Extended Abstract
Object Recognition via Robust Learning*

Horst Bischof
Inst. for Computer Graphics and Vision
Graz University of Technology
Inffeldgasse 16/II, A-8010 Graz
bischof@icg.tu-graz.ac.at

1 Introduction

Visual learning for object and scene recognition is a challenging task. Recently, many different methods mainly rooted in statistical pattern recognition have been proposed. Visual information is in most cases, treated in a direct manner, therefore these methods are not limited by object geometric complexity, texture, or surface markings. This direct representation and the link to statistical pattern recognition make these methods very suitable for learning.

One can characterize the methods if a local or a global image representation is used. Typical global approaches are Principal component analysis (PCA), linear discriminant analysis (LDA), Independent Component Analysis (ICA) etc. Recently proposed local approaches are based on detection of interest points (e.g., Harris, Difference of Gaussians, Maximally Stable Regions, etc.) and the description of their local grey-value neighborhood (e.g. SIFT, Shape context, Steerable filters, etc.). Another characterization of the learning approaches is, if they use a generative model (i.e., ability of reconstruction and generation of samples), or if a discriminative model is employed. Each of them offering distinctive advantages (e.g., generative models enable robustness in model construction, whereas discriminative methods achieve in general higher recognition rates).

Whereas both local and global methods have recently demonstrated how robust recognition can be obtained, the problem of robust learning has been hardly addressed. The main question is how we can learn an object model despite a significant amount of noise and/or occlusion and other disturbances in the training images. This extended abstract will briefly outline one approach and show how it can be used in a learning a pedestrian detection system.

2 Robust Learning

Appearance-based modeling of objects and scenes using subspace representations has become very popular in the vision community. Most of the approaches have used Principle Component Analysis (PCA) for building efficient representations and for subsequent recognition. However, the standard way to perform recognition, based on projections, is prone to errors in the case of non-Gaussian noise, e.g., occlusions, varying illumination conditions, and cluttered background in the input images. Therefore, different authors have proposed robust procedures [Rao, 1997; Black and Jepson, 1996; Leonardis and Bischof, 2000] to obtain reliable recognition also in these cases.

However, if the training images are taken under non-ideal conditions, the obtained representations encompass various non-desirable effects, which cannot be overcome at the recognition stage. This indicates that we need a method to perform *robust training* in order to obtain parametric representations insensitive to these effects. More specifically, we need a procedure which is able to detect inconsistencies in the input data, eliminate them, and then calculate the representation from the consistent data only. In the case of representations based on the PCA, this requires a novel way of calculating the eigenimages from a subset of data points.

We have proposed [Skocaj *et al.*, 2002] a novel robust PCA method to obtain a consistent subspace representation in the presence of outlying pixels in the training images. The method is based on the EM algorithm [Roweis, 1997; Tipping and Bishop, 1999], which enables the calculation of the eigenspaces, i.e., maximum likelihood solution of PCA, in the case of missing data. The fact that we can calculate the PCA on a subset of pixels in the input images, makes it possible to remove the outliers and treat them as missing pixels, arriving at a robust PCA representation. The outliers are determined by a consistency measure over the set of training images.

3 Object Detection

The outlined robust algorithm plays an integral part in a recently developed method for object detection. Starting with face detection [Rowley *et al.*, 1998; Viola and Jones, 2001] there has been a considerable interest in visual object detection in recent years, e.g., pedestrians [P. Viola, 2003], cars [Agarwal and Roth, 2002], bikes [Opelt *et al.*, 2004], etc. At the core of most object detection algorithms is usually a discriminative classifier, e.g., AdaBoost [Freund and Shapire, 1997], Winnow [Littlestone, 1987], Neural network [Rowley *et al.*, 1998] or support vector machine [Vapnik, 1995]. The

*This work is partly funded by the Austrian Joint Research Project Cognitive Vision under sub-projects S9103-N03 and S9104-N04 and by a grant from Federal Ministry for Education, Science and Culture of Austria under the CONEX program. Part of this work has been carried out within the K-plus Competence center ADVANCED COMPUTER VISION funded under the K plus program.

task of the classifier is to decide if the cropped window contains the object of interest or not. The search is repeated for all locations and scales, therefore, the classification has to be very fast.

A requirement of these methods is a representative training set which usually needs to be quite large (several thousands of scaled and aligned images). 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 significantly (e.g. car from the side and car from the back) the classifier needs to be retrained. Training data is most of the time obtained by hand labeling a large number of images which is a time consuming and tedious task. It is clear that this is not practicable for applications requiring a large number of different view-points (e.g. video surveillance by large camera networks).

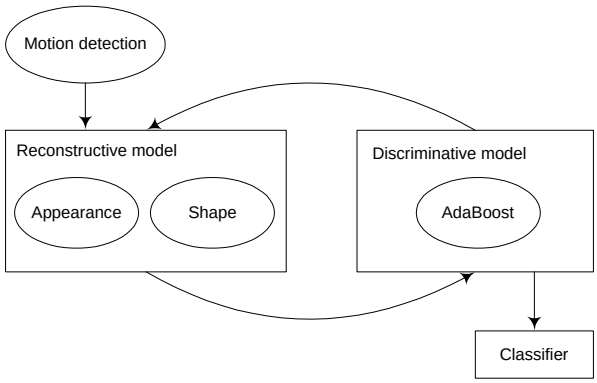

Figure 1: Cooperative learning framework. A reconstructive (generative) model is trained robustly on motion data and labels training data for a discriminative model.

