Conservative visual learning for object detection with minimal hand labeling effort *

Peter Roth[1], Helmut Grabner[1], Danijel Skočaj[2], Horst Bischof[1], and Aleš Leonardis[2]

[1] Inst. for Computer Graphics and Vision
Graz University of Technology, Austria
{pmroth,hgrabner,bischof}@icg.tu-graz.ac.at
[2] Faculty of Computer and Information Science
University of Ljubljana, Slovenia
{danijels,alesl}@fri.uni-lj.si

Abstract. We present a novel framework for unsupervised training of an object detection system. The basic idea is to (1) exploit a huge amount of unlabeled video data by being very conservative in selecting training examples; and (2) to start with a very simple object detection system and using generative and discriminative classifiers in an iterative co-training fashion to arrive at increasingly better object detectors. We demonstrate the framework on a surveillance task where we learn a person detector. We start with a simple moving object classifier and proceed with robust PCA (on shape and appearance) as a generative classifier which in turn generates a training set for a discriminative AdaBoost classifier. The results obtained by AdaBoost are again filtered by PCA which produces an even better training set. We demonstrate that by using this approach we avoid hand labeling training data and still achieve a state of the art detection rate.

1 Introduction

Starting with face detection [14, 19] there has been a considerable interest in visual object detection in recent years, e.g., pedestrians [20], cars [1], bikes [12], etc. This is sometimes also referred to as visual categorization as opposed to object recognition [4, 12]. At the core of most object detection algorithms is usually a classifier, e.g., AdaBoost [5], Winnow [9], neural network [14] or support vector machine [18]. The proposed approaches have achieved considerable success in the above mentioned applications.

* This work has been supported by the Austrian Joint Research Project Cognitive Vision under projects S9103-N04 and S9104-N04, by the Federal Ministry for Education, Science and Culture of Austria under the CONEX program, by the SI-A project, by the Federal Ministry of Transport, Innovation and Technology under P-Nr. I2-2-26p Vitus2, by the Research program Computer Vision P2-0214 (RS), by EU FP6-004250-IP project CoSy and by EU FP6-511051-2 project MOBVIS.

However, a requirement of all these methods is a training set which in some cases needs to be quite large. The problem of obtaining enough training data increases even further because the methods are view based, i.e., if the view-point of the camera changes the classifier needs to encompass this variability (e.g., car from the side and car from the back). Training data is usually obtained by hand labeling a large number of images which is a time consuming and tedious task. Clearly this is not practicable for applications requiring a large number of different view-points (e.g., video surveillance by large camera networks). Therefore, it is essential that a representative set of labeled object data is obtained. Negative examples (i.e., examples of images not containing the object) are usually obtained by a bootstrap approach [17]. One starts with a few negative examples and trains the classifier. The obtained classifier is applied to images not containing the object. Those sub-images where a (wrong) detection occurs are added to the set of negative examples and the classifier is retrained. This process can be repeated several times. Obtaining reliable positive examples is, however, a more difficult problem, since discriminant classifiers are very sensitive to false training data.

The main contribution of this paper is to propose a novel framework (depicted in Fig. 1) avoiding hand labeling of training data for object detection tasks. The basic idea is to use the huge amount of unlabeled data that is readily available for most detection task (i.e., just mount a video camera and observe the scene).

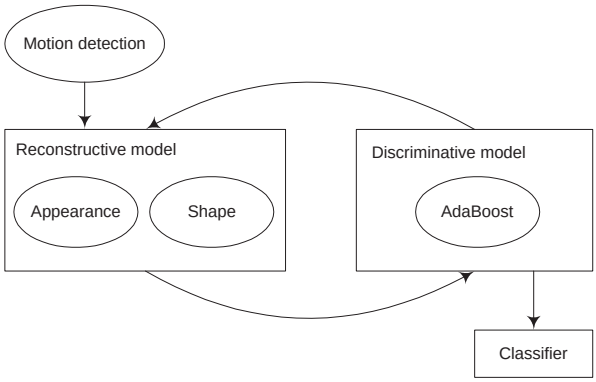


Fig. 1. The proposed conservative learning framework.

