Bag-of-Features Codebook Generation by Self-Organisation

Teemu Kinnunen, Joni-Kristian Kamarainen*, Lasse Lensu,
and Heikki Kälviäinen

Machine Vision and Pattern Recognition Laboratory (MVPR),
*MVPR/Computational Vision Group, Kouvola
Department of Information Technology
Lappeenranta University of Technology
Finland

Abstract. Bag of features is a well established technique for the visual categorisation of objects, categories of objects and textures. One of the most important part of this technique is codebook generation since its within-class and between-class discrimination power is the main factor in the categorisation accuracy. A codebook is generated from regions of interest extracted automatically from a set of labeled (supervised/semi-supervised) or unlabeled (unsupervised) images. A standard tool for the codebook generation is the c-means clustering algorithm, and the state-of-the-art results have been reported using generation schemes based on the c-means. In this work, we challenge this mainstream approach by demonstrating how the competitive learning principle in the self-organising map (SOM) is able to provide similar and often superior results to the c-means. Therefore, we claim that exploiting the self-organisation principle is an alternative research direction to the mainstream research in visual object categorisation and its importance for the ultimate challenge, unsupervised visual object categorisation, needs to be investigated.

1 Introduction

Visual object categorisation (VOC) means automatic detection of categories (e.g., "face", "motorbike", etc.) of objects in images. During the last decade, VOC has become an important and active research topic in computer vision. The motivation originates from the desire to automatically search the vast amount of digital image and video data distributed on the Internet. Researchers in this field have accepted the "Bag-of-Features" (BoF) approach (see, e.g., [1,2,3] and Fig. 1) as the main processing principle and it has achieved the mainstream status. In this work, we accept the main principle, but want also to revise one of its intrinsic parts: the visual feature codebook generation. A standard tool for the inter-category codebook generation is the c-means clustering. The state-of-the-art results have been achieved by enhancing the standard c-means with more sophisticated processing and optimisation.

J.C. Príncipe and R. Miikkulainen (Eds.): WSOM 2009, LNCS 5629, pp. 124–132, 2009.

Very recently, an ultimate challenge of visual object categorisation has been proposed [4,5]: unsupervised visual object categorisation. In the unsupervised problem, there is no training or validation sets with manually labeled ground truth, which, on the other hand, prevents using the most effective enhancements in the codebook generation. Now we need to revisit and revise the standard parts of the BoF approach. In this work we revisit the codebook generation part and investigate whether a self-organisation principle, especially self-organising map (SOM) [16], can provide novel or superior characteristics to the c-means.

2 Related Work

Due to the active past and current work in the field of supervised VOC, the reported results are now very incremental. For example, there are two main directions in the codebook generation algorithms: replacement of the c-means with another "more tailored" clustering method and enhancement of the c-means with application-specific parts. The latter one has been more successful.

Jurie and Triggs [6] have developed a clustering method which is more robust than the c-means. Their method avoids setting all cluster centres into high density areas, which is typical to the c-means. Their algorithm first chooses N samples randomly and then computes maximal density of the samples using mean-shift estimator. Then it assigns a cluster centre point to the maximal density and eliminates all samples that are within a certain radius from the cluster centre. Then the algorithm repeats these steps with remaining samples as long as there are too many samples left or the number of clusters is too low. Interestingly, this "topology preserving" enforcement is very similar to the main characteristic of self-organisation.

Gemert et al. [2] have developed a method based on the c-means. They replace the simple learning rule, which assigns a sample to the closest cluster, with uncertainty, plausibility and distance values. These values are used in the codebook generation. For example, if a data point is in the middle of two clusters, it will be assigned with the proportion of 50% to the both clusters.

Problem-specific clustering approaches have been developed as well. Leibe et al. [7] use hierarchical clustering to generate the codebook. Many other successful methods, however, use directly the c-means [8,9]. The main property in these enhancements is in locating the cluster centres to spread in a more intelligent manner than converging to few high density regions of the input samples.

