Journal of King Saud University – Science (2014) 26, 89–100

King Saud University
Journal of King Saud University – Science
www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Ensemble of different local descriptors, codebook generation methods and subwindow configurations for building a reliable computer vision system

Loris Nanni [a,*], Alessandra Lumini [b], Sheryl Brahnam [c]

[a] DEI, University of Padua, viale Gradenigo 6, Padua, Italy
[b] DISI, Università di Bologna, Via Venezia 52, 47521 Cesena, Italy
[c] Computer Information Systems, Missouri State University, 901 S. National, Springfield, MO 65804, USA

Received 24 June 2013; accepted 6 November 2013
Available online 18 November 2013

KEYWORDS
Object recognition;
Bag-of-features;
Texture descriptors;
Machine learning;
Support vector machine;
Usage scenarios

Abstract In the last few years, several ensemble approaches have been proposed for building high performance systems for computer vision. In this paper we propose a system that incorporates several perturbation approaches and descriptors for a generic computer vision system. Some of the approaches we investigate include using different global and bag-of-feature-based descriptors, different clusterings for codebook creations, and different subspace projections for reducing the dimensionality of the descriptors extracted from each region. The basic classifier used in our ensembles is the Support Vector Machine. The ensemble decisions are combined by sum rule. The robustness of our generic system is tested across several domains using popular benchmark datasets in object classification, scene recognition, and building recognition. Of particular interest are tests using the new VOC2012 database where we obtain an average precision of 88.7 (we submitted a simplified version of our system to the person classification-object contest to compare our approach with the true state-of-the-art in 2012). Our experimental section shows that we have succeeded in obtaining our goal of a high performing generic object classification system.

The MATLAB code of our system will be publicly available at http://www.dei.unipd.it/wdyn/?IDsezione=3314&IDgruppo_pass=124&preview=. Our free MATLAB toolbox can be used to verify the results of our system. We also hope that our toolbox will serve as the foundation for further explorations by other researchers in the computer vision field.

* Corresponding author. Tel.: +39 0547 339121; fax: +39 0547 338890.
E-mail addresses: loris.nanni@unibo.it, loris.nanni@unipd.it (L. Nanni).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

1. Introduction

Given the vast amount of data being collected machine analysis of image content is imperative (Müller et al., 2004; Lew et al., 2006), a key issue is finding effective feature representations for images. Early systems developed in the 1990s, e.g., Candid (Kelly et al., 1995), Photobook (Pentland et al., 1996), and

1018-3647
http://dx.doi.org/10.1016/j.jksus.2013.11.001

Nextra (Ma et al., 1997), exploited simple global features based on image color, texture, and shape. Approaches around the turn of the century, e.g. (Li et al., 2003; Fergus et al., 2004), focused on constellation models to locate distinctive object parts and to determine constraints on the spatial arrangement. The main drawback of these representations is that they typically are unable to handle significant deformations such as large rotations and occlusions. Moreover, they fail to consider objects, such as trees and buildings, with variable numbers of parts.

More recent systems have taken advantage of new developments in the application of local descriptors in pattern recognition, computer vision, and image retrieval. Of particular importance has been the use of such local features as keypoints and image patches, which have shown great promise in several application areas, including wide baseline matching for stereo pairs (Baumberg, 2000; Tuytelaars and Gool, 2004), object retrieval in videos (Sivic et al., 2004), object recognition (Lowe, 2004), texture recognition (Lazebnik et al., 2005), robot localization (Se et al., 2002), visual data mining (Sivic and Zisserman, 2004), and symmetry detection (Turina et al., 2001). A consensus has emerged from that literature supporting the value of the bag-of-words (BoW) technique for image representation (Lowe, 2004). BoW is based on powerful scale-invariant feature descriptors that are used to match identical regions between images by representing regions in a given image that are covariant to a class of transformations.