We use two types of models, a reconstructive one which assures robustness and serves for verification, and discriminative one, which actually performs the detection. To get the whole process started we use a simple motion detector to detect potential objects of interest. In fact, we miss a considerable amount of objects (which can be compensated by just using longer sequences) and we will get also a lot of miss-detections (which will be reduced in the subsequent steps). The output from the motion detector can be used to robustly build a first

initial reconstructive representation (to further increase the robustness we are using one representation on shape and the other on appearance). In particular, we use robust PCA [15] at this stage so that most of the miss-detections (background, false detections, over-segmentations, etc.) are not incorporated in the reconstructive model. This is very crucial as the discriminative classifier needs to be trained with "clean" images to produce good classification results. The discriminative model is then used to detect new objects in new images. The output of the discriminative classifier is then verified by the reconstructive model, and detected false positives can be fed back into the discriminative classifier as negative examples (and true positives as positive examples) to further improve the discriminative model. In fact, it has been shown in the active learning community [13], that it is more effective to sample the current estimate of the decision boundary than the unknown true boundary. This is exactly achieved by our combination of reconstructive and discriminative classifiers. Exploiting the huge amount of video data, this process can be iterated to produce a stable and robust classifier.

The outlined approach is similar to the recent work of Nair and Clark [11] and Levin et al. [8]. Nair and Clark propose to use motion detection for obtaining the initial training set and then Winnow as a final classifier. Their approach does not include generative classifiers, nor does it iterate the process to obtain more accurate results. In that sense our framework is more general. Levin et al. use the so called co-training framework to start with a small training set and to increase it by using a co-training of two classifiers operating on different features. We show that using a combination of generative and discriminative classifiers helps to increase the performance of the discriminative one.

The rest of the paper is organized as follows. In Section 2 we detail our approach. In order to make the discussion concrete we will use person detection from videos. The experimental results in Section 3 demonstrate the approach on some challenging video sequences with groups of people and occlusions. Finally, we present some conclusions and work in progress.

2 Our Approach

In this section we will explain the modules used in our implementation of the framework depicted in Fig. 1. We used a motion detection procedure based on a simple approximated median background model, a robust PCA as a reconstructive model, and AdaBoost as a discriminative classifier. But note that the particular methods are not crucial and other types of classifiers might be used as well.

2.1 Motion detection

Having a stationary camera a common approach to detect moving objects is to threshold the difference image between the current frame and a background model. A widely used and simple method for generating a background model is

a pixel-wise temporal median filter. To reduce computational costs and memory requirement McFarlane and Schofield [10] developed the approximated median a computationally more efficient method.

The obtained motion blobs can be labeled as persons if the aspect ratio of their bounding box is within the prespecified limits. We are very conservative in this step, i.e., we will miss many potential persons, but we nevertheless will obtain a few false negatives.

2.2 Reconstructive model

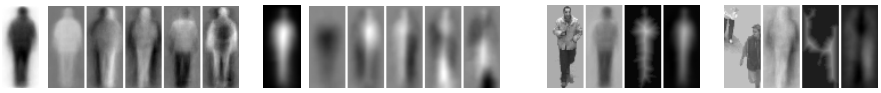
We use a PCA-based subspace representation as a reconstructive model. This low-dimensional representation captures the essential reconstructive characteristics by exploiting the redundancy in the visual data. As such, it enables "hallucinations" and comparison of the visual input with the stored model. In this way the inconsistent data can be rejected and the discriminative model can be trained from clear data only.

To be more specific, once the subspace representation has been built from the training images, we can verify if an input image can be modeled with this model simply by checking its reconstruction error. We can thus project the image into the eigenspace (using the standard projection or a robust procedure [6]), reconstruct the obtained coefficients and determine the reconstruction error, which is a good verification measure. Having a consistent model of training images, we can successfully evaluate new images by considering this measure.

It turns out that a similar approach can be used also in the learning stage, thus during the estimation of the principal subspace. By considering the reconstruction error, the robust learning procedure can discard inconsistencies in the input data and train the model from consistent data only [3, 15, 16]. We use a similar but simplified approach and by checking the consistency of the input images (patches) we accept or reject potential patches as positive or negative training examples for the discriminative learner.

