and optimal Bag of Words model is proposed. The main steps are summarized as follows: Firstly, before the image feature extraction, Region of Interest extraction based on the advantages of Shi–Tomasi corner and Itti Saliency Map makes the image SIFT features more representative, to a certain extent, eliminates the influence of image foreground and background confusion and cluttered scene. In fact, the experimental results show that the object recognition is more accurate after Region of Interest extraction. Secondly, we adopt k-means++ clustering algorithm in the Visual Codebook generation, which can overcome the shortcomings of k-means clustering algorithm that is heavily dependent on the initial cluster centers. After that, we use Gaussian mixture model (GMM) to model each cluster feature set as a visual word. The experimental results show that the method can more effectively improve the distinguish ability of different visual words and further measure uncertainty of the visual words than traditional visual codebook generation methods. Finally, in the image representation, we compute the similarities between each visual word and corresponding local features using posterior pseudo probabilities discriminative to construct a visual soft histogram for image representation. The traditional Term Frequency methods ignore the similarity between local features and visual words, simply put the distances of different eigenvectors from the clustering center as main foundation. The experimental results demonstrate that our approach has certain advantages over contrastive methods on image representation.

The detailed process is shown in Fig. 1; there are five steps in this approach: (1) ROI is extracted to select the interest image; (2) The SIFT feature descriptors are detected and described in the interest images. These highly distinctive features can distinguish the foreground from background of the image, and reduce the effects of the cluttered backgrounds; (3) a visual codebook is generated utilizing the Gaussian mixture models (GMM, which relies on k-means++ clustering results). It is more stable and effective than single k-means algorithm; (4) The similarities between each visual word and corresponding local feature is computed by posterior pseudo probabilities discriminative to construct a visual word soft histogram for image representation; (5) The Support vector machine (SVM) is used to perform image classification and recognition.



0	107	56	24	0	197	32	36	···	10
0	154	56	34	10	55	6	48	···	75
0	20	220	24	21	125	45	88	···	55

Fig. 1. Process of the method for object recognition based on the combination of ROI and optimal BoW model.