Quantized Wavelet Features and Support Vector Machines for On-Road Vehicle Detection

Zehang Sun[1], George Bebis[1] and Ronald Miller[2]

[1]Computer Vision Laboratory, Department of Computer Science, University of Nevada, Reno
[2]e-Technology Department, Ford Motor Company, Dearborn, MI
Email:(zehang,bebis)@cs.unr.edu,rmille47@ford.com

Abstract

The focus of this work is on the problem of feature selection and classification for on-road vehicle detection. In particular, we propose using quantized Haar wavelet features and Support Vector Machines (*SVMs*) for rear-view vehicle detection. Wavelet features are particularly attractive for vehicle detection because they form a compact representation, encode edge information, capture information from multiple scales, and can be computed efficiently. Traditionally, methods using wavelet features for classification truncate the coefficients by keeping only the ones with largest magnitude. We believe that the actual values of the wavelet coefficients are not very important for vehicle detection. In fact, the coefficient magnitudes indicate local oriented intensity differences, information that could be very different even for the same vehicle under different lighting conditions. Therefore, we argue and demonstrate experimentally that the actual coefficient values are less important compared to the simple presence or absence of those coefficients. Specifically, we propose quantizing large negative coefficients to -1, large positive coefficients to 1, and setting the rest coefficients to 0. The quantized coefficients seem to encode important information about the general shape and structure of vehicles, while ignoring fine details and allowing for sufficient inter-class variability. Experimental results and comparisons using real data demonstrate the superiority of the proposed approach which has achieved an average accuracy of 93.94% on completely novel test images.

1 Introduction

Robust and reliable vehicle detection in images acquired by a moving vehicle (on-road vehicle detection) is an important problem with application to driver assistance systems or autonomous, self-guided vehicles. Several factors make on-road vehicle detection very challenging including variability in scale, location, orientation, and pose. Vehicles, for example, come into view with different speeds and may vary in shape, size, and color. Vehicle appearance depends on its pose and is affected by nearby objects. In-class variability, occlussion, and lighting conditions also change the overall appearance of vehicles. Landscape along the road changes continuously while the lighting conditions depend on the time of the day and the weather. Last but not least, real-time processing is required.

Appearance-based methods represent a promising direction to vehicle detection. These methods learn the characteristics of the vehicle class from a set of training images which should capture the variability in vehicle appearance. To improve performance, many many methods also model the variability in the non-vehicle class. First, each training image is represented by a set of features which could be either local or global. Then, the decision boundary between the vehicle and non-vehicle classes is computed. In principle, this can be done using learning (e.g., Neural Network (*NN*)) or by modelling the probability distribution of the features in each class [1]. In Matthews et al. [2], feature extraction is based on *PCA*. Subwindows containing vehicle candidates were first scaled to a 20x20 subwindow. Each 20x20 subwindow was then divided into 25 4x4 subwindows and each 4x4 subwindow was subjected to *PCA*. The *PCA* features were then fed to a *NN* for classification. Goerick et al. [3] used a method called Local Orientation Coding (*LOC*) to extract edge information. The histogram of *LOC* within the area of interest was then fed to a *NN* for classification.

A statistical model for vehicle detection was investigated by Schneiderman et al. [4, 5]. First, a view-based approach with multiple detectors was used to cope with variation from different viewpoint. Second, a statistical model within each of these detectors was used to account for other variations. The statistics of both object appearance and "non-object" appearance were represented using the product of two histograms with each histogram representing the joint statistics of a subset of *PCA* features in [4] and Haar wavelet features in [5] and their position on the object. A different statistical model was investigated by Weber et al [6]. They represented each vehicle image as a constellation of local features and used the EM algorithm to learn the parameters of the probability distribution of the constellations. An interest operator, followed by clustering, is used to identify important local features in vehicle images. Papageorgiou et al. [7] have proposed using the Haar wavelet transform for feature extraction and *SVMs* for classification. A quadruple density dictionary was generated using redundant basis functions.

