Apprenticeship learning: transfer of knowledge via dataset augmentation

Miroslav Kobetski
and Josephine Sullivan

Computer Vision and Active Perception,
KTH, 114 28 Stockholm,
kobetski@kth.se,sullivan@nada.kth.se

Abstract. In visual category recognition there is often a trade-off between fast and powerful classifiers. Complex models often have superior performance to simple ones but are computationally too expensive for many applications. At the same time the performance of simple classifiers is not necessarily limited only by their flexibility but also by the amount of labelled data available for training. We propose a semi-supervised wrapper algorithm named apprenticeship learning, which leverages the strength of slow but powerful classification methods to improve the performance of simpler methods. The powerful classifier parses a large pool of unlabelled data, labelling positive examples to extend the dataset of the simple classifier. We demonstrate apprenticeship learning and its effectiveness by performing experiments on the VOC2007 dataset - one experiment improving detection performance on VOC2007, and one domain adaptation experiment, where the VOC2007 classifier is adapted to a new dataset, collected using a GoPro camera.

1 Introduction

Recent advances in visual object category recognition have resulted in powerful models for classification such as Deformable Part-based Models (DPM) and Multiple Kernel Learning (MKL) of a large number of different features [1,2]. These models often require much more computation than their linear competitors, and pay for their significantly better performance with test and training time costs. Object detection is particularly sensitive to test-time costs due to the large number of detection windows that need to be classified to localize an object in an image. MKL is so computationally expensive that it is unlikely to be fully applicable to such problems even in the medium term future.

Expensive methods are often to some degree sped up through approximations or various cascading schemes where classifiers progressively grow in power and complexity [1,3]. Ultimately such optimizations often still depend on a first stage that uses a simple and fast classifier such as a linear SVM. Even so, many applications that need to run in reasonable time, such as parsing of huge datasets, car or pedestrian detection, are unable to make use of the strongest available models. This is why improving the performance of simple classifiers is still an

important problem, even though simple classifiers often cannot reach the same performance as more complex ones.

We present a method, which we term *apprenticeship learning*, for transferring knowledge from a powerful classifier to a simpler via a large pool of unlabelled or weakly labelled data S_u. We refer to the stronger classifier H_m as the mentor classifier and the weaker classifier H_a as the apprentice classifier. H_m learns on a human-labelled dataset S_l, and then produces a new dataset S_e by searching in S_u for positive examples. The apprentice classifier is then trained on the extended set $S_l \cup S_e$. A good source for S_u is the vast number of images on the web, available through various search engines. Another good source could be unlabelled data collected in the target domain of the classifier. If that domain is different to that of the labelled data this can be seen as a technique for domain adaptation.

Apprenticeship learning is useful when the requirements of a problem forces one to use a less complex classifier H_a, and there exists an H_m that is more powerful than H_a. We formalize apprenticeship learning in terms of its necessary conditions and show that a far from perfect mentor classifier can be used to boost the performance of a weaker classifier.

Using apprenticeship learning we improve the performance of a HOG-SVM classifier on the VOC2007 dataset with 3.80% average precision (23.8% improvement). We also apply apprenticeship learning as a domain adaptation method to our own data, collected with a GoPro in an urban environment. In the domain adaptation experiment we see improvements of $3\% - 5\%$ AP ($15\% - 63\%$ improvement), depending on the strictness of the overlap criterion.

Fig. 1: **Mentor-labelled car detections and best overlapping apprentice detections.** The green bounding boxes show the labels that the mentor classifier provides for S_e. The purple bounding boxes show the best overlapping detection by the apprentice classifier.

