A General Framework for Object Detection

Constantine P. Papageorgiou   Michael Oren   Tomaso Poggio

Center for Biological and Computational Learning
Artificial Intelligence Laboratory
MIT
Cambridge, MA 02139
{cpapa,oren,tp}@ai.mit.edu

Abstract

This paper presents a general trainable framework for object detection in static images of cluttered scenes. The detection technique we develop is based on a wavelet representation of an object class derived from a statistical analysis of the class instances. By learning an object class in terms of a subset of an overcomplete dictionary of wavelet basis functions, we derive a compact representation of an object class which is used as an input to a support vector machine classifier. This representation overcomes both the problem of in-class variability and provides a low false detection rate in unconstrained environments.

We demonstrate the capabilities of the technique in two domains whose inherent information content differs significantly. The first system is face detection and the second is the domain of people which, in contrast to faces, vary greatly in color, texture, and patterns. Unlike previous approaches, this system learns from examples and does not rely on any a priori (hand-crafted) models or motion-based segmentation. The paper also presents a motion-based extension to enhance the performance of the detection algorithm over video sequences. The results presented here suggest that this architecture may well be quite general.

1 Introduction

This paper presents a novel framework for object detection in cluttered scenes, based on the use of an overcomplete dictionary of basis functions and combined with statistical learning techniques. The detection of real-world objects of interest, such as faces and people, poses challenging problems: these objects are difficult to model, there is significant variety in color and texture, and the backgrounds against which the objects lie are unconstrained. In contrast to the case of pattern classification where we need to decide between well-defined classes, the detection problem requires us to differentiate between the object class and the rest of the world. As a result, the class model must accommodate the intra-class variability without compromising the discriminative power in distinguishing the object within cluttered scenes. We also cannot assume that there are a certain number of objects, if any, in the image; MAP or maximum likelihood methods will not work since the classification of each pattern in an image is done independently. This paper also introduces an extension that uses motion cues to improve detection accuracy over video sequences. This motion module is a general one that can be used with many detection algorithms and does not compromise the ability of the system to detect non-moving objects.

Initial work on the detection of rigid objects in static images, such as street signs or faces, Betke & Makris[1], Yuille, et. al.[21], used template matching approaches with a set of rigid templates or hand-crafted parameterized curves. These approaches are difficult to extend to more complex objects such as people, since they involve a significant amount of prior information and domain knowledge. In recent research, more closely related to our system, the detection problem is solved using learning-based techniques that are data driven. This approach was used by Sung &Poggio[16] and Vaillant, et al.[18] for the detection of frontal faces in cluttered scenes, with similar architectures presented by Moghaddam and A. Pentland [9], Rowley, et. al.[14], and Osuna et al.[11].

Most previous systems that detect objects in video sequences focused on using motion and 3D models or constraints to find people: Tsukiyama & Shirai[17], Leung & Yang[6], Hogg[4], Rohr[13], Wren, et al.[20], Heisele, et. al.[3], McKenna & Gong[8]. These systems suffer from restrictive assumptions on the scene structure, for instance, a single object in the scene or a stationary camera and a sequence of frames. In some of these motion-based systems, the focus is on model fitting, tracking and motion interpretation. In contrast, our work addresses the issue of detection in single static images in unconstrained environments with cluttered backgrounds, while making no assumption on the scene structure.

One of the major issues in developing a system that will handle complex classes of objects is finding an appropriate image representation. To illustrate the importance of an appropriate visual coding, Figure 1 shows images of people and their corresponding edge maps. It is clear that both the pixel and edge-based representations are inadequate; the pedestrian images vary greatly in color and texture and the edge maps

Figure 1: The top row shows examples of images of people in the training database. The examples vary in color, texture, view point (either frontal or rear) and background. The bottom row show edge detection of the pedestrians. Edge information does not characterize the pedestrian class well.

lack consistency and have a lot of spurious information. Using these representations, it will be hard, if not impossible, to derive a class model. An important feature of our work is the use of an image coding that circumvents the variability in an object class.

