Example-Based Object Detection in Images by Components

Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio, *Member, IEEE*

Abstract—In this paper, we present a general example-based framework for detecting objects in static images by components. The technique is demonstrated by developing a system that locates people in cluttered scenes. The system is structured with four distinct example-based detectors that are trained to separately find the four components of the human body: the head, legs, left arm, and right arm. After ensuring that these components are present in the proper geometric configuration, a second example-based classifier combines the results of the component detectors to classify a pattern as either a "person" or a "nonperson." We call this type of hierarchical architecture, in which learning occurs at multiple stages, an Adaptive Combination of Classifiers (ACC). We present results that show that this system performs significantly better than a similar full-body person detector. This suggests that the improvement in performance is due to the component-based approach and the ACC data classification architecture. The algorithm is also more robust than the full-body person detection method in that it is capable of locating partially occluded views of people and people whose body parts have little contrast with the background.

Index Terms—Object detection, people detection, pattern recognition, machine learning, components.

1 Introduction

In this paper, we present a general example-based algorithm for detecting objects in images by first locating their constituent components and then combining the component detections with a classifier if their configuration is valid. We illustrate the method by applying it to the problem of locating people in complex and cluttered scenes. Since this technique is example-based, it can easily be used to locate any object composed of distinct identifiable parts that are arranged in a well-defined configuration, such as cars and faces.

The general problem of object detection in static images is a difficult one as the object detection system is required to distinguish a particular class of objects from all others. This calls for the system to possess a model of the object class that has high interclass and low intraclass variability. Further, a robust object detection system should be able to detect objects in uneven illumination, objects which are rotated into the plane of the image, and objects that are partially occluded or whose parts blend in with the background. Under all of the above conditions, the outline of an object is usually altered and its complete form may not be discernible. However, in many cases, the majority of the object's defining parts may still be identifiable. If an object detection system is designed to find objects in images by locating the various parts of the object, then it should be able to deal with such anomalies.

In this paper, we focus on the problem of detecting people in images; such a system could be used in surveillance systems, driver assistance systems, and image indexing. Detecting people in images is more challenging than detecting many other objects due to several reasons: First, people are articulate objects that can take on a variety of shapes and it is nontrivial to define a single model that captures all of these possibilities. The ability to detect people when the limbs are in different relative positions is a desirable trait of a robust person detection system. Second, people dress in a variety of colors and garment types (skirts, slacks, etc.), which leads to high intraclass variation in the people class, that would make it difficult for color or fine scale edge-based techniques to work well. The pictures of people in Fig. 1 illustrate the issues outlined above.

1.1 Previous Work

The approach we adopt builds on previous work in the fields of object detection and classifier combination algorithms. This section reviews relevant results in these fields.

1.1.1 Object Detection

The object detection systems that have been developed to date fall into one of three major categories. The first category consists of systems that are model-based, i.e., a model is defined for the object of interest and the system attempts to match this model to different parts of the image in order to find a fit [27]. The second type are image invariance methods which base a matching on a set of image pattern relationships (e.g., brightness levels) that, ideally, uniquely determine the objects being searched for [21]. The final set of object detection systems are characterized by their example-based learning algorithms [24], [22], [23], [18], [19], [16], [14]. These systems learn the salient

- *A. Mohan and C. Papageorgiou are with Kana Communications, 740 Bay Road, Redwood City, CA 94063. E-mail: amohan@alum.mit.edu, cpapa@ai.mit.com.*
- *T. Poggio is with the Brain Sciences Department and Artifical Intelligence Lab, Massachusetts Institute of Technology, 45 Carleton Street, E25-201, Cambridge, MA 02142. E-mail: tp@ai.mit.edu.*

Manuscript received 18 Aug. 1999; revised 31 Oct. 2000; accepted 5 Dec. 2000.
Recommended for acceptance by S.J. Dickinson.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 110448.

