Visual Object Detection Using Learned Features

Chris Stauffer & Eric Grimson

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Detecting classes of objects in images is a difficult problem in computer vision. Recent approaches have established that features based on local differences of regions (e.g., adaptations of Haar basis functions) contain useful information for classification of faces, pedestrians, and vehicles. This work involves deriving features specifically for detection of particular classes of objects.

Motivation: Determining an effective set of features for object detection is somewhat of a black art. Some have used an overcomplete Haar basis[1] while other have simply used a large set of "random" first and second order derivative approximations[3]. While these approaches appear to be general, they require parameters to define the size of the regions of integration and their relative location and the do not entertain non-contiguous features. In some cases, the absolute value of the coefficients is used.

Previous Work: Recently Viola et al.[3] have used a learning algorithm based on AdaBoost to find a subset of a large "random" set of features which effectively detect faces in images. The first two features which their algorithm determined to be most valuable in face detection corresponded to differences between forehead and eyes and between eyes and nose.

Oren et al. [1] used the abosolute value of the coefficients of an overcomplete Haar basis with similar features at two scales for pedestrian detection and determined that features which corresponded to differences between regions inside and outside the object were the most significant in detection.

Approach: It is our hypothesis that the reason the features mentioned above were effective in classification is that they take differences between two regions that each tend to be somewhat uniform but tend to be different from each other.

To investigate the validity of this hypothesis, we plan to experiment with object detection using Support Vector Machines (SVMs) on a varied set of feature coefficients derived from the same data including: the pixel values of the color image, the pixel values of grayscale images, a principle component analysis (PCA), the absolute value on an overcomplete Haar basis, and a new basis derived from recent work at our lab[2].

This new basis is a hierarchical basis representing regions of regularity. It is derived directly from the positive images in the training set. Figure 1 shows this decomposition for a set of pedestrian images. We believe that since these features represent the true regions of regularity, statistics derived from them will be more informative than differences of simple rectangular regions used currently. These regions are irregular in shape unlike the rectangular regions used in the other work.

The features will be distances between aggregate statistics on these image regions (e.g., the average color value). This allows us to take full advantage of color images. The features in other work often use grayscale values rather than color images. This work will also investigate which regions should be coupled into one feature.

Impact: Deriving better features without any supervision or ad hoc heuristics would be extremely useful to the field of computer vision. Also, this research should help to establish a better understanding about what properties of features are useful and why.

Future Work: If we can illustrate the effectiveness of these features in object detection, many interesting questions remain. We will investigate how to derive a feature set for a multi-class classification problem. We will also investigate whether the feature derivation process and learning the detector can be coupled.

Research Support: This work has been funded by DARPA under contract number N00014-00-1-0907 administered by the Office of Naval Research.

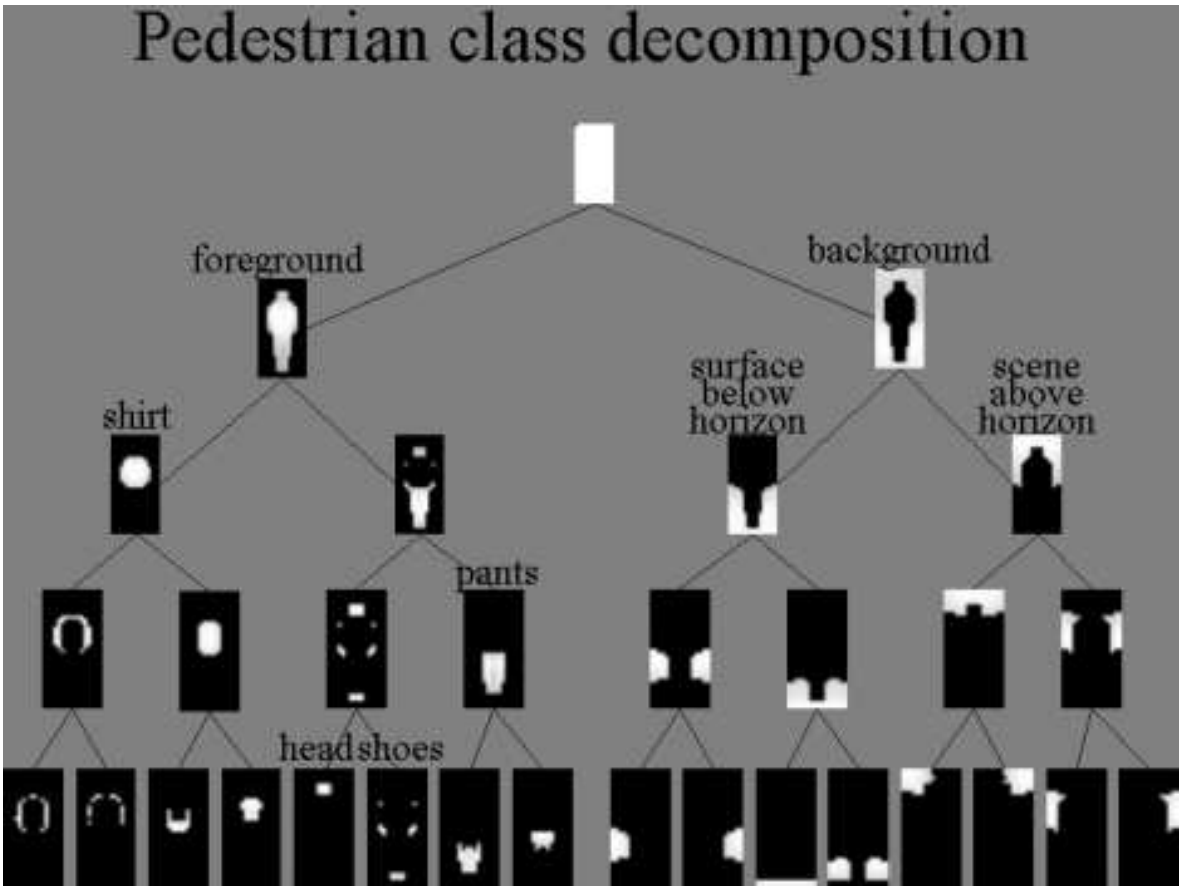

Figure 1: This figure shows the automatically generated binary decomposition of the image patch for the pedestrian data set. The root node represents every pixel in the image. The first branch represents foreground vs. background pixels. Further branches are discussed below.

References:

[1] M. Oren, C. Papageorgiou, E. Osuna P. Sinha, and T. Poggio. Pedestrian detection using wavelet templates. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 193–199, Puerto Rico, June 16–20 1997.

[2] C. Stauffer and W.E.L. Grimson. Similarity templates for detection and recognition. In *Proc. Computer Vision and Pattern Recognition*, Kauai, Hawaii, 2001.

[3] Paul Viola and Michael J. Jones. Robust real-time object recognition. *Compaq Research Tech Report 2001-01*, February 2001.