One problem-specific enhancement outside clustering is to utilise the spatial information in the codebook generation or probing. For example, Lazebnik et al. [10] reported a method which uses a spatial pyramid to organise descriptors based on their appearance and location. These enhancements, however, are particularly unsuitable for unsupervised methods.

In the recent work on unsupervised visual object categorisation, Sivic et al. [4] presented an unsupervised method utilising Latent Dirichlet Allocation (LDA) model. They improved the original LDA by introducing hierarchical LDA (hLDA). With the hierarchy, they were able to improve the categorisation

performance, but the results were reported only for a small number of categories and it is not clear if the approach generalises well.

3 Bag-of-Features Framework and Self-organisation in Codebook Generation

The general principle in the bag-of-features approach is very simple. First, interest points are automatically detected from the images, e.g., by using the SIFT [11], Maximal Stable Extremal Regions (MSER) [12] or salient region detector [13]. Then, invariant region descriptors are formed around these interest points (included to, e.g., the SIFT, Speeded Up Robust Features (SURF) [14] and Gradient Location and Orientation Histogram (GLOH) [15]). Then comes an important part: the descriptors are used to form a compact codebook. From any observed image, the interest point detection and descriptor formation parts are exactly the same, but then the contents in the image should be classified according to the "loads" in the codebook. Prior to the categorisation, spatial processing, such as segmentation, can be performed, but generally the main structure is obeyed. Now, it is clear that the codebook plays an essential role in this kind of system. The system is depicted in Fig. 1.

We are using bag-of-features approach, which is similar to the system which was presented by Dance et al. [1], to generate feature histograms for the images. These feature histograms are used to describe images. Let D be a set of descriptors which are extracted from an image using a local feature extractor such as SIFT, and CB be a codebook which contains M words. In practice, words in the CB are clusters' centre points. Let N be the number of descriptors extracted from the image. Then, an image feature image histogram F is generated according to the bag-of-features approach which is defined in Algorithm 1. The $Dist$ function calculates the Euclidean distance between two vectors. The smaller the distance, the greater similarity is between two vectors. Hence, a word that minimizes the distance from a descriptor is chosen as the best match, bm.

Algorithm 1. Feature generation using a bag-of-features approach

for $i = 1$ to N **do**
  $bm \leftarrow \min_j Dist(D_i, CB_j)$
  $F_{bm} \leftarrow F_{bm} + 1$
end for

Our main research question in this work is straightforward: what new or superior properties we can achieve by replacing the c-means in the codebook generation with the self-organising map [16] and how these properties can be quantitatively measured? We claim that a proper evaluation procedure is to perform a complete experiment on visual object categorisation and then test the effect of replacing different parts in the system. In our work, we apply the

