The devil is in the details: an evaluation of recent feature encoding methods

Ken Chatfield
http://www.robots.ox.ac.uk/~ken
Victor Lempitsky
http://www.robots.ox.ac.uk/~vilem
Andrea Vedaldi
http://www.vlfeat.org/~vedaldi
Andrew Zisserman
http://www.robots.ox.ac.uk/~az

Department of Engineering Science,
Oxford University

A large number of novel encodings for bag of visual words models have been proposed in the past two years to improve on the standard histogram of quantized local features. Examples include locality-constrained linear encoding [6], improved Fisher encoding [3], super vector encoding [7], and kernel codebook encoding [5]. While several authors have reported very good results on the challenging PASCAL VOC classification data by means of these new techniques, differences in the feature computation and learning algorithms, missing details in the description of the methods, and different tuning of the various components, make it impossible to compare directly these methods and hard to reproduce the results reported.

This paper addresses these shortcomings by carrying out a rigorous evaluation of these new techniques. The contribution offered over that of the original papers is that we: (1) conduct an evaluation of just the encoding methods themselves, fixing all other elements of the pipeline, namely the underlying representation (SIFT descriptors), learning framework (linear SVM), and their tuning, to allow the performance of the different methods to be compared directly; (2) disclose the source code used to generate the experimental results and describe all the implementation details (including some that were omitted in the original publications and that were obtained from personal communications with the authors); (3) analyse which aspects of the different constructions are important for performance and which are not. The overall picture that emerges cannot be inferred from the original publications alone. In particular, not all the methods performed as well as claimed in the original publications without the use of additional undocumented modifications.

We begin by fixing the details of the pipeline not directly related to the encoding. We consider: the computation of low level features, quantization by k-means, large scale k-means, Gaussian mixture models, spatial binning and learning using both linear and non-linear kernels. Following this, we describe the encoding methods themselves. In this paper, five recent encoding methods are considered:

Histogram encoding constitutes the baseline encoding upon which the other methods improve. Given a set of descriptors $\mathbf{x}_1,\ldots,\mathbf{x}_N$ sampled from an image, this method hard assigns each feature to a single visual word in a visual vocabulary pre-trained using k-means. If q_i are the assignments of each descriptor $\mathbf{x}_i$ then the histogram encoding is given by the non-negative vector $\mathbf{f}_{\text{hist}} \in \mathrm{R}^K$ such that $[\mathbf{f}_{\text{hist}}]_k = |\{i : q_i = k\}|$.

Kernel codebook encoding [4, 5] is a variant in which descriptors are assigned to visual words in a soft manner. Specifically, a descriptor is encoded as $[\mathbf{f}_{\text{kcb}}(\mathbf{x}_i)]_k = K(\mathbf{x}_i,\mu_k)/\sum_{j=1}^{K} K(\mathbf{x}_i,\mu_k)$ where $K(\mathbf{x},\mu) = \exp(-\frac{\gamma}{2}\|\mathbf{x}-\mu\|^2)$, and a set of N descriptors extracted from an image as $\mathbf{f}_{\text{kcb}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{f}_{\text{kcb}}(\mathbf{x}_i)$.

Fisher encoding [3] captures average first and second order differences between the image descriptors and the centres of a GMM, which can be thought of as a soft visual vocabulary. The construction of the encoding starts by learning a GMM model $\theta = (\pi_1,\mu_1,\Sigma_1,\ldots,\pi_K,\mu_K,\Sigma_K)$ where π_k are the prior probability values for each Gaussian component k, μ_k the means and Σ_k the positive definite covariance matrices. The Fisher encoding is then a concatenation of the gradient of the log likelihood of a given Gaussian component with respect to both the mean and standard deviation parameters, $\mathbf{u}_k$ and $\mathbf{v}_k$, given a novel descriptor $\mathbf{x}_i$ and is therefore $2KD$ dimensional.

Super vector encoding [7] is similar to the Fisher encoding, except that only the first order differences $\mathbf{u}_k$ between features and cluster centres are used, a component is added which represents the mass of each cluster and each cluster is normalized in a slightly different manner.

Locality-constrained linear (LLC) encoding [6] is a variant of soft encoding which assigns descriptors to visual words by minimizing reconstruction error. Given a set of descriptors $\mathbf{x}_1,\ldots,\mathbf{x}_N$ the visual word assignments α are given by $\operatorname{argmin}_\alpha \sum_{i=1}^{N}\|\mathbf{x}_i - B\alpha_i\|^2$ where B is a dictionary of visual words learned by using k-means. The assignments in this optimization are constrained to the local linear subspace spanned by the $M \ll K$ visual words closest to $\mathbf{x}_i$, with all other histogram bins set to zero. The LLC encoding of a set of descriptors $\mathbf{x}_1,\ldots,\mathbf{x}_N$ is then obtained by max-pooling: $[\mathbf{f}_{\text{LLC}}]_j = \max_{i=1,\ldots,N}[\mathbf{f}_{\text{LLC}}(\mathbf{x}_i)]_j$.

Method		*codebook size* 256	600	1500	2000
FK	Lin	77.78 ± 0.56	–	–	–
LLC	Lin	–	73.10 ± 1.09	74.84 ± 0.67	75.75 ± 0.71
LLC	Chi	–	72.30 ± 1.08	74.23 ± 0.62	75.24 ± 0.71
VQ	Chi	–	72.65 ± 0.77	73.62 ± 0.51	73.93 ± 0.79
KCB	Chi	–	73.38 ± 0.65	75.24 ± 0.63	75.50 ± 0.65

Table 1: Image classification results on Caltech-101 dataset (30 training images) **VQ** – baseline method; **FK** – Fisher kernel; **KCB** – kernel codebook; **LLC** – locality-constrained linear coding; **Lin/Chi** – linear/χ^2 kernel map

We evaluate each method over two standard image classification datasets: PASCAL VOC 2007 [1] and Caltech-101 [2]. A sample of the results over the Caltech-101 dataset are shown in Table 1. In each case, the exact parameters used for each experiment are described in detail. Following this, we analyse the comparative performance of the five methods. As expected, all the recently introduced methods improve the classification accuracy over the bag-of-words baseline. However, in some cases the advantage is not as dramatic as portrayed in the original papers. Indeed, other factors, such as vocabulary size, descriptor sampling density and the use of data augmentation techniques can often have as significant effect on the performance of the algorithms as the encoding method itself.

[1] M. Everingham, A. Zisserman, C. Williams, and L. Van Gool. The PASCAL visual obiect classes challenge 2007 (VOC2007) results. Technical report, Pascal Challenge, 2007.

[2] L. Fei-Fei, R. Fergus, and P. Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proc. ICCV*, 2003.

[3] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010.

[4] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. CVPR*, 2008.

[5] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In *Proc. ECCV*, 2008.

[6] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. *Proc. CVPR*, 2010.

[7] X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. In *Proc. ECCV*, 2010.