Incremental concept learning with few training examples and hierarchical classification

Henri Bouma, Pieter T. Eendebak, Klamer Schutte,
George Azzopardi and Gertjan J. Burghouts

TNO, Oude Waalsdorperweg 63, The Hague, The Netherlands

ABSTRACT

Object recognition and localization are important to automatically interpret video and allow better querying on its content. We propose a method for object localization that learns incrementally and addresses four key aspects. Firstly, we show that for certain applications, recognition is feasible with only a few training samples. Secondly, we show that novel objects can be added incrementally without retraining existing objects, which is important for fast interaction. Thirdly, we show that an unbalanced number of positive training samples leads to biased classifier scores that can be corrected by modifying weights. Fourthly, we show that the detector performance can deteriorate due to hard-negative mining for similar or closely related classes (e.g., for Barbie and dress, because the doll is wearing a dress). This can be solved by our hierarchical classification. We introduce a new dataset, which we call TOSO, and use it to demonstrate the effectiveness of the proposed method for the localization and recognition of multiple objects in images.

Keywords: Object recognition, deep learning, online learning, hierarchical clustering.

1. INTRODUCTION

The number of networked sensors – e.g., CCTV and smartphones – is growing exponentially and the amount of image data increases daily (e.g., YouTube, surveillance applications). Concept detection and localization is important to understand the content of video and allow flexible querying in a large number of cameras,[1,2] especially in the security and defence domain. Concepts may include objects, actions, scenes and events and in this work we will mainly focus on objects. Typically, the object detectors that perform well on public benchmarks are trained on large collections (e.g., ImageNet[3] or fine-grained datasets[4]) or annotated subsets (e.g., the Pascal Visual Object Challenge (VOC) challenge[5] and the ImageNet Large-Scale Recognition Challenge (ILSVRC)[6]). The deep convolutional neural network (CNN) has been demonstrated to be an effective approach of which several implementations have been proposed, such as Decaf/Caffe,[6–8] Overfeat[9] and R-CNN.[10]

The following problems, which are relevant for incremental concept learning and localization, have not yet been addressed. While CNNs have proven to be highly effective in challenges that contain huge training collections, it is not yet clear how they perform on practical applications with only a few training samples.[11] Furthermore, new classes may (initially) consist of a low number of positive examples, resulting in an unbalanced number of samples in each class. Finally, the user may start adding new classes that are very similar or closely related to existing classes. For example, in the case when 'Barbie' and 'dress' are two concepts and the Barbie is wearing a dress (similar for 'navy ship' and 'sea').

Our novel contribution is that we propose a concept-detection method with incremental learning that addresses these problems. Firstly, we show that for a focused application in a single domain, it is possible to reduce the number of training samples to a low number and that the performance benefits from a training set that is as specific as possible for the purpose. Secondly, we show that the addition of novel concepts is often possible without retraining existing concepts, which is important to minimize computational cost and to allow fast interaction. Thirdly, we show that an unbalanced number of positive training samples leads to biases in classifier scores that can be corrected by giving higher weight to the classes that occur less frequent. Fourthly, we show that deterioration of the detector performance due to hard-negative mining for similar or closely related classes can be solved by a hierarchical classification approach.

E-mail: henri.bouma@tno.nl, Phone: +31 888 66 4054

H. Bouma, P.T. Eendebak, K. Schutte, G. Azzopardi and G.J. Burghouts, "Incremental concept learning with few training examples and hierarchical classification," Proc. SPIE, Vol. 9652, (2015). http://dx.doi.org/10.1117/12.2194438

The outline of the paper is as follows. The method is presented in Section 2, experimental setup and results in Section 3 and 4, and conclusions in Section 5.