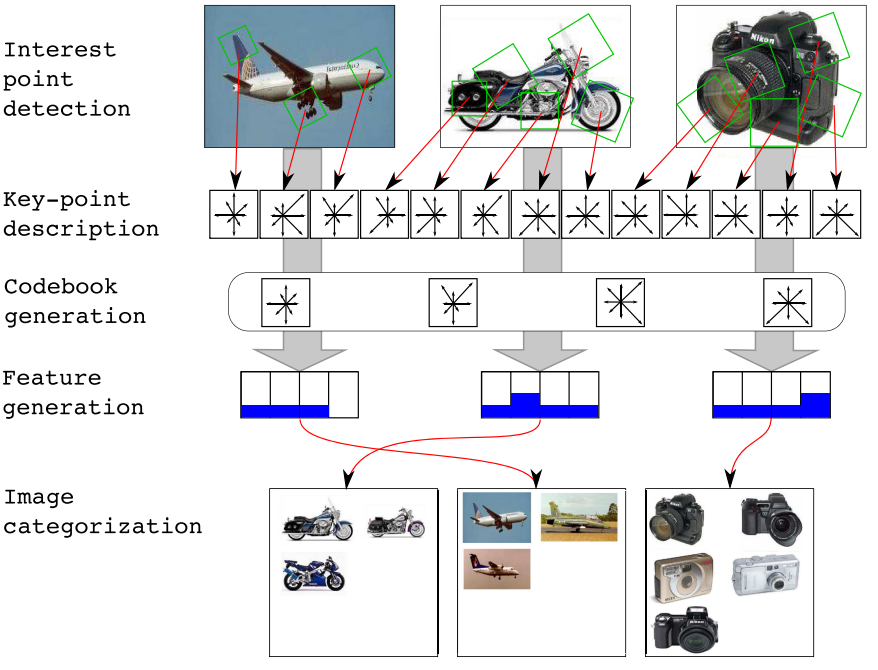


Fig. 1. General structure and information flow in the "bag-of-features" approach. Note that the codebook generation is performed only in the training phase for supervised methods.

simplest form of the BoF principle described, e.g., in [1]. We compare the two methods in supervised and unsupervised experiments with the same evaluation measures and data sets as in the recent state-of-the-art papers [1,4]. Moreover, we point out that their evaluation is lacking in some respect and claim that the evaluation should actually investigate the performance as a function of the number of categories. Only this asymptotic behaviour reveals information about generality and extensibility of a method.

4 Experiments

As discussed above, we do not support the idea that there would be a single evaluation criteria for the codebook selection, and therefore, we established the complete VOC framework and conducted experiments through the complete pipeline.

In the first experiment, we analyse BoF in its most typical structure, exactly the one depicted in Fig. 1, and supervised visual object categorisation. In the supervised VOC, we have a training set of labelled images (list of objects present in the images). In particular, we replicate the system and experiments in Dance et al. [1], except that we replace the support vector machine classifier with the

Fig. 2. Image set for the first test: CalTech 4 and side images of cars [17]. CalTech 4 contains of images of aeroplanes, cars (rear), faces and motorbikes.

Table 1. C-means vs. SOM generated codebooks in the VOC framework in Dance et al. [1] for the CalTech 4 + car side image set (optimal c-means codebook size is 100 and SOM 50)

Category	c-means w/ 1-NN	SOM w/ 1-NN	(c-means w/ SVM) Dance et. al. [1]
Aeroplanes	0.760	0.753	0.963
Cars (rear)	0.893	0.953	0.977
Cars (side)	0.980	0.953	0.996
Faces	0.787	0.833	0.940
Motorbikes	0.593	0.707	0.927
Average	0.803	0.840	0.961

simple 1-NN decision rule. In this experiment, Caltech 4 together with side images of cars were used. One example image from each category is shown in Fig. 2. The only tunable parameter for the SOM and c-means is the size of the codebook which was optimised for the best results to facilitate reliable comparison. The best results are given in Table 1 where it is evident that the basic SOM can easily match the performance of the c-means and, in this case, also outperform it. It should be noted that the results in [1] were achieved with tailored and heavily optimised support vector machine (SVM) classifier. However, the simple 1-NN classifier performed comparably with no special optimisation.

In the second experiment, we moved from the supervised VOC problem to the more recent challenge, unsupervised VOC. The unsupervised problem has been investigated hitherto only in a few papers, and we utilised the same data and the same performance measure as in Sivic et al. [4]. The performance of the system is defined in Eq. 2 as average performance of nodes. The node performance, p_t, is computed as

$$p_t = \max_i \frac{GT_i \cap P_t}{GT_i \cup P_t} \tag{1}$$

where GT_i is the number of ground truth images from the category i, P_t is the number of images assigned to the node t. The average performance, p, is then

$$p = \frac{1}{N_c} \sum_{i=1}^{N_c} \max_t p_{(t,i)} \tag{2}$$

where N_c is the number of categories. In the equation, the highest performing node is chosen for each category and then adds performances together and

Fig. 3. Image set for the second test: MSRC v1 [18]

