Computer Vision and Image Understanding 134 (2015) 22–32

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

GOLD: Gaussians of Local Descriptors for image representation ☆

Giuseppe Serra, Costantino Grana *, Marco Manfredi, Rita Cucchiara

Dipartimento di Ingegneria "Enzo Ferrari", Università degli Studi di Modena e Reggio Emilia, Modena, MO 41125, Italy

ARTICLE INFO

Article history:
Received 28 March 2014
Accepted 16 January 2015
Available online 17 February 2015

Keywords:
Image classification
Concept detection
Gaussian distribution
Stochastic Gradient Descent

ABSTRACT

The Bag of Words paradigm has been the baseline from which several successful image classification solutions were developed in the last decade. These represent images by quantizing local descriptors and summarizing their distribution. The quantization step introduces a dependency on the dataset, that even if in some contexts significantly boosts the performance, severely limits its generalization capabilities. Differently, in this paper, we propose to model the local features distribution with a multivariate Gaussian, without any quantization. The full rank covariance matrix, which lies on a Riemannian manifold, is projected on the tangent Euclidean space and concatenated to the mean vector. The resulting representation, a Gaussian of Local Descriptors (GOLD), allows to use the dot product to closely approximate a distance between distributions without the need for expensive kernel computations. We describe an image by an improved spatial pyramid, which avoids boundary effects with soft assignment: local descriptors contribute to neighboring Gaussians, forming a weighted spatial pyramid of GOLD descriptors. In addition, we extend the model leveraging dataset characteristics in a mixture of Gaussian formulation further improving the classification accuracy. To deal with large scale datasets and high dimensional feature spaces the Stochastic Gradient Descent solver is adopted. Experimental results on several publicly available datasets show that the proposed method obtains state-of-the-art performance.

1. Introduction

Object and Scene Recognition have been a major research direction in computer vision, and, in particular, the task of automatically annotating images has received considerable attention. Systems extract some description from a training set of images, train a classifier and then can be used to perform their task on new images. The current "standard" approach for this task is some derivation of the Bag of Words (BoW) [1], and consists mainly of three steps: (i) extract local features, (ii) generate a codebook and then encode the local features into codes, (iii) pool all the codes together to generate a global image representation. In this approach a key step is the codebook generation, because it is the base to define a high-dimensional Bag of Words histogram. Typically this is performed through clustering methods and the most common approach is the use of k-means clustering, because of its simplicity and convergence speed.

However, introducing a quantization of the feature space tightly ties dataset characteristics to the features representation, in the choice of both the position and the number of cluster centers to use. For the codewords positions, the quantization is learned from the training set, therefore the cluster centers reflect the training data distribution. The optimal number of cluster centers varies depending on the dataset. For example, in [2], the best accuracy using regular BoW is reached at 4k clusters for the Caltech-101 dataset, while, even if the improvement is progressively lower, in PASCAL VOC 2007 it does not reach saturation even with 25k cluster centers. Another example of this "hidden" dataset dependency inclusion may be found in many specializations of the BoW approach. [3,4] propose two different solutions to learn category specific codebooks and show how this is able to improve the descriptor ability to discriminate between similar categories.

The codebook generation step has been introduced in order to obtain a fixed length representation of the distribution of the local features of an image. This is not strictly necessary, since the descriptors distribution could be directly modeled with a parametric distribution [3,5], and the parameters obtained on the single image may provide a summary of the local descriptors. In some contexts though, the information coming from the specific dataset characteristics is able to significantly boost the performance of the classification system.

Based on these considerations, in this paper we propose a solution to allow the descriptors to be obtained either in a dataset

☆ This paper has been recommended for acceptance by Tinne Tuytelaars.

* Corresponding author.

E-mail addresses: giuseppe.serra@unimore.it (G. Serra), costantino.grana@unimore.it (C. Grana), marco.manfredi@unimore.it (M. Manfredi), rita.cucchaira@unimore.it (R. Cucchiara).

http://dx.doi.org/10.1016/j.cviu.2015.01.005
1077-3142/© 2015 Elsevier Inc. All rights reserved.

independent way or to leverage training information in their construction. Using a multivariate Gaussian distribution with full rank covariance matrix or a mixture of them it is possible to tune the system based on the context. We also show how to embed this descriptors in the Spatial Pyramid Representation [6] further removing border effects artifacts. The final image descriptor is then used both with an off-the-shelf batch classifier and with the Stochastic Gradient Descent on-line solver [7], which allows to deal with large scale datasets and high dimensional feature spaces.

We name our method Gaussian of Local Descriptors (GOLD) and demonstrate its effectiveness for automatic image annotation and object recognition. The main contributions of our work are:

- we provide a flexible local feature representation leveraging parametric probability density functions, that can be independent of the image archive (e.g. for collections that change dynamically) or specific to dataset characteristics;
- our method employs the projection of the full rank covariance matrix from the Riemannian manifold to the tangent Euclidean space to obtain a fixed length descriptor suitable for linear classifiers based on dot product;
- we conduct experiments on several public databases (Caltech-101, Caltech-256, ImageCLEF2011, ImageCLEF2013, PASCAL VOC07). Some examples are reported in Fig. 1. The results demonstrate the effectiveness of utilizing our descriptor over different types of local features, both in dataset dependent and independent settings.