2. METHOD

The framework that we propose can improve recognition pipelines that consist of feature computation, smart negative selection, and classification with support vector machines (SVMs), as is commonly used for recognition of concepts, such as objects[10] and actions.[12] For this paper, we selected the pipeline of R-CNN,[10] which has been demonstrated to be effective for object recognition and localization in several benchmark datasets.[10] The system takes an input image, extracts bottom-up candidate regions, computes features for each region using a large CNN and then classifies each region using class-specific linear SVMs.[13] We used the CNN model that generates 4096 first-stage features for each region[7] and – in some experiments – we used the pre-trained SVM models that generate 200 second-stage concept scores[10] as if they were first-stage features. The first-stage features need to be computed only once for each image, independent of the number of concepts, allowing scalability to a large number of concepts. The second-stage linear SVMs can be used for training new concepts on-the-fly. Non-maximum suppression is used to select the dominant concepts, and hard-negative mining[10] is used to boost the performance of the classifier. Our concept-detection method with incremental learning was recently integrated in an interactive demonstrator[14] that can translate natural language queries to structured queries,[15] find concepts in images, and rerank results for effective retrieval of images.[16]

The novel proposed concept-detection method allows incremental learning with a low and varying number of training examples. The system is able to retrain only one class without affecting the detectors of other classes, to make a random selection of other classes that are used to mine the negatives, and to make a random selection of images in each class. The SVM class weights (w in liblinear[13] or *pos_loss_weight* in R-CNN) are modified to give higher weight to small classes by setting $w = \max(2, 100/\mathrm{P})$ for positives and $w = 1$ for negatives, where P is the number of positive training images. Commonly, hard-negative mining performs better than random negative mining. However, for similar and closely related concepts, selection of hard negatives may lead to inferior quality. This mining should eliminate irrelevant negatives and focus on the most interesting negatives. However, for related concepts (such as 'navy ship' and 'sea'), it may be impossible to separate positives from negatives. In these cases, the mining of hard negatives becomes the mining of 'impossible negatives', which deteriorates the performance of the classifier. Highly similar concepts (those classes that retrieve a large portion of negatives from each other) are first merged in a super class. The super-classifier retrieves negatives outside the concerned super class. Then, sub-classifiers draw negatives from the other sub-classes within the same super class. The super-classifier detects and the sub-classifiers separates subgroups by assigning it to the most likely sub-class.

There are several approaches that perform hierarchical clustering and extraction of a taxonomy.[3,17–20] For the merging of related concepts, we have chosen an image-based approach, without the use of WordNet relations[19] or other external trees.[20] The hierarchical decision tree can avoid exhaustive testing of all C classifiers but reduce this to $\log_2(C)$ decisions, which enables scaling-up to many categories.[17] For the hierarchical merging of related concepts, a matrix is constructed with for each class the percentage of hard negatives from other classes. Hard-negative mining may deteriorate the detector quality, resulting in unexpected confusions. Therefore, we do not use a confusion matrix[17] or average SVM scores,[19] but a matrix with for each concept the number of hard negatives from another concept as a measure of relatedness of concepts. This matrix was made symmetric by averaging with its transpose and each row is divided by its maximum value for normalization and the diagonal was set to one. High values in the matrix give an indication of strong relations. A spectral clustering method[21] was applied to this matrix to create groups and a tree was created by recursive splitting in two groups until all concepts become leafs. The average Euclidean distance between the clusters in the spectral projection is the distance measure.

3. EXPERIMENTAL SETUP

The recognition of concepts is relevant for many domains, including the domain of defence and security. However, recordings of military material, such as details of ships and mines is often classified. Therefore, in our experiments,

we focussed on unclassified material of toys and office supplies, without loss of generality of the method. For the experiments, three different training sets were used. We selected 36 class labels (see Table 1). The first training set was downloaded from ImageNet (WordNet IDs are shown in same table). This dataset contains approximately 100 images per object. The second training set was retrieved with the Microsoft Bing search engine in November 2014 for the set of queries listed in Table 1. Note that for several objects, we performed multiple queries (e.g., for the 'car' we queried 'toy grey car lamborghini', 'toy red car range rover', etc.), which resulted in a dataset with domain knowledge. The search engine dataset – called 'Bing scrape' – contains approximately 20 images per query. The third set – called 'Toy and Office-Supply Objects' (TOSO) dataset* – was obtained by moving a video camera around office supplies and toy objects that match the class labels in Table 1. A selection of 100 frames is grabbed from these videos with uniform spacing to create the training set (Fig. 1, left). Similar to the queries, some objects contain multiple instances (e.g., different types of cars). The TOSO test set was created with a photo camera of the same objects, consisting of 145 images that include orientation, scale and light-source variation and a various number of objects per image (Fig. 1, right).

