Real-Time Face Detection

Bernd Heisele, Thomas Serre & Sam Prentice

Artificial Intelligence Laboratory and
The Center for Biological and Computational Learning
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: The problem is to develop a fast trainable system for face detection in a real-world application.

Motivation: Most object detection tasks in computer vision are computationally extensive because of a) the large amount of input data that has to be processed and b) the use of complex classifiers that are robust against pose and illumination changes. Speeding-up the classification is therefore of major concern when developing systems for real-world applications. We investigate two methods for speed-ups [8]: feature reduction and hierarchical classification.

Previous Work: In [4] we presented a system for detecting frontal and near-frontal views of faces in still gray images. The system achieved high detection accuracy by classifying 19×19 gray patterns using a non-linear SVM. However, searching an image for faces at different scales took several minutes on a PC—far too long for most real-world applications. One way to speed-up the system is to reduce the number of features. There are basically two types of feature selection methods in the literature: filter and wrapper methods [1]. Filter methods are preprocessing steps performed independent of the classification algorithm or its error criteria; PCA is an example of a filter method. Wrapper methods attempt to search through the space of feature subsets using the criterion of the classification algorithm to select the optimal feature subset. Wrapper methods can provide more accurate solutions than filter methods [6], but in general are more computationally expensive. We present a new wrapper method to reduce the dimensions of both input and feature space of an SVM.

Feature reduction is a generic tool that can be applied to any classification problem. When dealing with a specific classification task we can use prior knowledge about the type of data to speed-up classification. Two assumptions hold for most vision-based object detection tasks: a) The vast majority of the analyzed patterns in an image belongs to the background class and b) most of the background patterns can be easily distinguished from the objects. Based on these two assumptions it is sensible to apply a hierarchy of classifiers. Fast classifiers removes large parts of the background on the bottom and middle levels of the hierarchy and a more accurate but slower classifier performs the final detection on the top level. This idea falls into the framework of coarse-to-fine template matching [7, 2, 3] and is also related to biologically motivated work on attention-based vision [9, 5]. In contrast to classical coarse-to-fine template matching, the complexity of our detectors is not only controlled by the resolution of the input pattern but also by the classifier's decision function. The bottom level of our hierarchy consists of a linear classifier that operates on low resolution patterns while the top level consists of a non-linear classifier operating on higher resolution patterns.

Approach: We first apply feature selection to a 2^{nd} degree polynomial Support Vector Machine (SVM) that is at the top-level of our hierarchy. The problem of choosing the subset of features which minimizes the expected error probability of an SVM is NP-complete. To simplify the problem, we first rank the input features and then determine how many of the ranked features should be selected. Both ranking and selection is done by minimizing a bound on the expected error probability of the classifier. A similar strategy is used for reducing the features in the feature space. We first rank the features and then determine the number of features. In contrast to the input space technique, the number of features is determined based on the difference between the output of the SVM when using a subset of the features and the output of the SVM when using all features. In addition to feature selection, we implement a 3-layer hierarchy of SVM classifiers. The classifiers on the first and second level are a 9×9 linear SVM and a 19×19 linear SVM, respectively. They remove large parts of the background patterns so that only 1% of the image has to be analyzed by the top-level polynomial SVM.

Difficulty: The detection system must be fast and accurate.

Impact: We apply this technique to video stream indexing, i.e. detecting and tracking faces.

Future Work: We further want to experiment with different types of features, resolutions and complexities of classifiers. Also we want to apply feature selection to all levels of the hierarchy.

Research Support: Research at CBCL is supported by ONR, Darpa, NSF, Kodak, Siemens, DaimlerChrysler, ATR, ATT, Compaq, Honda, CRIEPI.

References:

[1] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 10:245–271, 1997.

[2] P. J. Burt. Smart sensing within a pyramid vision machine. *Proc. of the IEEE*, 76(8):1006–1015, 1988.

[3] J. Edwards and H. Murase. Appearance matching of occluded objects using coarse-to-fine adaptive masks. *Proc. of the IEEE on Computer Vision and Pattern Recognition*, pages 533–539, 1997.

[4] B. Heisele, T. Poggio, and M. Pontil. Face detection in still gray images. A.I. memo 1687, Center for Biological and Computational Learning, MIT, Cambridge, MA, 2000.

[5] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visaul attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Vision*, 20(11):1254–1259, 1998.

[6] R. Kohavi. Wrappers for feature subset selection. *Artificial Intelligence, special issue on relevance*, 97:273–324, 1995.

[7] A. Rosenfeld and G. J. Vanderbrug. Coarse-fine template matching. *IEEE Transactions on systems, man and cybernetics*, 2:104–107, 1977.

[8] T. Serre, B. Heisele, and T. Poggio. Feature selection for face detection. A.I. memo 1697, Center for Biological and Computational Learning, MIT, Cambridge, MA, 2000.

[9] A. Torralba and P. Sinha. Statistical context priming for object detection. *Proc. of the ICCV*, 2001.

System	Average time for a 320×240 image	Speed-up factor
Single second-degree polynomial SVM	768 s	–
Single second-degree polynomial SVM + Feature reduction	181 s	4.25
3-Level hierarchy + Feature reduction	2.95 s (1st: 33.5%, 2nd: 22.6%, 3rd: 43.9%)	260

Table 7: Computing time for face detection systems on a Pentium III with 500 MHz. The original image was rescaled in 5 steps to detect faces at resolutions between 27×27 and 63×63 pixels.

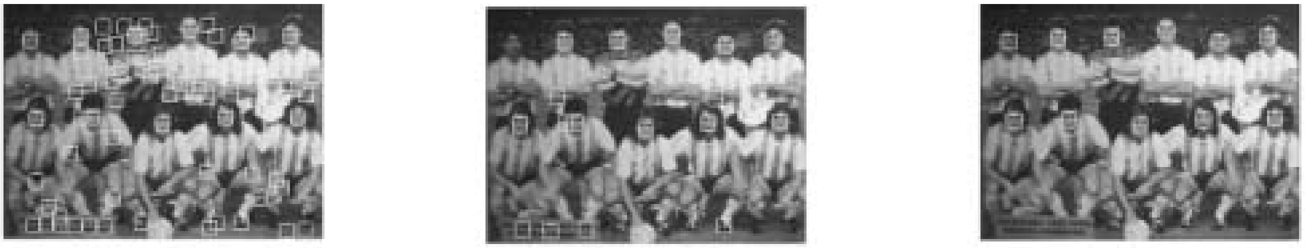

Figure 1: Detection results when the classifiers of the 3 layers (level 1 left, level 2 middle, level 3 right) are applied independently.