This work develops the idea of using an expressive, overcomplete set of basis functions to define the structure of an object class. We use a Haar wavelet representation to capture the structural similarities between instances of an object class. This idea of an overcomplete, or redundant, representation was introduced in [10] and we showed how this model is learnable from examples, using pedestrian detection as a testbed. Here, we expand on the original system and apply it to another domain, faces, with promising results.

2 The Wavelet Representation

The Haar wavelets are a natural set basis functions which encode differences in average intensities between different regions; for an in depth description of wavelets, see [7]. To achieve the spatial resolution necessary for detection and to increase the expressive power of the model, we introduce the *quadruple density* transform in [10], an extension of the 2D Haar wavelet (Figure 2-1), that yields an overcomplete set of basis functions. Whereas for a wavelet with size 2^n, the standard Haar transform shifts each wavelet by n, the quadruple density transform shifts the wavelet by $\frac{1}{4}2^n$ in each direction, shown in Figure 2-2. The use of this quadruple density transform results in an overcomplete dictionary of basis functions that facilitates the definition of complex constraints on the object patterns. In [10], we also show that we do not lose computational efficiency with respect to the standard wavelet transform.

3 Learning the Class Model

Given an object class, the central problem is how to learn which are the relevant coefficients that express structure common to the entire object class and which are the relationships that define the class. We divide

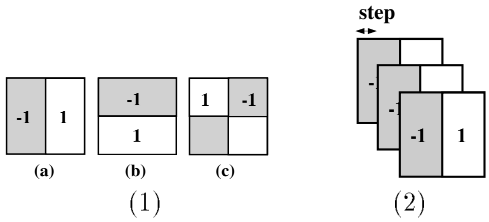


Figure 2: (1) The 3 types of 2-dimensional non-standard Haar wavelets; (a) "vertical", (b) "horizontal", (c) "diagonal". (2) Quadruple density 2D Haar basis.

the learning into a two-stage process: (1) identifying a small subset of basis functions that capture the structure of a class, and (2) using a classifier to derive a precise class model from the subset of basis functions. We illustrate the techniques on two different classes of objects: faces and pedestrians.

3.1 Stage 1: Learning the Significant Basis Functions

To develop our model for the face class, we use a set of 2429 grey-scale images of faces of size 19×19 consisting of a core set of faces with some small angular rotations to improve generalization; typical images from the database are shown in Figure 3. Databases of this size and composition have been used extensively in face detection [15] [14] [11] [12] and we keep this data format for comparison purposes. For the coefficient analysis, we use the wavelets at scales of 4×4 pixels and 2×2 pixels since their dimensions correspond to typical facial features for this size of face image. We have a total of 1734 coefficients.

The basic analysis in identifying the important coefficients consists of two steps. Since the power distribution of different types of coefficients may vary, the first step is to compute the class average of ($\{vertical, horizontal, diagonal\} \times \{2, 4\}$ for a total of 8 classes) and normalize every coefficient by its corresponding class average. The second step is to average

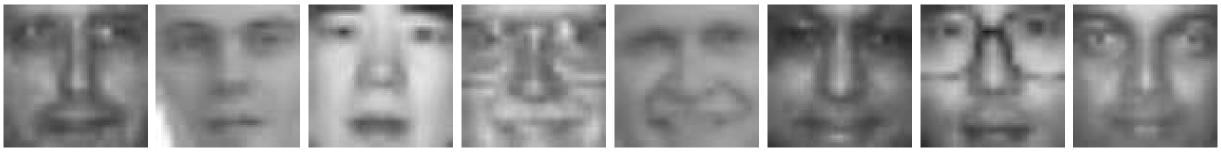
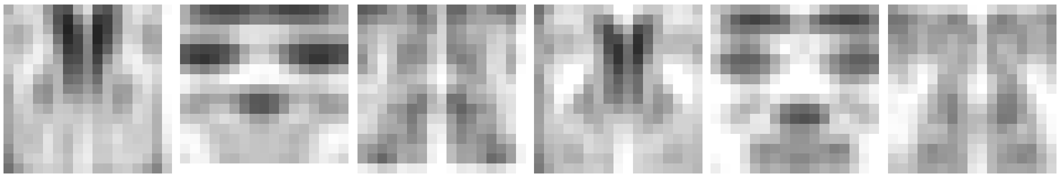


Figure 3: Examples of faces used for training. The images are gray level of size 19×19 pixels.