Table 1. Class labels, WordNet IDs and queries used for the Bing scrape.

Class (Wordnet id) Search engine query	Class (Wordnet id) Search engine query
1. airplane (n02691156) toy fighter airplane	18. keys (n03613294) keys {[], rsa token, car}
2. barbie (n03219135) barbie doll with {black, red, blue, white} dress	19. motorcycle (n03790512) motorcycle toy black
3. bolt (n03701191) black bolt screw	20. noisething (n/a) n/a
4. bumpinroad sign (n/a) speed (bump OR hump) sign red -yellow	21. pig (n02395406) pig toy plastic
5. bus (n02924116) toy red bus	22. plant (n11669921) plant artificial flower in pot
6. bus stop (n08517676) n/a	23. rollerskate (n04102618) pink roller skates blades sports barbie
7. camper (n02946348) VW red white camper van	24. screwdriver (n04154565) red screwdriver
8. car (n02958343) toy{grey car lamborghini, red car range rover, red sports car benz, yellow porsche}	25. sheep (n02411705) plastic toy sheep
9. computer mouse (n03793489) gray wireless computer mouse	26. ship (n04194289) toy rescue harbour boat
10. cow (n01887787) plastic toy cow black white	27. skateboard (n04225987) plastic finger skateboard
11. dinosaur (n01703569) plastic toy dark brown dinosaur	28. soccer ball (n04254680) orange soccer ball
12. donkey (n02389559) plastic toy brown donkey -horse	29. stapler (n04303497) {black, gray} stapler
13. donotenter sign (n/a) do not enter sign	30. stop sign (n/a) stop sign
14. dress (n03236735) dress {blue, green, white}	31. tool (n04451818) {tool flat file, plastic tool set benchvice pliers, combination wrench, red toy pipe wrench}
15. hamburger (n07697100) hamburger toy	32. traffic light (n06874185) traffic light
16. helmet (n03513137) helmet toy red	33. train (n04468005) tram yellow roof toy
17. horse (n02374451) horse toy plastic	34. turd (n/a) plastic turd
	35. water bottle (n04557648) spa water bottle reine -yourself
	36. zebra sign (n/a) pedestrian crossing zebra sign

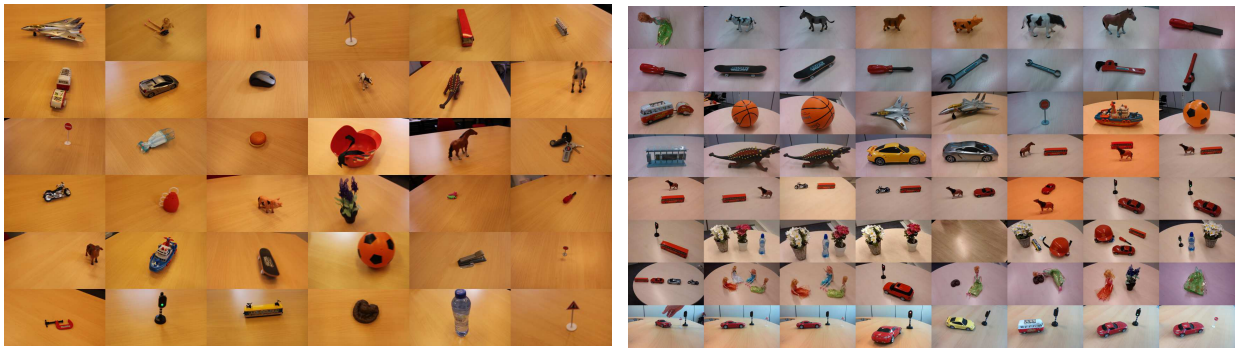

Figure 1. Left: Example images from the TOSO training set. Right: Example images from the TOSO test set (48 of the 145). The test images include various objects, various number of objects per image, and orientation, scale and light-source variation.