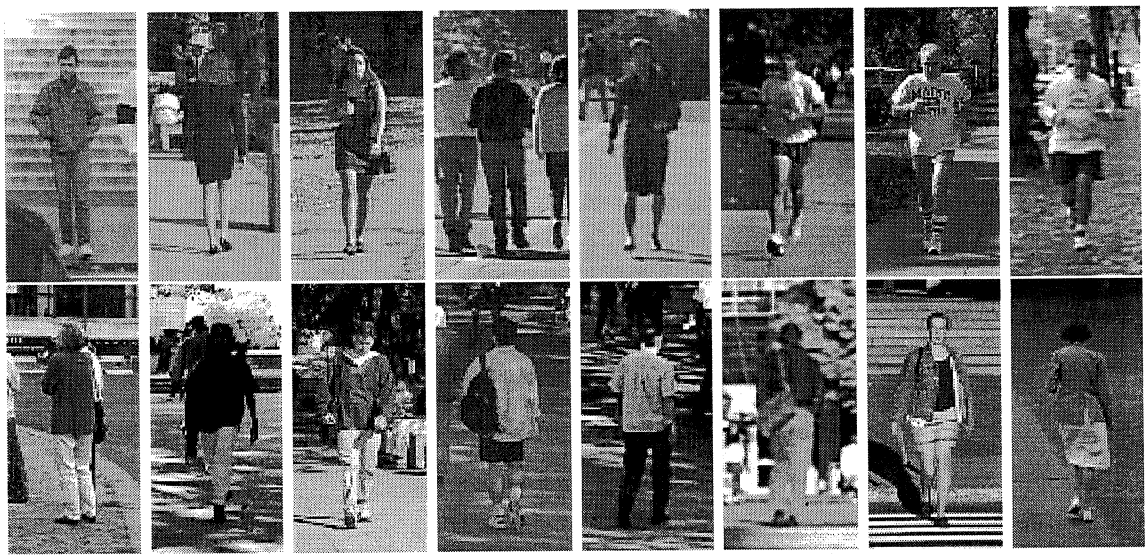


Fig. 1. These images demonstrate some of the challenges involved with detecting people in still images with cluttered backgrounds. People are nonrigid objects and dress in a wide variety of colors and garment types. Additionally, people may be rotated in depth, partially occluded, or in motion (i.e., running or walking).

features of a class from sets of labeled positive and negative examples. Example-based techniques have also been successfully used in other areas of computer vision, including object recognition [13].

People Detection in Images. Most people detection systems reported on in the literature either use motion information, explicit models, a static camera, assume a single person in the image, or implement tracking rather than pure detection; relevant work includes [8], [10], [7].

Papageorgiou et al. have successfully employed example-based learning techniques to detect people in complex static scenes without assuming any a priori scene structure or using any motion information. Their system detects the full body of a person. Haar wavelets [12] are used to represent the images and Support Vector Machine (SVM) classifiers [25] are used to classify the patterns. Details are presented in [16], [15], and [14].

Papageorgiou's system has reported successful results detecting frontal, rear, and side views of people, indicating that the wavelet-based image representation scheme and the SVM classifier are well-suited to this particular application. However, the system's ability to detect partially occluded people or people whose body parts have little contrast with the background is limited.

Component-Based Object Detection Systems. Previous research suggest that some of these problems associated with Papageorgiou's full-body detection system may be addressed by taking a component-based approach to detecting objects. A component-based object detection system is one that searches for an object by looking for its identifying components rather than the whole object. An example of such a system is a face detection system that finds a face when it locates a pair of eyes, a nose, and a mouth in the proper configuration.

Component-based approaches to object detection have been described in the past but their application to the problem of locating people in images is fairly limited. For component-based face detection systems see [20], [11], and [26]. Systems in [11] and [26] have the ability to explicitly deal with partial occlusions. These systems have two common features: They all have *component detectors* that identify candidate components in an image and they all have a means to integrate these components and determine if together they define a face In [4] and [5], the authors describe a system that uses color, texture, and geometry to localize horses and naked people in images. The system can be used to retrieve images satisfying certain criteria from image databases but is mainly targeted towards images containing one object. Methods of learning these "body plans" from examples are described in [4].

It is worth mentioning that a component-based object detection system for people is harder to realize than one for faces because the geometry of the human body is less constrained than that of the human face. This means that not only is there greater intraclass variation concerning the configuration of body parts, but also that it is more difficult to detect body parts in the first place since their appearance can change significantly when a person moves.

