Enhancing image classification with class-wise clustered vocabularies

Wojciech Wojcikiewicz[†‡], Alexander Binder[†], Motoaki Kawanabe[‡†]
[†]*Technical University of Berlin, Franklinstr. 28 / 29, 10587 Berlin, Germany*
[‡]*Fraunhofer Institute FIRST, Kekuléstr. 7, 12489 Berlin, Germany*
wojwoj@mail.tu-berlin.de binder@cs.tu-berlin.de nabe@first.fraunhofer.de

Abstract

In recent years bag-of-visual-words representations have gained increasing popularity in the field of image classification. Their performance highly relies on creating a good visual vocabulary from a set of image features (e.g. SIFT). For real-world photo archives such as Flicker, codebooks with larger than a few thousand words are desirable, which is infeasible by the standard k-means clustering. In this paper, we propose a two-step procedure which can generate more informative codebooks efficiently by class-wise k-means and a novel procedure for word selection. Our approach was compared favorably to the standard k-means procedure on the PASCAL VOC data sets.

1 Introduction

Bag-of-visual-words models [7, 2, 6] have been successfully applied to image classification problems in recent years. In the first step, image features e.g. SIFT descriptors are computed on a dense grid or from keypoints, then they are clustered into visual words, so that an image can be finally represented by a fixed-size histogram over the visual words. Among many proposals for codebook generation [4, 5], the standard approach is to cluster a subset of all descriptors from all training images with k-means clustering (KM). However, a couple of its drawbacks have been recognized: 1) it is not feasible to construct codebooks larger than a few thousands visual words, 2) no label information is used for constructing the generic k-means codebooks, and 3) KM chooses most cluster centers to be near high density regions, thus under-representing equally discriminant low-to-medium density ones [3].

In order to overcome these issues of KM, as the first step, we deploy the class-wise k-means (CWKM) procedure proposed by Farquhar *et al.* [2], i.e. we construct small vocabularies for each class and then aggregate them into one large vocabulary. This approach allows to construct large vocabularies very fast, since only small class-specific codebooks are generated and the process can be performed in parallel for different classes. To the best of our knowledge, Farquhar *et al.* [2] is the only work applying per-category clustering to image classification. The authors showed a performance increase, but did not provide empirical evidence for the claim that this is due to more discriminative words.

Although larger vocabularies in general achieve better performances, it takes longer to process new query images, especially in word assignment and kernel computation steps. Farquhar *et al.* [2] deployed dimensionality reduction (e.g. PCA and PLS) to get compact representations. However, this does not alleviate the word assignment and projection onto a low-dimensional subspace is hard to interpret. Instead, we consider feature selection as in [7] and propose novel methods for word selection which take into account informativeness of the visual words and coverage of the descriptor space. We show that codebooks by a combination of class-wise clustering and word selection improve classification performance compared to those of the same size by the standard all-in-one approach. We will also discuss the reasons of the performance gain.

This paper is organized as follows. In Section 2, we motivate the use of class-wise clustering and introduce the methods for word selection. After describing our experimental setup in Section 3, we compare the performance between KM and CWKM with and without word selection on PASCAL VOC 2007 and VOC 2008 data and analyse the differences between the vocabularies. Section 5 concludes with future research issues.

2 Codebook generation procedure

2.1 Class-wise clustered vocabularies

At first, for all classes $k = 1, \ldots, K$, we construct a small vocabulary $\mathbf{V}_k = \{v_{k1}, \ldots, v_{km}\}$

applying k-means clustering to the SIFT descriptors $\{s_{k1}, \ldots, s_{kd}\}$ selected from the training images containing category k. That is, V_k is obtained by minimizing $\sum_i \sum_j \delta_{ij} \|s_{ki} - v_{kj}\|^2$, where $\|\cdot\|$ denotes Euclidean distance and δ_{ij}'s are the indicator of the cluster assignment. Then, the resulting vocabularies are aggregated into one large overall vocabulary $\mathbf{V} = \{\mathbf{V}_1, \ldots, \mathbf{V}_K\}$ (size $K \times m$). In this way, we can produce large vocabularies which achieve better classification performance in a fraction of time (hours vs. days) compared to KM.

