Making Visual Object Categorization More Challenging: Randomized Caltech-101 Data Set

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

Abstract

Visual object categorization is one of the most active research topics in computer vision, and Caltech-101 data set is one of the standard benchmarks for evaluating the method performance. Despite of its wide use, the data set has certain weaknesses: i) the objects are practically in a standard pose and scale in the middle of the images and ii) background varies too little in certain categories making it more discriminative than the foreground objects. In this work, we demonstrate how these weaknesses bias the evaluation results in an undesired manner. In addition, we reduce the bias effect by replacing the backgrounds with random landscape images from Google and by applying random Euclidean transformations to the foreground objects. We demonstrate how the proposed randomization process makes visual object categorization more challenging improving the relative results of methods which categorize objects by their visual appearance and are invariant to pose changes. The new data set is made publicly available for other researchers.

1. Introduction

Visual object categorisation (VOC) has been one of the most active computer vision research topic during the recent years. The mainstream methods are based on the well-known "bag-of-features" (BoF) approach [1, 15, 10]. To evaluate the performance of the methods, researchers use public benchmarks which contain training and test images, the ground truth, that is, category labels and annotated object segments, and an evaluation protocol. The most important benchmarks are Caltech-101 [4], Caltech-256 [5] and LabelMe [13] data sets, and the annual Pascal VOC challenge [3, 2]. Caltech-256 [5] and LabelMe [13] provide the most difficult challenge, but Caltech-101 is still important for the basic research since they include 3D pose variations and contain of multiple objects in a single image. The images in Clatech-101 are of moderately good quality, the categories are well selected and annotated, and most importantly, its pose variation is controlled. Caltech-101 has been claimed to be too easy, but we argue that this partly results from the bias caused by data selection, which is studied in this work.

The images in Caltech-101 have many good properties, but there exists also certain undesirable properties due to the selection process of the images. This data selection bias may result distorted evaluation results and consequent misinterpretation of BoF methods' applicability. Specifically, i) the objects are mainly in a standard pose and scale in the middle of the images and ii) background variability is insufficient in certain categories making it a more characteristic feature than the object visual appearance. Our main contributions are i) quantitative analysis of the bias caused by data selection and ii) a new randomized version of Caltech-101, where the main factors resulting to the bias are reduced. In the randomized set, the backgrounds are replaced with random landscape images from Google and the strong prior of the object placement and pose is reduced by random Euclidean transformations. The randomization process is illustrated in Fig. 1 and the complete data set is available at http://www.it.lut.fi/project/visiq/.

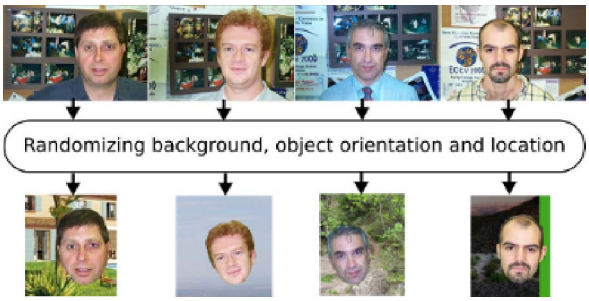

Figure 1. Randomized Caltech-101.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.124

1.1 Related work

The most important VOC data sets are Caltech-101 [4], its extension Caltech-256 [5], and the separate LabelMe [13]. The Caltech sets stimulated arranging the Pascal VOC challenges from 2005 to date [3, 2]. Prior to our work, the weaknesses in the collected images have been pointed out by Ponce et al. [12].

Fei-Fei et al. [4] have set the standards for VOC research by using Caltech-101. However, the published performance improvements saturated quite fast, and Ponce et al. [12] identified significant weaknesses in Caltech-101: the images are not challenging enough since the objects are captured from similar view points causing small variation in pose and scale, and often the object backgrounds are undesirably similar. These issues are visible in the original images in Fig. 1 where all the faces are frontal, their pose and location is very similar, and some similar background structures appear in every image. Ponce et al. proposed a new data set collected from Flickr images. The set was used in Pascal VOC 2005 and is continuously updated for the annual competition. In 2005, there were only four categories, but in the 2009 competition, the number was increased to 20 [2].

Our randomized Caltech-101 circumvents the problems related to the pose, location, and scale of the object, and to the background. The remaining difference to the other available sets is the fact that the randomized data is still intrinsically 2-D whereas the others contain images in all 3-D poses. Genuine 3-D data is extremely difficult for computer vision methods, but it is questionable whether the problem is learnable just from the data, or should the 3-D pose information be provided as well. Lately, state-of-the-art results have been reported in [14], but they used separate 3-D data in training. We agree that genuine 3-D data is the ultimate challenge, but we argue that 2-D data sets, such as Caltech-101, are still important for method development, and therefore, making them more challenging is important. On the other hand, categorization can be performed, in principle, using 2-D methods which are trained with objects in different poses separately (car front, car rear, car side, etc.)

