2010 International Conference on Pattern Recognition

Unsupervised visual object categorisation via self-organisation

Teemu Kinnunen, Joni-Kristian Kamarainen*, Lasse Lensu, Heikki Kälviäinen
Machine Vision and Pattern Recognition Laboratory
**Computational Vision Group, Kouvola*
Lappeenranta University of Technology

Abstract

Visual object categorisation (VOC) has become one of the most actively investigated topic in computer vision. In the mainstream studies, the topic is considered as a supervised problem, but recently, the ultimate challenge has been posed: Unsupervised visual object categorisation. Hitherto only a few methods have been published, all of them being computationally demanding successors of their supervised counterparts. In this study, we address this problem with a simple and effective method: competitive learning leading to self-organisation (self-categorisation). The unsupervised competitive learning approach is implemented using the Kohonen self-organising map algorithm (SOM). The SOM is used to perform the both unsupervised codebook generation and object categorisation. We present our method in detail and compare results to the supervised approach.

1. Introduction

Visual object categorisation (VOC) has been one of the most active computer vision research topic during the recent years. This activity has not only lead to new methods, but also novel concepts, such as the "bag-of-features" (BoF) [2, 13, 11], international competitions, e.g., the Pascal VOC challenge [3], and public databases for the method evaluation, e.g., Caltech-101 [4] and Caltech-256 [6].The vast majority of the proposed methods are based on the same generally accepted principles and structure: interest point/region detection, invariant region description, codebook generation and classification based on codebook features. The inputs for the codebook and classifier learning are category example images and ground truth, i.e. object labels and annotated outlines. The baseline comparison method is a performance-recall curve based on the receiver operating characteristic (ROC) analysis.

During the course of work, it has become apparent that collecting and annotating ever-growing databases sets restrictions for the future. The number of available images on the Internet is enormous, but the laborious annotation work cannot be extended beyond the limits which are now approaching. The supservised approach also biases development as state-of-the-art methods typically perform well with annotated sets, but fail to generalise for unseen categories. As a result, a new ultimate challenge has been posed: can the object categories be efficiently learnt in an unsupervised manner, i.e., without any information about image contents? This research topic is particularly new, and therefore, only a few methods have been proposed [12, 1]. These methods are computationally demanding and tested with only a small number of categories.

In this study, we propose a method which is unsupervised, but re-uses the most essential parts of the BoF approach: interest point detection, region (keypoint) description and codebook generation. For the categorisation part we simply replace any supervised method with the unsupervised learning principle occurring in the human brain: self-organisation. The principle is realised using the self-organising map by Kohonen [8]. The overall structure of our method is illustrated in Fig. 1.

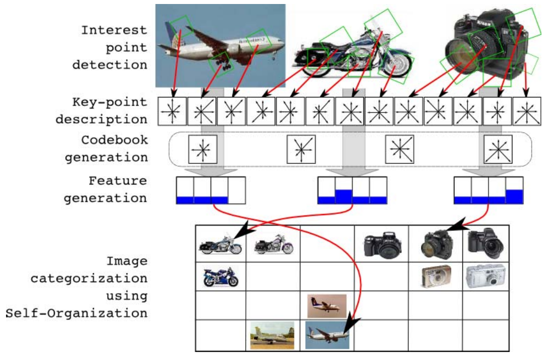


Figure 1. Bag-of-features based "self-categorisation".

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.116

Our main contribution is that we introduce the self-organisation principle and the Kohonen map (SOM) as a novel solution to unsupervised visual object categorisation problem. That completes our previous work where we showed that the SOM outperforms the baseline algorithm, k-means, for codebook generation [7]. Now, the self-organisation principle is exercised at the all levels of unsupervised BoF. In addition, we point out how the unsupervised approach rapidly collapses as the number of categories increases, which, in turn, brings up a very important consideration for the future research.

1.1. Related work

The interest point detection, region description and codebook generation parts of our method are similar to any other supervised BoF based method [2, 13, 11].

Only a few unsupervised VOC methods have been proposed. Sivic et al. [12] presented a method which is able to automatically build a hierarchical model of the visual appearance. The method utilises hierarchical latent Dirichlet allocation (hLDA), which produces a tree, where the root represents an average over all images and its subnodes more specific visual appearance (leaf nodes being the most specific categories). Bart et al. [1] compute statistics of the images utilising co-occurrence of the same codebook codes and they produce a "taxonomy" tree as well. As compared to our method, the main difference is that we are not trying to implement an explicit model to discriminate the categories, but allow input data to automatically organise via the self-organisation principle. The categorisation occurs naturally as it is the essence of competitive learning. The advantage of our approach is that it is capable to scale up to thousands of categories since it is not computationally as demanding as the Sivic et al. or Bart et al. methods.

The most relevant work related to ours is the PicSOM system developed by Laaksonen et al. [9]. The PicSOM uses the SOM for unsupervised categorisation, but does not otherwise utilise the BoF processing stages. The PicSOM uses more traditional features for the categorisation. Therefore, our work can be seen as an extension of the PicSOM method to meet the current state-of-the-art with the BoF approach.