2.2 Visual word selection

Unfortunately, large vocabularies slow down the classification process in test phase[1]. Therefore, we reduce the codebook size by novel methods for word selection which are based on informativeness of visual words and coverage of the descriptor space. We will show that we can construct better vocabularies of the same size compared to the standard ones by KM.

Consider a vocabulary $\mathbf{V}$ and a set of labeled input features $J = \{(s_i, c_i), i = 1, \ldots, N\}$, where s_i denotes a SIFT descriptor and c_i one of the class labels assigned to the image which s_i comes from. We propose to use the distribution of the labels in the cluster assigned to each word v as a measure of informativeness, namely

$$h_v(k) := \frac{n(v,k)}{n(v)}, \tag{1}$$

where $n(v)$ is the size of the cluster assigned to the word v and $n(v,k)$ is the number of the descriptors with the label k in the cluster. This is an estimator of the conditional probability $p(k \mid v)$ of the labels given the word. Ideally, a visual word v is discriminative, when $h_v(k)$ is peaked at a class k or a small number of classes, or in other words, when it occurs more frequently in the class k (or a small number of classes) but is rare in the other classes. However, in reality, we cannot expect such clear-cut situations. Therefore, for word selection we deploy the entropy

$$\tau(v) = -\sum_k h_v(k) \log(h_v(k)), \tag{2}$$

which measures informativeness about the labels. A small τ value indicate a word with little uncertainty about the label. We note that the set J should be balanced, so that it contains the same number of SIFTs for each class, otherwise we need to use other criteria to cope with unbalanced class sizes.

[1]Not only vocabulary construction which is created only once takes longer, but also the word assignment and kernel computation step which is required everytime we classify a test image.

Apart from the informativeness of a visual word, we want to make sure that the descriptor space is covered by the reduced codebook. Otherwise, all selected words can come from the same region in the descriptor space, i.e. represent only a similar pattern in images and cannot capture wide variety of other patterns. Such a non-representative vocabulary result potentially in performance drop. In order to assure that the selected subset of words still provide a good coverage, we propose to recluster the vocabulary i.e. to apply k-means to the words of the large codebook. In our experiment, we will compare the following selection schemes.

- `selEntropy`: Select the words v with smallest entropies $\tau(v)$.
- `selCW`: Rank the words v for each class k according to $h_v(k)$ in descending order. Rotate over the classes and select the word with the highest rank which has not been selected yet.
- `recluNew`: Recluster vocabulary and take cluster centers as new words.
- `recluEntropy`: Recluster vocabulary and from each cluster select the word with minimal $\tau(v)$.
- `recluMax`: Recluster vocabulary and from each cluster select the word with highest $\max_k\{h_v(k)\}$.

3 Experimental Setup

We experimented on the PASCAL VOC 2007 and VOC 2008 data sets [1] containing 9963 and 8780 images of 20 object classes. We used SIFT descriptors over the grey channels and whole images as the base feature for the bag-of-words representation. Our choice of kernel function is the χ^2 kernel, which has proved to be a suitable similarity measure for histograms. The kernel parameter was set to be the mean of the χ^2 distances between all pairs of training samples.

The evaluation criterion is the average precision (AP) over all recall values calculated from the precision-recall curves. The regularization parameter C of the SVM was optimized on the validation data. For VOC 2007 test results will be reported, whereas for VOC 2008 we show only validation performances, as the labels for the test images are not available.

4 Results

At first, we checked that the class-wise clustered vocabulary with 20000 words (`CWKM20000`), the maximal size we tested, performs much better than the standard 4000 words codebook (`KM4000`). The mean AP

(MAP) gain over all classes in VOC 2007 is 0.0205 (or 4.5 %), however, the maximal increase in AP is 0.0838 (or 33.2 %) for the class 'diningtable' and no class performs worse. In the case of VOC 2008, we obtain a mean increase in validation performance of 0.0277 (or 7.7 %), the maximal gain is 0.0824 (or 34 %) and no class does significantly worse. In both cases the performance gain is non-uniform for the classes (see Figure 2). We remark that the time needed for clustering is much less for CWKM, i.e. few hours vs. one week.

