

# StreetScenes

Stanley M. Bileschi and Lior Wolf

**The Problem:** In the StreetScenes project we study how natural scenes can be processed computationally to produce meaningful semantic information, such as the location and properties of certain object classes, actions or interactions of objects, and the nature or category of the scene itself. Previous work suggests that accurate and powerful scene understanding may be built in a hierarchical manner, where less complex detectors first detect the primitive parts of an object (such as the eyes and nose of a face) and the full detector combines lower level outputs into a final detection. Simultaneously, top-down feedback influences the lower levels by providing category level prior information. It is still an open question how feedback works best in these situations, and what role it plays most naturally. This year, the StreetScenes project will be focusing on these inter-layer interactions so as to increase accuracy and speed of the understanding network.

**Motivation:** Intelligent surveillance and scene understanding systems are in high demand in the marketplace for surveillance, both in civilian and municipal markets. Furthermore, it is in the interest of the biological vision community to be exposed to prototypes capable of the types of difficult processing that humans seem to be able to do so effortlessly.

Hierarchical detection and recognition systems have been shown to outperform systems which treat the entire object region in a homogeneous way [4]. The street scenes project has produced a single algorithm capable of detecting not only strictly geometric objects like cars and faces, but also amorphous objects such as streets and buildings. Our further implementations should be able to also recognize actions and interactions between objects. It is our goal to investigate an elegant solution which is able to produce useful semantic information from all visual input stimuli.

**Previous Work:** Scene understanding is a common area of study in computer vision. There are many published systems which adopt a hierarchical methodology [1, 3, 4, 7, 8, 9, 11], spanning a history of over 10 years. Each of these systems can be divided into a training and a testing process. In the training phase, a corpus of labeled examples is presented to the algorithm; in general these examples are of both positive and negative examples of the object to be detected (i.e. images of faces and of anything-but-faces), but systems which leverage unlabeled examples (unsupervised data) also exist. The key differences in these algorithms involve the network structure, the learning method, and the representation of the visual stimulus.

In the testing phase, a novel stimulus is fed to the detector. The detector must then produce semantic information about this image, including, but not limited to, the location and properties of any recognized objects. In each of the systems referenced above, this search takes place in two steps.

In [1] Support Vector Machines (SVMs) are trained to find the part examples, as well as perform the second level classification, whereas in [3] this is done with probability density modeling techniques. More recent systems use more advanced object-specific features, and more complicated combination strategies [12].

**Approach:** We have chosen to limit our study to the domain of street-level images and video. Our goal is to study scene understanding in a framework general enough to detect examples of roads, buildings, and the sky, but accurate enough to compete with previously built systems for detecting heavily-constrained objects like cars and pedestrians. Currently we have built a corpus of over 3,000 labeled still images for training and testing data. These images are taken from the streets of Cambridge and Boston taken with a portable digital camera at 1200 x 960 pixels, and labeled by hand. The classifier will have information only from the image and the label; no stereo information will be available for this database. In the future we would want to extend our system to StreetScenes video which will require a new database. Furthermore, we may be interested in developing a synthetic 3d world for a much larger corpus of (synthetic) training data.

So far in this project we have designed and tested algorithms which will automatically generate hierarchal object detectors. We have studied the effects of parameters involved with the number of constituent parts, as well as the effect of different types of representations and different learning algorithms. We have also developed a foveated version of the project which mimics the human retina, including the recognition of complicated objects at the retina and only coarser stimuli near the periphery. The next stage of this project will involve developing feed-back mechanisms capable of improving the speed and accuracy of the system. We would like to develop a framework in which computation is spent efficiently at location which merit the most attention, and complicated object models are double checked within a detection-verification loop.

Finally, the study of the performance of the algorithms on these objects is complex, with a number of competing measures each with weaknesses. Performance measures with all the necessary properties for inter-database and inter-classifier performance comparison have not yet been perfected. We would like to in the course of this project provide a common framework for the comparison of scene understanding systems.

**Difficulty:** The role of feedback in the visual biology is not completely understood. Similarly in computer vision, many methods use completely feed-forward algorithms. By investigating a number of different plausible roles, we hope to reach some of the goals described above. As always, in unknown territory, progress will require discipline and perseverance.

**Impact:** It is hoped that our experiments will aide future researchers in the intelligent design of scene understanding algorithms specific to their problem. **Future Work:** The next phase of the street scenes project, in short, will involve moving from images to video, developing feedback mechanisms, and perfecting measures of performance.

**Research Support:** Research at CBCL is supported by ONR, Darpa, NSF, Kodak, Sienmens, Daimler-Chrysler, ATR, ATT, Compaq, Honda, CRIEPI.

## References:

- [1] B. Heisele, P. Ho, and T. Poggio. Face Detection with Support Vector Machines, Global vs. Component-based Approach. In: Proc. International Joint Conference on Neural Networks.
- [2] H. Schneiderman, T. Kanade. Object Detection Using the Statistics of Parts In International Journal of Computer Vision 2002.
- [3] H. Schneiderman, T. Kanade. A Statistical Approach to 3D Object Detection Applied to Faces and Cars IEEE Conference on Computer Vision and Pattern Recognition, IEEE, June 2000.
- [4] B. Heisele, T. Poggio, M. Pontil. Face Detection in Still Gray Images A.I. Memo No. 1687, C.B.C.L. Paper No. 187 Center for Biological and Computational Learning, M.I.T., Cambridge MA. 2000.
- [5] B. Heisele, T. Serre, M. Pontil, T. Poggio. Component-based Face Detection. In: Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), IEEE Computer Society Press, Kauai, Hawaii, Vol. 1, 657-662, December 2001.
- [6] B. Heisele, T. Serre, M. Pontil, T. Poggio. Categorization by Learning and Combining Object Parts. In: Advances in Neural Information Processing Systems (NIPS'01), Vancouver, Canada, 2002, to appear.
- [7] S. Agarwal, D. Roth. Learning a Sparse Representation for Object Detection Presented in ECCV '02.
- [8] T. Leung, M. Burl, P. Perona. Finding Faces in Cluttered Scenes using Random Labeled Graph Matching Fifth Int. Conf. on Comp. Vision, Cambridge, MA (June 1995).
- [9] S. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, H. Shum. Statistical Learning of Multi-view Face Detection, ECCV 2002, LNCS 2353, pp. 67-81, 2002.
- [10] C. Papageorgiou, T. Poggio. A Trainable System for Object Detection IJCV 2000, pp. 15-33.
- [11] E. Sudderth, A. Torralba, W. Freeman, A. Willsky. Learning Hierarchal Models of Scenes, Objects and Parts, ICCV 2005.
- [12] S. Bileschi, L. Wolf. Perception Strategies in Hierarchal Vision Systems, CVPR 2006.