The focus of this work is on the problem of feature selection and classification for vehicle detection from rear views. In particular, we investigate the performance of several different feature selection schemes using Haar wavelets. Features based on Haar wavelets (e.g., coefficients) have yielded promising results in various applications including vehicle detection [5] [7]. Several reasons make these features attractive for vehicle detection. First, they form a compact representation. Second, they encode edge information, an important feature for vehicle detection. Third, they capture information from multiple resolution levels. Finally, there exist fast algorithms, especially in the case of Haar wavelets, for computing these features.

We have implemented and compared three different feature selection schemes using Haar wavelet coefficients in this work. Given an input image, first we compute its Haar wavelet decomposition. In the first feature selection scheme, the features consist of all the coefficients except the ones in *HH* sub-band of the first level [5]. In the second and most traditional scheme, we keep only the coefficients with largest magnitude. In the last scheme, we keep the largest magnitude coefficients as before, however, we quantize their values to -1, 0, 1. This last scheme has been motivated by the work of Jacobs et al. [8] on image retrieval. In all cases, classification is performed using *SVMs*. Our experimental results demonstrate that using quantized wavelet coefficients leads to improved performance both in terms of error rate as well as false positive/false negative rate.

The rest of the paper is organized as follows: In Section 2, we provide brief review of Haar wavelet and *SVMs*. The different feature extraction schemes used in this study are described in detail in Section 3. Our real dataset is described in Section 4. Our experimental results and comparisons are presented in Section 5. Finally, Section 6 contains our conclusions and directions for future research.

2 Haar Wavelet Transform and SVMs Reviews

2.1 Haar Wavelet Transform

Wavelets are essentially a type of multiresolution function approximation that allow for the hierarchical decomposition of a signal or image. They have been applied successfully to various problems including object detection [7, 5], face recognition [9] and image retrieval [8]. Any given decomposition of a signal into wavelets involves just a pair of waveforms (mother wavelets). The two shapes are translated and scaled to produce wavelets (wavelet basis) at different locations (positions) and on different scales (durations). We formulate the basic requirement of multiresolution analysis by requiring a nesting of the spanned spaces as:

$$\cdots V_{-1} \subset V_0 \subset V_1 \cdots \subset L^2 \tag{1}$$

In space V_{j+1}, we can describe finer details than in space V_j. In order to construct a multi-resolution analysis, a scaling function ϕ is necessary, together with the dilated and translated version of it:

$$\phi_i^j(x) = 2^{\frac{j}{2}}\phi(2^j x - i). \quad i = 0, \cdots, 2^j - 1. \tag{2}$$

The important features of a signal can be better described or parameterized, not by using $\phi_i^j(x)$ and increasing j to increase the size of the subspace spanned by the scaling functions, but by defining a slightly different set of function $\psi_i^j(x)$ that span the difference between the spaces spanned by various scales of the scale function. These functions are the wavelets, which spanned the wavelet space W_j such that $V_{j+1} = V_j \bigoplus W_j$,and can be described as:

$$\psi_i^j(x) = 2^{\frac{j}{2}}\psi(2^j x - i). \quad i = 0, \cdots, 2^j - 1. \tag{3}$$

Different scaling function $\phi_i^j(x)$ and wavelets $\psi_i^j(x)$ determines different wavelet transform. In this paper, we use Haar wavelet. Haar wavelet is the simplest to implement and computationally the least demanding. Furthermore, since Haar basis forms an orthogonal basis, the transform provides a non-redundant representation of the input images. The Haar scaling function is defined as:

$$\phi(x) = \begin{cases} 1 & for \;\; 0 \le x < 1 \\ 0 & otherwise \end{cases} \tag{4}$$

and the Haar wavelet is defined as:

$$\psi(x) = \begin{cases} 1 & for \;\; 0 \le x < \frac{1}{2} \\ -1 & for \;\; \frac{1}{2} \le x < 1 \\ 0 & otherwise \end{cases} \tag{5}$$

Wavelet features capture visually plausible features of the shape and interior structure of objects. Features at different scales capture different levels of detail. Coarse scale features encode large regions while fine scale features describe smaller, local regions. All these features together disclose the structure of an object in different resolutions.