As large vocabularies have disadvantages in terms of computation time, it is fair to compare vocabularies of the same size. We created smaller vocabularies directly with CWKM, i.e. we constructed very small class-specific codebooks (e.g. 200 words) and aggregated them into mid-size vocabulary (e.g. 4000 words). However, as shown in Table 1 (first and second row), the performances of CWKM is on par or even worse than those by the standard KM. Thus, the only advantage is a faster computation. The reason is that the class-specific codebooks created by CWKM are simply too small (e.g. 40 words in the 800 words vocabulary) to capture all relevant information.

This claim is supported by Figure 1. We plot histograms of distances from visual words in one codebook to the nearest neighbors in the other vocabulary. The upper panel shows the distances between `CWKM800` and `KM800` and in the lower panel the distances between words of `recluEntropy` (800 words) and `KM800` are presented. We clearly see that in the first case there are some regions in descriptor space which are captured by `KM800`, but not by `CWKM800`, since the red histogram visualizing the distances from `CWKM800` to `KM800` is shifted to the right. In contrast to that the `recluEntropy` vocabulary covers the descriptor space much better than `KM800`.

Therefore, we constructed mid-size vocabularies by selecting words from the largest CWKM codebook (`CWKM20000`). The results for different vocabulary sizes and the five word selection methods described above are summarized in Table 1. As can be seen, the methods `selEntropy` and `recluNew` perform worst, i.e. there is no performance gain when compared to the corresponding KM vocabulary, whereas the other three methods show a clear gain e.g. the absolute difference between `recluEntropy` and KM on a 800 words vocabulary is 0.0211 or 5.11 % for VOC 2007.

The selection schemes resulting in worse performances neglect either coverage (`selEntropy`) or informativeness (`recluNew`). We also remark that the new cluster centers by `recluNew` do not coincide with the original CWKM words. The reclustering methods `recluEntropy` and `recluMax` perform much better, as they take into account both coverage and informativeness. We conjecture that `selCW` performs not that bad, because it retains sufficient coverage indirectly by selecting same numbers of informative words from all classes. It seems that the coverage of the descriptor space is crucial for good performance and we introduced a new way of evaluating it, namely by using distance histograms as in Figure 1.

As we can see in Figure 1, there are more outlier words in the `recluEntropy` vocabularies which are not well-represented by the KM codebooks compared to the other way around. Table 2 shows the median of p-values of the label uncertainty τ for the top $\alpha\%$ words being outliers by distance to the nearest neighbor[2]. In other words we compute the τ values of the top $\alpha\%$ outlier words in `recluEntropy` (as done in figure 1), compute the p-value for them and report the median. The small values compared to 'all' indicate that the outlier words have significantly small uncertainty τ, i.e. are very informative. This property can

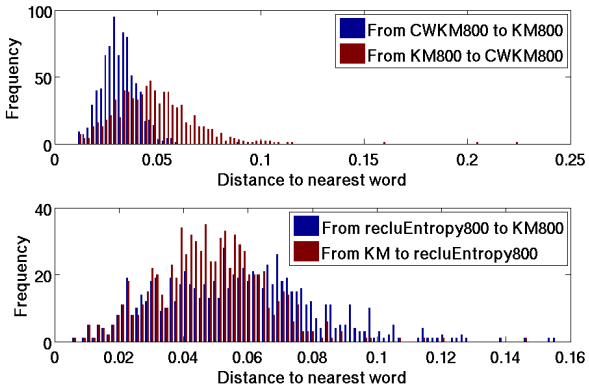

Figure 1. Distance distribution to nearest word between `CWKM800` - `KM800` (top) and `recluEntropy` - `KM800` (bottom).

Method	4000 words		2000 words		800 words	
VOC	2007	2008	2007	2008	2007	2008
KM	45.45	35.92	44.42	35.08	41.23	32.65
CWKM	45.15	35.76	44.01	35.76	40.23	32.33
`selEnt`	45.55	36.68	44.30	34.56	40.86	31.22
`selCW`	**46.63**	37.26	44.76	34.60	42.43	32.81
`recluNew`	46.18	36.73	44.14	35.40	41.18	32.85
`recluEnt`	46.56	**37.48**	**44.96**	35.91	**43.33**	34.15
`recluMax`	46.58	36.80	44.94	**35.97**	42.86	**34.18**

Table 1. Results (MAP $\times$ 100) for VOC images. `CWKM20000` result is 47.49 (38.69).