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

Figure 4: Ensemble average values of the wavelet coefficients for faces coded using color. Each basis function is displayed as a single square in the images above. Coefficients whose values are close to the average value of 1 are coded gray, the ones which are above the average are coded using red and below the average are coded using blue. We can observed strong features in the eye areas and the nose. Also, the cheek area is an area of almost uniform intensity, ie. below average coefficients. (a)-(c) vertical, horizontal and diagonal coefficients of scale 4×4 of images of faces. (d)-(f) vertical, horizontal and diagonal coefficients of scale 2×2 of images of faces.

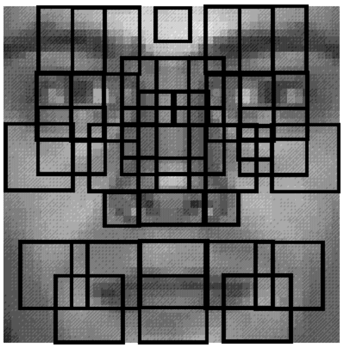


Figure 5: The significant basis functions for face detection that are uncovered through our learning strategy, overlayed on an example image of a face.

the normalized coefficients over the entire set of examples. The normalization has the property that the average value of coefficients of random patterns will be 1. If the average value of a coefficient is much greater than 1, this indicates that the coefficient is encoding a boundary between two regions that is consistent along the examples of the class; similarly, if the average value of a coefficient is much smaller than 1, that coefficient encodes a uniform region.

To illustrate this analysis, we code the coefficients' values using grey-scale in Figure 4, where each coefficient, or basis function, is drawn as a distinct square in the image. The arrangement of the squares corresponds to the spatial location of the basis functions, where strong coefficients (large average values) are coded by darker grey levels and weak coefficients (small average values) are coded by lighter grey levels. It is important to note that in Figure 4, a basis function corresponds to a single square in each image and not the entire image. It is interesting to observe how the different types of wavelets – vertical, horizontal, and diagonal – capture various facial features, such as the eyes, nose, and mouth.

From this statistical analysis, we derive a set of 37 coefficients, from both the coarse and finer scales, that capture the significant features of the face. These significant bases consist of 12 vertical, 14 horizontal, and 3 diagonal coefficients at the scale of 4×4 and 3 vertical, 2 horizontal, and 3 corner coefficients at the scale of 2×2. Figure 5 shows a typical human face from our training database with the significant 37 coefficients drawn in the proper configuration.

For the task of pedestrian detection, we use a database of 924 color images of people (Figure 1). A similar analysis of the average values of the coefficients was done for the pedestrian class and Figure 6 shows the grey-scale coding similar to Figure 4. We refer the interested reader to [10] for the details. It is interesting to observe that for the pedestrian class, there are no strong internal patterns as in the face class; rather, the significant basis functions are along the exterior boundary of the class, indicating a different type of significant visual information. Through the same type of analysis, we choose 29 significant coefficients from the initial, overcomplete set of 1326 wavelet coefficients. These basis functions are shown overlayed on an example pedestrian in Figure 7.

It should be observed, that from the viewpoint of the classification task, we could use the whole set of coefficients as a feature vector. However, using all the wavelet functions that describe a window of 128×64 pixels in the case of pedestrians would yield vectors of very high dimensionality, as we mentioned earlier. The training of a classifier with such a high dimensionality, on the order of 1000, would in turn require too large