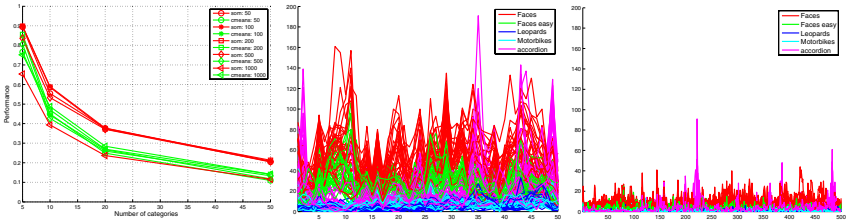
Table 2. C-means vs. SOM codebook generation & c-means vs. SOM unsupervised classification for the MSRC V1 image set

Category	c-means w/ c-means	c-means w/ SOM	SOM w/ c-means	SOM w/ SOM
Aeroplanes	0.248	0.263	0.430	0.485
Bikes	0.246	0.165	0.339	0.605
Buildings	0.258	0.165	0.149	0.251
Cars	0.123	0.187	0.211	0.271
Cows	0.217	0.261	0.159	0.263
Grass	0.203	0.356	0.174	0.461
Faces	0.252	0.178	0.250	0.245
Sky	0.196	0.206	0.170	0.209
Trees	0.265	0.233	0.346	0.450
Average	0.223	0.224	0.247	0.360

divided sum by the number of categories which give average categorisation accuracy over all categories.

In the unsupervised scheme, the 1-NN rule must be omitted and replaced with an unsupervised approach. As a simple approach, we fed the extracted codebook loadings (histograms) again to the clustering method, and assigned to each cluster the most representative category label afterwards. Knowledge about labels of the images is not used in learning phase, but they are needed for performance evaluation. Hence, data must have labels otherwise; it is not possible to measure performance. The data used in Sivic et al. consists of nine manually segmented object categories from the MSRC v1 image set [18]. Examples of the images are shown in Fig. 3. We also adopted their performance measure despite the fact that it is intended for measuring consistency of object hierarchies.

We tested all four combinations (c-means/SOM codebook generation & c-means/SOM "category clustering") and optimised the codebook sizes to report the best performances. The results are shown in Table 2 where it is evident



(a) (b) (c)

Fig. 4. Test results and feature histograms. (a) C-means vs. SOM codebook generation & 1-NN classification for the CalTech 101 database. Note that the SOM graphs are coded with red and c-means with green colour. (b) Feature histograms from five categories using 50 words and SOM. (c) Feature histograms from five categories using 500 words and SOM.

that the SOM-SOM combination provided distinctly better results than any other combination, and again the SOM generated codebooks outperformed the c-means. It should be noted that Sivic et al. reported the performance as high as 0.72, but it describes accuracy of object hierarchy which is not included to our method at all. Moreover, the actual level of supervision is not very clear from their report.

The previous two experiments demonstrated the superiority of the SOM in the two previously reported test cases. However, we claim that in those test cases the used performance measure and the amount of data were not adequate for a reliable evaluation of unsupervised VOC performance. The important factor is actually the asymptotic behaviour of the performance as a function of the number of categories. After all, it is more important to know how a method performs with hundreds and thousands of categories. Performance with some specific number of categories can tell only about performance of the system with a specific test set. When the performance of the system is tested with different number of categories, it can tell overall performance of the system more completely. To initiate a better practise, we performed the last experiment using a more proper performance measure and with the well-known Caltech 101 [17] database. Our evaluation procedure was adopted from Fei-Fei et al. [17], where 5 iterations were computed for 30 random images in the training set and another 20 random images in the testing set. We can observe two important results from this experiment (see Fig. 4(a)). At first, collapse of the performance occurs quite rapidly if more than 10 categories are used. Secondly, the SOM systematically outperforms the c-means algorithm. The best overall performances for the SOM were (codebook size 100) 0.898 accuracy for 5 categories, 0.589 for 10, 0.377 for 20 and 0.208 for 50 categories. The best performances for the c-means were 0.856, 0.471, 0.269 and 0.141 respectively. This experiment, we believe, is the strongest proof of superiority of the SOM algorithm in the codebook generation.