2 Relation to previous work

Apprenticeship learning is a semi-supervised wrapper method. Other wrapper methods that make use of several classifiers are Co-Training [4,5,6] and Multiview

learning [7,8]. In co-training several classifiers with different view of the data iteratively extend each others training set until convergence. Multview learning does not explicitly label data but rather adds an additional regularizer to enforce agreement between the different views or classifiers. Our method differs from these approaches as we do not assume that we have several independent views of the data, but rather that we have two overlapping views of which one is more complete - the mentor classifier. This also leads to the difference that the information exchange is one-way since the mentor classifier is not updated based on output from the apprentice classifier, since it is unlikely to provide any better examples than the mentor classifier could provide itself using self-training.

Apprenticeship learning has similarities to self-training [9,10], where the learnt classifier creates S_e from examples that it can confidently classify. Since discriminative learning is interested in the optimal decision boundary according to some criterion - such as low classification error - examples with low or negative margin are the most interesting ones, which is also evident for most of the successful loss functions for classification. In self-training the dataset is augmented with confident examples. These examples have a large positive margin and will therefore have low or zero loss and will not have a significant effect on the learning.

Active learning applies the opposite approach of self-training, focusing on examples that are close to the decision boundary and letting an oracle (i.e. human labeller) provide the label of these potentially informative examples [11]. In active learning H_a parses through S_u and selects examples that H_a is uncertain about. The oracle is then queried for the label of these examples, H_a retrained and the process repeats. In apprenticeship learning the mentor classifier H_m is not a perfect oracle, but instead it has the ability to automatically parse through the unlabelled data, labelling new examples for the apprentice classifier. The examples generated this way are not directly dependent on their relation to the decision boundary of H_a. In this way both examples that H_a is uncertain about and those it is incorrectly certain about can be added to the extended dataset S_e.

A similar method to apprenticeship learning has been used for model compression [12] on ensemble methods. Bucila et al. present MUNGE, an algorithm for generating synthetic examples and use it to compress ensemble models to compact neural network models. Our approach focuses less on the generation of data, as visual examples are notoriously hard to generate faithfully, and more on the applicability to object detection.

Training set augmentation using weakly labelled images from web-searches is not new and has been used to improve image classification results [13,14,15,16,17,9]. These methods treat the web images as examples with uncertain positive labels and in various ways apply robust learning techniques. The robust learning is necessary due to the large number of mislabelled or outlier examples generated through web searches. Outlier pruning [14,17] is therefore an important part of several of the methods. Other methods include the outlier pruning implicitly, like the domain adaptation approach of Bergamo and Torresani [13]. They show

good classification results using a transductive SVM [18,13] where the labels of the unlabelled images are included as optimization variables. Another approach to label uncertainty is Multiple instance learning [15], where each image query results in a positive bag (where at least one example is assumed to be positive) rather than a number of labelled examples.

These methods have been applied to image classification rather than detection due to the lack of bounding box annotation, and since detection requires bounding-box output, the above methods cannot be directly applied to it. This shows a strength of apprenticeship learning - weak annotation only increases the likelihood of finding positive examples in the parsed images, but is not necessary. Even the complete lack of annotation does not prevent apprenticeship learning from being applied, which we show in our domain adaptation experiment.

3 Apprenticeship learning

The idea of apprenticeship learning is to improve the performance of a simple classifier by letting a more complex classifier mine for additional positive training examples from which the simple classifier can learn. The algorithm is applicable when the most powerful available classifier is not suitable for a task due to limitations on test-time speed or memory.

Apprenticeship learning requires a specific (although quite general) setting to be effective. We now formalize this with three necessary conditions. As knowledge is transferred via informative positive examples, and since informative negatives can be successfully mined using standard hard negative mining, the training sets S_l, S_e and S_u only refer to positive examples.

Unlabelled diversity condition
Apprenticeship learning depends on a large pool of unlabelled data, which has to be sufficiently diverse. The probability p of a random example in S_u containing any new information is as $p = \mid (S_l \cup S_e) \setminus S_u \mid / \mid S_u \mid$. In the extreme case $(S_l \cup S_e) \setminus S_u = \emptyset$, apprenticeship learning can no longer bring any potential gain.