1.1.2 Classifier Combination Algorithms

Recently, a great deal of interest has been shown in hierarchical classification structures, i.e., data classification devices that are a combination of several other classifiers. In particular, two methods have received considerable attention—*bagging* and *boosting*. Both of these algorithms have been shown to increase the performance of certain classifiers for a variety of data sets [2], [6], [17], [1]. Despite the well-documented practical success of these algorithms, the reasons why they work so well is still open to debate.

1.2 Component-Based People Detection—Our Approach

The approach we take to detecting people in static images borrows ideas from the fields of object detection in images and data classification. In particular, the system detects the components of a person's body in an image, i.e., the head, the left and right arms, and the legs, instead of the full body. The

system then checks to ensure that the detected components are in the proper geometric configuration and then combines them using a classifier. This approach of integrating components using a classifier promises to increase accuracy based on the results of previous work in the field.

We introduce a new hierarchical classification architecture where example-based learning is conducted at multiple levels, called an Adaptive Combination of Classifiers (ACC). Specifically, it is composed of *distinct* example-based *component classifiers* trained to detect different object parts, i.e., heads, legs, and left and right arms, at one level and a similar example-based *combination classifier* at the next. The combination classifier takes the output of the component classifiers as its input and classifies the entire pattern under examination as either a "person" or a "nonperson." It bears repeating that since the classifiers are example-based, this system can easily be modified to detect objects other than people.

A component-based approach to detecting people is appealing and has the following advantages over existing techniques:

- It allows for the use of the geometric information concerning the human body to supplement the visual information present in an image and thereby improve the overall performance of the system. More specifically, the visual data in an image is used to detect body components and knowledge of the structure of the human body allows us to determine if the detected components are proportioned correctly and arranged in a permissible configuration. In contrast, a full-body person detector relies solely on visual information and does not take full advantage of the known geometric properties of the human body. In particular, it employs an implicit and fixed representation of the human form and does not explicitly allow for variations in limb positions [16], [15], [14].
- Sometimes it is difficult to detect the human body pattern as a whole due to variations in lighting and orientation. The effect of uneven illumination and varying viewpoint on body components (like the head, arms, and legs) is less pronounced and, hence, they are comparatively easier to identify.
- The component-based framework directly addresses the issue of detecting people that are partially occluded or whose body parts have little contrast with the background. This is accomplished by designing the system, using an appropriate classifier combination algorithm, so that it detects people even if all of their components are not detected.
- The structure of the component-based solution allows for the convenient use of hierarchical classification machines to classify patterns which have been shown to perform better than similar single layer devices for certain data classification tasks [2], [6], [17], [1].

The rest of the paper is organized as follows: Section 2 describes the system in detail. Section 3 reports on the performance of our system. In Section 4, we present conclusions along with suggestions for future research in this area.

2 System Details

2.1 Overview of System Architecture

The section explains the overall architecture and operation of the system by tracing the detection process when the system is applied to an image; Fig. 2 is a graphical representation of this procedure.

The system starts detecting people in images by selecting a 128×64 pixel window from the top left corner of the image as an input. This input is then classified as either a "person" or a "nonperson," a process which begins by determining where and at which scales the components of a person, i.e., the head, legs, left arm, and right arm, may be found within the window. All of these candidate regions are processed by the respective component detectors to find the strongest candidate components.

The component detectors process the candidate regions by applying the Haar wavelet transform to them and then classifying the resultant data vector. The component classifiers are quadratic Support Vector Machines (SVM) which are trained prior to use in the detection process (see Section 2.2). The strongest candidate component is the one that produces the highest positive raw output, referred to in this paper as the *component score*, when classified by the component classifiers. If the highest component score for a particular component is negative, i.e., the component detector in question did not find a component in the geometrically permissible area, then a component score of zero is used instead. The raw output of an SVM is a rough measure of how well a classified data point fits in with its designated class and is defined in Section 2.2.1. The highest component score for each component is fed into the combination classifier which is a linear SVM. The combination classifier processes the scores to determine if the pattern is a person.

This process of classifying patterns is repeated at all locations in an image by shifting the 128×64 pixel window across and down the image. The image itself is processed at several sizes, ranging from 0.2 to 1.5 times its original size. This allows the system to detect various sizes of people at any location in an image.

