Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

An experimental study on the universality of visual vocabularies

Jian Hou[a,*], Wei-Xue Liu[a], Xu E[a], Qi Xia[b], Nai-Ming Qi[b]

[a] *School of Information Science and Technology, Bohai University, Jinzhou 121013, China*
[b] *School of Astronautics, Harbin Institute of Technology, Harbin 150001, China*

ARTICLE INFO

Article history:
Received 30 November 2012
Accepted 2 August 2013
Available online 14 August 2013

Keywords:
Bag-of-visual-words
Visual vocabulary
Universal
Bag-of-words
Classification
Experimental study
Image representation
Visual features

ABSTRACT

Bag-of-visual-words has been shown to be a powerful image representation and attained success in many computer vision and pattern recognition applications. Usually for a given classification task, researchers choose to build a specific visual vocabulary, and the problem of building a universal visual vocabulary is rarely addressed. In this paper we conduct extensive classification experiments with three features on four image datasets and show that the visual vocabularies built from different datasets can be exchanged without apparent performance loss. Furthermore, we investigate the correlation between the visual vocabularies built from different datasets and find that they are nearly identical, which explains why they are universal across classification tasks. We believe that this work reveals what is behind the universality of visual vocabularies and narrows the gap between bag-of-visual-words and bag-of-words in text domain.

1. Introduction

Bag-of-visual-words is a popular image representation and widely used in various computer vision and pattern recognition applications. Salient image regions (keypoints) are detected from images and described with descriptors, e.g., SIFT [14]. All the descriptors from images are then pooled together and clustered into groups. We treat the centroid of each group as a visual word and obtain a visual vocabulary consisting of all visual words. An image can then be represented as a distribution of all the visual words, i.e., a bag-of-visual-words [21,8].

While various works have been published surrounding bag-of-visual-words, the topic of building a universal visual vocabulary is rarely touched. In this paper by *universal* we mean that a visual vocabulary can be used in different classification tasks, and the classification performances are comparable to the ones obtained with the specifically built vocabularies from the target datasets. For a given classification task, researchers usually need to build a specific visual vocabulary and then use this vocabulary in classification. This is quite different from bag-of-words in text domain, where a universal vocabulary can be used in different classification tasks [24]. Noticing that bag-of-words in text domain is the counterpart of bag-of-visual-words in image domain, in this paper we are interested to find out if it is possible to eliminate this difference and build universal visual vocabularies.

Our work on the universality of visual vocabulary is as follows. Firstly, we conduct extensive classification experiments and find that the visual vocabularies built from different datasets can be exchanged without apparently harming the classification performance. Our experiments with three popular features on four image datasets indicate that the visual vocabularies built from different datasets is universal, and this observation applies to different features with different keypoint strategies. Secondly, we investigate the correlation between the visual vocabularies built from different datasets in order to find out what is behind the universality of these visual vocabularies. As a result, we find that these vocabularies from different datasets are nearly identical, only if the number of images used to build them is large enough. This result explains why the visual vocabularies built from different datasets can be exchanged without apparent performance loss.

This paper is organized as follows. In Section 2 we briefly review some related works on bag-of-visual-words and explain the differences between our work and existing works. Section 3 details our experiments on the universality of visual vocabularies from different datasets. In Section 4 we empirically show that when the number of images is large enough, the vocabularies from different datasets are nearly identical, which explains why these vocabularies are universal across classification tasks. Finally, Section 5 concludes the paper.

2. Related works

Noticing that each visual word describes one image pattern, a bag-of-visual-words is actually a histogram of salient image pat-

[*] Corresponding author.
E-mail address: dr.houjian@gmail.com (J. Hou).