2. Visual object categorisation

2.1. General bag-of-features framework

In Fig. 1, our bag-of-features approach is illustrated. First, interest points or regions are detected from images. Second, these regions are converted to scale and rotation invariant descriptors in the keypoint description step. In the third step, a codebook is constructed using the descriptors. In the original model the codebook generation is performed during the training phase using clustering algorithms, such as the k-means [2] or SOM [7]. In the best methods, however, the training ground truth is used to refine and probe more efficient codebooks [5, 10]. In the feature generation step, extracted image keypoints are assigned with codes from the generated codebook. A standard feature is the frequency vector over the codebook codes - "a bag of features". Finally, a category is assigned by feeding the feature vector to a classifier, such as the support vector machine (SVM) [2].

2.2. Changing framework from supervised to unsupervised

In supervised methods, the codebook generation step is performed in the training stage prior to category classifier training. Therefore, the best results can be achieved by utilising the ground truth information both in the codebook generation and classifier desing and optimisation stages. The ground truth is not available for unsupervised methods, and therefore, the best methods which couple categorisation and codebook generation are not anymore usable (e.g., [5, 10]). No classifier can be trained either. For the unsupervised methods, the codebook generation and image categorisation steps need to be removed. In Fig. 1 they are replaced by the SOM. An unsupervised method is expected to output the same "labels" for images which represent objects from the same category. The unsupervised performance can be evaluated using the formula proposed by Sivic et al. [12]. In the second, the performance is measured by computing accuracy of each SOM node and selecting the best to represent each category. The node performance, $p_{t,i}$, is computed as

$$p_{t,i} = \frac{GT_i \cap P_t}{GT_i \cup P_t} \tag{1}$$

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, $perf$, is then

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

where N_c is the number of categories. Our measure is analogous to the Sivic's approach, except that each SOM node is considered as a leaf node, and thus, there is only one "level" in our hierarchy. In the first experiment, we computed the performance of the system in similar manner with supervised methods, since Sivics measure is more laborious to categorization errors than the commonly used performance measure defined

$$\mu_{acc} = \frac{1}{N_c} \sum_{i=1}^{N_c} Acc(i) \tag{3}$$

Where $Acc(i)$ is categorization accuracy of class i and the performance, μ_{acc}, becomes average categorization accuracy over all classes. Class categorization accuracy $Acc(i)$ is then

$$Acc(i) = \frac{\#Correctly\ categorized}{\#Test\ images} \tag{4}$$

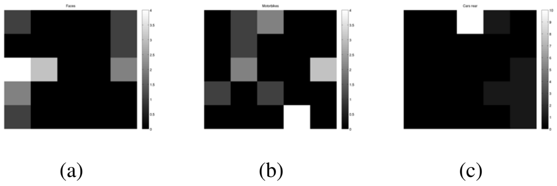
i.e. it is ratio between correctly categorized images and all test images from one class. We used this method in the first experiment by building a category map using training images. We used the category map to predict categories of unknown test images. However, we did not use any supervision when we build the category map. We used labelled information only when we computed the accuracy of categorization using Eq. 3 and 4.

2.2.1 Embedding self-organisation into the general framework

Changing the codebook generation step in Fig. 1 to unsupervised is easy: we just do not couple generation and classification, but utilise an unsupervised clustering. Advanced techniques, such as object-specific labelling or removal of rarities [13], cannot be performed.

A more complicated problem is the image categorisation part which requires some form of classification. Fortunately, categorisation is the essence of self-organisation where similar feature vectors are organised close to each other [8]. This results from the competitive learning principle where units compete for inputs and only the most similar unit, best matching unit (BMU), receives the reward (winner takes all). Typically the BMU spreads reward to its neighbours, i.e., the BMU units and their neighbours adapt to their best matching inputs, which leads to a topology-preserving mapping (features close to each other in the SOM are close to each other in the original space).

In our case, a 2-D SOM is first randomly initialised, and then, codebook histograms are fed to the SOM algorithm. The SOM algorithm iteratively trains the map by randomly picking a vector, finding its best matching unit (BMU) and adjusts the BMU weights according to the SOM learning rule. In addition to the BMU, also its neighbourhood units are accordingly adjusted. Finally, images assigned to the same unit are considered to belong to the same category. In Fig. 2 is demonstrated how the three different categories from Caltech101 database are located on the same SOM of size 5×5. The mapped locations are clearly distinct leading to a successful automatic self-categorisation. The selected size of the SOM affects to the accuracy, but in this study we experimentally selected optimal values and postponed this issue to the future research. It was found out, however, that the effect is not drastic as also demonstrated in the experimental part.



(a) (b) (c)

Figure 2. Result of self-categorisation (intensity denotes the number of instances assigned to a specific map unit): (a) faces, (b) motorbikes, (c) cars.