2.2 Details of System Architecture

2.2.1 First Stage—Identifying Components of People in an Image

When a 128×64 pixel window is evaluated by the system, the individual component detectors are applied only to specific areas of the window and only at particular scales, since the relative proportions must match and the approximate configuration of body parts is known a priori. This is necessary because even though a component detection is the strongest in a particular window under examination (it has the highest component score), it does not imply that it is in the correct position, as illustrated in Fig. 3. The centroid and boundary of the allowable rectangular area for a component detection (relative to the upper left-hand corner of the 128×64 pattern) determine the location of the component and the width of the rectangle is a measure of a component's scale.

We calculated the geometric constraints for each component from a sample of the training images, tabulated in

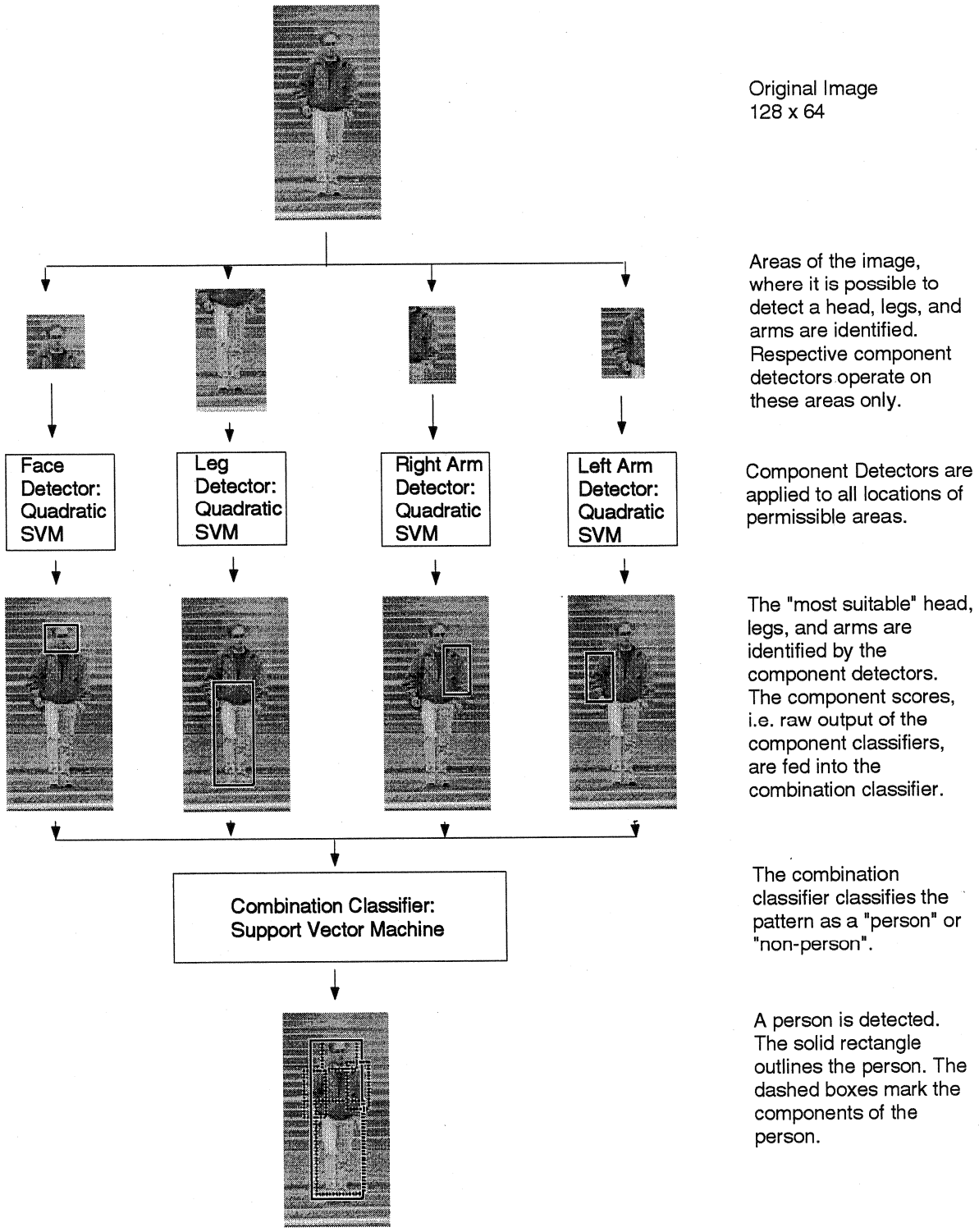


Fig. 2. Diagrammatic description of the operation of the system.

