Unsupervised visual object categorisation with BoF and spatial matching

Teemu Kinnunen[1], Jukka Lankinen[2], Joni-Kristian Kämäräinen[3], Lasse Lensu[2], and Heikki Kälviäinen[2]

[1] Department of Media Technology, Aalto University, Finland
[2] Machine Vision and Pattern Recognition Laboratory, Lappeenranta University of Technology, Finland
[3] Department of Signal Processing, Tampere University of Technology, Finland

Abstract. The ultimate challenge of image categorisation is unsupervised object discovery, where the selection of categories and the assignments of given images to these categories are performed automatically. The unsupervised setting prohibits the use of the best discriminative methods, and in Tuytelaars *et al.* [30] the standard Bag-of-Features (BoF) approach performed the best. The downside of the BoF is that it omits spatial information of local features. In this work, we propose a novel unsupervised image categorisation method which uses the BoF to find initial matches for each image (pre-filter) and then refines and ranks them using spatial matching of local features. Unsupervised visual object discovery is performed by the normalised cuts algorithm which produces the clusterings from a similarity matrix representing the spatial match scores. In our experiments, the proposed approach outperforms the best method in Tuytelaars *et al.* with the Caltech-101, randomised Caltech-101, and Caltech-256 data sets. Especially for a large number of classes, clear and statistically significant improvements are achieved.

1 Introduction

Visual object categorisation (VOC) has been one of the most active computer vision research topics for the last 5-7 years. The topic is important due to the enormous amount of visual data, images and videos, in the Internet and on personal devices, which we wish to search and index automatically based on the visual content.

The state of the art supervised VOC has improved remarkably in the last few years due to numerous new methods, increased computation power, and the amount of labelled data available for training [6]. Humans can recognise more than 30,000 object categories [1], but for the computational methods it becomes troublesome to learn and use such a large number of classes [5] and it would be extremely laborious to obtain sufficient amount of validated and labelled data. There are two possible solutions to the data problem: harvest new data automatically (*e.g.*, [25]) or perform clustering without the labels, i.e., use unsupervised image categorisation.

The state-of-the-art VOC methods for visual object categorisation are strongly supervised and cannot be easily adopted for the unsupervised setting. Tuytelaars *et al.* [30] showed that the popular Bag-of-Features approach with modern interest point detectors and descriptors, and a proper codebook and feature normalisation procedure achieves the best performance in the unsupervised setting. The main problem of the BoF is the complete omittance of spatial information. However, spatial "verification" of local features has been shown to be very advantageous for visual indexing and search [23,3,10]. Our main contribution in this work is a novel method for the unsupervised image categorisation using the BoF as a pre-processing stage to rank image matches, and then a random sampling based spatial matching procedure to re-rank the best matches. Our "spatial BoF" provides better unsupervised categorisation (Fig. 1) and performance superior to the state-of-the-art [30].

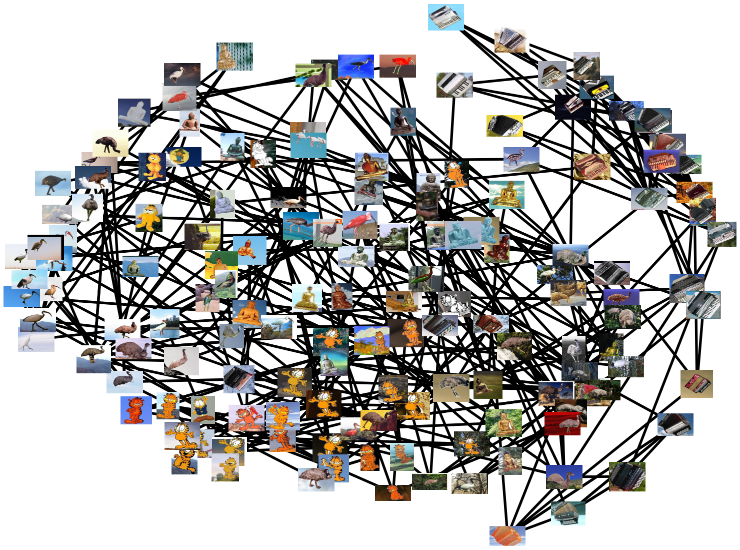


Fig. 1. A match graph. Edges represent similarity between a pair of images. Similarity is computed using spatial matching.