This paper is organized as follows. We introduce the state of the art on image descriptors focusing on encodings, normalizations and pooling strategies in Section 2. Then we elaborate the formulation of the GOLD descriptor in Section 3, and its combination with the spatial pyramid representation in Section 4. In Section 5 the extension to the mixture of Gaussian distributions is presented. We conduct extensive experiments in Section 6 to verify the advantage of our method for automatic image annotation and object recognition. Conclusions are drawn in Section 7.

2. Related work

The basic component of all object recognition and scene understanding systems are local descriptors [8]. The most famous and effective ones are SIFT [9], and all their color variations [10].

After describing images with unordered sets of local descriptors, we would like to directly compare them in order to get information on the images similarities. The problem could be tackled with solutions inspired by the assignment problem, but this would be infeasible as soon as we move away from tiny problems. For this reason, research has focused on finding a fixed length summary of local descriptors density distribution.

The original solution, named Bag of Words, consists in finding a set of *codewords* (obtained by the *k-means* algorithm) and assigning each local feature to a codeword. The final descriptor is given by a histogram counting the number of local features assigned to every codeword (cluster center) [1]. This last strategy was later referred to as *hard-assignment*.

A histogram is obviously a crude representation of the local features continuous density profile, it introduces quantization errors and it is sensitive to noise and outliers [11]. Thus, it would appear that by improving this density representation to more accurately represent the input feature set the classifiers performance could be improved as well [3]. For example, in [12] the hard-assignment of features is replaced with soft-assignment, which distributes an appropriate amount of probability mass to all codewords, depending on the relative similarity with each of them. The Locality-constrained Linear Coding [13] projects each descriptor on the space formed by its k-nearest neighbors (with small k; they propose $k = 5$). This procedure corresponds to performing the first two steps of the locally linear embedding algorithm [14], except that the neighbors are selected among the codewords of a dictionary rather than actual descriptors, and the weights are used as features instead of being mere tools to learn an embedding.

In [15] two supervised nonnegative matrix factorizations are combined together to identify latent image bases, and represent the images in this bases space; in [16] the authors propose to combine structures of input features and output multiple tags into one regression framework for multitag image annotation.

Fisher encoding [17], models the codewords with a Gaussian Mixture Model (GMM), restricted to diagonal covariance matrices for each of the k components of the mixture. Then, they capture the average first and second order differences between the image descriptors and the centers of the GMM.

The Vector of Locally Aggregated Descriptors [18] (VLAD) can be seen as a simplification of the Fisher kernel. Each local descriptor is associated to its nearest visual word. The idea of the VLAD descriptor is to accumulate, for each visual word, the differences of the vectors assigned to it, thus characterizing the distribution of the vectors with respect to the center. As for Fisher encoding, the descriptors are pooled together with averaging. Recently a comprehensive study concerning feature coding methods that summarizes their main characteristics including motivations and mathematical representations has been presented in [19].

The techniques discussed so far have all focused on improving the local descriptors encoding, relaying on training data for codewords generation. Given that there are a great number of unlabeled images available, some works focused on semi-supervised learning in order to leverage unlabeled data for large-scale image annotation [20].

In order to overcome the dataset dependency, some authors tried to build a codebook in a fully data-independent way. In [21] the feature space is directly discretized using a regular lattice. With four subdivisions for each dimension, the number of bins is in the order of 10^{77}, most of which are obviously empty. They thus employ a hash table and store only the non-empty bins. Constant time table lookup, i.e., independent of the size of the visual vocabulary, can then be guaranteed. In [22] it is shown that this fixed quantization method performs significantly worse then other techniques, probably due to the fact that it splits dense regions of the descriptor space arbitrarily along dimension axes, and the bins do not equally split the unit hypersphere which SIFT covers, resulting in a wildly uneven distribution of points. Moreover they further highlight on Oxford [23] and Paris [24] datasets that the performance on drop of quantization approaches when generating codewords from a dataset and using them on another. Similar conclusions were also found in [25]. In short, referring to a configuration as dataset1/dataset2 (meaning that codewords are generated by dataset1 and used them for retrieval on dataset2), the Oxford/Oxford combination provides a mAP value of 0.673, against a Paris/Oxford mAP of 0.494. In a recent work [26], to avoid to recompute codewords at every dataset change, a particularly effective solution for cluster center adaptation, applicable to VLAD descriptors, is proposed. This, combined with an appropriate normalization step, shows a remarkable improvement when the codewords are generated from a different dataset. It is significant to note that the more different the codeword generation dataset is, the worse the performance are. Although the proposed adaptation is particularly efficient, it still requires to apply a transformation to all VLAD descriptors of the dataset.

A different strategy was proposed in [3], in order to avoid codeword generation completely, and in this way intrinsically remove