As a performance measure, we use the mean average precision (MAP), which is the mean of the average

*https://www.researchgate.net/publication/275347805_TOSO-dataset

precision (AP) scores for each query (Eq. 1):

$$\mathrm{AP} = \frac{\sum_{k=1}^{n} P(k) * \mathrm{rel}(k)}{\mathrm{numberOfRelevantDocuments}} \tag{1}$$

where k is the rank in the sorted results, n is the number of retrieved items, $P(k)$ is the precision at k and $\mathrm{rel}(k)$ is 1 if the item at k is correct and 0 otherwise. Missing WordNet IDs or queries are ignored in the computation of the MAP.

4. EXPERIMENTS AND RESULTS

In the Introduction (Sec. 1), we mentioned four key aspects of concept detection with incremental learning and few training examples and similar concepts. Each of these aspects is analyzed and the results are presented in the same order. We analyzed in-domain classification with few training samples (Sec. 4.1), the incremental learning of novel objects (Sec. 4.2), unbalanced data sets (Sec. 4.3), and finally, the hierarchical classification of similar objects (Sec. 4.4).

4.1 In-domain classification with few training examples

In this first experiment, the quality of the three different training sets is analyzed. The 4096 CNN features[7] were computed and linear SVMs were trained on the three training sets and applied to the test data. The results in Table 2 show that the performance of generic out-domain ImageNet is worst, the Bing scrape is better and the in-domain TOSO training set performs best. The Bing scrape performs better than ImageNet, because we included 'in-domain' knowledge in the query to describe the specific objects. For example, we included the word 'toy' or color information (e.g., 'red bus'). This result shows that the performance benefits from a training set that is as specific as possible for the purpose.

Table 2. MAP values on TOSO testset for the three trainsets using 4096 CNN features.

Number of train images	ImageNet	Bing scrape	TOSO trainset
8	15%	46%	86%
16	15%	52%	88%

Being able to use a small number of training images per class is extremely important for incremental learning, where novel objects are introduced that may initially have a only few examples. Therefore, the size of the training set is analyzed and we also analyzed the size of the feature descriptor. The CNN creates 4096 first-stage features and R-CNN creates 200 object probabilities.[10] Results are shown in Fig. 2 (Left) for different amount of training images per object. The approach based on 4096 features requires only 4 images per class to obtain a MAP $\geq 80\%$. The results also show that training a second-stage SVM on 4096 features performs better than on 200 features, especially for small number of training images. This may be related to the abstraction level of the features[22] and to the number of values in the feature vector.

4.2 Incremental concept learning

For online incremental learning, it is important to be able to add one class without retraining all the others to minimize the computational cost and allow fast interaction. Therefore, we performed an experiment and randomly varied the number of 'other' classes, i.e. classes that are used to select negatives. For this experiment, we used 64 training images from TOSO per class. The results are shown in Figure 2 (Right). Of course, it is beneficial to include as many negative classes as possible. However, if the number of other classes is already high (e.g. 32), and one class is incrementally added, it is not necessary to retrain all other concept detectors, since the performance hardly increases from 32 to 35. This allows rapid addition of novel concepts in an interactive system and retraining is only needed if many new concepts are added (or if concepts are closely related; see Sec. 4.4 for this special case).