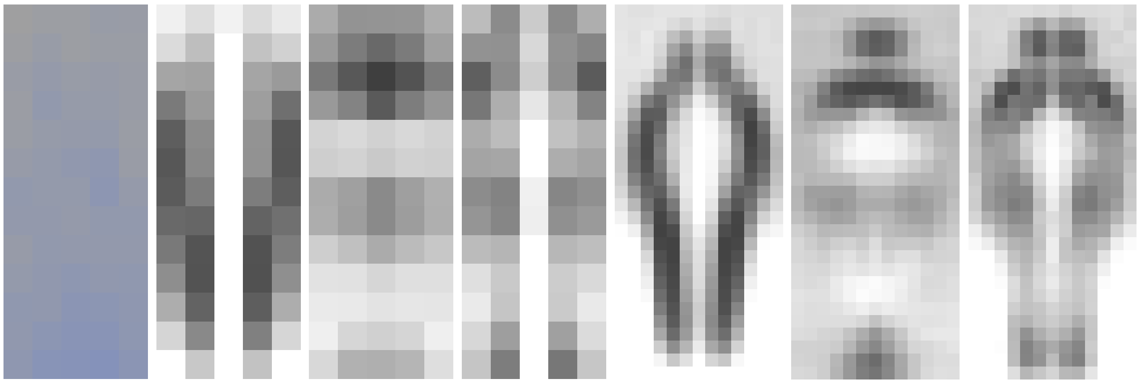


Figure 6: Ensemble average values of the wavelet coefficients coded using gray level. Coefficients whose values are above the template average are darker, those below the average are lighter. (a) vertical coefficients of random scenes. (b)-(d) vertical, horizontal and corner coefficients of scale 32×32 of images of people. (e)-(g) vertical, horizontal and corner coefficients of scale 16×16 of images of people.

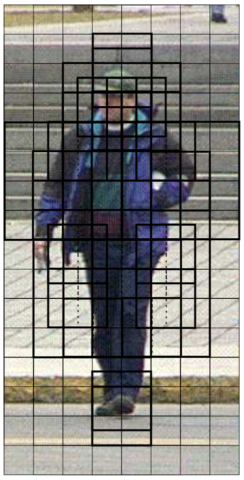


Figure 7: The significant basis functions for pedestrian detection that are uncovered through our learning strategy, overlayed on an example image of a pedestrian.

an example set. This dimensionality reduction stage serves to select the basis functions relevant for this task and to reduce their number considerably.

3.2 Stage 2: Learning the Class Model

Once we have identified the important basis functions we can use various classification techniques to learn the relationships between the wavelet coefficients that define the object class. The classification technique we use is the support vector machine (SVM) developed by Vapnik et al.[2][19]. This recently developed technique has the appealing features of having very few tunable parameters and using structural risk minimization which minimizes a bound on the generalization error (see [11] [12]).

We train our systems using databases of positive examples gathered from outdoor and indoor scenes. The initial negative examples in the training database are patterns from natural scenes not containing people or faces. While the target class is well-defined, there are no typical examples of the negative class. To overcome this problem of defining this extremely large negative class, we use the idea of "bootstrapping" training [16]. In the context of the pedestrian detection system, after the initial training, we run the system over arbitrary images that do not contain any people, adding false detections into the training set as examples of the negative class, and retraining the classifier (Figure 8). This incremental refinement of the decision surface is iterated until satisfactory performance is achieved.