1.1 Related work

Some early works of image categorisation were misleadingly considered as "unsupervised" (*e.g.*, [31]), but were actually supervised/semi-supervised classification since training sets with class labels were used, and often also the object bounding boxes, provided for the most popular benchmarks, were used. A labelled training set is a general assumption in the mainstream works. For this supervised setting the methods' performance has improved remarkably and the current state of the art can be spotted from the results of the annual Pascal VOC challenge [6]. At the same time, the methods have decreasingly less in common with the original Bag-of-Features approach (BoF) which dates back to the seminal works of Sivic and Zisserman in 2003 [29] and Csurka *et al.* in 2004 [4]. Also critical arguments against the canonical mainstream direction have been raised. For example, Deng *et al.* [5] showed how the best methods for a small number of classes are not necessarily best for a large number of classes. In their work, Tuytelaars *et al.* [30] investigated the unsupervised setting and in their experiments the basic BoF using the k-means clustering performed better than the more complex methods, such as the latent Dirichlet allocation (LDA) [27].

Prior to our work, an improved BoF for unsupervised categorisation was proposed by Lou *et al.* [17] who used information bottleneck clustering. They reported slightly better results to Tuytelaars *et al.* , but as their method detected more categories than actually available, comparing the results is somewhat questionable. Our work was motivated by results in specific object recognition (not categorisation) where co-occurence of matching descriptors in a similar spatial "constellation" was used to match two images [23,3]. In our case, this is not straightforward as with categories the number of matches is very low, only a few, and the match quality is much worse, i.e. the best matching descriptors are rarely the spatially correct matches.

One of the first approaches to introduce spatial information to the bag-of-features model was introduced by Ponce et al. [24]. They introduced a spatial pyramid method that computes bag-of-features histograms on different levels of the pyramid. The idea is to encode a local appearance of the image within a bin of the pyramid to capture spatial information.

Krapac*et al.* [15] introduced a method that uses a probabilistic model to represent BOF histograms and then learns probabilistic models for matched local features. The method heavily relies on supervised learning and thus it is not suitable for unsupervised visual object category detection.

2 Unsupervised Bag-of-Features

The BoF approach originates from text document retrieval where the Bag-of-Words (BoW) approach was used to describe the contents of a text document [2]. The core idea is to describe the contents of a document by computing the frequencies of different words and use the histogram as a feature. The match similarity between two documents is straightforward to compute, for example, using the Euclidean distance or any other vector or histogram distance or similarity measure. The norm based distances are preferred due to their low computational cost. In image classification the idea is similar, but since no natural "visual dictionary" is available, one is constructed by clustering automatically extracted local image features (e.g., SIFT). For image classification based on the word/code histograms efficient machine learning methods, such as support vector machines, can be used [4]. In the unsupervised setting, the supervised classifier must be replaced with an unsupervised clustering method.

A variant of the above unsupervised BoF can be made by replacing any of its components with another similar method. For example, Tuytelaars *et al.* [30] tested k-means, LDA and spectral clustering for the categorisation stage. They also tested different methods for local feature detection (Hessian-Laplace [21], Harris-Laplace [20] and dense sampling). Detector combinations performed the best which is in correspondence with the findings in [22]. One more component is the histogram normalisation procedure for which Tuytelaars *et al.* found that the L2-norm performs best in the most cases.

In our work, the unsupervised BoF is used as a "pre-filter" before a random spatial scoring procedure described next. In our implementation, we replace k-means with the self-organising map (SOM) algorithm [14]. In our experiments, the SOM provides comparable results, but is less sensitive to the optimal codebook size selection than k-means. The pre-filtering stage using SOM was introduced by the authors in [12] and this work extends the method with the spatial scoring procedure. The spatial matching updates the order of best matching images and provides rank based pair-wise similarities for which we apply the normalised cuts algorithm [26] to form the final graph of image categories (Fig. 1).