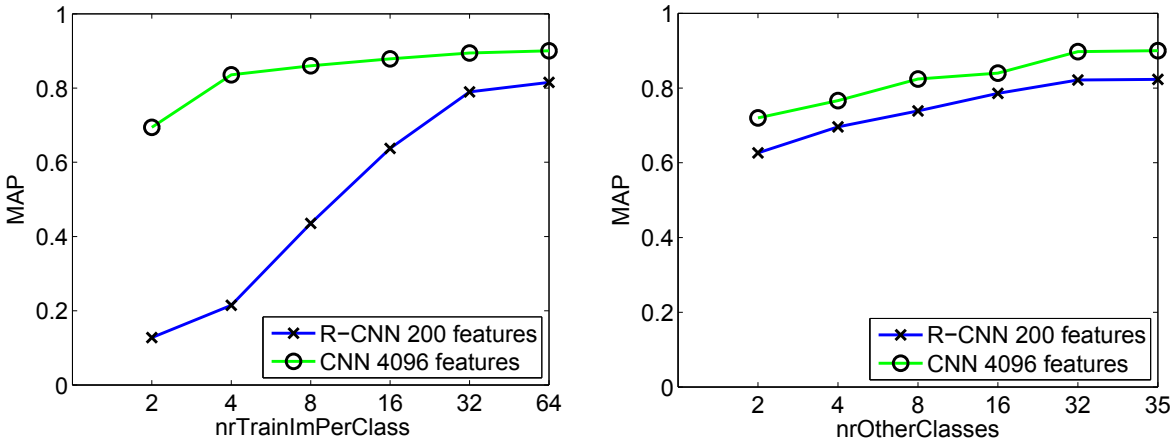

Figure 2. Left: MAP values on the TOSO dataset for different number of training images per class (nrTrainImPerClass) for the cases with 4096 CNN or 200 R-CNN features. Right: MAP with 64 number of training images per class for different number of other classes (nrOtherClasses) for the cases with 4096 CNN or 200 R-CNN features.

4.3 Modified weights for unbalanced classes

Incrementally added novel classes may (initially) consist of a low number of positive examples, resulting in an unbalanced number of samples in each class. A low number of training examples in a class may lead to lower SVM scores and hence biased results when used together with larger classes. To analyze the effect, we trained a linear SVM once on all (approximately 100 per class) images and once on only 8 training images. The SVM-scores of the higher-scoring class are shown in Figure 3. The left figure clearly shows a bias effect with lower scores for the SVM with less training images. The bias can be compensated by modifying SVM class weights (see Sec. 2 for details). Figure 3 (Right) was created with modified weights and it shows that the bias was reduced. Figure 3 also shows correctness of classifications. The SVM that was trained on many images hardly makes mistakes for an SVM-score ≥ -0.2. The main causes of misclassification are related to traffic signs (they are very small in the test images) and the screwdriver. Manual inspection of the data showed that the incorrect sample in the figure with extremely low SVM score (below -0.6) appeared to be a screwdriver. This will be further analyzed in the next subsection.

4.4 Hierarchical classification of similar concepts

The screwdriver that was mentioned in the previous subsection has an extremely low SVM output score. This occurs because the screwdriver is very similar to one of the tools ('file'), since both have a red grip and a gray blade. If classes are similar or related, the mining of hard negatives can become the mining of 'impossible negatives', which deteriorates the overall performance of the classifier.

We first inspect the distribution of negatives over the classes to see the effects of hard-negative mining. The percentage of samples from the most dominant negative class, appears to be 22%($\pm$9.0) on average over all 36 classes. Some examples with a high percentage of one dominant group are summarized. The screwdriver has many negatives from the tool and vice versa (45% resp. 40%). Both objects are very similar. The dress has many negatives from Barbie (i.e. fashion doll) and vice versa (30% resp. 17%). This can be understood, because the Barbie is always wearing a dress. The bump-in-road sign has many negatives from the zebra sign (37%). They are similar, because both are traffic signs. The camper and the bus have many negatives from the car (39% resp. 35%). Apparently, the vehicles are similar. The horse has most of its negatives from the sheep and vice versa (23% resp. 21%). Apparently, animals are similar.

The screwdriver appeared to have the lowest SVM score (below -0.6) and it appears to have the highest number of negatives from one class (45%). Apparently, the screwdriver and tool are mining many hard (or impossible) negatives from each other, which may deteriorate the performance of the detector.