2.2 SVMs

SVMs are primarily two-class classifiers that have been shown to be an attractive and more systematic approach to learning linear or non-linear decision boundaries [10] [11]. Given a set of points, which belong to either of two classes, *SVM* finds the hyper-plane leaving the largest possible fraction of points of the same class on the same side, while maximizing the distance of either class from the hyper-plane. This is equivalent to performing structural risk minimization to achieve good generalization [10] [11]. Assuming l examples from two classes

$$(x_1, y_1)(x_2, y_2)...(x_l, y_l), \quad x_i \in R^N, y_i \in \{-1, +1\} \tag{6}$$

finding the optimal hyper-plane implies solving a constrained optimization problem using quadratic programming. The optimization criterion is the width of the margin between the classes. The discriminate hyper-plane is defined as:

$$f(x) = \sum_{i=1}^{l} y_i a_i k(x, x_i) + b \tag{7}$$

where $k(x, x_i)$ is a kernel function and the sign of $f(x)$ indicates the membership of x. Constructing the optimal hyper-plane is equivalent to find all the nonzero a_i. Any data point x_i corresponding to a nonzero a_i is a support vector of the optimal hyper-plane.

Suitable kernel functions can be expressed as a dot product in some space and satisfy the Mercer's condition [10]. By using different kernels, *SVMs* implement a variety of learning machines (e.g., a sigmoidal kernel corresponding to a two-layer sigmoidal neural network while a Gaussian kernel corresponding to a radial basis function (*RBF*) neural network). The Gaussian radial basis kernel is given by

$$k(x, x_i) = \exp(-\frac{\| x - x_i \|^2}{2\delta^2}) \tag{8}$$

The Gaussian kernel is used in this study (i.e., our experiments have shown that the Gaussian kernel outperforms other kernels in the context of our application).

3 Feature Selection Schemes

The input to the vehicle detection system is a 32×32 image which is decomposed into 5 levels. In the first feature selection scheme, we keep all the coefficients except the ones in the *HH* sub-band of the first level[5]. Our motivation here is keeping as much information as possible while rejecting coefficients that are likely to encode noise. In the second feature selection scheme, we reject all small magnitude coefficients. Small magnitude coefficients encode mostly noise or fine details that are not essential for vehicle detection. Figure 1 (2nd row) shows examples of reconstructed vehicle/non-vehicle images using only the 50 largest coefficients. It should be clear from Figure 1 that these coefficients convey important shape information, a very important feature for vehicle detection, while unimportant details have been removed.

The third feature selection scheme is based on the observation that the actual values of the wavelet coefficients might not be very important since we are interested in the general shape of vehicles only. In fact, the magnitudes indicate local oriented intensity differences, information that could be very different even for the same vehicle under different lighting conditions. Therefore, the actual coefficient values might be less important or less reliable compared to the simple presence or absence of those coefficients. Similar observations have been made in [8] assuming an image retrieval application. The third feature selection thus, contains the quantized truncated coefficients selected in the second scheme. We use three quantization levels: -1, 0, and +1 (i.e., -1 representing large negative coefficients, +1 representing large positive coefficients, and 0 representing everything else). The images in the third row of Figure 1 illustrate the quantized wavelet coefficients of the vehicle images shown in the first row. For comparison purposes, the last row of Figure 1 shows the quantized wavelet coefficients of the non-vehicle images shown in the fourth row.

