Fisher Kernels on Visual Vocabularies for Image Categorization

Zhang, Liqiang

May 27, 2018

Abstract

Within the field of pattern classification, the Fisher kernel is a powerful framework which combines the strengths of generative and discriminative approaches. The idea is to characterize a signal with a gradient vector derived from a generative probability model and to subsequently feed this representation to a discriminative classifier. The author propose to apply this framework to image categorization where the input signals are images and where the underlying generative model is a visual vocabulary: a Gaussian mixture model which approximates the distribution of low-level features in images.

1. Introduction

twocolumn Image categorization is the pattern classification problem which consists in assigning one or multiple labels to an image based on its semantic content. This is a very challenging task as one has to cope with inherent object/scene variations as well as changes in viewpoint, lighting and occlusion. Several approaches consist in modeling the distribution of low-level features contained in imagesirrespective of their absolute or relative locations within the image. Despite their relative simplicity, such approaches have shown state-of-the-art performance in a recent evaluation.

The most popular approach, which was inspired by the bag-of-words used in text categorization, is referred to as the bag-of-keypatches or bag-of-visterms. In the following, the author use the latter denomination which is more general (the term keypatches assumes the use of an interest point detector for the extraction of low-level feature vectors). Given a visual vocabulary, the idea is to characterize an image with the number of occurrences of each visual word.

As the two objectives of having a truly universal and compact vocabulary seem irreconcilable, some researchers have departed from the idea of having one unique visual vocabulary across images and proposed to have one (much smaller) per-image vocabulary. In [1], K-means clustering is applied to estimate 40 visual words per image and the similarity between image signatures is measured with the Earth Movers Distance.

2. Fisher Kernels Principle

Pattern classification techniques can be divided into the classes of generative approaches and discriminative approaches. While the first class focuses on the modeling of class-conditional probability density functions, the second one focuses directly on the problem of interest: classification. This explains the theoretical superiority of discriminative methods over generative ones. However, generative approaches have a number of properties which still make them attractive, including the possibility to handle variable length data.

Fisher kernels have been introduced to combine the benefits of generative and discriminative approaches [3]. Let p be a pdf whose parameters are denoted λ. Then one can characterize the samples $X = \{x_t, t = 1...T\}$ with the following gradient vector:

$$\nabla_\lambda logp(X|\lambda) \tag{1}$$

Intuitively, the gradient of the log-likelihood describes the direction in which parameters should be modified to best fit the data. It transforms a variable length sample X into a fixed length vector whose size is only dependent on the number of parameters in the model.

This gradient vector can then be classified using any discriminative classifier. For those discriminative classifiers which use an inner product term it is important to normalize the input vectors. In [2], the Fisher information matrix F is suggested for this purpose:

$$F_\lambda = E_X[\nabla_\lambda logp(X|\lambda)\nabla_\lambda logp(X|\lambda)'] \tag{2}$$

Because of the cost associated with its computation and inversion, F_λ is often approximated by the identity matrix and no normalization is performed. In the next section, the author will derive a diagonal approximation of F_λ (this corresponds to a dimension-wise normalization of the dynamic range), the author will show that using such a normalization impacts favorably the performance.

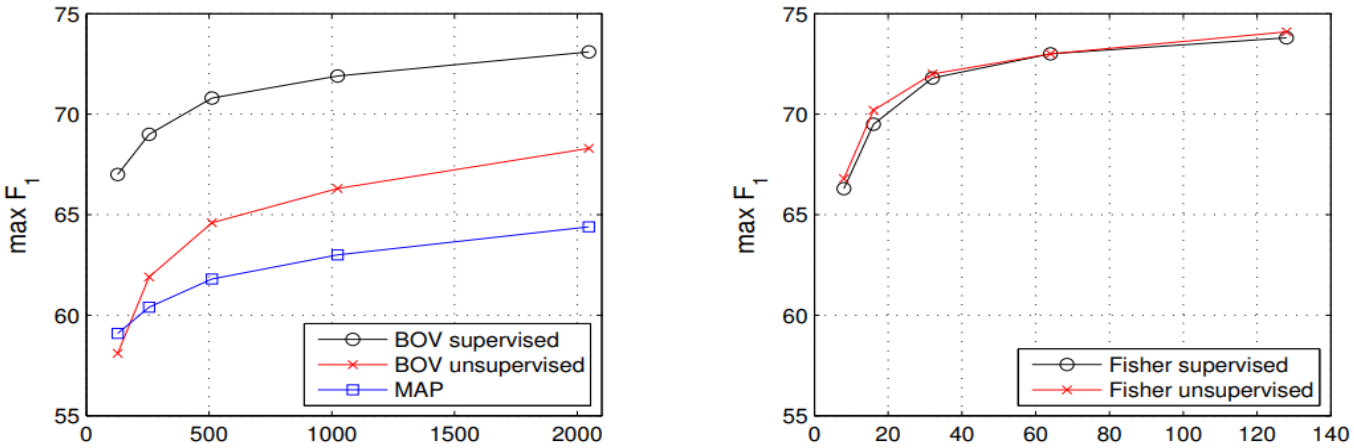

Figure 1. Number of Gaussian components.

Fig. 1 is the performance of the three baseline systems as a function of the number of Gaussian components on our in-house dataset. Performance plateaus after 2048 or 128 Gaussians.

3. In-house database

gradient	max $F1$ (in %)	gradient dimension
w	58.1	127
μ	69.4	6,400
σ	70.4	6,400
$\mu\sigma$	74.1	12,800
$w\mu\sigma$	74.1	12,927

Table 1. Contribution of each parameter (w = weights, μ= mean and σ= standard deviation) to the classification accuracy and to the dimensionality of the gradient space for a GMM with 128 Gaussians.

Results are shown in Table. 1, along with the dimensionality of the gradient representation. When one takes the gradient with respect to weights only (equivalent to the traditional BOV histogram), one obtains a much poorer performance than when taking the gradient with respect to means or standard deviations only.

References

[1] G. Csurka. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision Eccv*, 44(247):1–22, 2004. 1

[2] F. Moosman, B. Triggs, and F. Jurie. Randomized clustering forests for building fast and discriminative visual vocabularies. *Advances in Neural Information Processing Systems*, 2006. 1

[3] S. Ullman, M. Vidalnaquet, and E. Sali. Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, 5(7):682–687, 2002. 1