Tracking-Based Automatic Object Recognition

Chris Stauffer & Eric Grimson

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: While a tracking system is unaware of the identity of any object it tracks, the identity remains the same for the entire tracking sequence. Our system leverages this information by using accumulated joint co-occurrences of the representations within the sequence to create a hierarchical binary-tree classifier of the representations. This classifier is useful to classify sequences as well as individual instances. This work classifies tracked objects based on their shape using the binary motion silhouettes produced by our tracker[5].

Motivation: Infants have an innate ability to do primitive tracking. As they track objects, they are unaware of the identity of the object. But as long as they observe a stable input signal, they can be relatively certain that the object is the same object. We are investigating using the information gained by simple, primitive tracking behavior to aid in visual tasks- in particular, unsupervised hierarchical classification.

Our goal is to produce a system that can be situated in a new environment and, without intervention, produce models of the objects in that environment. Because this work depends on tracking moving objects to classify, it will be most useful in tasks which involve tracking because it can first learn to classify and then use the classification in its task. For instance, it may be capable of determining a vocabulary of object types for activity monitoring, interactive environments, home security, outdoor wildlife monitoring, or many other video sources.

Previous Work: Baumberg and Hogg have done some early work on learning parametric models of shape from sequences of images[1]. More complex, hand-crafted 3D models have also been explored[4]. In contrast, our approach is completely non-parametric. Our hierarchical decomposition method is similar to the recent non-negative matrix Factorization (NMF)[3] and its predecessor Thomas Hofmann's Aspect model[2].

Approach: This new approach to object classification involves extracting representations from a large number automatically-obtained, unlabeled sequences of binary silhouettes of tracked objects. The primary goals are to determine the number of object classes in the scene and to classify new sequences of images effectively into those classes.

To acquire the image sequences, a static camera is directed towards a relatively static scene in which objects are moving. A tracking algorithm determines the moving objects and extracts the motion silhouettes of the objects[5].

A second, independent process selects sequences at random and attempts to determine a set of prototypes which adequately cover the possible inputs. A co-occurrence matrix, C, is approximated in which $C_{i,j}$ is the probability that an image sequence will contain two instances representing both prototype i and j. The prototypes are clustered into equivalency classes based on C.

Figure 1 illustrates a single example and shows the prototypes, the co-occurrence matrix, and resulting classifier.

Impact: Because of the noise inherent in the tracking process, the silhouettes can be extremely noisy making classification difficult. Despite the difficulty of the general classification problem, the greatest challenge arises from trying to automatically generate classifiers using the invariances observed in the tracking data.

Future Work: The next step in development is to extend the use and stability of this method to include larger sets of objects under widely varying conditions. Also, using prototypes which depend only on the features which have been reliable for classification would provide a weak segmentation. Ultimately, this method will adapt low-level features based on its particular domain and use those features for classification.

Research Support: This work has been funded by DARPA under contract number N00014-97-0363 administered by the Office of Naval Research and by NTT.

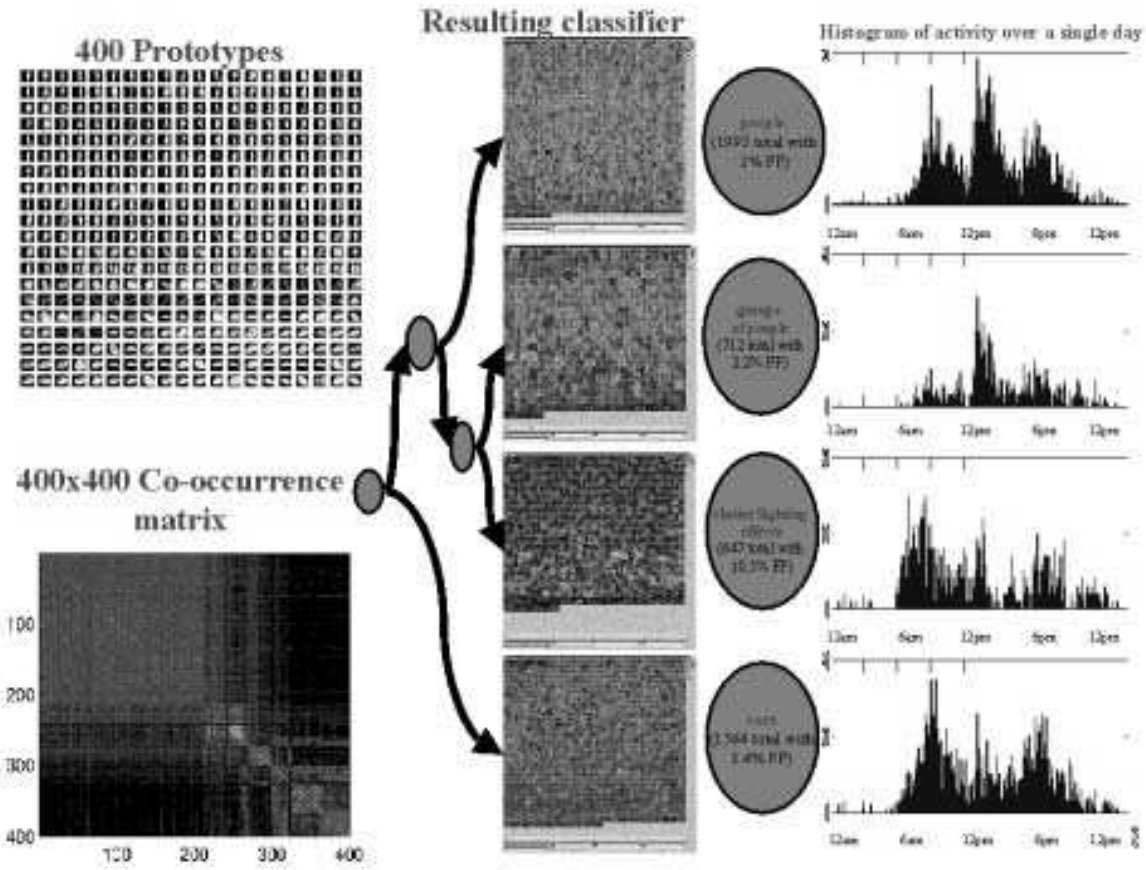

Figure 1: This figure shows the 400 silhouette prototypes and the corresponding 400x400 co-occurrence matrix. The resulting classifier classifies tracked objects as individuals, groups of people, clutter/lighting effects, and vehicles. There are relatively few false positives. The histograms on the right show the activity over a single day. Morning and evening rush hours are evident. A midday rush hour is prominent for both individuals and groups of pedestrians.

References:

[1] A Baumberg and D. C. Hogg. Learning flexible models from image sequences. *European Conference on Computer Vision(ECCV94)*, May 1994.

[2] Thomas Hofmann. Probabilistic latent semantic analysis. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*, 1999.

[3] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[4] K. Rohr. Incremental recognition of pedestrians from image sequences. *Image and Vision Computing*, 1(1):5–20, 1983.

[5] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. Computer Vision and Pattern Recognition*, pages 246–252, 1999.