Sufficient flexibility condition
The performance of H_a must improve as the training set increases in size. If H_a^* is obtained by training on a completely and perfectly annotated, hypothetical version of $S_u \cup S_l$, and $\hat{H}_a$ is obtained by training on S_l, then the sufficient flexibility condition can be stated as $E[L(Y, H_a^*(X))] < E[L(Y, \hat{H}_a(X))]$, for a loss function $L(Y, f(X)$ and test data (X, Y). Intuitively this condition only holds if the added examples include new information about the class that the feature/learner combination is able to learn. Due to the high-dimensional data in vision and the limited datasets available, we believe this condition to be true for the popular features and common simple classifiers.

Adept mentor condition
The final condition is that there exists a mentor classifier H_m that is able

to reach higher generalization performance than the apprentice classifier H_a or $E[L(Y, H_m(X))] < E[L(Y, \hat{H}_a(X))]$. This does not assume that H_m is an oracle, only that it performs better than H_a and is therefore able to correctly classify additional examples that H_a cannot.

4 Experiments

Object detection enforces stringent requirements on the speed of a classifier, due to the highly imbalanced distribution of positive and negative subwindows in an image. We therefore select the detection problem as a suitable experiment to demonstrate and test apprenticeship learning.

4.1 Experiment details

We base our experiments on Pascal VOC2007 [19], as it is a popular object recognition dataset with relatively good generalizations properties [20]. VOC2007 has 20 classes with manually annotated bounding boxes around each object. These bounding boxes populate S_l. We perform two experiments - one where we try to improve the performance of the simple classifier on the VOC2007 test set, by using unlabelled Flickr images as S_u. We also perform a second experiment on our own dataset, collected using a chest-mounted GoPro camera. In the second experiment the simple classifier trained on VOC2007 is adapted to the domain of the new dataset.

In both experiments we use a mixture of deformable part-based models (DPM) [2] for the mentor classifier H_m. DPM is used due to its state-of-the art performance and (for moderate datasets) feasible computational time. For H_a we use a single HOG-template linear classifier as in [21], but with the modified HOG implementation from [2]. Although DPM can also be seen as a linear classifier in an extended HOG-space, it has to compute three more HOG-pyramids and convolve 54 filters with the HOG pyramids instead of one. In our implementation this results in an order of magnitude speed difference. In a hardware-specific implementation, where cache size and similar parameters are considered, this speedup could be even more.

VOC2007 - dataset augmentation experiment To obtain S_u we download images from Flickr using the class names as queries. This weak annotation increases p and therefore limits the number of images that H_m needs to parse, but is not a necessary condition. To avoid potential overlap with the VOC2007 test-set we restrict ourselves to images uploaded after 2007. We try to download 3000 images for each class, but for some classes much fewer images are obtained - probably due irregular class names like *tvmonitor*.

After obtaining S_u we run a multi-scale sliding window detection using H_m on the new images, using detections as bounding boxes to create S_e. The linear SVM classifier is afterwards trained on the extended training set $S_l \cup S_e$ - the

VOC2007 images and the newly acquired ones. The extended apprentice-trained linear SVM classifier and the original one are evaluated on the VOC2007 test set. Since we use detections as new annotations, the mentor classifier does not explicitly try to add any negative examples, but we perform hard negative mining until convergence in the iterative fashion described in [2].

The images obtained from Flickr can be seen as having noisy weak labels, since they lack bounding box annotation and only a fraction of them even contain objects from the correct category. Some contain cartoon-like or abstract representations of the queried category. Also, H_m has different performance for different classes and the distribution of the downloaded data differs from that of the labelled training data. For these reasons we tune the thresholds on the classification score required to add new examples, by cross-validating on the labelled training set. The noisy and wrong-domain examples are therefore implicitly filtered by the mentor classifier since such examples receive lower scores than examples commonly found in S_l.