Figure 1: 1st row: vehicle sub-images used for training; 2nd row: reconstructed sub-images using the top 50 coefficients; 3rd row: illustration of the top 50 quantized coefficients; 4th and 5th rows: similar results for some non-vehicle sub-images.

4 Dataset

The images used in our experiments were collected in Dearborn, Michigan during two different sessions, one in the Summer of 2001 and one in the Fall of 2001, using Ford's proprietary low-light camera. To ensure a good variety of data in each session, the images were caught during different times, different days, and on five different highways. The training set contains sub-images of rear vehicle views and non-vehicles which were extracted manually from the Fall 2001 data set. A total of 1051 vehicle sub-images and 1051 non-vehicle sub-images were extracted by several students in our lab. Although specific instructions were given to the students, there is some variability in the way the sub-images were extracted. For example, certain sub-images cover the whole vehicle, others cover the vehicle partially, and others contain the vehicle and some

background (see Figure 1). In [7], the sub-images were aligned by wrapping the bumpers to approximately the same position. We have not attempted to align the data in our case since alignment requires detecting certain features on the vehicle accurately. Moreover, we believe that some variability in the extraction of the sub-images can actually improve performance. Each sub-image in the training and test sets was scaled to 32×32 and preprocessed to account for different lighting conditions and contrast [12]. First, a linear function was fit to the intensity of the image. The result was subtracted out from the original image to correct for lighting differences. Then, histogram equalization was performed to improve contrast.

To evaluate the performance of the proposed approach, the average accuracy (AR), false positives (FPs), and false negatives (FNs), were recorded using a three-fold cross-validation procedure. Specifically, we split the training dataset randomly three times (*Set1*, *Set2* and *Set3*) by keeping 80% of the vehicle sub-images and 80% of the non-vehicle sub-images (i.e., 841 vehicle sub-images and 841 non-vehicle sub-images) for training. The rest 20% of the data was used for validation during the training of the neural network classifier which was used for comparison purposes. For testing, we used a fixed set of 231 vehicle and non-vehicle sub-images which were extracted from the Summer 2001 data set.

5 Experimental Results and Comparisons

We have performed a number of experiments and comparisons to evaluate the performance of the three feature selection schemes. In all cases, classification was performed using a *SVM* classifier with Gaussian kernel. First, we considered the original Haar Wavelet coefficients, without using the coefficients from the *HH* sub-band of the first level as discussed in Section 3. We will be referring to this approach as (*OSVM*). Table 1 shows the results we obtained in this case. Using 768 coefficients, the AR rate was 91.49%, the average FP rate was 6.50%, and the average FN rate was 2.02%. Besides the relatively high FP rate, these results could be considered reasonable given that we used a very simple feature selection scheme. The last column of Table 1 shows the number of support vectors created in each case. On the average, this method creates 496 support vectors.

Next, we considered the second selection scheme, that is, using coefficients with large magnitude. We will be referring to this approach as *TSVM*. We run several experiments keeping the top 25, 50, 100, 125, 150, and 200 coefficients. Figures 2.(a-c) shows the AR, FP, and FN rates obtained for each case. The best results were obtained using 125 coefficients *T125SVM*. In this case, the AR rate was 92.06%, the average FP rate was 4.33%, and the average FN rate was 3.61%.

In terms of accuracy, the *T125SVM* approach

Table 1: Performance of *OSVM*

	Set1	Set2	Set3	Aver
AR	92.21%	90.48%	91.78%	91.49%
FP	5.63%	7.36%	6.50%	6.50%
FN	2.16%	2.16%	1.73%	2.02%
SV	349	459	680	496

Table 2: Performance of *T125SVM*

	Set1	Set2	Set3	Aver
AR	92.21%	92.64%	91.34%	92.06%
FP	3.90%	3.90%	5.20%	4.33%
FN	3.90%	3.46%	3.46%	3.61%
SV	313	309	302	308

