Information Sciences 329 (2016) 33–51

Contents lists available at ScienceDirect

Information Sciences

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

Keypoint selection for efficient bag-of-words feature generation and effective image classification

Wei-Chao Lin [a], Chih-Fong Tsai [b,*], Zong-Yao Chen [b], Shih-Wen Ke [c]

[a] Department of Computer Science and Information Engineering, Hwa Hsia University of Technology, Taiwan
[b] Department of Information Management, National Central University, Taiwan
[c] Department of Information and Computer Engineering, Chung Yuan Christian University, Taiwan

ARTICLE INFO

Article history:
Received 19 May 2014
Revised 25 June 2015
Accepted 16 August 2015
Available online 21 August 2015

Keywords:
Bag-of-words
Object categorization
Image classification
Keypoint selection

ABSTRACT

One of the most popular image representations for image classification is based on the bag-of-words (BoW) features. However, the number of keypoints that need to be detected from images to generate the BoW features is usually very large, which causes two problems. First, the computational cost during the vector quantization step is high. Second, some of the detected keypoints are not helpful for recognition. To resolve these limitations, we introduce a framework, called iterative keypoint selection (IKS), with which to select representative keypoints for accelerating the computational time to generate the BoW features, leading to more discriminative feature representation. Each iteration in IKS is comprised of two steps. In the first step some representative keypoint(s) are identified from each image. Then, the keypoints are filtered out if the distances between them and the identified representative keypoint(s) are less than a pre-defined distance. The iteration process continues until no unrepresentative keypoints can be found. Two specific approaches are proposed to perform the first step of IKS. IKS1 focuses on randomly selecting one representative keypoint and IKS2 is based on a clustering algorithm in which the representative keypoints are the closest points to their cluster centers. Experiments carried out based on the Caltech 101, Caltech 256, and PASCAL 2007 datasets demonstrate that performing keypoint selection using IKS1 and IKS2 to generate both the BoW and spatial-based BoW features allows the support vector machine (SVM) classifier to provide better classification accuracy than with the baseline features without keypoint selection. However, it is found that the computational cost of IKS1 is larger than the baseline methods. On the other hand, IKS2 is able to not only efficiently generate the BoW and spatial-based features that reduce the computational time for vector quantization over these datasets, but also provides better classification results than IKS1 over the PASCAL 2007 and Caltech 256 datasets.

1. Introduction

In computer vision, object recognition (or categorization) is a major research problem whose aim is to classify an individual object into a specific category, such as a face, a car, a dog, or the like. This can also be regarded as an image classification problem. In order to effectively find a given object in an image or video sequence, image features should be well extracted and represented as image descriptors to describe image objects.

* Corresponding author. Tel.: +886 3 422 7151; fax: +886 3 425 4604.
E-mail address: cftsai@mgt.ncu.edu.tw (C.-F. Tsai).

http://dx.doi.org/10.1016/j.ins.2015.08.021
0020-0255/

The bag-of-words (BoW) model, a well-known and popular feature representation method for document representation in information retrieval, was first applied in the field of image and video retrieval by Sivic and Zisserman [36]. This method has generally shown promise for object categorization [10,17,23,35] as well as image annotation and retrieval tasks [9,13,16,41].

The BoW feature is usually based on the utilization of a tokenizing keypoint-based feature, e.g., scale-invariant feature transform (SIFT) [21], to generate a visual-word vocabulary (or codebook). The visual-word vector of an image conveys the presence or absence of the information for each visual word in the image (e.g., the number of keypoints in the corresponding cluster).

The visual word can be defined as follows. Given a training dataset D containing n images represented by $D = d_1, d_2,\ldots,$ and d_n where d represents the extracted visual features. A specific unsupervised learning algorithm, such as the k-means, is used to group D based on a fixed number of visual words W (or categories) represented by $W = w_1, w_2,\ldots,$ and w_V, where V is the cluster number. Then, we can summarize the data in a $V \times N$ co-occurrence table of counts $N_{ij} = n(w_i, d_j)$, where $n(w_i, d_j)$ denotes how often the word w_i occurs in an image d_i.