Domain adaptation experiment We also apply apprenticeship learning to a dataset collected by walking through a city center with a forward looking GoPro camera mounted on the chest. Cars are frequent in the recorded videos, so we focus on the car category. One qualitatively observed difference between the cars in VOC2007 and the GoPro data is that the frequency of occluded and truncated cars is much higher in the latter. Weather and lighting conditions could also be responsible for large appearance differences (such as snow-covered cars - although not observed in these particular recordings). There is also likely to be a number of smaller differences such as color space, distortion, distribution of car brands etc.

As an unlabelled set S_u we use a non-annotated video sequence containing 11315 frames of an urban city center environment. The video sequence was collected with a chest-mounted GoPro camera, while walking in Florence. We use a DPM H_m to parse through the video sequence to generate bounding box annotation S_e as in the previous experiment. H_a is a simple HOG template, but we also learn a slightly more complex apprentice classifier $H_a^\dagger$ to get additional characterization of apprenticeship learning. $H_a^\dagger$ is a mixture model of 3 simple HOG templates, where each captures different aspect ratios of car. This is a particularly suitable classifier for cars, since different aspect ratios typically correspond to different views.

The test set is a video with manually annotated bounding-boxes around cars. The test set is disjoint from the training set - this is ensured by collecting the test set on a different day and walking down a different street.

5 Results

Overall we see that apprenticeship learning improves performance both for the domain adaptation experiment and the VOC2007 detection experiment. VOC2007 performance is improved by 3.80% mean AP, with high gains for particular

classes such as 14.80% AP gain for the person class. The domain adaptation experiment also shows improvement similar to that of the VOC2007 experiment ($+3-5\%$ AP). For both experiments the performance is still far from the mentor classifier, which is to be expected due to the much lower flexibility of the apprentice classifier. The main point is that the performance of the apprentice classifier is improved, which is valuable in the argued case where the mentor classifier is not applicable due to speed or memory restrictions. In the setting of our experiments the apprentice classifier can also typically be used as a first cascade when implementing a cascaded version of the mentor classifier [3]. Improving the performance of the first linear classifier then increases the speed of the cascaded mentor classifier, due to the first cascade's higher ability to discard negatives.

5.1 VOC2007 detection results

	size of training set		**AP of classifier**				
Category	S_l	S_e	$H_m^{S_l}$	$H_a^{S_l}$	H_a^{st}	$H_a^{S_l \cup S_e}$	improvement
aeroplane	612	57	28.9	17.06	18.77	**19.89**	2.83
bicycle	706	422	59.5	34.61	**38.48**	38.16	3.55
bird	972	1002	10.0	6.31	9.30	**9.52**	3.21
boat	580	107	15.2	0.64	0.64	9.50	8.86
bottle	1010	382	25.5	14.37	16.61	**18.77**	4.40
bus	458	341	49.6	28.26	30.02	31.63	3.37
car	2500	454	57.9	31.32	30.70	**32.12**	0.80
cat	752	198	19.3	1.60	3.07	**11.36**	9.76
chair	1596	1094	22.4	11.31	11.18	**12.25**	0.94
cow	518	83	25.2	14.05	10.68	**15.77**	1.72
diningtable	430	44	23.3	9.63	14.96	**15.53**	5.90
dog	1020	500	11.1	5.58	2.23	**9.95**	4.37
horse	724	575	56.8	22.24	**23.18**	22.80	0.56
motorbike	678	979	48.7	24.89	24.86	**25.78**	0.89
person	9380	12 844	41.9	10.69	9.96	**25.49**	14.80
pottedplant	1028	286	12.2	11.04	6.16	**11.75**	0.71
sheep	514	330	17.8	13.76	12.97	**19.63**	5.87
sofa	496	57	33.6	11.83	6.37	**12.90**	1.07
train	594	439	45.1	16.13	17.1	**19.87**	3.74
tvmonitor	648	127	41.6	**33.73**	31.66	32.44	−1.28
mean	2693	1016	32.3	15.95	15.94	**19.76**	3.80

Table 1: **Detection performance using HOG+linear SVM classifier on VOC2007.** Displayed is the average precision of the apprenticeship-trained classifier $H_a^{S_l \cup S_e}$, and the regular one $H_a^{S_l}$. The last column displays the difference between the two classifiers. We also show the number of examples added to S_e, the performance of the mentor classifier and we provide a self-trained apprentice classifier H_a^{st} as reference.