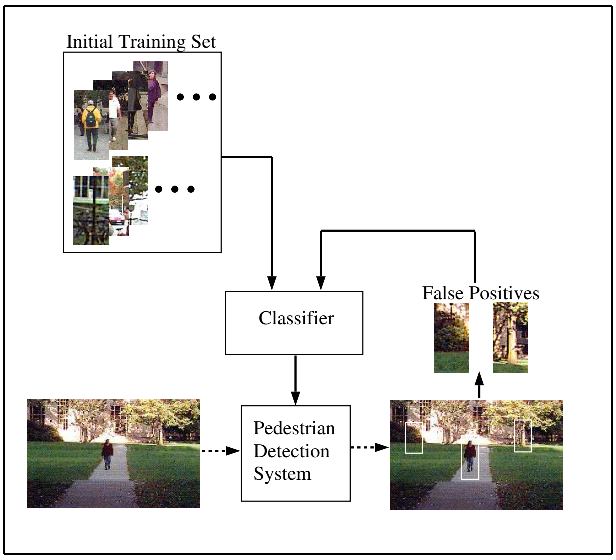
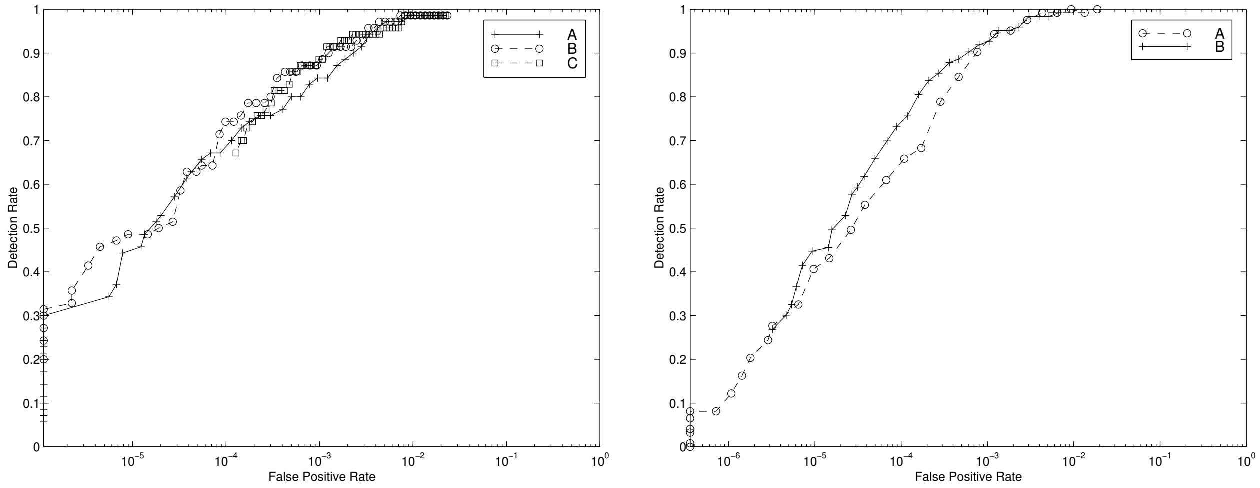


Figure 8: Incremental bootstrapping to improve the system performance.



(a) Face Detection System

(b) People Detection System

Figure 9: ROC curves for the detection systems. The detection rate is plotted against the false detection rate, measured on a logarithmic scale. The false detection rate is defined as the number of false detections per inspected window; (a) Face Detection: System A was trained with equal penalty for missed positive examples and false detections; systems B and C were trained with penalties for missed positive examples that were 1 and 2 orders of magnitude greater than the penalty for false detections, (b) People Detection: System A penalizes incorrect classifications of positive and negative examples equally, system B penalizes incorrectly classified positive examples 5 times more than negative examples.

4 The Experimental Results

The system detects objects in arbitrary positions in the image and in different scales. Once the training phase in Section 3 is complete, the system can detect objects at arbitrary positions by scanning all possible locations in the image by shifting the detection window. This is combined with iteratively resizing the image to achieve multi-scale detection. For our experiments with faces, we detected faces from the minimal size of 19×19 to 5 times this size by scaling the novel image from 0.2 to 1.0 times its original size, at increments of 0.1. For pedestrians, the image is scaled from 0.2 to 2.0 times its original size, again in increments of 0.1. At any given scale, instead of recomputing the wavelet coefficients for every window in the image, we compute the transform for the whole image and do the shifting in the coefficient space.

