MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1673 October 1999
C.B.C.L Paper No. 180

A Trainable Object Detection System: Car Detection in Static Images

Constantine P. Papageorgiou and **Tomaso Poggio**

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1500-1999/AIM-1673.ps

Abstract

This paper describes a general, trainable architecture for object detection that has previously been applied to face and people detection with a new application to car detection in static images. Our technique is a learning based approach that uses a set of labeled training data from which an implicit model of an object class – here, cars – is learned. Instead of pixel representations that may be noisy and therefore not provide a compact representation for learning, our training images are transformed from pixel space to that of Haar wavelets that respond to local, oriented, multiscale intensity differences. These feature vectors are then used to train a support vector machine classifier. The detection of cars in images is an important step in applications such as traffic monitoring, driver assistance systems, and surveillance, among others. We show several examples of car detection on out-of-sample images and show an ROC curve that highlights the performance of our system.

This report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and at the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. This research is sponsored by ONR/MURI grant N00014-95-1-0600. Additional support is provided by Eastman Kodak Company, DaimlerChrysler, Siemens Corporate Research, Inc. and AT&T.

1 Introduction

This paper describes a trainable system for object detection in static images with a particular application to car detection. This system has previously been applied to both face and people detection with success; we highlight the generality of the system with this new domain of car detection. The detection of cars in images is an important step in applications such as traffic monitoring, driver assistance systems, and surveillance, among others. Our approach is to use *example based learning*; we provide the system with a set of training data and it *learns* what a car looks like. Rather than using "stock" images of single cars, we consistently use images gathered from real world scenes. This system currently identifies frontal and rear views of cars.

While it is possible to construct simple models for identifying and tracking cars in constrained domains – for instance, if we know that our camera will always be mounted at a fixed location over a highway – these types of systems will have limited use in more general applications and conditions. We avoid any handcrafting and present a learning based approach to car detection that uses a set of labeled training data to derive an implicit model of cars. Since the pixel images may be noisy and therefore not provide a compact representation for learning, we use features that respond to local, oriented intensity differences in the images; specifically, we use a Haar wavelet representation. The car images we use for training are transformed from pixel space to wavelet space and are then used to train a support vector machine classifier. Support vector machines are capable of finding optimal separating hyperplanes in high dimensional spaces with very few training examples.

The previous work in car detection can be divided into approaches that find cars in static images and techniques that process video sequences; we first look at some static approaches. Bregler & Malik, 1996 [4] describe a system that uses mixtures of experts to identify different classes of cars. The inputs for classification are a large number of second order Gaussian features projected onto a smaller dimensional space. This system assumes that the cars are already segmented and scaled; it is not a detection system, but shares some inspiration (large number of intensity difference operators to represent classes of objects) with ours. Lipson, 1996 [7] describes a system that uses a deformable template for side view car detection. In this system, the wheels, mid-body region, and regions above the wheels are roughly detected based on photometric and geometric relations. The wheels are then more precisely localized using a Hausdorf match. Processing is confined to high resolution images, possibly a restriction for more general detection tasks. This system has been applied to scene classification [8] and shares some conceptual similarity with that of Sinha [16, 17]. Rajagopalan *et al.*, 1999 [15] have recently developed a trainable car detection system that clusters the positive data in a high dimensional space and, to classify an unknown pattern, computes and thresholds a distance measure based on the higher order statistics of the distribution. This technique has a good deal in common with the face detection system of [19, 20].

Motion, or at least the use of multiple frames of relevant information, contains information that can be used to better identify cars; the following systems use dynamical information in one way or another. Beymer *et al.*, 1997 [3] present a traffic monitoring system that has a car detection module. This portion of the system locates corner features in highway sequences and groups feature for single cars together by integrating information over time. Since the system operates in a fairly restricted domain, their detection requirements are not as stringent as our own. Betke *et al.*, 1997 [1] and Betke & Nguyen, 1998 [2] use corner features and edge maps combined with template matching to detect cars in highway video scenes. This system can afford to rely on motion since it is designed for a fairly narrow domain, that of highway scene analysis from a vehicle.

2 System Overview

The core system we use is a general, trainable object detection system that has previously been described in [13, 11, 14, 12]. This paper provides further evidence that this system is indeed a general architecture by adding car detection to the existing applications of face and people detection. We note that each of these instances uses a single framework with no specialized modifications to the code; only the training data is different.