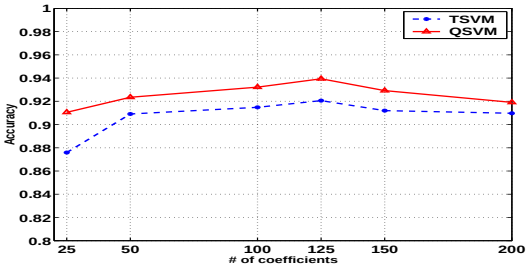
Table 3: Performance of *Q125SVM*

	Set1	Set2	Set3	Aver
AR	93.94%	94.37%	93.51%	93.94%
FP	2.60%	1.73%	2.60%	2.31%
FN	3.46%	3.90%	3.90%	3.75%
SV	409	379	421	403

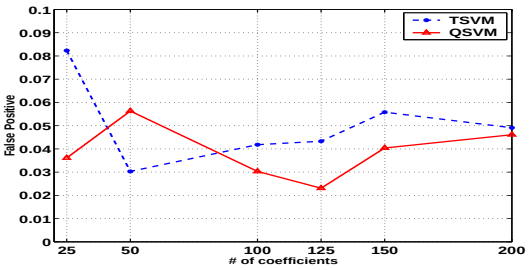
Table 4: Performance of *Q125NN*, with 80 hidden nodes

	Set1	Set2	Set3	Aver
AR	84.41%	83.55%	83.98%	83.98%
FP	9.68%	13.85%	12.12%	11.88%
FN	5.91%	2.60%	3.90%	4.14%

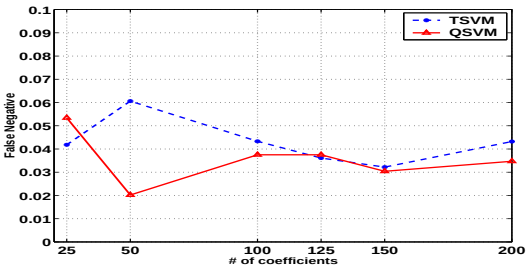
yielded higher AR than the *OSVM* approach, however, the difference is not that significant. However, the *T125SVM* approach yielded lower *FPs* but higher *FNs*. In terms of support vectors, the *T125SVM* approach created quite less support vectors. The third selection scheme was considered last, that is, using quantized truncated coefficients. We will be referring to this approach as *QSVM*. We run several experiments again by quantizing the top 25, 50, 100, 125, 150, and 200 coefficients as described in Section 3. Figures 2.(a-c) show the AR, FP, and FN rates obtained for each case. As can be observed from Figure 2.(a), the *QSVM* approach demonstrated higher AR than the *TSVM* approach in all cases. In terms of *FPs*, the performance of the *QSVM* approach was consistently better or equal to the performance of the *TSVM* approach when keeping 100 coefficients or more (see Figure 2.(b)). In terms of *FNs*, the performance of the *QSVM* approach was consistently better or equal to that of the *TSVM* approach when keeping 25 coefficients or more (see Figure 2.(c)). Our best results were



(a)



(b)