4.1 Face Detection

To evaluate the face detection system performance, we start with a database of 2429 positive examples and 1000 negative examples. To understand the effect of different penalties in the Support Vector training (see [11] [12]), we train several systems using different penalties for misclassification. The systems undergo the bootstrapping cycle detailed in Section 3, and end up with between 4500 and 9500 negative examples. Out-of-sample performance is evaluated using a set of 131 faces and the rate of false detections is determined by running the system over approximately $900,000$ patterns from images of natural scenes that do not contain either faces or people. To give a complete characterization of the systems, we generate ROC curves that illustrate the accuracy/false detection rate tradeoffs, rather than give a single performance result. This is accomplished by varying the classification threshold in the support vector machine. The ROC curves are shown in Figure 9a and indicate that even higher penalties for missed positive examples may result in better performance. We can see that, if we allow one false detection per 7,500 windows examined, the rate of correctly detected faces reaches 75%.

In Figure 10 we show the results of running the face detection system over example images. The missed detections are due to higher degrees of rotations than were present in the training database; with further training on an appropriate set of rotated examples, these types of rotations could be detected. In the image in the lower right, there are several incorrect detections. Again, we expect that with further training, this can be eliminated.

4.2 People Detection

The frontal and rear pedestrian detection system starts with 924 positive examples and 789 negative examples and goes through 9 bootstrapping steps ending up with a set of 9726 patterns that define the non-pedestrian class. We measure performance on novel data using a set of 105 pedestrian images that are close to frontal or rear views; it should be emphasized that we do not choose test images of pedestrians in perfect frontal or rear poses, rather, many of these test images represent slightly rotated or walking views of pedestrians. We use a set of $2,800,000$ patterns from natural scenes to measure the false detection rate. We give the ROC curves for the pedestrian detection system in Figure 9b; as with faces, these curves indicate

Figure 10: Results from the face detection system. The missed instances are due to higher degrees of rotation than were present in the training database; false detections can be eliminated with additional training.

that even larger penalty terms for missed positive examples may improve accuracy significantly. From the curve, we can see, for example, that if we have a tolerance of one false positive for every 15,000 windows examined, we can achieve a detection rate of 70%. Figure 11 exhibits some typical images that are processed by the pedestrian detection system; the images are very cluttered scenes crowded with complex patterns. These images show that the architecture is able to effectively handle detection of people with different clothing under varying illumination conditions.

Considering the complexity of these scenes and the difficulties of object detection in cluttered scenes, we consider the above detection rates to be high. We believe that additional training and refinement of the current systems will reduce the false detection rates further.

5 Motion Extension

In the case of video sequences, we can utilize motion information to enhance the robustness of the detection; we use the pedestrian detection system as a testbed. We compute the optical flow between consecutive images and detect discontinuities in the flow field that indicate probable motion of objects relative to the background. We then grow these regions of discontinuity using morphological operators, to define the full regions of interest. In these regions of motion, the likely class of objects is limited, so we can relax the strictness of the classifier. It is important to observe that, unlike most person detection systems, we do not assume a static camera nor do we need to recover camera ego-motion, but rather we use the dynamic motion information to assist the classifier. Additionally, the use of motion information does not compromise the ability of the system to detect non-moving people. Figure 12 demonstrates how the motion cues enhance the performance of the system.

We test the system over a sequence of 208 frames; the detection results are shown in Table 1. Out of a possible 827 pedestrians in the video sequence – including side views for which the system is not trained – the base system correctly detects 360 (43.5%) of them with a false detection rate of 1 per 236,500 windows. The system enhanced with the motion module detects 445 (53.8%) of the pedestrians, a 23.7 % increase in detection accuracy, while maintaining a false detection rate of 1 per 90,000 windows. It is important to iterate that the detection accuracy for non-moving objects is not compromised; in the areas of the image where there is no motion, the classifier simply runs as before. Furthermore, the majority of the false positives in the motion enhanced system were partial body detections, ie. a detection with the head cut off, which were still counted as false detections. Taking this factor into account, the false detection rate is even lower.

This relaxation paradigm has difficulties when there are a large number of moving bodies in the frame or when the pedestrian motion is very small when compared to the camera motion. Based on our results, though, we feel that this integration of a trained classifier with the module that provides motion cues could be extended to other systems as well.