Region matching using local image features handles illumination changes, blurring, zoom effects, and many degrees of occlusion and of distortions in perspective. Approaches for region description have been proposed that analyze different aspects of images, such as color, texture, edges, and pixel intensities. Some of the most promising descriptors are those based on histogram distributions (Mikolajczyk and Schmid, 2005). Some important examples of these descriptors include the intensity-domain spin image (Lazebnik et al., 2006), an histogram approach that represents regions using the distance from the center point and intensity values; the SIFT descriptor (Lowe, 2004), an histogram that takes the weighed gradient locations and orientations; and the geodesic intensity histogram (Ling and Jacobs, 2005), a histogram that provides a deformation invariant local descriptor. Other descriptors of this type include PCA-SIFT (Ke and Sukthankar, 2004), moment invariants (Gool et al., 1996), and complex filters (Schaffalitzky and Zisserman, 2002). Some powerful texture descriptors include center-symmetric local binary patterns (CS-LBP) Heikkilä et al., 2009, a LBP-based texture descriptor which is computationally simpler than SIFT and more robust to illumination problems. Another interesting result in region description is reported in Nowak et al. (2006), where it is shown that random sampling, in the case where a large number of regions is available, gives equal or better classification rates than the other more complex operators that are in common use. Some recent effort on visual recognition for very large databases are (Lin et al., 2011; Krizhevsky et al., 2012; Perronnin et al., 2010).

Some recent advances in the problem of building recognition are also noteworthy (Hutchings and Mayol-Cuevas, 2005; Jing and Allinson, 2009). The specific difficulties of this task are the various forms of occlusions encountered (e.g., trees and moving vehicles) and the varying viewpoints in the images. In Hutchings and Mayol-Cuevas, (2005); and Jing and Allinson, (2009) global features (intensity and color information at different scales) and local features (Gabor features at several different scales and orientations) were extracted from a database of building images and used as a powerful feature vector. Moreover, in Jing and Allinson (2009) several subspace learning-based dimensionality reductions were tested and compared to improve performance and to alleviate computational complexity.

Starting from these and other results, we report improvements of our previously published generic system for object recognition (Nanni et al., 2012, 2013). The new system reported in this paper is based on the following ideas:

- The utilization of both local and global descriptors to represent images; we fuse several texture descriptors.
- Dimensionality reduction of the texture descriptors using principal component analysis (PCA) according to the PCA-SIFT approach (Ke and Sukthankar, 2004); PCA handles the problems of high correlation among the features as well as the curse of dimensionality. Different projections are performed retaining different training subsets for building different projection matrices. In this way it is possible to build an ensemble of classifiers by varying the projection matrix. For each projection matrix a different classifier is trained.
- The utilization of the BoW approach by computing textons considering different clusterings; each cluster is performed separately using a subset of the images of each class. In this way different global texton vocabularies are created, and for each vocabulary a different SVM is trained.
- A new method proposed in this paper that is based on cloud of features where all the subwindows extracted from a given region of the image are used to train a one-class support vector machine.

The strength of this paper lies in the detailed experiments that, together with the shared code, may provide helpful bases for researchers interested in image classification, especially for students who are new to the topic. Different local descriptors, codebook generation methods, subwindow configurations, etc. are combined together and state-of-the-art results are obtained in the tested datasets.

Our new generic system is compared with other approaches using several well-known and widely used datasets: a 15-class scene dataset (Xiao et al., 2010), a building recognition dataset (Amato et al., 2010), the caltech-256 dataset (Griffin et al.), and the person classification dataset of the object classification contest of VOC2012. The new VOC2012 is the last of a very famous series of computer vision competitions, where our system was submitted as a participant so that we could report a comparison of our system with the true state-of-the-art of 2012. In 2001 the accuracy in the 15-class scene dataset was only 73.3%; by 2012 it had become 88.1% (Xiao et al., 2010). The system proposed in this paper obtains an accuracy of 88.3% in the scene dataset; 95.6% in the building recognition; 40% in the caltech-256 dataset, and 88.7% in the person-classification VOC2012 dataset.

A full-feature MATLAB toolbox containing all the source codes used in our proposed system is available at http://www.dei.unipd.it/wdyn/?IDsezione=3314&IDgruppo_pass=124&preview=. We plan on maintaining this toolbox and