To further increase the robustness of the reconstructive model, we build two subspace representations in parallel: appearance-based and shape-based representation. The former is created from the cropped and resized appearance patches, which are detected by the motion detector. Since the output of this detector is also a binary segmentation mask, this mask is used to calculate the shape images based on the Euclidean distance transform [2]. The mean and the first five eigenvectors of the appearance-based and shape-based model are shown in Figs. 2(a,b).

Having these models, each image can be checked whether is consistent with them or not. Figs. 2(c,d) depict an image and its appearance and shape reconstructions in the case of a correct and a false detection. In the latter case, the reconstruction error is significantly larger (i.e., the original image and its reconstruction differ significantly), thus the patch, which encompasses parts of two pedestrians instead of a single one, gets discarded. Since the main idea of conservative learning is to consider only the images (patches), which are sufficiently consistent with the current model (and would not change it significantly), we



(a) (b) (c) (d)

Fig. 2. Mean and first five principal vectors: (a) appearance, (b) shape. Appearance image, its reconstruction, shape image, its reconstruction: (c) in case of correct detection, (d) in case of false detection.

accept only the images, which are close enough to both, the appearance and the shape model. We thus assure that the discriminative learner gets only the clean data.

2.3 Discriminative Model

In principle our proposed concept can be used with any discriminative classifier, but due to its popularity we have used the classical AdaBoost classifier from Viola and Jones [19]. It allows a very fast processing while achieving a high detection rate. The main assumption from Viola and Jones is that a small set of important features can separate the object classes from the background. This feature selection is done by boosting.

To improve the performance we use in addition to the Haar Wavelets local edge oriented histograms, similar to Levi and Weiss [7]. To detect an object the classifier is evaluated at many possible positions and scales on the image. Since both feature types can be calculated with integral images, this can be done very efficiently for each sub-window.

3 Experimental Results

We have created challenging indoor surveillance video sequences showing a corridor in a public building. We have recorded images over several days. A simple motion detector triggers the camera and then each second one image is recorded. In total we have recorded over 35000 images.

For training the classifiers a sequence containing approximately 4500 frames has been used. In order to have a challenging test situation we created an independent test set containing groups of persons, persons partially occluding each other and persons walking in different directions. The test set consists of 300 frames (235 persons) and was manually annotated.

3.1 Description of Experiments

We applied the conservative learning framework as outlined above. To demonstrate the success of the individual steps, we trained an AdaBoost classifier after each step, and applied it to the test set. The AdaBoost classifiers have all the same VC-dimension (i.e., we used 60 weak classifiers). For evaluation we used non-local maximum suppression of the detections.

AdaBoost1: On the training sequence the motion detection produced 412 detections that where considered as persons (approximately 10% false positives), in addition 1000 negative examples are created by randomly sampling image regions where no motion was detected.

AdaBoost2: A robust reconstructive representation obtained by PCA is computed from the output of the motion detector. This representation is used to verify the output of motion detection to create the set of positive examples. From the 412 detections only 140 positive examples are extracted. The set of negative examples is the same as for AdaBoost1.

AdaBoost3: The detections of AdaBoost2 are verified by the reconstructive model (subdivided into 3 groups: true positives, false positives, any others). Detected false positives are fed back into the AdaBoost as negative examples and true positives as positive examples (76 patches were added to positive examples, 209 to negative examples). Note that these are extremely valuable examples because they sample the current decision boundary.

3.2 Results

As an evaluation criterion we used similar to [1], precision, recall and the F-measure that can be considered as tradeoff between recall and precision. The results of the experiments on the test set are summarized in Table 1:

method	true-pos.	false-pos.	recall	precision	F-measure
AdaBoost1	229	605	97.4 %	27.5 %	42.9 %
AdaBoost2	216	160	91.9 %	57.4 %	70.7 %
AdaBoost3	220	12	93.6 %	94.8 %	94.2 %

Table 1. Experimental results on the test set.

From the experimental results one can clearly see that the number of false positives is considerably reduced by the different stages of the classifier (this is exactly what is to be expected from conservative learning). The F-measure improves from initially 42% to more than 90%. To show the benefit of our approach an AdaBoost classifier was trained with hand labeled positive examples. Using this classifier we detected 224 persons and got 126 false positives (F-measure: 77%). Thus, the result is comparable to AdaBoost2.

