Int J Comput Vis (2013) 105:222–245
DOI 10.1007/s11263-013-0636-x

Image Classification with the Fisher Vector: Theory and Practice

Jorge Sánchez · Florent Perronnin · Thomas Mensink · Jakob Verbeek

Received: 18 January 2013 / Accepted: 28 May 2013 / Published online: 12 June 2013

Abstract A standard approach to describe an image for classification and retrieval purposes is to extract a set of local patch descriptors, encode them into a high dimensional vector and pool them into an image-level signature. The most common patch encoding strategy consists in quantizing the local descriptors into a finite set of prototypical elements. This leads to the popular Bag-of-Visual words representation. In this work, we propose to use the Fisher Kernel framework as an alternative patch encoding strategy: we describe patches by their deviation from an "universal" generative Gaussian mixture model. This representation, which we call Fisher vector has many advantages: it is efficient to compute, it leads to excellent results even with efficient linear classifiers, and it can be compressed with a minimal loss of accuracy using product quantization. We report experimental results on five standard datasets—PASCAL VOC 2007, Caltech 256, SUN 397, ILSVRC 2010 and ImageNet10K—with up to 9M images and 10K classes, showing that the FV framework is a state-of-the-art patch encoding technique.

J. Sánchez (✉)
CIEM-CONICET, FaMAF,
Universidad Nacional de Córdoba,
X5000HUA Córdoba, Argentina
e-mail: jsanchez@famaf.unc.edu.ar

F. Perronnin
Xerox Research Centre Europe, 6 Chemin de Maupertuis,
38240 Meylan, France
e-mail: florent.perronnin@xrce.xerox.com

T. Mensink
Inteligent Systems Lab Amsterdam, University of Amsterdam,
Science Park 904, 1098 XH Amsterdam, The Netherlands
e-mail: thomas.mensink@uva.nl

J. Verbeek
LEAR Team, INRIA Grenoble, 655 Avenue de l'Europe,
38330 Montbonnot, France
e-mail: jakob.verbeek@inria.fr

Keywords Image classification · Large-scale classification · Bag-of-Visual words · Fisher vector · Fisher kernel · Product quantization

1 Introduction

This article considers the image classification problem: given an image, we wish to annotate it with one or multiple keywords corresponding to different semantic classes. We are especially interested in the large-scale setting where one has to deal with a large number of images and classes. Large-scale image classification is a problem which has received an increasing amount of attention over the past few years as larger labeled images datasets have become available to the research community. For instance, as of today, ImageNet[1] consists of more than 14M images of 22K concepts (Deng et al. 2009) and Flickr contains thousands of groups[2]—some of which with hundreds of thousands of pictures—which can be exploited to learn object classifiers (Perronnin et al. 2010c; Wang et al. 2009).

In this work, we describe an image representation which yields high classification accuracy and, yet, is sufficiently efficient for large-scale processing. Here, the term "efficient" includes the cost of computing the representations, the cost of learning the classifiers on these representations as well as the cost of classifying a new image.

By far, the most popular image representation for classification has been the Bag-of-Visual words (BoV) (Csurka et al. 2004). In a nutshell, the BoV consists in extracting a set of local descriptors, such as SIFT descriptors (Lowe 2004), in an image and in assigning each descriptor

[1] http://www.image-net.org

[2] http://www.flickr.com/groups