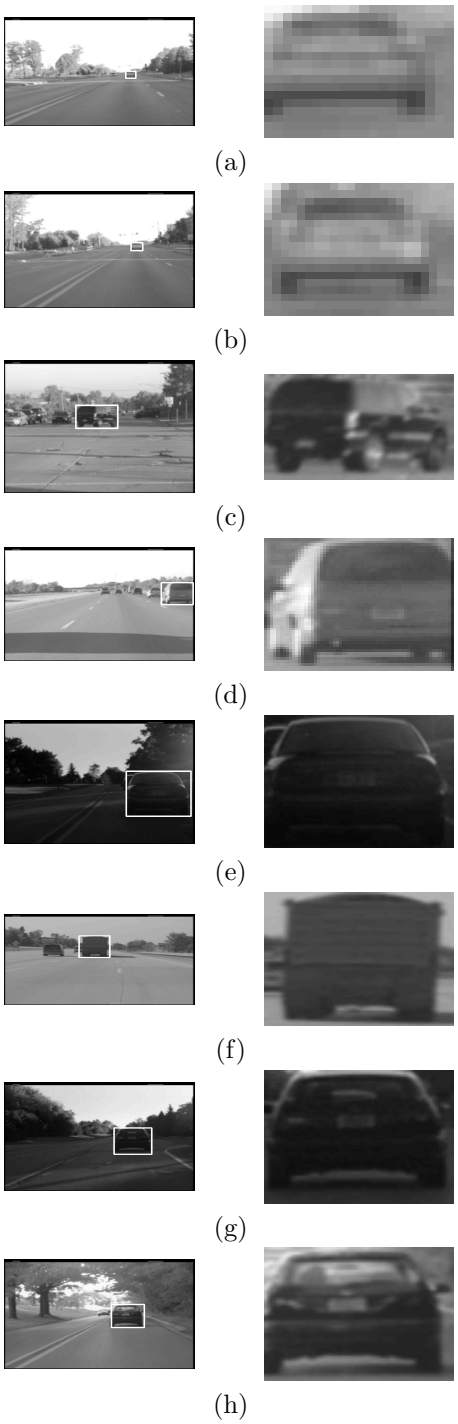


(c)

Figure 2: Performances v.s. number of coefficients kept. (a). Detection accuracy. (b). False positive. (c). False negative

obtained using 125 coefficients (see Table 3). The *AR* rate obtained in this case was 93.94%, the average *FP* rate was 2.31%, and the average *FN* rate was 3.75%. In terms of support vectors, the *Q125SVM* approach created more support vectors than the *T125SVM* approach, in general however, the *QSVM* created less number of support vectors than the *TSVM* approach. Overall, the *Q125SVM* approach demonstrated better performance compared both to the *TSVM* and *OSVM* approaches. For comparison purposes, we tested the three best feature sets (*Q100 Q125*, *Q150*) using a *NN* classifier. The *NN* classifier used was a fully connected, two-layer, feed-forward neural network trained by the back-propagation algorithm. We varied the number of hidden nodes to obtain optimum performance. We obtained our best results using 125 coefficients and 80 hidden nodes (see Table 4). Specifically, the average *AR* rate was 83.98%, the average *FP* rate was 11.88%, and the average *FN* rate was 4.14%. Obviously, the performance of the *SVM* classifier outperforms that of the *NN* classifier in every respect.

Figure 3 shows some successful detection examples using the *Q125SVM* approach. The results illustrate some interesting points. Figure 3(a-b) shows a case



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

Figure 3: Some successful vehicle detection examples.

where only the general shape of the vehicle is available (i.e., no details) due to its far distance from the camera. The *Q125SVM* method seems to discard irrelevant details, leading to improved robustness. The vehicles in Figures 3(c-d), were detected successfully from their side view, although we have not included side views in the training set. This demonstrates good generalization properties. Also, the proposed method can tolerate some illumination changes as can be seen from Figures 3(e-h). Some *FP* and *FN* examples are