Fig. 3 shows some example detections obtained by the different classifiers. Since the persons are not moving there is no motion detected (a). The AdaBoost classifier trained with the noisy data (b) yields a lot of false positives. The classifiers trained with the clean data (c) and with the verified false positives (d) provide much better results.

Fig. 4 shows examples of correctly detected persons applying the final classifier. The bright clothed women (a) is as well detected (while the cleaning cart is not detected), so is the dark man (b), the man with the knapsack (c) and the persons close together (d).

(a) (b) (c) (d)

Fig. 3. Detected persons: (a) Motion, (b) AdaBoost1, (c) AdaBoost2, (d) AdaBoost3.

(a) (b) (c) (d)

Fig. 4. Examples of detected persons applying the final classifier.

4 Conclusion

We have presented a novel framework for unsupervised training of an object detection system. The basic idea is to start with a very simple object detection system and then using reconstructive and discriminative classifiers in an iterative fashion (by being very conservative in accepting when a training sample should be added to the training set) to generate better object detectors. We have demonstrated the framework on a surveillance task where we have learned a pedestrian detector. We have started with a simple moving object classifier and then used PCA (on shape and appearance) as a reconstructive classifier which in turn was used to generate a training set for an discriminative AdaBoost classifier. The results obtained by AdaBoost are again filtered by PCA which produces an even better training set. In fact, using this strategy we produce a training set for the AdaBoost classifier which is optimal in the sense that we always sample at the current estimate of the decision surface [13] and not at the unknown theoretic decision boundary.

The framework we have presented is quite general and can be extended in several directions. Our next step is to use online classifiers. In fact, for PCA we have already on-line algorithms, using also on-line AdaBoost will avoid collecting

training data in batches and training the system off-line in different phases. In addition, we plan to increase the diversity of different classifiers and to include also voting in the process.

References

1. S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. PAMI*, 26(11):1475–1490, 2004.
2. H. Breu, J. Gil, D. Kirkpatrick, and M. Werman. Linear time euclidean distance transform algorithms. *IEEE Trans. PAMI*, 17(5):529–533, 1995.
3. F. De la Torre and M. J. Black. A framework for robust subspace learning. *IJCV*, 54(1):117–142, 2003.
4. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. CVPR 2003*, pages 264–271, 2003.
5. Y. Freund and R. Shapire. A decision-theoretic generalization of online learning and an application to boosting. *J. of Computer and System Sciences*, 55:119–139, 1997.
6. A. Leonardis and H. Bischof. Robust recognition using eigenimages. *Computer Vision and Image Understanding*, 78:99–118, 2000.
7. K. Levi and Y. Weiss. Learning Object Detection from a Small Number of Examples: The Importance of Good Features. In *Proc. CVPR 2004*, 2004.
8. A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using co-training. In *Proc. ICCV*, pages 626–633, 2003.
9. N. Littlestone. Learning quickly when irrelevant attributes abound. *Machine Learning*, 2:285–318, 1987.
10. N.J.B. McFarlane and C.P. Schofield. Segmentation and tracking of piglets. *Machine Vision and Applications*, 8(3):187–193, 1995.
11. V. Nair and J.J. Clark. An unsupervised, online learning framework for moving object detection. In *Proc. CVPR 2004*, pages 317–324, 2004.
12. A. Opelt, M. Fussenegger, Axel Pinz, and Peter Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proc. ECCV 2004*, volume II, pages 71–84, 2004.
13. Jin-Hyun Park and Young-Kiu Choi. On-line learning for active pattern recognition. *IEEE Signal Processing Letters*, 3(11):301–303, 1996.
14. H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. PAMI*, 20(1):23–38, 1998.
15. D. Skočaj, H. Bischof, and A. Leonardis. A robust PCA algorithm for building representations from panoramic images. In *Proc. ECCV 2002*, volume IV, pages 761–775, 2002.
16. D. Skočaj and A. Leonardis. Weighted and robust incremental method for subspace learning. In *Proc. ICCV 2003*, volume II, pages 1494–1501, 2003.
17. K. Sung and T. Poggio. Example-based learning for view-based face detection. *IEEE Trans. PAMI*, 20:39–51, 1998.
18. V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
19. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR 2001*, pages 511–518, 2001.
20. P. Viola, M.J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc. ICCV 2003*, volume 2, pages 734–741, 2003.