3 Spatial matching using random sampling

Our sub-task is to find the best matches for a query image I given a database or ensemble of images to be categorised, i.e. short list output by the pre-filtering stage. We formulate the best match search as a scoring procedure and thus scores can be used in the next stage of clustering with the normalised cuts algorithm [26]. Let us assume that we have a "pre-filter" (BoF in our case) which provides N best matching candidates and we choose a sufficiently large N (100 in our case) in order to guarantee that at least one image from the same category appears with some probability. Next, we explain the spatial matching procedure which re-ranks the best N with the ultimate aim to bring the same category image to the first place. For spatial scoring, we adopt the local feature scoring procedure used at the core of the unsupervised image alignment method in [16]. Full unsupervised categorisation is performed by using each image at time as the query image. This can be achieved with the minimal computational cost by pooling the similarities over the previous iterations.

The main task of the spatial matching is to find a set of local image features which match under some pre-defined transformation. We chose to use similarity transformation because it is capable of detecting object in different locations, orientations and scales and has less free variables than affine and projective transformations, and thus, is faster to estimate. Trying all possible combinations is not feasible and thus we utilise random sampling. The number of random samples is the computational bottleneck of our method and thus needs to be carefully set. The rule of thumb is that we should sample as many as possible within the given time slot for each image. We randomly sample local features of the query image, select their matches in the candidate image, and re-compute similarity by descriptor distances using only the K best spatially matching features.

We describe the algorithm for a single query image I_q and a single candidate image I_c. These two images are represented by their local feature descriptors, $\mathbf{d}_q$ ($N_q \times 128$) and $\mathbf{d}_c$ ($N_c \times 128$), and the spatial locations of the features, $\mathbf{x}_q$ ($N_q \times 4$) and $\mathbf{x}_c$ ($N_c \times 4$). Since testing all possible combinations would be computationally expensive, we use random sampling and pick a sufficiently large number of samples R (100 in our case). In every sample, we select a random pair of features from the query image and their correspondences in the candidate image. Next, we estimate homography (similarity) using the direct linear transform (DLT) [9], which maps candidate features to the query space. DLT was chosen, because it has been used successfully for finding stable landmarks from a collection of images [16]. We check which features are within a pre-set distance (5% of the image diagonal in our case). We accept the correspondences if they are within $L = 5$ best descriptor matches. This is important as between two examples of a class the best matches are rarely the correct ones. The final match distance is the sum of descriptor distances $\mathbf{D}$ of K best matching features. The spatial matching can be seen as a spatial verification step for the local features. The procedure is sketched in Algorithm 1. For clarity, the algorithm is given for a single query and a single candidate image. In the experiments, it was executed for all images in a given dataset and for N=100 best BoF pre-filtered candidates of each.

Algorithm 1 Spatial matching algorithm.

```
Require: $\mathbf{d}_q$, $\mathbf{d}_c$, $\mathbf{x}_q$, $\mathbf{x}_c$
Compute the descriptor (SIFT) distance matrix: $\mathbf{D} \leftarrow \mathit{SIFTDist}(\mathbf{d}_q, \mathbf{d}_c)$.
for $R$ random iterations do
    Select two random query features $\mathbf{x} = \{\mathbf{x}_{q1}, \mathbf{x}_{q2}\}$ and their random correspondences $\mathbf{x}'$ within the $L = 5$ best matches in $\mathbf{D}$.
    Estimate the homography $\mathbf{H} \leftarrow \mathit{homography}(\mathbf{x}', \mathbf{x})$ using the DLT method [9].
    Transform all candidate image points to the query image space using $\mathbf{H}$.
    Select query points $\{\mathbf{p}\}$ which have a spatial match within the image diagonal normalised distance $\tau = 0.05$ and a descriptor match within the $L = 5$ best in $\mathbf{D}$.
    From the points $\mathbf{p}$ select the $K$ best $\mathbf{p}_K$ according to the descriptor distances $\mathbf{D}$
    Form the spatial $\mathit{score} \leftarrow \mathit{sum}(\mathbf{D}_{p_k})$ and update $\mathit{bestScore}_{q,c}$ if necessary.
end for
return $\mathit{bestScore}_{q,c}$.
```

The work flow of our method is as follows: for each image 1) find the best BoF matches, 2) for the best matches, find the sub-set of features which match also spatially, 3) re-rank the best matches using the sum of distances of the descriptors of the K best spatial feature matches. The parameter values N, K, R, L and τ were selected according to our preliminary tests and the algorithm seems to be stable also for other values. N and R are mainly affected by the available computation resources.