2. Bag-of-Features

We use our own implementation of the general BoF approach illustrated in Fig. 2. In the first stage, we use the standard SIFT procedure [9] to detect and describe interest points. In the second stage, we cluster the descriptors with the self-organising map (SOM) [7] instead of, for example, K-Means [11, 17]. The SOM nodes form the codebook. The codebook is used to describe the image content by forming a codebook histogram, that is, the occurrences of codes in an input image. The histogram is used for the classification using, for example, the 1-NN rule. The SOM-generated codebooks provide similar and often superior results to the K-means. More details of our system and its performance can be found in [6].

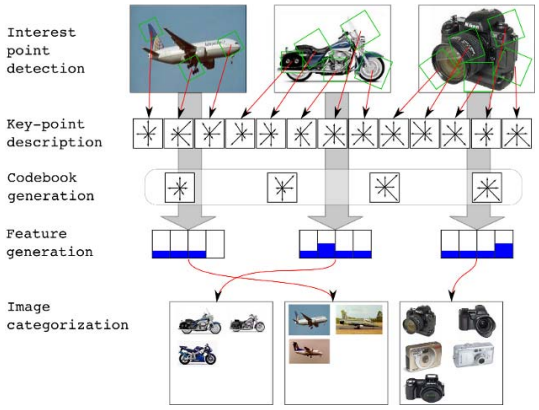

Figure 2. The applied Bag-of-features approach.

3. Randomized Caltech-101

Borders of the foreground objects have been annotated for all the images in Caltech-101. Using the annotation data, the foreground regions can be cropped, geometrically transformed, and drawn onto other backgrounds. The standard Euclidean transformation options are translation, scale, and rotation. In our randomization process, we apply random rotations of $\pm 20°$. The range of angles was selected to limit the variations below the direction sensitivity of the human visual system [16]. Random translations were achieved by positioning the transformed regions randomly onto the random background images from Google. The scale was not explicitly changed, but the varying size of the random backgrounds implicitly changed the proportional object scale.

The minor pose and alignment variance of the original images is visible in the middle column in Fig. 3, where the selected categories are clearly recognizable from their average images. On the other hand, the average images become blurry when the averages have been computed after random rotations only. This can be seen clearly for the natural objects in the rightmost column of the figure, while the two simple human-made objects, stop sign and ying-yang symbol, are still recognisible due to the rotation limits. It is evident, that the randomized rotations and translations, and the implicit scale changes prevent the utilization of the strong prior related to the object alignment and pose in the original Caltech-101 images for VOC learning.

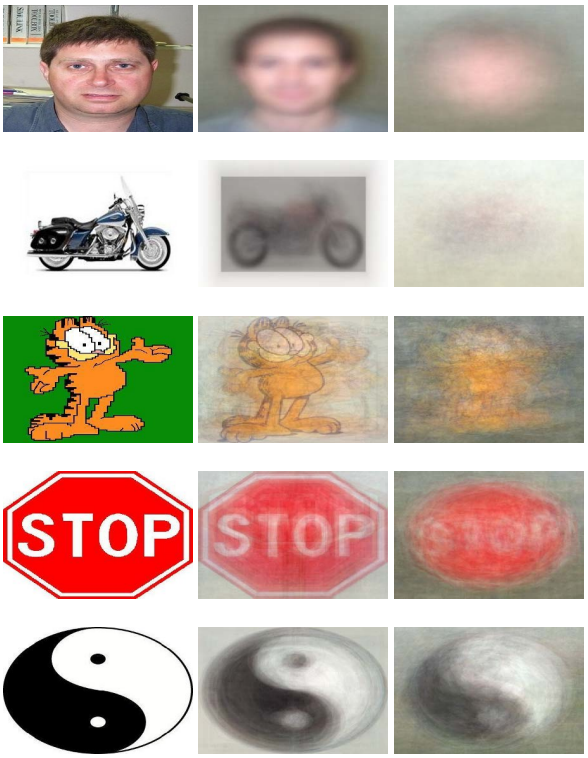

Figure 3. Examples of original Caltech-101 images (left), average images of the original ones (middle), and the average of randomized images (right).

The importance of background randomization is not evident from the average images in Fig. 3, but is quantitatively verified by the experiments in the next section. We gathered some natural scenery and landscape images from the Internet using Google and embedded the foreground objects onto these randomly selected backgrounds at random locations. It is noteworthy, that the images cannot be considered as "natural" anymore because the objects do not appear in their typical scene. However, methods based purely on the object appearance and tolerating geometric variance should remain unaffected while methods which exploit the insufficient background variation in Caltech-101 may severely fail.