We can see in table 1 that the apprenticeship-trained classifier outperforms the baseline classifier on all categories except tvmonitor. Manual inspection of the downloaded "tvmonitor" images shows that only a few of the added detection actually have tv monitors in them, with many of the others being laptop or desktop monitors. In order to not bias the results we did not want to supervise the web searches. This, together with the varying performance of H_m for different classes resulted in a large variance in the size of the extended data sets S_e. Flickr images contain a large number of people and the DPM person classifier has relatively good performance, which results in a relatively large number of detections. A single HOG template is also a reasonably well-suited feature for upright people, so the "sufficient flexibility" condition is well fulfilled and we see a great boost in performance. Horse, chair, car and motorbikes on the other hand have limited improvement, despite a good size of S_e. This could be due to the limited flexibility of a single HOG. We also see that self-training does not give any notable improvement on this dataset. For many classes self-training achieves a small improvement, but for several poorly performing classes it instead results in a performance drop. This is likely due to addition of poor examples, and might be improved by an additional view or filter. Figure 2 show a number of detection results of the simple classifier before and after apprenticeship learning. Since no changes have been made to the HOG feature or the linear learner model this improvement has come at a zero increase in complexity at test-time.

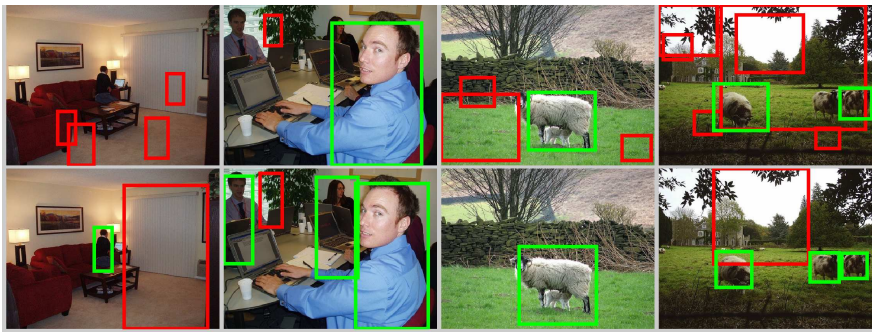


Fig. 2: **Examples of person and sheep detections on VOC2007.** The top row shows a detector using the baseline simple classifier learned and the bottom row shows the simple classifier learnt using apprenticeship learning. Green boxes are detections that are considered as true positives and red boxes are false positives.

5.2 Domain adaptation results

Figure 5a shows the average precision of the apprentice car detectors on the GoPro dataset. The x-axis shows the size of S_e, i.e. number of examples provided by the mentor classifier. Examples are added to S_e in order of their classification

score $H_m(x)$, so the most certain examples are added first. This means that the x-axis also represents a threshold on $H_m(x)$, for adding examples to S_e. 1503 of the added examples are confidently positive ($H_m(x) > 1$) and 7324 have positive margin ($H_m(x) > 0$). Although none of the added examples is outside the margin of negative support vectors it is likely that, after example 7324, we start adding a reasonable number of outliers and poorly localized bounding boxes. We see that adding examples beyond this threshold is more likely to reduce performance than improve it.