The car detection system uses a database of 516 frontal and rear color images of cars, normalized to 128×128 and aligned such that the front or rear bumper is 64 pixels across. For training, we use the mirror images as well for a total of 1,032 positive patterns and 5,166 negative patterns; a few examples from our training database are shown in Figure 1. From the images, it should be easy to see that the pixel based representations have a significant amount of variability that may lead to difficulties in learning; for instance, a dark body on a white background and a white body on a dark background would have significantly different characteristics under a pixel representation.

To avoid these difficulties and provide a compact representation, we use an overcomplete dictionary of Haar wavelets in which there is a large set of features that respond to local intensity differences at several orientations. We present an overview of this representation here; details can be found in [9] [18].

For a given pattern, the wavelet transform computes the responses of the wavelet filters over the image. Each of the three oriented wavelets – vertical, horizontal, and diagonal – are computed at several different scales allowing the system to represent coarse scale features all the way down to fine scale features. In our system for car detection, we use the scales 32×32 and 16×16. In the traditional wavelet transform, the wavelets do not overlap; they are shifted by the size of the support of the wavelet in x and y. To achieve better spatial resolution and a richer set of features, our transform shifts by $\frac{1}{4}$ of the size of the support of each wavelet, yielding an overcomplete dictionary of wavelet features. The resulting high dimensional feature vectors are used as training data for our classification engine.

There is certain *a priori* knowledge embedded in our

Figure 1: Examples from the database of cars used for training. The images are color of size 128×128 pixels, normalized so that the front or rear bumper is 64 pixels wide.

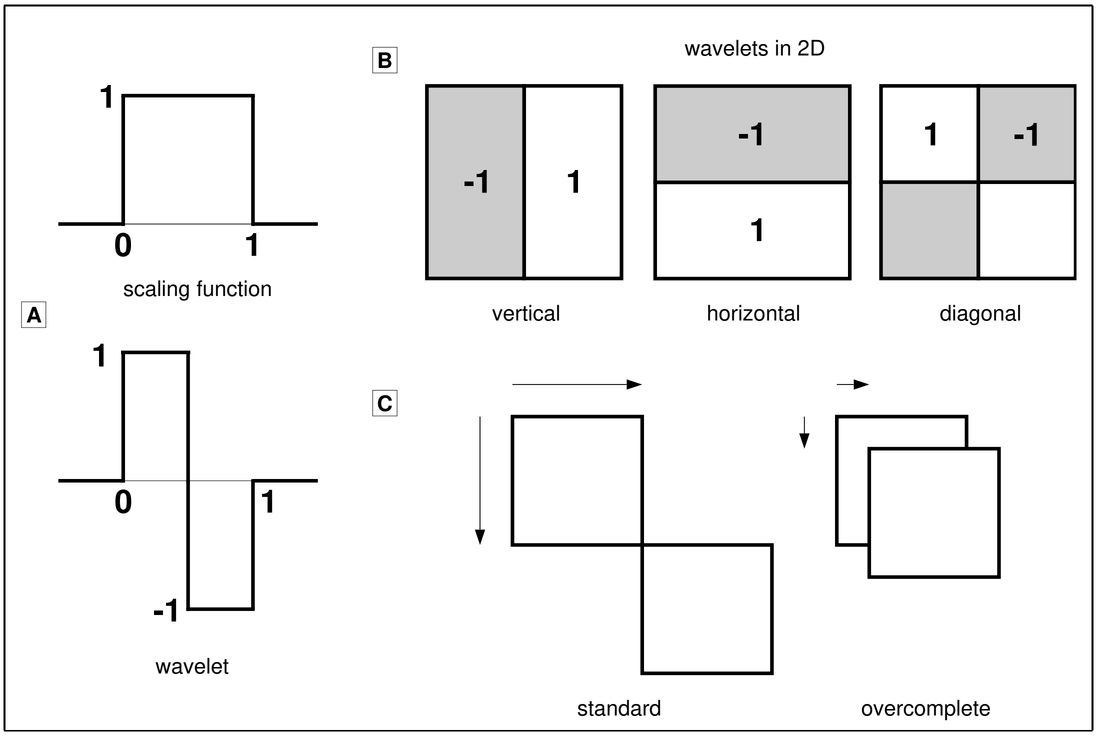
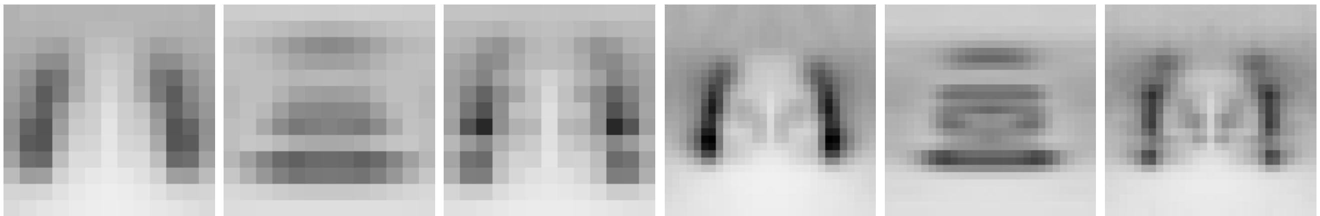