3.1 Experimental validation

To experimentally validate our approach, we tested it on the randomized Caltech 101 images [13] for which the object are randomly translated, rotated, and scaled and the backgrounds replaced with random Google landscape images. The randomisation removes some of the undesired biases of the original dataset. We tested every image as a query image, computed the BoF distance to every other image, and then, for the spatial matching selected the best BoF matches and re-computed the distances using increasing number of spatial matches (x-axis). The results are shown in Fig. 2 where the x-axis represents the number of best candidates and y-axis represents the probability that the correct class example appears within the candidates. The spatial matching clearly improves the rank of the correct matches in the beginning which is exactly desired behaviour. The significance becomes clear in the next section where also the ±2× standard deviations are plotted for the used performance measures. After the 20 best matches, the improvement of the image re-ranking based on spatial matching starts to get smaller and after the 60 best matches, spatial matching improves results only slightly. The reason can be the fact that the visual appearance of the seed image and candidate image differs considerably, and thus, local feature matches are mostly incorrect matches which cause problems for spatial matching.

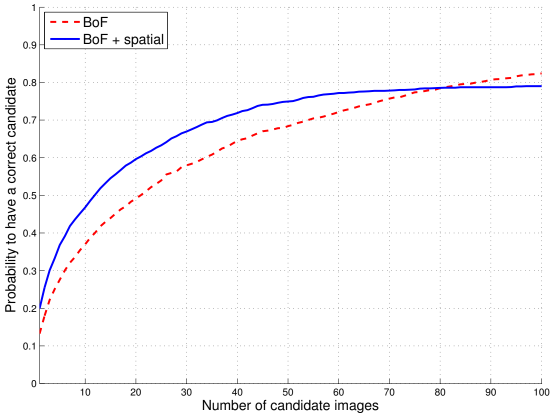


Fig. 2. Retrieval performances of BoF [12] and BoF + spatial matching. The red dashed line denotes the BoF results and the blue solid line BoF + spatial matching.

3.2 Unsupervised categorisation

For the unsupervised categorisation step, we follow the approach by Kim *et al.* [11] who construct a similarity graph of images using pair-wise image similarities and the normalised cuts algorithm [26]. We replace their nearest neighbour BoF similarity with our spatial matching similarity ranks, i.e. 1 for the best spatial match, 2 for the second best and so on. The rank values improved the accuracy in our experiments as compared to the plain distances. The ranks need to be converted similarity values where larger value is better. From the spatial matching step, we have distances from each image to N images provided by the BoF pre-filtering. Using the best score outputs of Algorithm 1 the image matches can be sorted into the ascending order and the ranks assigned for N

best. To convert the rank to a similarity we use the simple formula:

$$S(i,j) = \frac{N}{rank(i,j)} \tag{1}$$

and assign 0 to the unknown image pairs. Asymmetries in the similarity matrix **S** are fixed by setting $S(i,j) = max(S(i,j), S(j,i))$. The final clustering result is computed by the normalised cuts algorithm [26] which takes the similarity matrix as an input.

4 Experiments

In this section we compare our method of BoF+spatial matching+normalised cuts (NC) to the BoF [12] (BoF+NC in our case) which outperformed the best method in Tuytelaars *et al.* [30]. Other parts of our algorithm are the Hessian-Affine [19] interest point detector and the SIFT descriptor [18]. Extracted features are matched to the codebook codes using the SIFT distances and a histogram is constructed from all codes. The histograms are normalised by the L2-norm. The size of the codebook is 10k. We use $N = 100$ best BoF matches for the spatial matching. The spatial scores are converted to a similarity matrix and the final categorisation is obtained by the normalised cuts algorithm as described in Sec. 3.2 by fixing the number of nodes to the number of ground truth categories – this is justified for the evaluation purposes and to compare the results with [30].

4.1 Performance evaluation

Performance evaluation and comparison of the unsupervised methods is not straightforward. That is due to the model selection problem, i.e. selection of the optimal number of categories: how to compare two methods which produce different number of categories. For that reason, alternative performance evaluation metrics have been introduced. We briefly explain the two most prominent: Sivic *et al.* [28] and Tuytelaars *et al.* [30]. We report our results with the both measures since they have distinct advantages and disadvantages. Sivic's performance is more intuitive and can be used to evaluate clustering results also when the number of clusters is unequal to the number of ground truth categories. On the other hand, the conditional entropy by Tuytelaars *et al.* does not penalise so dramatically for the false detections. Moreover, the two methods do not necessarily produce the same results since Sivic *et al.* method computes the average value over categories making category contributions equal while in Tuytelaars *et al.* every image contributes equally which biases results toward categories with more examples. In our case, it turned out that the both methods provide consistent and comparable results in our experiments.