3. Experiments

In the experimental part of this work, we first demonstrate how the unsupervised approach is able to learn object categories without any prior knowledge. For comparison, results of a more sophisticated method [2] are also included. Secondly, we utilise the Sivic's performance measure for unsupervised methods and use the Caltech-101 data set to test our approach. The second experiment demonstrates how the system performance rapidly decreases, collapses, as the number of categories increases. That can appear for other unsupervised methods as well, but cannot be proved since results only for a few categories have so far been reported.

In the first experiment, we utilised the same data, Caltech4 and car side images, as in [2]. We generated various size codebooks, first using the k-means (kmeans - SOM) as in the original study, and then using the SOM as in [7] for the codebook generation (SOM - SOM). In our method, the classification was done unsupervised using the SOM instead of a SVM classifier. The size of the SOM was experimentally set to 30 × 1 in k-means - SOM and 20 × 1 in SOM - SOM case. The results are reported in Table 1. This experiment revealed the fact that unsupervised learning is not as much beyond the strongly supervised approach as would be expected. By replacing the supervised SVM with the SOM, 74.3 % accuracy was achieved and slightly improved to 75.1 % by replacing the k-means with the SOM. This is only 21.0 % beyond the sophisticated supervised approach in [2]. The results are reported in Table 1.

Next, we used the Caltech101 data set to study how well our method generalises to an increasing number of categories. The categorisation performance as a funcion

Table 1. Performance for Caltech-4 + car side.

Category	k-means-SOM	SOM-SOM	Dance et al. [2]
average	74.3	75.1	96.1

of number of categories is shown in Fig. 3, where it is clear that the method performance collapsed rapidly as the number of categories increases. This was not explicitly studied in any of the cited studies of unsupervised categorisation and they reported results only for a small number of categories.

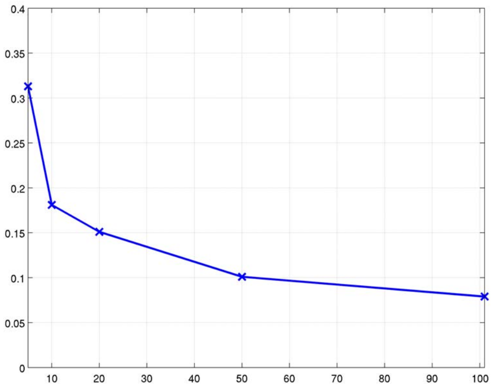


Figure 3. Performance for Caltech-101 as the function of the number of categories.

The performance of the system is 31.3 % when the number of categories is 5 with the SOM size 7 (1 × 7). The highest performance (18.13 %) with 10 categories is achieved using a SOM with 13 (1 × 13) units. With 20 categories 15.1 % performance is achieved using 51 (3 × 17) units. When the number of categories is increased to 50 categories, the best performance (10.1 %) is achieved with 171 (9×19) units. With 101 categories, the highest performance (7.9) % is achieved with 323 (17 × 19) units. This experiment illustrates the problem of many categories. The performance with a large number of categories can be improved slightly by increasing the size of the SOM that is used to categorise the image feature vectors. However, the differences between the performances with a small SOM and a large SOM are not significant.

4. Conclusions

We proposed a novel approach for the unsupervised visual object categorisation and demonstrated its power by applying it to a task where the supervised methods perform very well in a similar framework, but under heavy supervised tuning. The results were very promising, and it is expected that the results can be significantly improved in the future. In addition, we demonstrated an undesired property, fast collapse of the unsupervised approach as the number of categories increases. We claim that the most important challenge in the future is to develop approaches which are able to avoid such behaviour. Considering the premature nature of this research topic, it is understandable that much more work is needed.

Acknowledgements

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

References

[1] E. Bart, I. Porteous, P. Perona, and M. Welling. Unsupervised learning of visual taxonomies. In *CVPR*, 2008.

[2] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *ECCV Workshop*, 2004.

[3] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, 2008.

[4] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR Workshop*, 2004.

[5] B. Fulkerson, A. Vedaldi, and S.Soatto. Localizing objects with smart dictionaries. In *ECCV*, 2008.

[6] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[7] T. Kinnunen, J.-K. Kamarainen, L. Lensu, and H. Kälviäinen. Bag-of-features codebook generation by self-organisation. In *Workshop on Self-Organizing Maps (WSOM)*, 2009.

[8] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[9] J. Laaksonen, S. L. M. Koskela, and E. Oja. Picsom - content-based image retrieval with self-organizing maps. *Pattern Recognition Letters*, 21(13-14):1199–1207, 2000.

[10] B. Leibe, A. Ettlin, and B. Schiele. Learning semantic object parts for object categorization. *Image and Vision Computing*, 26:15–26, 2008.

[11] M. Marszałek and C. Schmid. Constructing category hierarchies for visual recognition. In *ECCV*, 2008.

[12] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros. Unsupervised discovery of visual object class hierarchies. In *CVPR*, 2008.

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