[2]The p-value in this case is the percentage of the words in the corresponding KM vocabulary which have smaller uncertainty τ.

top $\alpha\%$	4000 words	2000 words	800 words
0.5%	6.3×10^{-3}	3.0×10^{-3}	7.5×10^{-3}
1%	1.3×10^{-2}	5.0×10^{-3}	7.5×10^{-3}
2%	2.0×10^{-2}	5.5×10^{-3}	1.2×10^{-2}
5%	6.6×10^{-2}	2.1×10^{-2}	1.0×10^{-2}
all	0.37	0.32	0.30

Table 2. Median p-values of the label uncertainty τ for outlier words from `recluEntropy` on VOC 2007.

lead to the performance gains by the two-step approach `recluEntropy` reported here.

Another interesting observation is that the performance gain for CWKM is non-uniformly distributed over classes. Figure 2 shows the absolute performance gain of `CWKM20000` (blue), `recluEntropy` with 4000 words (green) and `selCW` with 4000 words (red) over `KM4000`. There is a correlation between the performance gain of `CWKM20000` and `recluEntropy` or `selCW`, the correlation coefficients are 0.4654 and 0.7176 respectively. This means that the advantages of the `CWKM20000` codebook are retained in the smaller vocabularies. Classes showing a performance gain seem to be better represented when clustering class-wise. This can be due to more discriminant words as assumed by Farquhar *et al.* [2] or by a better coverage of the descriptor space as shown in this paper.

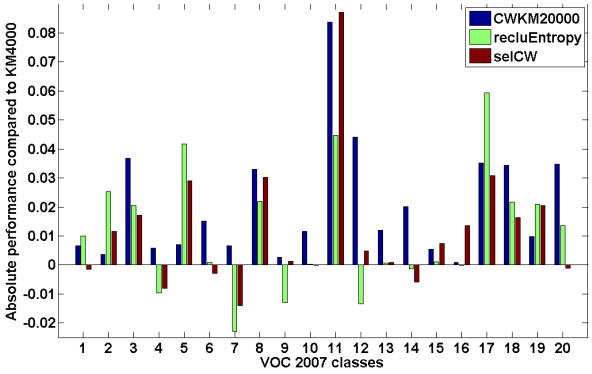

Figure 2. Absolute performance gain per class over `KM4000`.

5 Conclusion

In this paper we showed empirically with challenging data sets that using class-wise clustering produces better results while allowing to generate large vocabularies in a fraction of time compared to KM (hours vs. days). Furthermore, in order to shorten processing time of test query images, we proposed to construct smaller codebooks by new methods for word selection which take into account informativeness of visual words and coverage of the descriptor space. Our experimental results showed that selecting informative representatives after reclustering works best, which implies importance of retaining good coverage of the descriptor space. We want to remark that our method outperforms the KM baseline for codebooks of the same size, we could not achieve comparable performance with other efficient alternatives like Extremely Randomized Clustering Forests (ERCF, [5]) or Hierarchical k-means (HKM, [6]), e.g. with 4000 words on VOC 2007 they had an mean AP score of 0.4435 and 0.4501 respectively.

Since CWKM does not scale with the number of classes, we are investigating the use of a group-wise clustering approach, i.e. groups of similar object classes are created by the user and vocabularies are clustered for each group separately.

Acknowledgements We thank Klaus-Robert Müller for valuable discussions. This work was supported by the Federal Ministry of Economics and Technology of Germany under the project THESEUS (01MQ07018), by the German Research Foundation (GRK 1589/1) and the FP7-ICT Programme of the European Community, under the PASCAL2 Network of Excellence, ICT-216886.

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (2008) Results.

[2] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving bag-of-keypoints image categorisation. Technical report, University of Southampton, 2005.

[3] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV '05: Proceedings of International Conference on Computer Vision*, 2005.

[4] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization. In *British Machine Vision Conference*, 2006.

[5] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Advances in Neural Information Processing Systems*, 2006.

[6] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR '06: Proceedings of Conference on Computer Vision and Pattern Recognition*, 2006.

[7] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *MIR '07: Proceedings of workshop on multimedia information retrieval*, 2007.