Sivic *et al.* [28] performance measure

Sivic *et al.* proposed a performance evaluation method that computes categorisation accuracy of each node for each category and chooses the node with the highest accuracy for each category. Then, the categorisation performance is computed by computing the mean over the categories. The categorisation accuracy of a single node, $p(t, i)$, is computed as

$$p(t,i) = \max_i \frac{|GT_i \cap P_t|}{|GT_i \cup P_t|} \ , \tag{2}$$

where GT_i are the ground truth images of the category i, P_t are the images assigned to node t, and $|\cdot|$ denotes the number of images. The average performance is computed as

$$perf = \frac{1}{N_c} \sum_{i=1}^{N_c} \max_t \ p(t,i) \ , \tag{3}$$

where N_c is the number of categories. The method ultimately chooses nodes that give the best categorisation performance per each object category, and then computes the average over these nodes. Large values represent better performance.

Tuytelaars *et al.* [30] performance measure
Tuytelaars *et al.* proposed a performance evaluation method based on the conditional entropy defined as

$$H(X \mid Y) = \sum_{y \in Y} p(y) \sum_{x \in X} p(x \mid y) \ log \frac{1}{p(x \mid y)} \ , \tag{4}$$

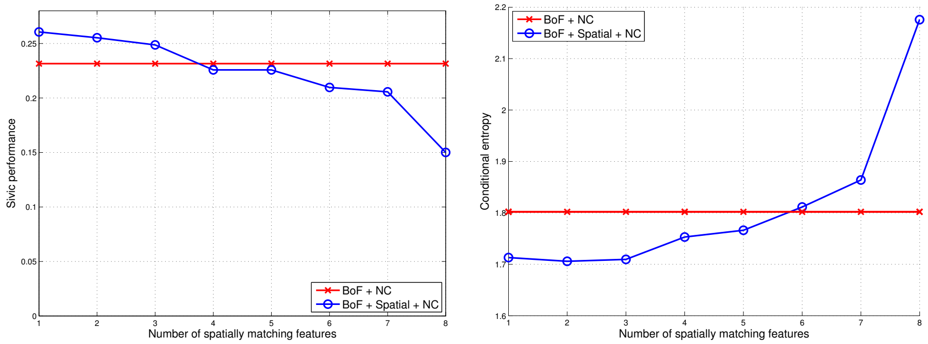
where Y stands for the cluster labels and X for the ground truth labels. Conditional entropy measures how certain one can be that the image actually belongs to the cluster and the measure is justified by information theory. Smaller values represent better performance.

4.2 Caltech-256(20) from Tuytelaars *et al.* [30]

In the first experiment, the performance of the BoF approach with and without spatial matching were compared to the state-of-the-art in Tuytelaars *et al.* [30] using the same data. That is, we selected the same 20 categories from the Caltech-256 [8] image set and followed their procedure to select the all images from the categories. The amount of images varies between categories and thus affects differently to the two performance measures.

We tested our method by varying the number of the best spatially verified features K. The results are shown in Fig. 3. The spatial matching clearly improves categorisation accuracy for $K = 1, 2, 3$ spatially verified features according to the Sivic *et al.* measure and for $K = 1, \ldots, 5$ according to the Tuytelaars *et al.* measure. The difference is due to the uneven amount of images per category. However, the curves behave consistently and at least for the values $K = 1, 2, 3$ the spatial matching improves the unsupervised categorisation performance. Our SOM-based BoF is nearly the same to the best conditional entropy reported by Tuytelaars *et al.* (1.78) while the BoF + spatial matching is clearly better

(1.71). The statistical significance of the difference between the BoF and its spatial extension becomes clear in the next experiment where also the two standard deviations are plotted.



