Pattern Recognition 45 (2012) 897–907

Contents lists available at ScienceDirect

Pattern Recognition

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

Supervised learning of Gaussian mixture models for visual vocabulary generation ☆

Basura Fernando [a,b,c], Elisa Fromont [a,b,c], Damien Muselet [a,b,c], Marc Sebban [a,b,c,*]

[a] *Université de Lyon, F-42023, Saint-Étienne, France*
[b] *CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000 Saint-Étienne, France*
[c] *Université de Saint-Étienne, Jean-Monnet, F-42000 Saint-Étienne, France*

ARTICLE INFO

Article history:
Received 18 February 2011
Received in revised form
12 July 2011
Accepted 19 July 2011
Available online 27 July 2011

Keywords:
Bags of visual words
Supervised Gaussian mixture model
Dictionary generation
Expectation-Maximization algorithm

ABSTRACT

The creation of semantically relevant clusters is vital in bag-of-visual words models which are known to be very successful to achieve image classification tasks. Generally, unsupervised clustering algorithms, such as *K*-means, are employed to create such clusters from which visual dictionaries are deduced. *K*-means achieves a *hard assignment* by associating each image descriptor to the cluster with the nearest mean. By this way, the within-cluster sum of squares of distances is minimized. A limitation of this approach in the context of image classification is that it usually does not use any supervision that limits the discriminative power of the resulting visual words (typically the centroids of the clusters). More recently, some supervised dictionary creation methods based on both supervised information and data fitting were proposed leading to more discriminative visual words. But, none of them consider the *uncertainty* present at both image descriptor and cluster levels. In this paper, we propose a supervised learning algorithm based on a Gaussian mixture model which not only generalizes the *K*-means algorithm by allowing *soft assignments*, but also exploits supervised information to improve the discriminative power of the clusters. Technically, our algorithm aims at optimizing, using an EM-based approach, a convex combination of two criteria: the first one is unsupervised and based on the likelihood of the training data; the second is supervised and takes into account the purity of the clusters. We show on two well-known datasets that our method is able to create more relevant clusters by comparing its behavior with the state of the art dictionary creation methods.

1. Introduction

Many of the object recognition and scene classification algorithms first aggregate local statistics computed from images to induce object models before classifying images using supervised techniques such as SVM [1]. Bag of visual words (BoW) approaches have indisputably become a reference in the image processing community [1–3]. In this context, visual dictionary creation constitutes the first crucial step in BoW methods usually known as *vector quantization*. Generally, clustering algorithms are used to achieve this task in the descriptor space. Each cluster representative (typically the centroid) is considered as a visual word of the visual dictionary. The *K*-means clustering algorithm [1,4] is the most common method to create such visual dictionaries even though other unsupervised methods such as K-median clustering [5], mean-shift clustering [6], hierarchical *K*-means [7], agglomerative clustering [8], radius based clustering [6,9], or regular lattice-based strategies [10] have also been used. One of the common features of these unsupervised methods is that they only optimize an objective function fitting to the data but ignoring their class information. Therefore, this reduces the discriminative power of the resulting visual dictionaries. For example, the *K*-means algorithm minimizes the within-cluster sum of squares of distances without considering the class of the data (*i.e.* the label of the image the descriptor has been extracted from). Without any supervision, only one dictionary can thus be created for all the categories in the dataset, usually called *universal dictionary* or *universal vocabulary*.

To create more discriminative visual words, one solution consists in using supervised approaches. In this context, some methods have been proposed to create class specific or concept specific multiple dictionaries. For instance, in [11,12], an image is characterized by a set of histograms—one per class—where each histogram describes whether the image content is well modeled by the dictionary of that specific class. But one limitation of these methods is that they ignore the correlation in the *D*-dimensional space representing the descriptors. For instance, descriptors of a cat's eye and a dog's eye are highly correlated and are likely complementary to learn the generic concept of *eye*. In [13], a dictionary creation method based on the

☆ This work is a part of the ongoing ANR SATTIC 07-1_184534 research project.

[*] Corresponding author at: CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000 Saint-Étienne, France.

E-mail addresses:
palamandadige.basura.fernando@etu.univ-st-etienne.fr (B. Fernando),
elisa.fromont@univ-st-etienne.fr (E. Fromont),
damien.muselet@univ-st-etienne.fr (D. Muselet),
marc.sebban@univ-st-etienne.fr (M. Sebban).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.07.021

learning of a set of classifiers—one per category—is presented, but they are learned independently. In [14], class specific dictionaries are first created and then merged. Despite the fact that, once again, this is achieved without considering the correlations, visual words are redundant because they can occur in different class specific dictionaries. In [15], Zhang et al. solve this problem of redundancy using a boosting procedure and learn multiple dictionaries with complementary discriminative power. A disadvantage of this approach is that the number of visual dictionaries corresponds to the number of boosting iterations. Too many iterations (dictionaries) obviously lead to overfitting and the authors do not provide any theoretical result to determine the optimum number of visual dictionaries.