Table 1 and shown in Fig. 4, by taking the means of the centroid and top and bottom boundary edges of each component over positive detections in the training set. The tolerances were set to include all positive detections in the training set. Permissible scales were also estimated from the training images. There are two sets of constraints for the arms, one intended for extended arms and the other for bent arms.

Wavelet functions are used to represent the components in the images. Wavelets are a type of multiresolution function approximation that allow for the hierarchical decomposition of a signal [12]. When applied at different scales, wavelets encode information about an image from the coarse approximation all the way down to the fine details. The Haar basis is the simplest wavelet basis and provides a mathematically sound extension to an image

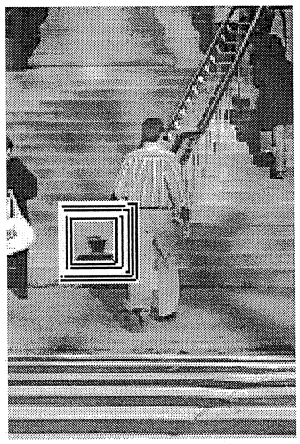


Fig. 3. It is very important to place geometric constraints on the location and scale of component detections. Even though a detection may be the strongest in a particular window examined, it might not be at the proper location. In this figure, the shadow of the person's head is detected with a higher score than the head itself. If we did not check for proper configuration and scale, component detections like these would lead to false alarms and/or missed detections of people.

invariance scheme [21]. Haar wavelets of two different scales (16 × 16 pixels and 8 × 8 pixels) are used to generate a multiscale representation of the images. The wavelets are applied to the image such that they overlap 75 percent with the neighboring wavelets in the vertical and horizontal directions; this is done to increase the spatial resolution of our system and to yield richer representation. At each scale, three different orientations of Haar wavelets are used, each of which responds to differences in intensities across different axes. In this manner, information about how intensity varies in each color channel (red, green, and blue) in the horizontal, vertical, and diagonal directions is obtained. The information streams from the three color channels are combined and collapsed into one by taking the wavelet coefficient for the color channel that exhibits the greatest variation in intensity at each location and for each orientation. At these scales of wavelets there are 582 features for the 32 × 32 pixel window for the head and shoulders and 954 features for the 48 × 32 pixel windows representing the lower body and the left and right arms. This method results in a thorough and compact representation of the components, with high interclass and low intraclass variation.

We use support vector machines (SVM) to classify the data vectors resulting from the Haar wavelet representation of the components. SVMs were proposed by Vapnik [25] and have yielded excellent results in various data classification tasks, including people detection [16], [14] and text classification [9]. Traditional training techniques for classifiers like multilayer perceptrons use empirical risk minimization and lack a solid mathematical justification. The SVM algorithm uses structural risk minimization to find the hyperplane that optimally separates two classes of objects. This is equivalent to minimizing a bound on generalization error. The optimal hyperplane is computed as a decision surface of the form:

$$f(\mathbf{x}) = sgn(g(\mathbf{x})), \tag{1}$$

where

$$g(\mathbf{x}) = \left(\sum_{i=1}^{l^*} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i^*) + b\right). \tag{2}$$

In (2), K is one of many possible kernel functions, $y_i \in \{-1, 1\}$ is the class label of the data point $\mathbf{x}_i^*$, and $\{\mathbf{x}_i^*\}_{i=1}^{l^*}$ is a subset of the training data set. The $\mathbf{x}_i^*$ are called *support vectors* and are the points from the data set that define the separating hyperplane. Finally, the coefficients α_i and b are determined by solving a large-scale quadratic programming problem. One of the appealing characteristics of SVMs is that there are just two tunable parameters, *Cpos* and *Cneg*, which are penalty terms for positive and negative pattern misclassifications, respectively. The kernel function K that is used in the component classifiers is a quadratic polynomial and is $K(\mathbf{x}, \mathbf{x}_i^*) = (\mathbf{x} \cdot \mathbf{x}_i^* + 1)^2$.