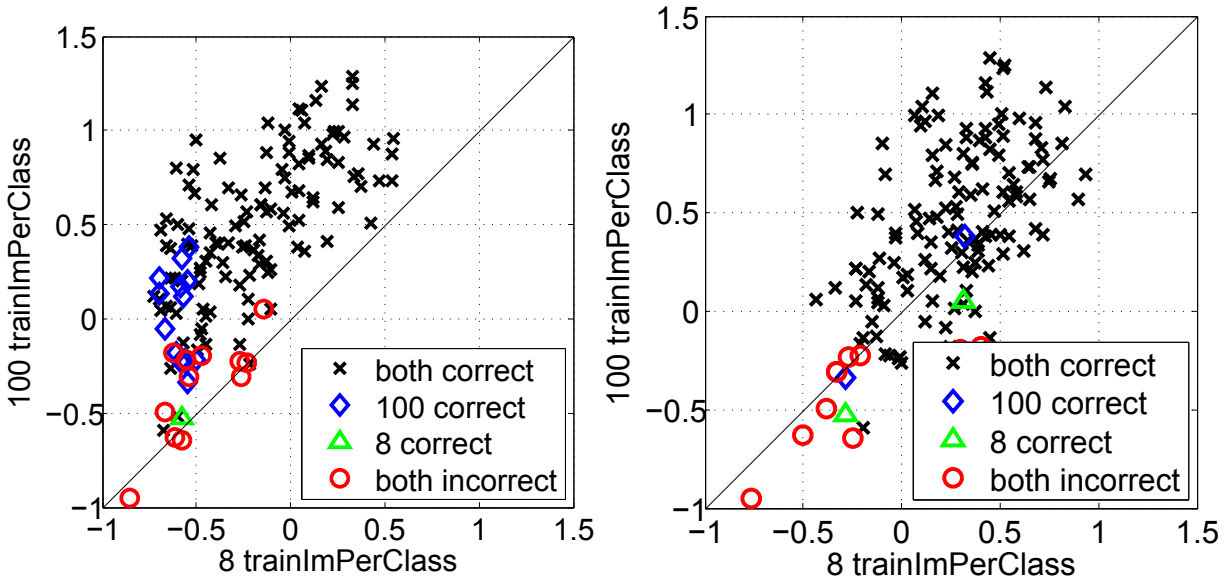

Figure 3. Highest SVM-scores for each image in the test data, where the concept detector is based on the maximal number of training images per class (approximately 100) or 8 training images per class; for equal SVM class weights (left) and increased weight for the 8-training images (right). Colors and marker type indicate when the highest scores correspond to the ground truth.

To avoid deterioration of the detector for similar or related classes, it would be better use a two-step approach. The first step detects the group of related concepts together. The second step separates the subclasses.

The related concepts can be automatically grouped with a spectral clustering of the selected negatives (see Sec. 2 for more details) resulting in a hierarchical representation. Figure 4 shows a dendrogram to visualize nodes, leafs and distances in polar coordinates. We do not only see the grouping of dominant relations in this figure (e.g., horse+sheep, tool+screwdriver, car+camper, dress+Barbie), but we also see that high-level concepts are grouped together (e.g., many vehicles are on the red branch and animals on the green).

For clarity, we now focus on the performance of 'screwdriver+tool' and 'dress+Barbie' . The first step detects the group of related concepts (i.e. the super-class) together and the second step separates the sub-classes. For 'dress' we use only the merged detector because there are no Barbies without dress in the TOSO test set. The preliminary results are shown in Table 3. The initial results indicate an improvement of the average precision when the hierarchical approach is used. Further research is needed to confirm and generalize this approach.

Table 3. Average precision (AP) of the direct classification and the hierarchical detection and classification process.