Our novel framework proposes a method to avoid the 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). In particular, the framework is depicted in Fig. 1. We use two types of models a reconstructive (generative) 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. In fact the motion detector will miss a considerable amount of objects (which can be compensated by just using longer sequences) and we will produce also a lot of miss-detections (which will be reduced in the subsequent steps). The output from the motion detector can be used to build a first initial reconstructive representation (in fact to increase the robustness we are using one representation on shape and the other on appearance). Since we get also miss-detections it is of particular importance to use a robust method, otherwise the miss-detections (background, false detections, over-segmentations, etc.) will be incorporated in the model which would severely deteriorate the performance of the whole system. This is very crucial as the discriminative classifier needs to be trained with "clean" images to produce good classification results. The discriminative classifier (at the moment we are using Adaboost) is then used to detect new objects on new images. The output of the discriminative classifier is verified by reconstructive model, and detected false positives can be fed back into the discriminative method 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 [Park and Choi, 1996], 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.

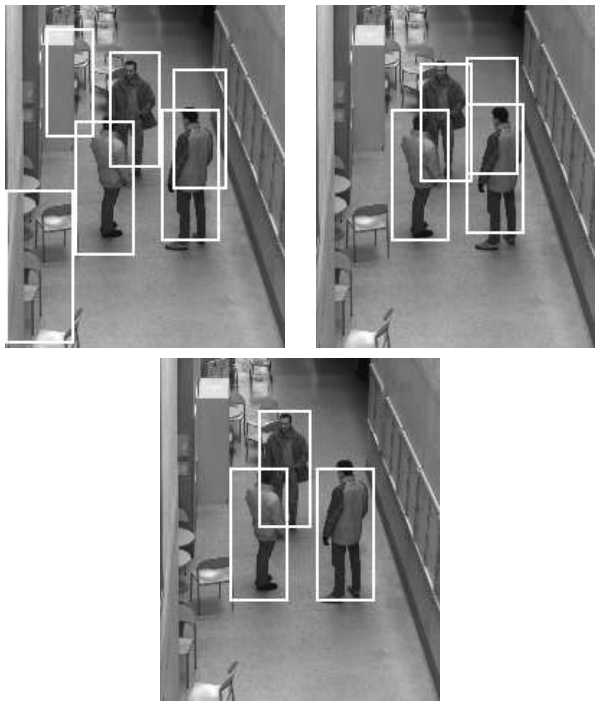

Figure 2: Detected persons at various stages of training.

Fig. 2 depicts how the system iteratively improves. Note that the false positives get considerably by iterating the training.

The whole system shows that by exploiting the huge amount of video data that is available we can produce a stable and robust object detection system.

4 Conclusion

Robust learning is of considerable interest in many applications. In fact, robust building of representations is an unavoidable step in all realistic learning scenarios when the environment can not be specifically tailored for the training phase. The example of learning an object detection system has demonstrated the importance of robust learning. In fact, having a robust learner available would facilitate many new applications, e.g., in the medical domain where the hand labeling efforts of doctors can be considerably reduced.

References

[Agarwal and Roth, 2002] Shivani Agarwal and Dan Roth. Learning a sparse representation for object detection. In *Proceedings of the 7th European Conference on Computer Vision*, volume 4, pages 113–130, 2002.

[Black and Jepson, 1996] M. Black and A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In *ECCV96*, pages 329–342. Springer, 1996.

[Fergus *et al.*, 2003] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *In Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2003.*, pages 264–271, 2003.

[Freund and Shapire, 1997] 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.

[Leonardis and Bischof, 2000] Aleš Leonardis and Horst Bischof. Robust recognition using eigenimages. *Computer Vision and Image Understanding*, 78(1):99–118, 2000.

[Littlestone, 1987] N. Littlestone. Learning quickly when irrelevant attributes abound. *Machine Learning*, 2:285–318, 1987. Winnow algorithm.

[Opelt *et al.*, 2004] A. Opelt, M. Fussenegger, Axel Pinz, and Peter Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proc. ECCV2004*, volume II, pages 71–84, 2004.

[P. Viola, 2003] D. Snow P. Viola, M.J. Jones. Detecting pedestrians using patterns of motion and appearance. In *Proceedings of the Ninth IEEE Conference on Computer Vision (ICCV'03)*, volume 2, pages 734–741, 2003. Pedestria dedection using AdaBoost.

[Park and Choi, 1996] Jin-Hyun Park and Young-Kiu Choi. On-line learning for active pattern recognition. In *IEEE Signal Processing Letters*, pages 301–303, 1996.

[Rao, 1997] R. Rao. Dynamic appearance-based recognition. In *CVPR'97*, pages 540–546. IEEE Computer Society, 1997.

[Roweis, 1997] S. Roweis. Em algorithms for pca and spca. In *NIPS*, pages 626–632, 1997.

[Rowley *et al.*, 1998] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[Skocaj *et al.*, 2002] D. Skocaj, H. Bischof, and A. Leonardis. A robust PCA algorithm for building representations from panoramic images. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proc. ECCV02*, volume IV, pages 761–775. Springer, 2002.

[Sung and Poggio, 1998] K. Sung and T. Poggio. Example-based learning for view-based face detection. *IEEE Trans. PAMI*, 20:39–51, 1998.

[Tipping and Bishop, 1999] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. Technical report, Microsoft Research, 1999.

[Vapnik, 1995] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[Viola and Jones, 2001] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Conference on Computer Vision and Pattern Recognition*, pages 511–518. IEEE CS, 2001.