l

(a) Sivic *et al.* performance

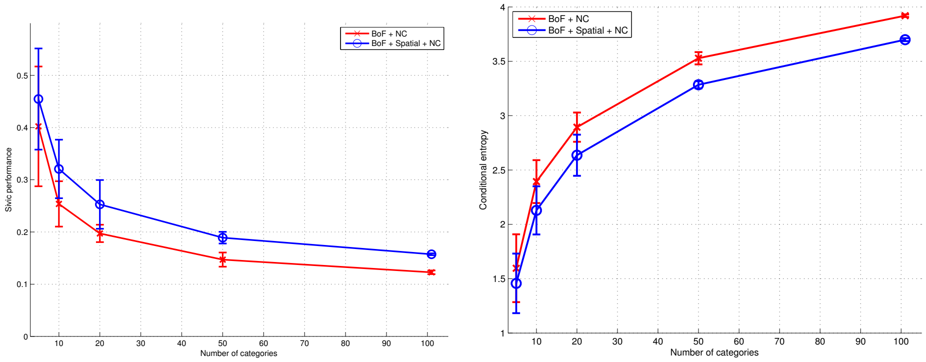
(b) Tuytelaars *et al.* conditional entropy

Fig. 3. Results for the Caltech-256(20) experiment in [30]. Unsupervised categorisation performances for the BoF approach (red crosses) and BoF with spatial matching (blue circles). X-axis is the number of spatially matching features used K. Note that for Sivic *et al.* larger values are better and for Tuytelaars *et al.* measure the smaller values are better.

4.3 Caltech-101

In this experiment, we investigated BoF vs. BoF + spatial matching as the function of number of categories, i.e. how well the methods scale up. Moreover, we included the same number of images per category to make the performance measures more comparable. The Caltech-101 benchmark [7] data was used. For 10 random iterations, 30 images were randomly chosen from each category for the unsupervised categorisation task. This was repeated for 5, 10, 20, 50, and 101 randomly selected categories. The results are shown in Fig. 4.

In this experiment, the two performance measures provided identical results and the BoF + spatial matching outperforms the BoF consistently for any number of categories. The ±2× standard deviations over the random iterations are also plotted to demonstrate the performance variation distributions and their overlap with 95% confidence. The statistical significance of the results is less evident for small numbers of classes, but is very clear as the number of categories increases - the overlap is practically zero for more than 50 classes. Results show that the spatial matching can significantly improve the performance of the standard BoF in unsupervised image categorisation.



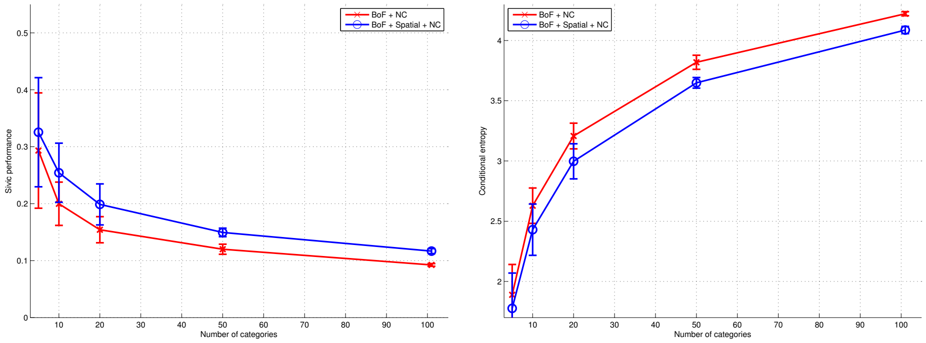
(a) Sivic *et al.* performance (b) Tuytelaars *et al.* conditional entropy

Fig. 4. Average performance and ±2× standard deviations (bars) over random iterations for the Caltech-101 experiment. X-axis is the number of categories and $K = 4$.

4.4 Randomised Caltech-101

We repeated the previous Caltech-101 experiment, but changed the images to the randomised Caltech-101 dataset [13]. In r-Caltech-101 the backgrounds of the original Caltech-101 are replaced with random landscape images, and the foregrounds are translated and rotated randomly making the image set considerably more difficult than the original.

Results of this experiment are shown in Fig. 5. The performances are expectedly weaker than for the original data in Fig. 4, but the conclusion that BoF + spatial matching is superior is verified and is also statistically reliable.



(a) Sivic *et al.* performance (b) Tuytelaars *et al.* conditional entropy

Fig. 5. Performance curves for the randomised Caltech-101 experiment.