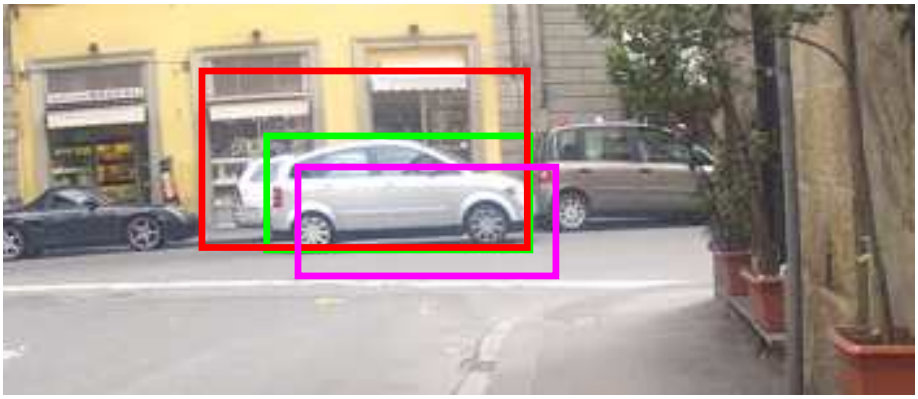


Fig. 3: **Illustration of the Pascal detection criterion.** Ground truth (green box) and two simulated detections (red and purple boxes) that yield 50% overlap with the ground truth.

The standard Pascal detection criterion requires an overlap of $area(B_p \cap B_{gt}/area(B_p \cup B_{gt}) \geq 50\%$ [19], where B_{gt} is the bounding box of the ground truth and B_p the bounding box of the detection. This is not a very strict overlap criterion as can be seen in figure 3. The blue plots in figure 5 show the performance using the 50 % overlap criterion, the red plots use a slightly stricter criterion of 65 % and the purple ones correspond to 80% overlap. We observe that apprenticeship learning improves performance more when stricter detection conditions are imposed. Figure 4 also shows that the detections from the apprenticeship learnt classifiers are slightly better localized. The detections from the mentor classifier can be seen as good latent detections, which make the model sharper and therefore more likely to be well localized.

6 Conclusions and future work

We have proposed a semi-supervised wrapper algorithm named apprenticeship learning. The algorithm transfers knowledge from a more powerful classifier to a

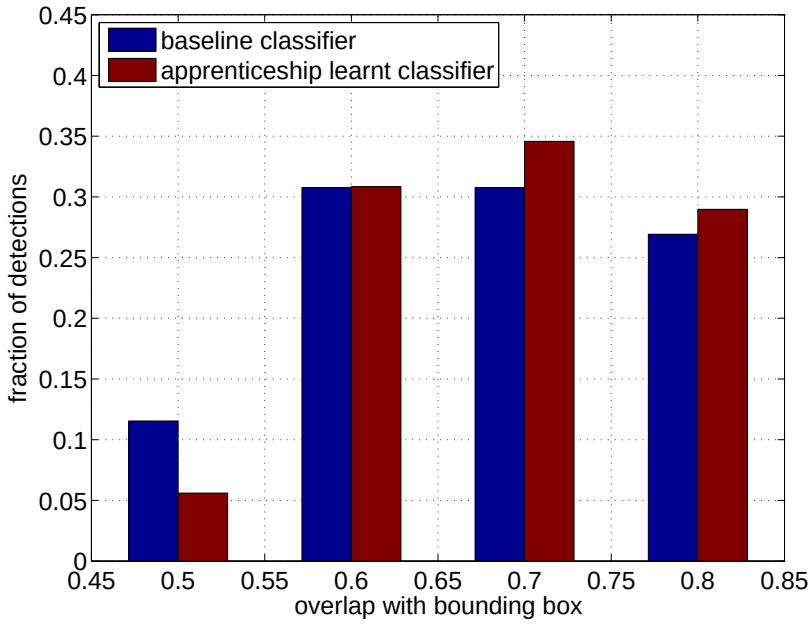
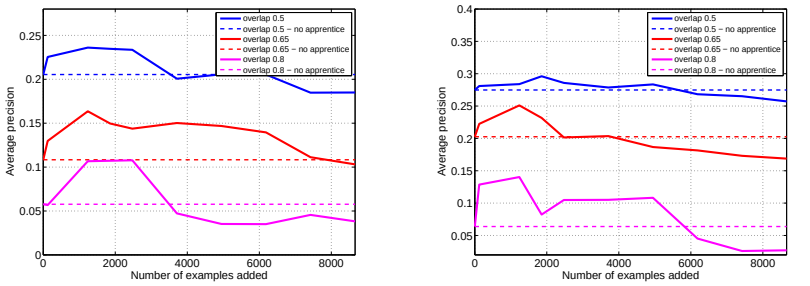