Table 3. Results in Fig. 4(a) listed in the table for different codebook sizes

	SOM					c-means				
Num. of categer.	50 words	100 words	200 words	500 words	1000 words	50 words	100 words	200 words	500 words	1000 words
5	0.894	**0.898**	**0.898**	0.836	0.654	0.766	0.846	0.856	0.806	0.752
10	0.554	**0.589**	0.586	0.533	0.393	0.450	0.444	0.471	0.426	0.487
20	0.376	**0.377**	0.372	0.371	0.236	0.249	0.267	0.269	0.260	0.284
50	0.204	0.208	**0.217**	0.206	0.112	0.109	0.117	0.141	0.133	0.141

Figs. 4(b) and 4(c) shows feature histograms. These two figures discover the fact that when the size of the codebook increases, feature histograms gets less distinctive to each other and thus it is more difficult to separate different images and image categories.

5 Conclusions

In this work, we studied whether the self-organisation principle and especially the self-organising map algorithm could provide novel or superior properties in the codebook generation for the visual object categorisation problem. In all the performed experiments, it was shown how the SOM matches, and in the most of the cases, outperforms the c-means algorithm which is the standard in this task. Lower performance of the c-means is a result of poor clustering. C-means sets most of the cluster centre points near to density areas and thus centre points cover well only a fraction of the data. SOM assigns cluster centre points more evenly and thus they cover most of the data. It leads to better codebooks which increases the performance of VOC system. Quantization error could be decreased by increasing the size of the codebook, but it does not lead always to good performance of the system. When the size of the codebook increases, feature histograms get less distinctive and hence it is more difficult to separate images from each other. This affects to the performance in negative manner. This phenomenon is illustrated in Figures 4(b) and 4(c). The results motivate us in the future work to further study the self-organising principle as the predominant principle for realising visual object categorisation and especially unsupervised visual object categorisation.

Acknowledgements

The authors would like to thank the Academy of Finland and partners of the VisiQ project (no. 123210) for support.

References

1. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision (2004)

2. van Gemert, J., Geusebroek, J., Veenman, C., Smeulders, A.: Kernel codebooks for scene categorization. In: Proc. of the European Conf. on Computer Vision, pp. 696–709 (2008)
3. Marszałek, M., Schmid, C.: Constructing category hierarchies for visual recognition. In: Proc. of the European Conf. on Computer Vision (2008)
4. Sivic, J., Russell, B.C., Zisserman, A., Freeman, W.T., Efros, A.A.: Unsupervised discovery of visual object class hierarchies. In: Proc. of the Computer Vision and Pattern Recognition, pp. 1–8 (2008)
5. Bart, E., Porteous, I., Perona, P., Welling, M.: Unsupervised learning of visual taxonomies. In: Proc. of the Computer Vision and Pattern Recognition (2008)
6. Jurie, F., Triggs, B.: Creating efficient codebooks for visual recognition. In: Int. Conf. on Computer Vision, pp. 604–610 (October 2005)
7. Leibe, B., Ettlin, A., Schiele, B.: Learning semantic object parts for object categorization. Image and Vision Computing 26, 15–26 (2008)
8. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 490–503. Springer, Heidelberg (2006)
9. Willamowski, J., Arregui, D., Csurka, G., Dance, C., Fan, L.: Categorizing nine visual classes using local appearance descriptor. In: ICPR Workshop Learning for Adaptable Visual Systems (2004)
10. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Conf. on Computer Vision and Pattern Recognition, pp. 2169–2178 (2006)
11. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. Journal of Computer Vision 20, 91–110 (2004)
12. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. In: Proc. of the British Machine Vision Conf., pp. 384–393 (2002)
13. Kadir, T., Zisserman, A., Brady, M.: An affine invariant salient region detector. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 228–241. Springer, Heidelberg (2004)
14. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
15. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. Int. Journal of Computer Vision 65(1/2), 43–72 (2005)
16. Kohonen, T.: The self-organizing map. Proc. of the IEEE 78(9), 1464–1480 (1990)
17. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In: CVPR Workshop on Generative-Model Based Vision (2004)
18. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. In: Int. Conf. on Computer Vision, pp. 1800–1807 (2005)