4. Experiments

To quantitatively support our observations, we conducted several experiments which are described in the following. According to the standard VOC evaluation procedure, we utilised categorisation performance as the quantitative evaluation measure. The performance was computed as average classification accuracy over classes as it is presented by Lazebnik et al. [8]. We computed the performance values as a function of the number of categories. The asymptotic VOC behaviour is important since the methods should ultimately cope with thousands or even hundreds of thousands of categories.

The experimental procedure is randomized itself: for each number of categories, 10 independent iterations were performed by first selecting random categories and 30 random training images for each category. 20 images, or what was left out from the training process, were used in testing. The optimal codebook size was experimentally selected by trying different sizes from 50 to 10 000, and choosing the codebook that performed the best on average.

There were five data configurations for which the BoF method described in Section 2 was tested and the codebook size optimised: i) the original Caltech-101 data (*Original*), ii) only the Caltech-101 foreground objects (*Objects only*), iii) only the Caltech-101 backgrounds (*Backgrounds only*), iv) the original images with random backgrounds (*Random backgrounds*), and v) the full randomized images according to Section 3 (*Random bg + rot + trans*).

The results from all the experiments are shown in the single graph in Fig. 4. The surprising result is that, on average, the backgrounds provide more discriminative features than the foreground objects themselves. This reflects both the importance of background (scene) information for the recognition, but also the limited variability of background in the data set. Caltech-101 images reflect more the process of how the data was selected than how the objects appear in the natural scenes – the bias from data selection. The foreground information performed only slightly better than the full images, which reflects the importance of scene analysis for VOC. As an important result, the randomization of the backgrounds yielded to a significant collapse in the performance. The collapse was even more severe with the random transformation. The result can be explained by the SIFT features, which are invariant to scale and rotation, but in practise not perfect due to their limited amount of discrete "bins".

5. Conclusions

In this work, we studied the effects of data selection to training of visual object categorisation methods. Specifically, we pointed out the findings similar to Ponce et al. [12], that the widely-used Caltech-101 data set has certain weaknesses for VOC research: insufficient variation in object pose and alignment and unrealistically strong dependence between the background and objects. Ponce et al. solved these issues by collecting a new data set, which also introduced a new source

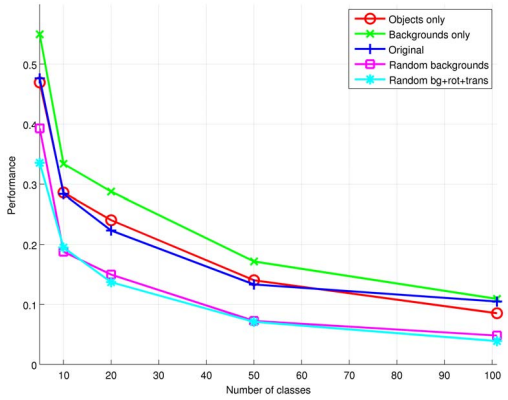

Figure 4. Performance when using different types of data.

of variation: the full 3-D pose view point change. Our contribution is an alternative solution where Caltech-101 was made more challenging by randomization. We demonstrated the advantages of randomization qualitatively by plotting the examples of average images and quantitatively by running performance evaluations with our BoF algorithm for various data configurations. For Caltech-101, the bias caused by data selection is very strong, which is evident from the fact that the best results were achieved using the image backgrounds only. The bias was reduced by the randomization procedure proposed in this work.

We believe that Caltech-101 is still useful for research since it provides good-quality category data with controlled 2-D variation. To facilitate future work, we have published the new randomized data at our web site: http://www.it.lut.fi/project/visiq/.

Acknowledgements

The authors wish to thank the Academy of Finland and partners of the VisiQ project (no. 123210) for support (URL: http://www2.it.lut.fi/project/visiq/).

References

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

[2] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results.

[3] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results.

[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 on Generative-Model Based Vision*, 2004.

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

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

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

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Conf. on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 20:91–110, January 2004.

[10] M. Marszałek and C. Schmid. Constructing category hierarchies for visual recognition. In *Proc. of the ECCV*, 2008.

[11] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *Proc. of the ECCV*, 2006.

[12] J. Ponce, T. Berg, M. Everingham, D. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. Russell, A. Torralba, C. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In *Workshop on Category Level Object Recognition*, pages 29–48, 2006.

[13] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. of Comp. Vision*, 77(1-3):157–173, 2008.

[14] H. Su, M. Sun, L. Fei-Fei, and S. Savarese. Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In *Proc. of the ICCV*, 2009.

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

[16] B. Wandell. *Foundations of Vision*. Sinauer Associates, Inc., Sunderland, Massachusetts, USA, 1995.

[17] J. Willamowski, D. Arregui, G. Csurka, C. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptor. In *ICPR Workshop on Learning for Adaptable Visual Systems Cambridge*, 2004.