In (1), $f(\mathbf{x}) \in \{-1, 1\}$ is referred to as the *binary class* of the data point $\mathbf{x}$ which is being classified by the SVM. As (1) shows, the binary class of a data point is the sign of the *raw output* $g(\mathbf{x})$ of the SVM classifier. The raw output of an SVM classifier is the distance of a data point from the decision hyperplane. In general, the greater the magnitude of the raw output, the more likely a classified data point belongs to the binary class it is grouped into by the SVM classifier.

TABLE 1
Geometric Constraints Placed on Each Component

Component	*Centroid*		*Scale*		*Other Criteria*
	Row	*Column*	*Minimum*	*Maximum*	
Head and Shoulders	23 ± 3	32 ± 2	28×28	42×42	
Lower Body		32 ± 3	42×28	69×46	*Bottom Edge:* Row: 124 ± 4
Right Arm Extended	54 ± 5	46 ± 3	31×25	47×31	
Right Arm Bent		46 ± 3	31×25	47×31	*Top Edge:* Row: 31 ± 3
Left Arm Extended	54 ± 5	17 ± 3	31×25	47×31	
Left Arm Bent		17 ± 3	31×25	47×31	*Top Edge:* Row: 31 ± 3

All coordinates are relative to the upper left-hand corner of a 128×64 *rectangle.*