Figure 2: The Haar wavelet framework; (a) the Haar scaling function and wavelet, (b) the three types of 2-dimensional non-standard Haar wavelets: vertical, horizontal, and diagonal, and (c) the shift in the standard transform as compared to our quadruply dense shift resulting in an overcomplete dictionary of wavelets.



(a) (b) (c) (d) (e) (f)

Figure 3: Ensemble average values of the wavelet features of cars coded using gray level. Coefficients whose values are above the average are darker, those below the average are lighter; (a)-(c) are the vertical, horizontal, and diagonal wavelets at scale 32×32, (d)-(f) are the vertical, horizontal, and diagonal wavelets at scale 16×16.

choice of the wavelets. First, we use the absolute values of the magnitudes of the wavelets; this tells the system that a dark body on a light background and a light body on a dark background have the same information content. Second, we compute the wavelet transform for a given pattern in each of the three color channels and then, for a wavelet of a specific location and orientation, we use the one that is largest in magnitude. This allows the system to use the most visually significant features. We note that for car detection we are using exactly the same set of prior assumptions as for our people detection system; this is contrasted with our face detection system which operates over grey level images. The characteristics of this wavelet representation are depicted in Figure 2.

The two scales of wavelets we use for detection are 16×16 and 32×32. We collapse the three color channel features into a single channel by using the maximum wavelet response of each channel at a specific location, orientation, and scale. This gives us a total of 3,030 wavelet features that are used to train the SVM.

The average wavelet feature values are coded in gray level in Figure 3. The grey level coding of the average feature values show that the wavelets respond to the significant visual characteristics of cars: the vertical wavelets respond to the sides of the car, the horizontal wavelets respond to the roof, underside, top of the grille and bumper area, and the diagonal wavelets respond to the corners of the car's body. At the scale 16×16, we can even see evidence of what seems to be license plate and headlight structures in the average responses.

Once we have computed the feature vectors for our positive and negative patterns, we use these to train a support vector machine (SVM) classifier. SVMs are a principled technique to train classifiers that is well-founded in statistical learning theory; for details, see [21] [5]. Unlike traditional training algorithms like back propagation that only minimizes training set error, one of the main attractions of using SVMs is that they minimize a bound on the empirical error and the complexity of the classifier, at the same time. In this way, they are capable of learning in high dimensional spaces with relatively few training examples.

This controlling of both the training set error *and* the classifier's complexity has allowed support vector machines to be successfully applied to very high dimensional learning tasks; [6] presents results on SVMs applied to a 10,000 dimensional text categorization problem.

3 Experimental Results

To detect cars in out-of-sample images, we shift the 128×128 window over all locations in the image, compute the wavelet representation for each pattern, and feed it into the SVM classifier to tell us whether or not it is a car. To achieve multiscale detection, we iteratively resize the entire image and at each step run the fixed size window over the resized images. The shifting and wavelet computation can be done more efficiently be computing the wavelet representation for an entire image once and then shifting in wavelet space. Figure 4 shows some examples of our system running over out-of-sample images gathered from the internet.

To obtain a proper characterization of the performance of our system, we present the ROC curve which quantifies the tradeoff in detection accuracy and rate of false positives in Figure 5. The false positive rate is measured as number of false positives per window processed. For an average sized image in our test set (around 240×360) there are approximately 100,000 patterns the system processes. For a 90% detection rate, we would have to tolerate 1 false positive for every 10,000 patterns, or 10 false positives per image. There has been very little formal characterization the performance of car detection systems in the literature, making it difficult to compare our approach to others'. Given the performance we have achieved, however, we believe our system will compare favorably to existing car detection systems.

