Int J Comput Vis (2013) 103:163–175
DOI 10.1007/s11263-012-0600-1

Learning Vocabularies over a Fine Quantization

Andrej Mikulik · Michal Perdoch · Ondřej Chum · Jiří Matas

Received: 28 March 2012 / Accepted: 17 November 2012 / Published online: 8 December 2012

Abstract A novel similarity measure for bag-of-words type large scale image retrieval is presented. The similarity function is learned in an unsupervised manner, requires no extra space over the standard bag-of-words method and is more discriminative than both L2-based soft assignment and Hamming embedding. The novel similarity function achieves mean average precision that is superior to any result published in the literature on the standard Oxford 5k, Oxford 105k and Paris datasets/protocols. We study the effect of a fine quantization and very large vocabularies (up to 64 million words) and show that the performance of specific object retrieval increases with the size of the vocabulary. This observation is in contradiction with previously published results. We further demonstrate that the large vocabularies increase the speed of the tf-idf scoring step.

Keywords Image retrieval · Vocabulary · Feature track

A. Mikulik (✉)· M. Perdoch · O. Chum · J. Matas
CMP, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
e-mail: mikulik@cmp.felk.cvut.cz

M. Perdoch
e-mail: predom1@cmp.felk.cvut.cz

O. Chum
e-mail: chum@cmp.felk.cvut.cz

J. Matas
e-mail: matas@cmp.felk.cvut.cz

1 Introduction

Recently, large collections of images have become readily available (Google Street View. http://books.google.com/help/maps/streetview/. Panoramio. http://www.panoramio.com/. Flickr. http://www.flickr.com/) and image-based search in such collections has attracted significant attention of the computer vision community (Sivic and Zisserman 2003; Nister and Stewenius 2006; Chum et al. 2007; Jegou et al. 2008; Perdoch et al. 2009). Most, if not all, recent state-of-the-art methods extend the bag-of-words representation introduced by Sivic and Zisserman (Sivic and Zisserman 2003) who represented the image by a histogram of "visual words", *i.e.*, discretized SIFT descriptors (Lowe 2004). The bag-of-words representation possesses many desirable properties required in large scale retrieval. If implemented as an inverted file, it is compact and supports fast search. It is sufficiently discriminative and yet robust to acquisition "nuisance parameters" like illumination and viewpoint change as well as occlusion [1] .

Discretization of SIFT features is necessary in large scale problems as it is neither possible to compute distances on descriptors efficiently nor feasible to store all the descriptors. Instead, only the identifier of the vector quantized prototype for visual word is kept. After quantization, Euclidean distance in a high (128) dimensional space is approximated by a 0–∞ pseudo-metric—features represented by the same visual word are deemed identical, while the others are treated as "totally different".

[1] We only consider and compare with methods that support queries that cover only a (small) part of the test image. Global methods like GIST (Oliva and Torralba 2006) achieve a much smaller memory footprint at the cost of allowing whole image queries only.