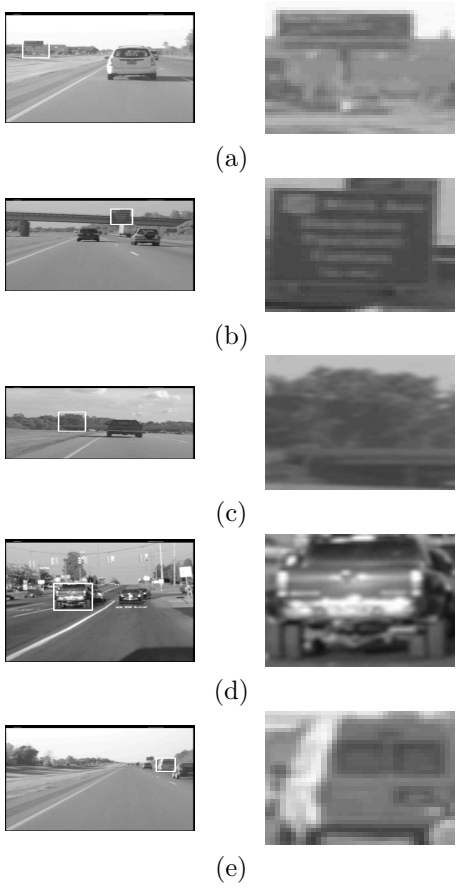


Figure 4: Examples of FPs (a-c) and FNs (c,d)

shown in Figure 4. The majority of the FNs were due to the lack of representative examples in the training set and due to some extreme rotations. We believe that some of the *FPs* were due to the relatively small number of "non-vehicle" examples we used for training. Given that the "non-vehicle" class is much larger than the "vehicle" class, it would be necessary to include more "non-vehicle" examples in the training set. Bootstrapping [13] would definitely be very useful for choosing good "non-vehicle" examples to improve generalization.

6 Conclusion and Future Work

We have proposed the use of quantized Haar wavelet features and *SVMs* for rear-view vehicle detection. Our experimental results and comparisons have shown that quantizing the largest Haar wavelet features offers improved performance compared to using the original values of the same coefficients or of a larger set. These results confirm the fact that feature selection is an important issue for vehicle detection. For future work, we will consider the problem for feature selection for vehicle detection in more detail. In particular, we plan to investigate the potential of Genetic Algorithms for feature selection, an approach that has yielded very promising results in a gender classification application [14].

Acknowledgements

This research was supported by Ford Motor Company under grant No. 2001332R and in part by NSF under CRCD grant No.0088086. The authors would like to thank Dave DiMeo and Perry MacNeille from Ford Research Lab for their help with the data collection.

References

[1] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Jon-Wiley, 2001.

[2] N. Matthews, P. An, D. Charnley, and C. Harris, "Vehicle detection and recognition in greyscale imagery," *Control Engineering Practice*, vol. 4, pp. 473–479, 1996.

[3] C. Goerick, N. Detlev and M. Werner, "Artificial neural networks in real-time car detection and tracking applications," *Pattern Recognition Letters*, vol. 17, pp. 335–343, 1996.

[4] H. Schneiderman and T. Kanade, "Probabilistic modeling of local appearance and spatial relationships for object recognition," *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 45–51, 1998.

[5] H. Schneiderman, *A statistical approach to 3D object detection applied to faces and cars*. CMU-RI-TR-00-06, 2000.

[6] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," *European Conference on Comptuer vision*, pp. 18–32, 2000.

[7] C. Papageorgiou and T. Poggio, "A trainable system for object detection," *International Journal of Computer Vision*, vol. 38, no. 1, pp. 15–33, 2000.

[8] C. Jacobs, A. Finkelstein and D. Salesin, "Fast multiresolution image querying," *Proceedings of SIGGRAPH*, pp. 277–286, 1995.

[9] G. Garcia, G. Zikos, and G. Tziritas, "Wavelet packet analysis for face recognition," *Image and Vision Computing*, vol. 18, pp. 289–297, 2000.

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

[11] C. Burges, "Tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 955–974, 1998.

[12] G. Bebis, S. Uthiram, and M. Georgiopoulos, "Face detection and verification using genetic search," *International Journal on Artificial Intelligence Tools*, vol. 9, no. 2, pp. 225–246, 2000.

[13] H. Rowley, S. Baluja and T. Kanade, "Neural network-based face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998.

[14] Z. Sun, X. Yuan, G. Bebis, and S. Louis, "Neural-network-based gender classification using genetic eigen-feature extraction,," *IEEE International Joint Conference on Neural Networks*, 2002(accepted, available from http://www.cs.unr.edu/ bebis/Publications.html).