Int J Comput Vis
DOI 10.1007/s11263-013-0660-x

Object Bank: An Object-Level Image Representation for High-Level Visual Recognition

Li-Jia Li · Hao Su · Yongwhan Lim · Li Fei-Fei

Received: 2 January 2012 / Accepted: 11 September 2013

Abstract It is a remarkable fact that images are related to objects constituting them. In this paper, we propose to represent images by using objects appearing in them. We introduce the novel concept of object bank (OB), a high-level image representation encoding object appearance and spatial location information in images. OB represents an image based on its response to a large number of pre-trained object detectors, or 'object filters', blind to the testing dataset and visual recognition task. Our OB representation demonstrates promising potential in high level image recognition tasks. It significantly outperforms traditional low level image representations in image classification on various benchmark image datasets by using simple, off-the-shelf classification algorithms such as linear SVM and logistic regression. In this paper, we analyze OB in detail, explaining our design choice of OB for achieving its best potential on different types of datasets. We demonstrate that object bank is a high level representation, from which we can easily discover semantic information of unknown images. We provide guidelines for effectively applying OB to high level image recognition tasks where it could be easily compressed for efficient computation in practice and is very robust to various classifiers.

L.-J. Li (✉)
Yahoo! Research, 701 First Avenue, Sunnyvale, CA 94089, USA
e-mail: lijiali@yahoo-inc.com

H. Su · Y. Lim · L. Fei-Fei
Computer Science Department, Stanford University,
353 Serra Mall St., Stanford, CA 94305, USA
e-mail: haosu@cs.stanford.edu

Y. Lim
e-mail: yongwhan@cs.stanford.edu

L. Fei-Fei
e-mail: feifeili@cs.stanford.edu

Keywords Scene classification · Image representation · Object recognition · Image classification · Image feature

1 Introduction

High-level image recognition is one of the the most challenging domains in the field of computer vision. Any high-level image recognition task using computer vision algorithms starts with image representation, the process of turning pixels into a vector of numbers for further computation and inference. Of all the modules for a robust high-level image understanding system, the design of robust image representation is of fundamental importance and has been attracting many vision researchers. Compared to other data modalities, visual data is particularly challenging because of the extreme richness and diversity of the contents being captured in real world, and the large variability in photometric and geometric changes the real world could map onto the 2D pixel world. Pixels, unlike other information carriers such as words, carry very little meaning themselves, and are extremely volatile to noise. In the past decade, a great amount of research has been conducted on developing robust image representation. Among the image representations widely adopted so far, most of them are low level image representations focusing on describing images by using some variant of image gradients, textures and/or colors [e.g. SIFT (Lowe 1999), filterbanks (Freeman and Adelson 1991; Perona and Malik 1990), GIST (Oliva and Torralba 2001), etc.]. However, there exists a large discrepancy between these low level image representations and the ultimate high level image recognition goals, which is the so called 'Semantic gap'. One way to bridge the semantic gap is by deploying increasingly sophisticated models, such as the probabilistic grammar model (Zhu et al. 2007), compositional random fields (Jin and Geman 2006),

Published online: 22 September 2013