4 Future Work and Conclusion

This paper has presented a trainable framework for object detection as applied to the domain of car detection. While there has been considerable work in face and people detection, much of which uses motion, there has been relatively little work in car detection in static images. The framework we describe is indeed general and has successfully been applied to both face, people, and now car detection. This success can be attributed to our use of an effective representation that smooths away noise and while capturing the important aspects of our object class and the use of the support vector machine classifier that allows the system to learn in a 3,030 dimensional space with only 6,198 examples (1,032 positive and 5,166 negative).

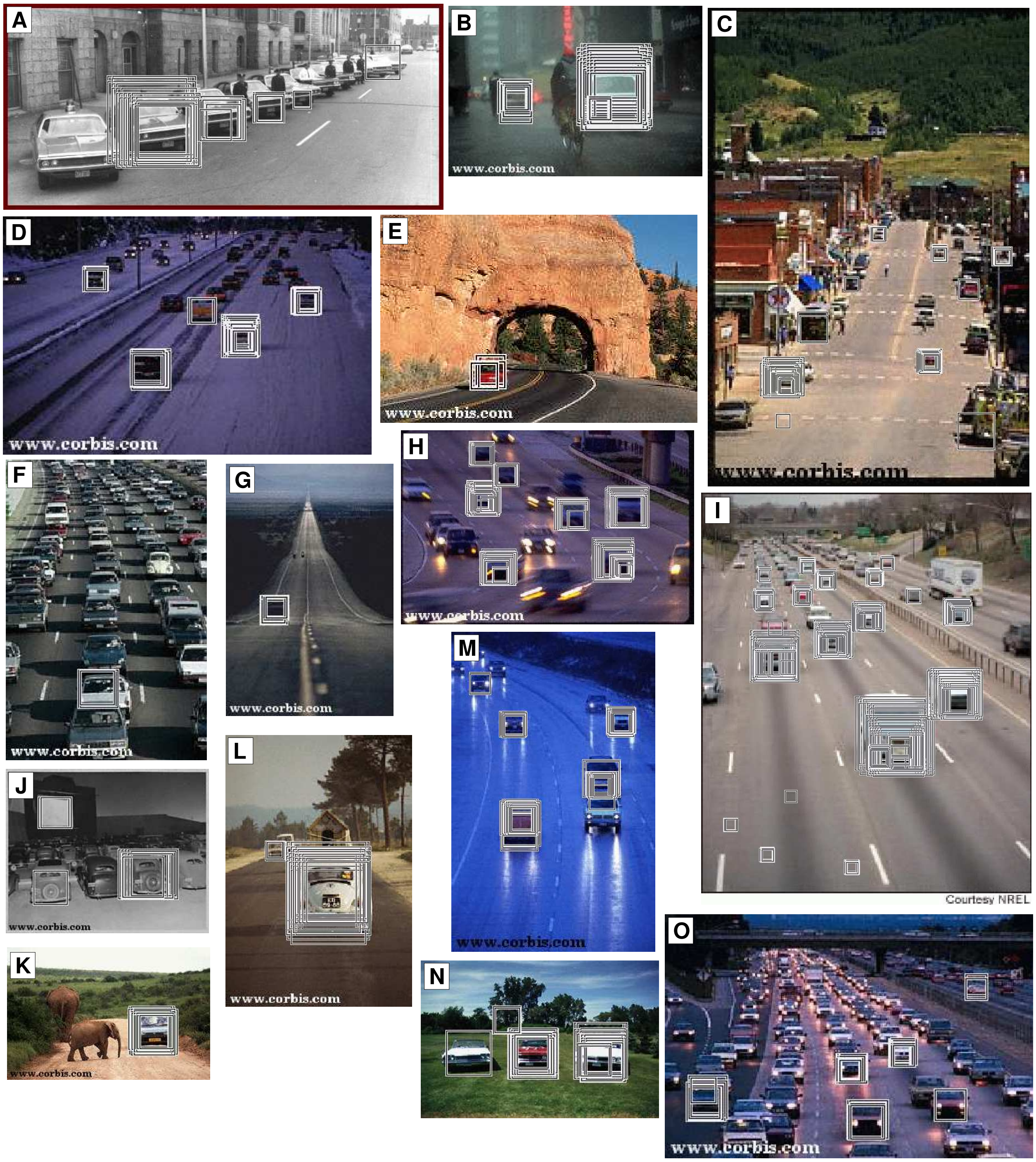


Figure 4: Results of car detection on out-of-sample images. A is from www.lewistonpd.com; B, C, D, E, F, G, H, J, K, L, M, O are from www.corbis.com; I is from www.enn.com; N is from www.foxglove.com. Missed positive examples are due to occlusions (A, F, O) or where a car is too close to the edge of the image (A). False positives (C, J, I, N) are due to insufficient training and can be eliminated with more negative training patterns.