Concept	Direct classification	Two step classification
Barbie (fashion doll)	95.7	91.9
Dress	83.5	96.3
Screwdriver	89.2	100.0
Tool	87.0	94.8
Mean AP over these four classes	88.8%	95.8% (+7.0%)

5. CONCLUSIONS

In this paper, we proposed a method for incremental object learning that can handle few training examples and similar objects. Firstly, we showed that for a focused application in a single domain, it is possible to reduce the number of training samples to a low number (e.g., 4 images per class to obtain a MAP larger than 80%) and significant performance gains from a training set that is as specific as possible for the purpose. Secondly, we showed that the incremental addition of novel objects is often possible without retraining existing object

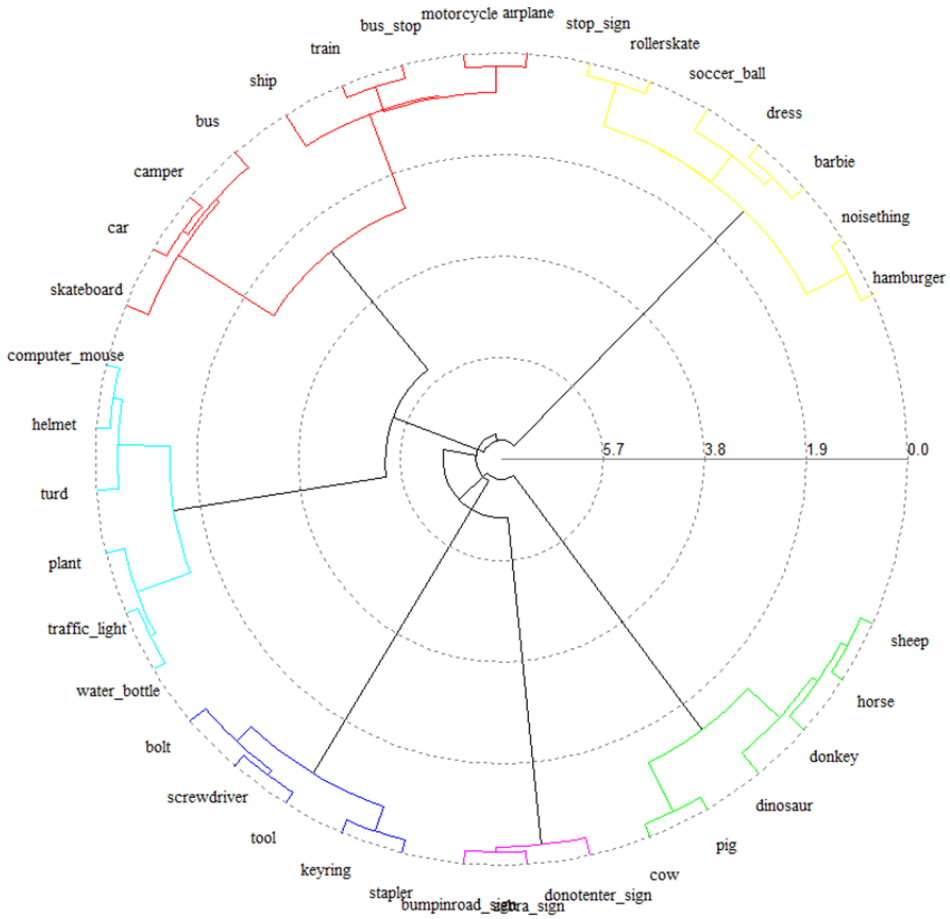

Figure 4. Hierarchical merging of concepts

detectors, which is important to minimize computational cost and to allow fast interaction. Thirdly, we showed that an unbalanced number of positive training samples – which will always occur in an online learning system – leads to biases in classifier scores that can be corrected by giving higher weight to the classes that occur less frequent. Fourthly, we showed that deterioration of the detector performance due to hard-negative mining for similar or closely related classes can be solved by a hierarchical classification approach, where we first detect the group of related concepts together and subsequently separate subclasses. We demonstrated the effectiveness and efficiency of the proposed approach in experiments to locate and recognize multiple objects in the unclassified TOSO dataset.