5 Conclusion

In this work, we improved the state-of-the-art in unsupervised image categorisation by combining the Bag-of-Features (BoF) approach with the novel spatial matching procedure. In our method, the BoF is an effective pre-filter which produces a set of best matches for each image. Then, the spatial matching is executed to re-rank the candidate set using only the best local features which match by their spatial configuration. A random sampling based matching procedure handles this part. The re-ranked values are used to construct a similarity matrix for the normalised cuts algorithm which produces the required number of classes. To study the performance, we performed a set of experiments on the popular visual object categorisation benchmarks and showed statistically significant superiority to the BoF based unsupervised categorisation and a state-of-the-art method in the literature. In the future work, we will address the problem of automatic selection of the number of categories and computational issues.

Acknowledgements

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

References

1. Biederman, I.: Recognition-by-components: A theory of human image understanding. Psychological Review 94(2), 115–147 (1987)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3(4-5), 993–1022 (May 2003)
3. Chum, O., Matas, J.: Unsupervised discovery of co-occurrence in sparse high dimensional data. In: CVPR (2010)
4. Csurka, G., Dance, C., Willamowski, J., Fan, L., Bray, C.: Visual categorization with bags of keypoints. In: ECCV Workshop on Statistical Learning in Computer Vision (2004)
5. Deng, J., Berg, A., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? In: ECCV (2010)
6. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC) challenge. IJCV 88(2) (2010)
7. 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)
8. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology (2007)
9. Hartley, R., Zisserman, A.: Multiple View Geometry in computer vision. Cambridge press (2003)
10. Jégou, H., Douze, M., Schmid, C.: Improving bag-of-features for large scale image search. International Journal of Computer Vision 87(3), 316–336 (feb 2010)
11. Kim, G., Faloutsos, C., Hebert, M.: Unsupervised Modeling of Object Categories Using Link Analysis Techniques. In: CVPR (2008)

12. Kinnunen, T., Kamarainen, J.K., Lensu, L., Kälviäinen, H.: Unsupervised object discovery via self-organisation. Pattern Recognition Letters 33(16), 2102–2112 (2012)
13. Kinnunen, T., Kamarainen, J.K., Lensu, L., Lankinen, J., Kälviäinen, H.: Making visual object categorization more challenging: Randomized caltech 101 data set. In: ICPR (2010)
14. Kohonen, T.: The self-organizing map. Proc. of the IEEE 78(9), 1464–1480 (September 1990)
15. Krapac, J., Verbeek, J., Jurie, F.: Modeling spatial layout with fisher vectors for image categorization. In: Proc. of International Conference on Computer Vision. pp. 1487–1494 (2011)
16. Lankinen, J., Kamarainen, J.K.: Local feature based unsupervised alignment of object class images. In: Proc. of British Machine Vision Conference (2011)
17. Lou, Z., Ye, Y., Liu, D.: Unsupervised object category discovery via information bottleneck method. In: Proc. of the Int. Conf. on Multimedia (2010)
18. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJVC 20, 91–110 (2004)
19. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. IJCV 65(1/2) (2005)
20. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: ICCV (2001)
21. Mikolajczyk, K., Schmid, C.: An affine invariant interest point detector. In: ECCV (2002)
22. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: ECCV (2006)
23. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR (2007)
24. Ponce, J., Berg, T., Everingham, M., Forsyth, D., Hebert, M., Lazebnik, S., Marszalek, M., Schmid, C., Russell, B., Torralba, A., Williams, C., Zhang, J., Zisserman, A.: Dataset issues in object recognition. In: Workshop on Category Level Object Recognition. pp. 29–48 (2006)
25. Schroff, F., Criminisi, A., Zisserman, A.: Harvesting image databases from the web. T-PAMI 33(4) (2011)
26. Shi, J., Malik, J.: Normalized cuts and image segmentation. T-PAMI 22(8) (2000)
27. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering object categories in image collections. In: ICCV (2005)
28. Sivic, J., Russell, B.C., Zisserman, A., Freeman, W.T., Efros, A.A.: Unsupervised discovery of visual object class hierarchies. In: CVPR. pp. 1–8 (2008)
29. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: CVPR (2003)
30. Tuytelaars, T., Lampert, C., Blaschko, M., Buntine, W.: Unsupervised object discovery: A comparison. IJCV 88(2) (2010)
31. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. In: ECCV (2000)