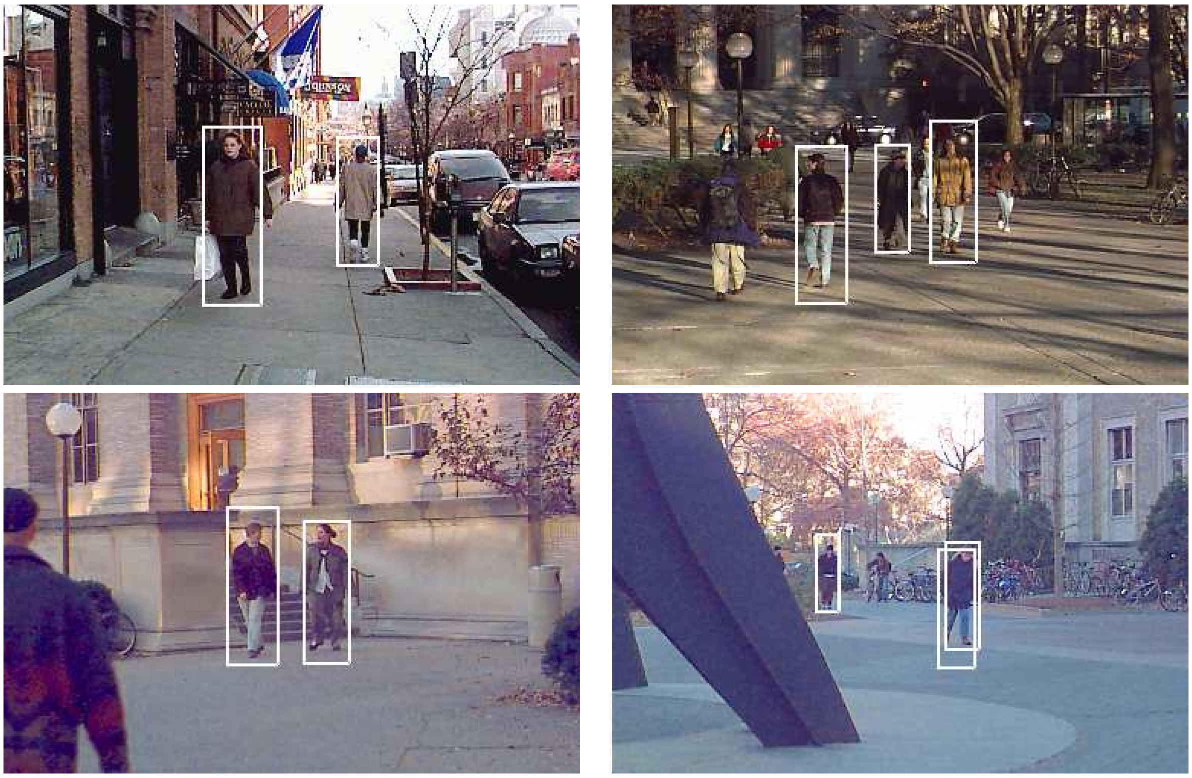


Figure 11: Results from the pedestrian detection system. These are typical images of relatively complex scenes that are used to test the system. Missed examples of pedestrians are usually due to the figure being merged with the background.

	Detection Rate	*False Positive Rate (per window)*
Base system	43.5%	1:236,500
Motion extension	53.8%	1:90,000

Table 1: Performance of the pedestrian detection system with the motion-based extensions, compared to the base system.

6 Conclusion

In this paper, we describe the idea of an overcomplete wavelet representation and demonstrate how it can be learned and used for object detection in a cluttered scene. This representation yields not only a computationally efficient algorithm but an effective learning scheme as well.

We have decomposed the learning of an object class into a two-stage learning process. In the first stage, we perform a dimensionality reduction where we identify the most important basis functions from an original overcomplete set of basis functions. The relationships between the basis functions which define the class model are learned in the second stage using a support vector machine (SVM). Without this dimensionality reduction stage, the training on the original overcomplete set would be difficult, if not intractable. Most of the basis functions in the original full set do not necessarily convey relevant information about the object class we are learning, but, by starting with a large overcomplete dictionary, we would not sacrifice details or spatial accuracy. The learning step extracts the most prominent features and results in a significant dimensionality reduction.