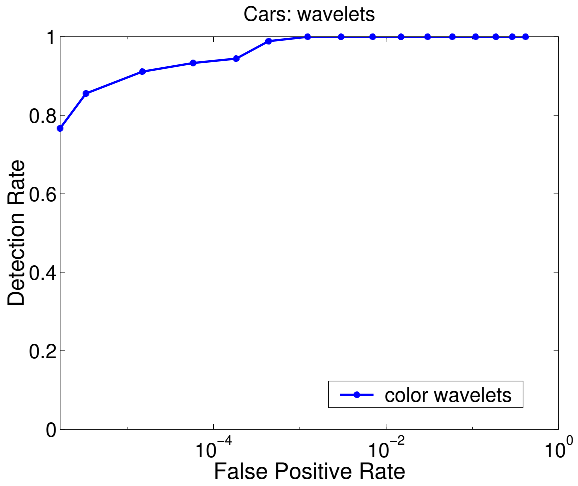


Figure 5: ROC curve for car detection using wavelet features over color images.

Extending this to detect arbitrary poses of cars may be difficult as side view poses have much higher variability frontal and rear views. One possible solution to this is to use a component-based approach that identifies wheels, windows, and other identifiable parts in the proper geometric configuration; a version of this approach is applied to people detection in [10].

References

[1] M. Betke, E. Haritaoglu, and L. Davis. Highway scene analysis in hard real-time. In *Intelligent Transportation Systems*. IEEE, July 1997.

[2] M. Betke and H. Nguyen. Highway scene analysis form a moving vehicle under reduced visibility conditions. In *Intelligent Vehicles*, pages 131–136, October 1998.

[3] D. Beymer, P. McLauchlan, B. Coifman, and J. Malik. A Real-time Computer Vision System for Measuring Traffic Parameters. In *Computer Vision and Pattern Recognition*, pages 495–501. IEEE Computer Society Press, 1997.

[4] C. Bregler and J. Malik. Learning appearance based models: Mixtures of second moment experts. In *Advances in Neural Information Processing Systems*, 1997.

[5] C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. In Usama Fayyad, editor, *Proceedings of Data Mining and Knowledge Discovery*, pages 1–43, 1998.

[6] T. Joachims. Text Categorization with Support Vector Machines. Technical Report LS-8 Report 23, University of Dortmund, November 1997.

[7] P. Lipson. *Context and Configuration Based Scene Classification*. PhD thesis, Massachusetts Institute of Technology, 1996.

[8] P. Lipson, W. Grimson, and P. Sinha. Configuration Based Scene Classification and Image Indexing. In *Computer Vision and Pattern Recognition*, pages 1007–1013, 1997.

[9] S.G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–93, July 1989.

[10] A. Mohan. Robust Object Detection in Images by Components. Master's thesis, Massachusetts Institute of Technology, May 1999.

[11] M. Oren, C.P. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *Computer Vision and Pattern Recognition*, pages 193–99, 1997.

[12] C. Papageorgiou, T. Evgeniou, and T. Poggio. A trainable pedestrian detection system. In *Intelligent Vehicles*, pages 241–246, October 1998.

[13] C.P. Papageorgiou. Object and Pattern Detection in Video Sequences. Master's thesis, MIT, 1997.

[14] C.P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proceedings of 6th International Conference on Computer Vision*, 1998.

[15] A. Rajagopalan, P. Burlina, and R. Chellappa. Higher order statistical learning for vehicle detection in images. In *Proceedings of 7th International Conference on Computer Vision*, pages 1204–1209, 1999.

[16] P. Sinha. Object Recognition via Image Invariants: A Case Study. In *Investigative Ophthalmology and Visual Science*, volume 35, pages 1735–1740. Sarasota, Florida, May 1994.

[17] P. Sinha. Qualitative image-based representations for object recognition. A.I. Memo 1505, MIT Artificial Intelligence Laboratory, 1994.

[18] E.J. Stollnitz, T.D. DeRose, and D.H. Salesin. Wavelets for computer graphics: A primer. Technical Report 94-09-11, Department of Computer Science and Engineering, University of Washington, September 1994.

[19] K-K. Sung and T. Poggio. Example-based learning for view-based human face detection. A.I. Memo 1521, MIT Artificial Intelligence Laboratory, December 1994.

[20] K-K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, January 1998.

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