Fig. 4: **Distribution of detection overlap with ground truth.** The apprenticeship learnt classifier produces slightly better overlapping detections.



(a) Single template HOG (b) Simple mixture of 3 HOG templates

Fig. 5: **Average precision for car detections in GoPro data.** Average precision vs. number of examples added by the mentor classifier.

weaker one by extending its training set from a pool of unlabelled data. Apprenticeship learning is designed for the scenario where one has a well performing classifier that does not conform to existing constraints such as a memory or test time budget, a common scenario in computer vision. We show that our algorithm does improve the performance of a linear SVM and HOG classifier by performing experiments using web-mined images as our unlabelled pool. Self training does not seem to improve performance at all in the same scenario. We also show how apprenticeship learning can be used for domain adaptation. Although our simple classifier still performs worse than the mentor classifier, the comparison of the two is moot if the mentor classifier is infeasible for a problem at hand. Also, since simple classifiers are used as early cascades together with complex classifiers, improving the performance of the early cascades speeds up the whole cascade.

Acknowledgements

This work has been funded by the Swedish Foundation for Strategic Research (SSF); within the project VINST.

References

1. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: Proceedings of the International Conference on Computer Vision. (2009)
2. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010) 1627–1645
3. Felzenszwalb, P., Girshick, R., D., M.: Cascade object detection with deformable part models. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. (2010)
4. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the Conference on Computational Learning Theory. (1998)
5. Feng, H., Chua, T.S.: A bootstrapping approach to annotating large image collection. In: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval. (2003)
6. Levin, A., Viola, P., Freund, Y.: Unsupervised improvement of visual detectors using cotraining. In: Proceedings of the International Conference on Computer Vision. (2003)
7. Cowan, I., Tesauro, G., De Sa, V.: Learning classification with unlabeled data. In: Proceedings of the Advances in Neural Information Processing Systems. (1993)
8. Saffari, A., Leistner, C., Godec, M., H., B.: Robust multi-view boosting with priors. In: Proceedings of the European Conference on Computer Vision. (2010)
9. Li, L.J., Niebles, J., Fei-Fei, L.: OPTIMOL: Automatic Online Picture Collection via Incremental Model Learning. International Journal of Computer Vision **88** (2010) 147–168
10. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-Supervised Self-Training of Object Detection Models. In: Seventh IEEE Workshop on Applications of Computer Vision. (2005)

11. Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning (2012)
12. Bucila, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: International Conference on Knowledge Discovery and Data Mining. (2006)
13. Bergamo, A., Torresani, L.: Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In: Proceedings of the Advances in Neural Information Processing Systems. (2010)
14. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from Google's image search. In: Proceedings of the International Conference on Computer Vision. (2005)
15. Vijayanarasimhan, S., K., G.: Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. (2008)
16. Fei-Fei, L., Fergus, R., P., P.: One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006) 594–611
17. Schroff, F., Criminisi, A., Zisserman, A.: Harvesting image databases from the web. IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2011) 754–766
18. Joachims, T.: Transductive inference for text classification using support vector machines. In: Proceedings of the International Conference on Machine Learning. (1999)
19. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88** (2010) 303–338
20. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. (2011)
21. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the Conference on Computer Vision and Pattern Recognition. (2005)