We also present an extension that uses motion cues to improve pedestrian detection accuracy over video sequences. This module is appealing in that, unlike most systems, it does not totally rely on motion to accomplish detection; rather, it takes advantage of the a priori knowledge that the class of moving objects is limited while not compromising performance in detecting non-moving pedestrians.

The strength of our system comes from the expressive power of the overcomplete set of basis functions – this representation effectively encodes the intensity relationships of certain pattern regions that define a complex object class. The encouraging results of our system in two different domains, faces and people, suggest that the approach described in this paper may well generalize to several other object detection tasks.

References

[1] M. Betke and N. Makris. Fast object recognition in

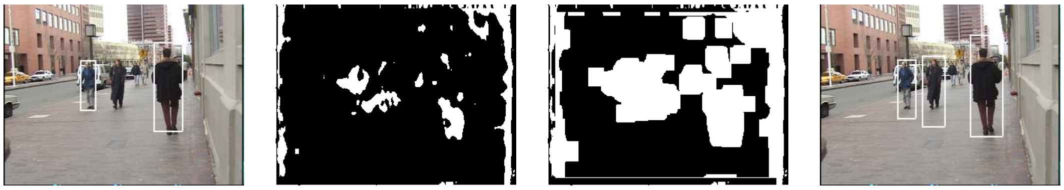


Figure 12: The sequence of steps in the motion-based module showing, from left to right, static detection results, motion discontinuities, full motion regions, and improved detection results.

noisy images using simulated annealing. In *Proceedings of the Fifth International Conference on Computer Vision*, pages 523–20, 1995.

[2] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optim margin classifier. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–52. ACM, 1992.

[3] B. Heisele, U. Kressel, and W. Ritter. Tracking non-rigid, moving objects based on color cluster flow. In *CVPR '97*, 1997. to appear.

[4] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.

[5] M. Leung and Y.-H. Yang. Human body motion segmentation in a complex scene. *Pattern Recognition*, 20(1):55–64, 1987.

[6] M. Leung and Y.-H. Yang. A region based approach for human body analysis. *Pattern Recognition*, 20(3):321–39, 1987.

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

[8] S. McKenna and S. Gong. Non-intrusive person authentication for access control by visual tracking and face recognition. In J. Bigun, G. Chollet, and G. Borgefors, editors, *Audio- and Video-based Biometric Person Authentication*, pages 177–183. IAPR, Springer, 1997.

[9] B. Moghaddam and A. Pentland. Probabilistic visual learning for object detection. Technical Report 326, Media Laboratory, Massachusetts Institute of Technology, 1995.

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

[11] E. Osuna, R. Freund, and F. Girosi. Support vector machines: Training and applications. A.I. Memo 1602, MIT A. I. Lab., 1997.

[12] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: An application to face detection. In *Computer Vision and Pattern Recognition*, pages 130–36, 1997.

[13] K. Rohr. Incremental recognition of pedestrians from image sequences. *Computer Vision and Pattern Recognition*, pages 8–13, 1993.

[14] H. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. Technical Report CMU-CS-95-158, School of Computer Science, Carnegie Mellon University, July/November 1995.

[15] K.-K. Sung. *Learning and Example Selection for Object and Pattern Detection.* PhD thesis, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, December 1995.

[16] K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. A.I. Memo 1521, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, December 1994.

[17] T. Tsukiyama and Y. Shirai. Detection of the movements of persons from a sparse sequence of tv images. *Pattern Recognition*, 18(3/4):207–13, 1985.

[18] R. Vaillant, C. Monrocq, and Y. L. Cun. Original approach for the localisation of objects in images. *IEE Proc.-Vis. Image Signal Processing*, 141(4), August 1994.

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

[20] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. Technical Report 353, Media Laboratory, Massachusetts Institute of Technology, 1995.

[21] A. Yuille, P. Hallinan, and D. Cohen. Feature Extraction from Faces using Deformable Templates. *International Journal of Computer Vision*, 8(2):99–111, 1992.