1 Review: Fisher Kernels on Visual Vocabularies for Image Categorization

Murat Ambarkutuk {murata@vt.edu}, 1/26/2016

Image categorization is the effort of extracting semantic level information from images and labeling them based on the extracted information. In their paper [1], the authors propose an approach in which Fisher Kernels (FK) and Gaussian Mixture Models (GMM) are utilizied to solve image categorization problems. By doing so, the authors provide a unified framework for two different kinds of pattern classification methods, namely, generative and discriminative approaches. The intuitive explanation of the proposed method is to approximate low-level feature vectors extracted from the images with Gaussian functions, while representing them with FK.

The proposed approach contains two major steps in constructing the vocabulary: the approximation of the low-level feature vectors with a number of Gaussian functions, and the representation of the derived Gaussian mixtures with the gradients of log-likelihood functions. The first contribution of this paper is to employ the pipeline of GMM and FK to solve image categorization problems, providing a combined framework for two seperate approaches of categorization. Each N-dimensional feature vector is formulated as the sum of log-likelihood functions given a set number of Gaussian functions. One could object to the idea of representing any arbitrary distribution with a set number of Gaussian functions, given that one may end up having highly non-Gaussian (multi-modal) distributions. To solve the problem, the GMM step can be handled in a sequential manner. KL-divergence formula can be used as a mean of measuring to make sure the number of Gaussian used is good enough to represent the distrubition. Given the influence of this parameter in the whole process, choosing a constant is not trivial. The second step of the approach is to calculate the gradient vectors of the log-likelihood functions of feature vectors. Along with utilization of FK in image categorization problems, the derivation of the closed-form formulation of the Fisher information matrix is the second contribution of the paper. After that, the dimensionality of the feature vectors are reduced to 50. In the training (unsupervised) step, Maximum Likelihood estimation was used to construct a "universal vocabulary"[1], from which the class-vocabulary is established by using Maximum a Posteirori estimation.

The proposed method was compherensively tested with two different databases, in-house and VOC 2006, under supervised and unsupervised learning schemes with two different low-level feature vectors to validate the derived approach. Both of the datasets contained more than 5000 images, for testing and training combined for less than 20 categories. Thus, it inferred that FK is a tailored solution for particular problems, given the amount of effort and data in the training process. The results provided by the authors validate the overall performance of FK, showing at least 10% better performance than median of the VOC 2006 applicants. FK seems to prove its success in categorization problems, which may result from having higher dimensionality in representation of the words. Along with its relatively higher success rate, FK is 25 times more computationally efficient than the conventional BOV representation, albeit the aforementioned increase in dimensionality.

References

[1] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.