Recently, some methods have been proposed to create a unique universal dictionary while using supervised information. In [16], the authors use a Gaussian mixture model to take advantage of a soft assignment (unlike *K*-means) and try to maximize the discriminative ability of visual words using image labels. They use a supervised logistic regression model to modify the parameters of the Gaussian mixture. In [17], the authors optimize jointly a single sparse dictionary (using the L1 norm) and a classification model in a mixed generative and discriminative formulation. In both methods [16,17], each image descriptor is assumed to have been generated from a single object category or a single class. We claim that this is a too strong assumption that limits the efficiency of the resulting vocabularies. We will show experimental evidences of this limitation in this paper. In [18], the authors present a general approach to vector quantization that tries to minimize the information loss under the assumption that each descriptor approximates a sufficient statistic for its class label. Once again, this is a strong assumption notably for images described by local descriptors (*e.g.* SIFT descriptors [19]). In [20], randomized clustering forests are used to build visual dictionaries. This approach is first based on the supervised learning of small ensembles of trees which contain a lot of valuable information about locality in descriptor space. Then, ignoring the class labels, these trees are used as simple spatial *partitioners* that assign a distinct region label to each leaf. The disadvantage of this method is that the resulting clusters suffer from a lack of generalization ability since only normalized mutual information is used as splitting criterion during the construction process. Moreover, the method totally ignores the likelihood of the data which is important in BoW image representation. In [21], the authors present an incremental gradient descent-based clustering algorithm which optimizes the visual word creation by the use of the class label of training examples. This method also assumes that each descriptor is generated from a single class and ignores the correlation in the *D*-dimensional space representing the descriptors. Even though all the previous supervised methods allow us to significantly outperform traditional dictionary creation methods, they often assume that each local descriptor belongs to a single object category. Moreover, they do not try to optimize at the same time the *likelihood* of the training data and the *purity* of the clusters.

By integrating both criteria in the objective function to optimize, we claim that it is possible to jointly manage the two kinds of uncertainty the descriptors are usually subject to: the *cluster uncertainty* and the *class uncertainty*. The *cluster uncertainty* expresses the fact that it is something of an over-simplification to achieve a hard assignment (like *K*-means) during the construction of the clusters. For instance, a wheel can contribute to the construction of a visual word representing either a wheel of a bicycle or a wheel of a stroller, with different probabilities of membership. Taking into account this uncertainty during the creation of the visual dictionary can be realized using soft clustering such as Gaussian mixture (GM) models, which have already been shown to outperform hard assignment-based approaches [9]. The *class uncertainty* can be illustrated by the following example: a brown patch descriptor may have been generated from both dog and cow classes. So given a brown patch descriptor, it would be short-sighted to label it by only one of these two classes. This type of uncertainty is usually ignored at the image descriptor level in most of the supervised dictionary creation algorithms. To overcome this limitation, we propose to exploit the probability for each descriptor to belong to each class. The estimation of these probabilities can be achieved by resorting to learned classifiers and approximating the Bayesian rule.

It is important to note that the discriminative power of a cluster in the context of image classification is a function of its purity which depends on both the *cluster uncertainty* and the *class uncertainty*. Although several supervised or semi-supervised GM models have been proposed in various domains [22–25] and in visual dictionary creation [12,16,26,27], none of them addresses the problem of this two-fold uncertainty and none jointly optimizes generalization and discriminative abilities of clusters. To solve these limitations, we present in this paper a new dictionary learning method which optimizes a convex combination of the likelihood of the (labeled and unlabeled) training data and the purity of the clusters. We show that our GM-based method has the ability to quantify both uncertainties during the dictionary creation process and leads to semantically relevant clusters.

The rest of this paper is organized as follows: after having introduced some notations and definitions in Section 2, we present in Section 3 our GM model and the corresponding objective function which will be optimized using an Expectation-Maximization algorithm. Section 4 is devoted to an experimental analysis. We evaluate our method using the *PASCAL VOC-2006* dataset [28] and the *Caltech-101* dataset [29] and show that it significantly outperforms the state of the art dictionary creation approaches. We conclude this paper in Section 5 by outlining promising lines of research on dictionary learning.

2. Notations and definitions

Let $X=\{x_k|k=1\dots n,x_k\in\mathbb{R}^D\}$ be the set of training examples, *i.e.* descriptors extracted from images and living in a *D*-dimensional space (*e.g.* SIFT descriptors usually live in a 128-dimensional space). Let $C=\{c_j|j=1\dots R\}$ be the set of classes (*i.e.* the labels of the original images). Since labeling data can be very expensive, we assume that X may contain both labeled and non-labeled data. Let $S=\{s_i|i=1\dots I\}$ be the set of clusters, where $I>1$.

A Gaussian mixture (GM) model is a generative model where it is assumed that data are *i.i.d* from an unknown probability density function [30]. In our approach, the distribution over the set of clusters is modeled using a GM model $\Theta=\{\theta_i,i=1\dots I\}$ where $\theta_i=\{\mu_i,\Sigma_i,w_i\}$ are the model parameters of the ith Gaussian (corresponding to the cluster s_i). Here, μ_i is the mean, Σ_i is the covariance matrix and w_i is the weight of the ith Gaussian. Given the GM model defined by its parameters Θ, the probability of the descriptor $x_k\in X$ is computed as follows:

$$p(x_k|\Theta)=\sum_{i=1}^{I}w_i\times N_{\mu_i,\Sigma_i}(x_k), \tag{1}$$

where $N_{\mu,\Sigma}(x)$ is the multivariate Gaussian distribution, such that

$$N_{\mu,\Sigma}(x)=\frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right). \tag{2}$$

In a GM model, the posterior probability $p(s_i|x_k,\Theta)$ is calculated as follows:

$$p(s_i|x_k,\Theta)=\frac{w_i\times p(x_k|s_i,\theta_i)}{\sum_t w_t\times p(x_k|s_t,\theta_t)} \tag{3}$$