However, Van de Sande et al. [38] have shown a severe drawback to the bag-of-words model, that is its high computational cost in the quantization step. In other words, the most expensive part in a state-of-the-art setup of the bag-of-words model is the vector quantization step, i.e., finding the closest cluster for each data point in the k-means algorithm.

For example, with n SIFT descriptors of length d in an image, the quantization against a codebook with m elements requires the full $(n \times m)$ distance matrix between all descriptors and codebook elements. For values which are common for visual categorization, $n = 10{,}000$, and $d = 128$ and codebook size $m = 4000$, it takes approximately 5 s per image for CPU implementation, as the complexity is $O(ndm)$ per image [38].

Van de Sande et al. [38] has already studied how to accelerate the vector quantization step using normal codebooks for the graphics processing unit (GPU). The problem we focus on here is how to select more representative keypoints from each image so as to reduce the total number of original keypoints detected in a given training dataset. The number of keypoint descriptors is usually very large over existing benchmarks. For example, there are over 20 million keypoint descriptors in the MIRFLICKR-25000 dataset.[1]

Selecting detected keypoints from images is not new in object recognition, but very few studies have focused on the BoW scenario (c.f., Section 2). It is also a known that many detected keypoints are not helpful for better recognition. There are two advantages if the most discriminative keypoints could be identified. First, the vector quantization step is likely to be accelerated by the same codebook size with m elements. In addition, the codebook size could even be reduced, e.g., to half of m, which would result in much less computational complexity, because the total number of selected keypoints would be much smaller than with the total number of original keypoints. Therefore, the codebook size is not necessarily the same as m which, as a result, would make the model learning more efficient. Second, image feature representation after keypoint selection can provide more discriminative power than the process without keypoint selection. This is because many unrepresentative keypoints are filtered out, such as the keypoints belonging to the background and ones that are common to many images and objects.

In this paper, we introduce two approaches for keypoint selection, based on the proposed framework for iterative keypoint selection (IKS), which we call IKS1 and IKS2. IKS is very simple and easy to implement. Simply speaking, given a set of keypoint descriptors extracted from an image, the aim of IKS is, first of all, to identify some representative keypoints through random selection (IKS1) or k-means clustering (IKS2). Keypoints are filtered out if their distance from an identified keypoint is less than a pre-defined threshold. This occurs because they are close together in the feature space. We argue that many of these similar keypoints do not need to be considered when generating the BoW features. Next, these two steps are performed over the remaining set of keypoint descriptors. The representative keypoints identified previously are repeatedly examined until no more keypoints can be filtered out (c.f. Section 3.2).

The main contributions of this paper are summarized below:

- We present a novel keypoint selection scheme that enables us to formulate the BoW feature generation problem as an instance selection-like problem, the aim of which is to reduce the size of the training set by filtering out redundant data. In keypoint selection, we aim to select representative keypoints from a given training image. (Please see the discussion about the difference between keypoint selection and instance selection in Section 3.1.)
- The experimental results, based on two different image datasets (i.e., Caltech 101 and Caltech 256), demonstrate that this method of performing keypoint selection by IKS to generate the BoW features outperforms that which uses the baseline BoW features without keypoint selection. In particular, we show that keeping too many similar keypoints per image, as in the non-keypoint selection baseline approach, not only increases the computational cost to generate the BoW features, but also degrades the classification accuracy compared to the BoW features with keypoint selection.

The rest of this paper is organized as follows. A review of related studies is given in Section 2. Section 3 defines keypoint selection and introduces our proposed approaches. Section 4 presents the experimental setup and results. The conclusion and suggestions for future work are provided in Section 5.

[1] http://press.liacs.nl/mirflickr/.