Acknowledgment

This research was performed in the GOOSE project, which is jointly funded by the enabling technology program Adaptive Multi Sensor Networks (AMSN) and the MIST research program of the Dutch Ministry of Defense. This publication was supported by the research program Making Sense of Big Data (MSoBD).

REFERENCES

[1] Bouma, H., Azzopardi, G., Spitters, M., et al., "TNO at TRECVID 2013: multimedia event detection and instance search," *Proc. TRECVID*, (2009).

[2] Schutte, K., Bomhof, F., Burghouts, G., et al., "GOOSE: semantic search on internet connected sensors," *Proc. SPIE 8758*, (2013).

[3] Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L., "Imagenet: a large-scale hierarchical image database," *IEEE CVPR*, 248–255 (2009).
[4] Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S., "CNN features off-the-shelf: an astounding baseline for recognition," *arXiv 1403.6382*, (2014).
[5] Everingham, M., van Gool, L., Williams, C., Winn, J., and Zisserman, A., "The PASCAL visual object classes (VOC) challenge," *IJCV* 88, 303–338 (2010).
[6] Krizhevsky, A., Sutskever, I., and Hinton, G., "Imagenet classification with deep convolutional neural networks," *Adv. neural inf. proces. sys.*, (2012).
[7] Jia, Y., Shelhamer, E., Donahue, J., et al., "Caffe: convolutional architecture for fast feature embedding," *Proc. ACM Multimedia*, 675–678 (2014).
[8] Donahue, J., Jia, Y., Vinyals, O., et al., "Decaf: A deep convolutional activation feature for generic visual recognition," *ICML*, (2014).
[9] Sermanet, P., Eigen, D., Zhang, X., et al., "Overfeat: integrated recognition, localization and detection using convolutional networks," *ICLR*, (2014).
[10] Girshick, R., Donahue, J., Darrell, T., and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE CVPR*, 580–587 (2014).
[11] Habibian, A., Mensink, T., and Snoek, C., "Videostory: a new multimedia embedding for few-example recognition," *Proc. ACM Multimedia*, 17–26 (2014).
[12] Burghouts, G., Schutte, K., Bouma, H., and den Hollander, R., "Selection of negative samples and two-stage combination of multiple features for action detection in thousands of videos," *Machine Vision and Applications* 25, 85–98 (2014).
[13] Fan, R., Chang, K., Hsieh, C., Wang, X., and Lin, C., "Liblinear: A library for large linear classification," *J. Machine Learning Research* 9, 1871–1874 (2008).
[14] Schutte, K., Bouma, H., Schavemaker, J., et al., "Interactive detection of incrementally learned concepts in images with ranking and semantic query interpretation," *IEEE CBMI*, (2015).
[15] de Boer, M., Daniele, L., Brandt, P., and Sappelli, M., "Applying semantic reasoning in image retrieval," *Proc. ALLDATA*, (2015).
[16] Schavemaker, J., Spitters, M., Koot, G., and de Boer, M., "Fast re-ranking of visual search results by example selection," *Int. Conf. Computer Analysis of Images and Patterns*, (2015).
[17] Griffin, G. and Perona, P., "Learning and using taxonomies for fast visual categorization," *IEEE CVPR*, (2008).
[18] Li, L., Wang, C., Lim, Y., Blei, D., and Fei-Fei, L., "Building and using a semantivisual image hierarchy," *IEEE CVPR*, (2010).
[19] Bannour, H. and Hudelot, C., "Hierarchical image annotation using semantic hierarchies," *Proc. ACM CIKM*, (2012).
[20] Binder, A., Muller, K., and Kawanabe, M., "On taxonomies for multi-class image categorization," *IJCV* 99, 281–301 (2011).
[21] Ng, A., Jordan, M., and Weiss, Y., "On spectral clustering: Analysis and an algorithm," *Adv. neural inf. processing systems*, 849–856 (2002).
[22] Singh, S., Gupta, A., and Efros, A., "Unsupervised discovery of mid-level discriminative patches," *ECCV*, (2012).