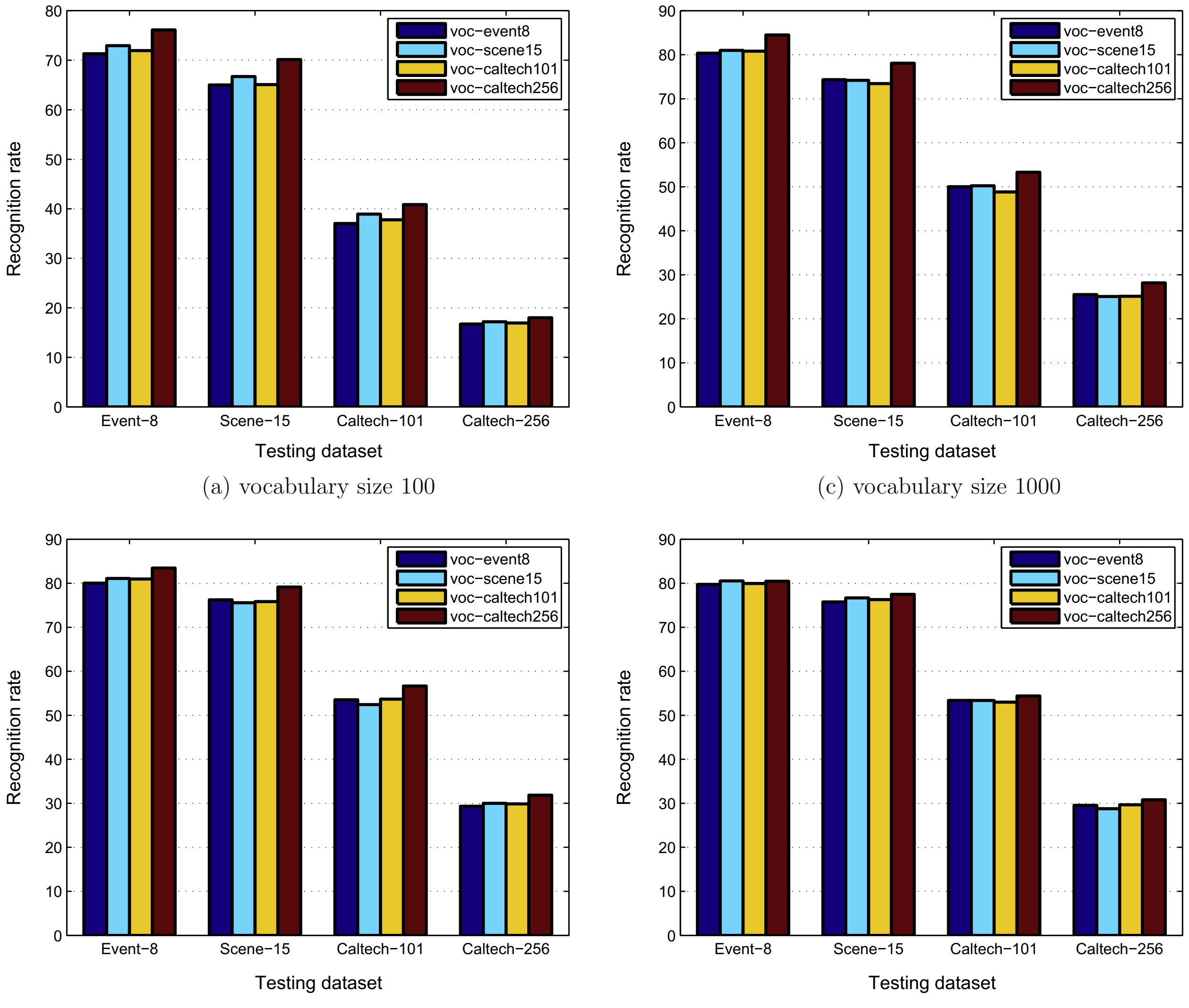
1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.08.002

Table 1
The characteristics of the four image datasets used in experiments.

	# of classes	# of images	# of training	# of testing
Event-8	8	1579	70	60
Scene-15	15	4485	100	The rest
Caltech-101	101	9144	30	15
Caltech-256	256	30,607	40	25

terns in an image. The basic bag-of-visual-words representation captures the distribution of image patterns in the whole image and thus ignores the spatial relationships among keypoints. However, the experiments in [11,24] showed that the spatial distribution of keypoints is very helpful in improving object recognition and classification precision. In order to encode spatial information, [12] proposed to equally partition an image into rectangle regions in a pyramidal manner and compute a bag-of-visual-words histogram in each region. The histograms of all regions are then concentrated into one final description. This spatial pyramid representation is shown to be very discriminative in object classification experiments and has become a standard paradigm in bag-of-visual-words. Besides spatial pyramid, some other approaches have also been proposed to make use of the spatial information [16,22]. Although each visual word in a visual vocabulary describes a certain image pattern, some may be more informative than the others for a certain application. This feature has been exploited to design novel weighting schemes for visual words [17,24,3] and to reduce the vocabulary size for better efficiency [13,15]. In order to adapt the bag-of-visual-words to be used with a large vocabulary and a large dataset, [17] proposed to build a vocabulary tree by hierarchical *k*-means clustering for efficient lookup of visual words.

In literature, the most related works to ours are the experiments in [20,9]. In Ref. [20] the authors investigated the possibility of building a visual vocabulary from generic images and using this vocabulary in different classification tasks. They found that the vocabularies built from different datasets can be exchanged without apparently harming the performance, only if the number of images used to build the vocabularies is large enough. In Ref. [9] we further proposed a method to derive the optimal visual vocabulary for a given dataset, and showed that the optimal visual vocabularies from different datasets are exchangeable. While [20,9] provided strong evidences of the existence of universal visual vocabularies, they left some important problems unsolved.



(a) vocabulary size 100

(c) vocabulary size 1000

(e) vocabulary size 10000

(f) vocabulary size 50000

Fig. 1. Recognition rates using vocabularies built from different datasets, with the SIFT descriptor. x-axis represents different testing datasets, and different bars indicate vocabularies built from different datasets.