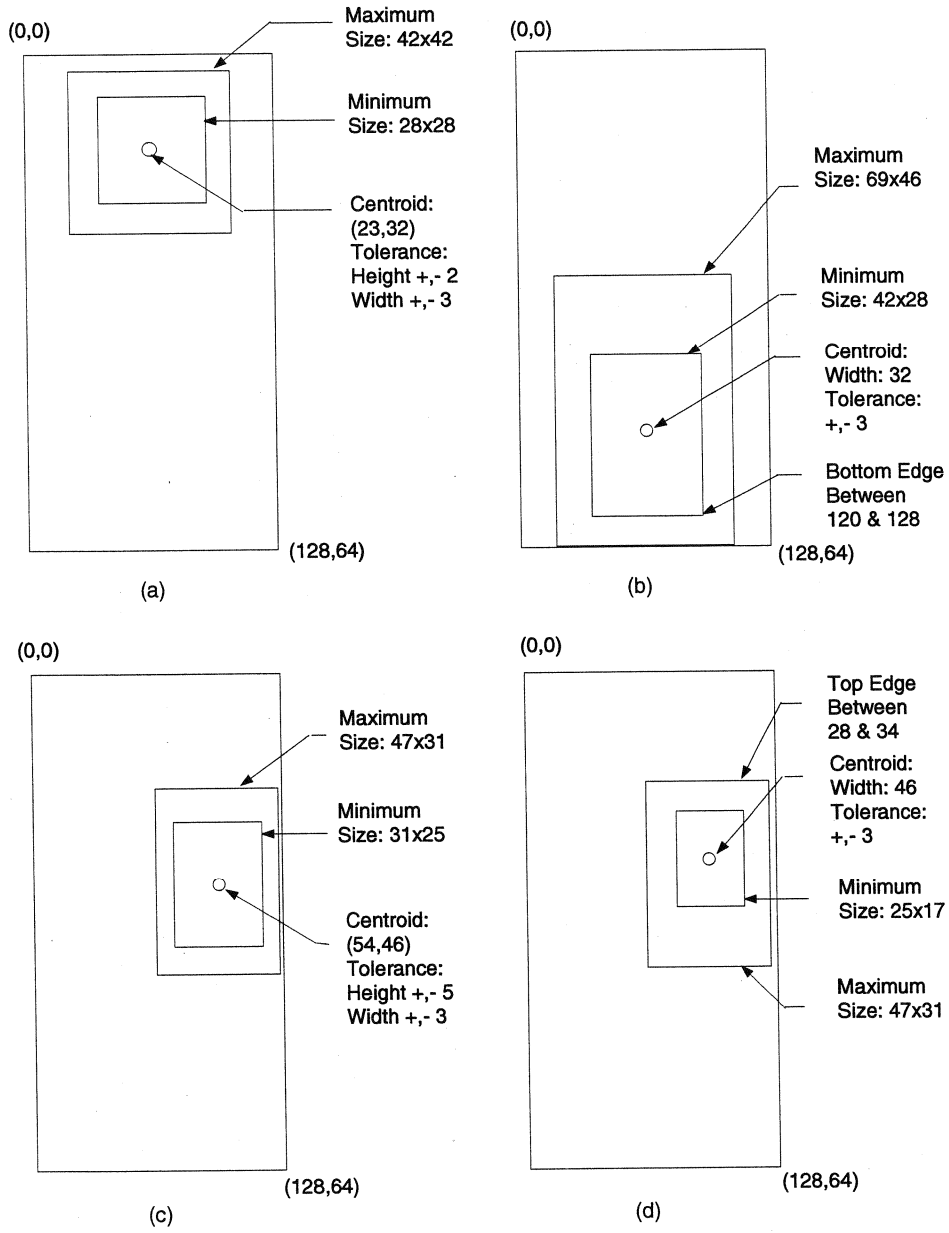


Fig. 4. Geometric constraints that are placed on the different components. All coordinates are relative to the upper left-hand corner of a 128 × 64 rectangle. (a) Illustrates the geometric constraints on the head, (b) the lower body, (c) an extended right arm, and (d) a bent right arm.

The component classifiers are trained on positive images and negative images for their respective classes. The positive examples are of arms, legs, and heads of people in various environments, both indoors and outdoors and under various lighting conditions. The negative examples are taken from scenes that do not contain any people. Examples of positive images used to train the component classifiers are shown in Fig. 5.

2.2.2 *Second Stage—Combining the Component Classifiers*

Once the component detectors have been applied to all geometrically permissible areas within the 128 × 64 pixel window, the highest component score for each component type is entered into a data vector that serves as the input to the combination classifier. The component score is the raw output of the component classifier and is the distance of the test point from the decision hyperplane, a rough measure of how "well" a test point fits into its designated class. If the component detector does not find a component in the designated area of the 128 × 64 pixel window, then zero is placed in the data vector.

The combination classifier is a linear SVM classifier. The kernel K that is used in the SVM classifier and shown in (2) has the form $K(\mathbf{x}, \mathbf{x}_i^*) = (\mathbf{x} \cdot \mathbf{x}_i^* + 1)$. This type of hierarchical classification architecture where learning occurs at multiple stages is termed an Adaptive Combination of Classifiers (ACC). Positive examples were generated by processing 128 × 64 pixel images of people at one scale and taking the highest component score (from detections that are geometrically allowed) for each component type.

Fig. 5. The top row shows examples of "heads and shoulders" and "lower bodies" of people that were used to train the respective component detectors. Similarly, the bottom row shows examples of "left arms" and "right arms" that were used for training purposes.

3 Results

We compare the performance of our component-based person detection system to that of other component-based person detection systems that combine the component classifiers in different ways and the full-body person detection system that is described in [16] and [14] and reviewed in Section 1.1.1.

3.1 Experimental Setup

All of the component-based detection systems that were tested in this experiment are two tiered systems. Specifically, they detect heads, legs, and arms at one level and at the next they combine the results of the component detectors to determine if the pattern in question is a person or not. The component detectors that were used in all of the component-based people detection systems are identical and are described in Section 2.2.1. The positive examples for training these detectors were obtained from a database of pictures of people taken in Boston and Cambridge, Massachusetts, with different cameras, under different lighting conditions, and in different seasons. This database includes images of people who are rotated in depth and who are walking, in addition to frontal and rear views of stationary people. The positive examples of the lower body include images of women in skirts and people wearing full length overcoats as well as people dressed in pants. Similarly, the database of positive examples for the arms were varied in content, including arms at various positions in relation to the body. The negative examples were obtained from images of natural scenery and buildings that did not contain any people. The head and shoulders classifier was trained with 856 positive and 9,315 negative examples, the lower body with 866 positive and 9,260 negative examples, the left arm with 835 positive and 9,260 negative examples, and the right arm with 838 positive and 9,260 negative examples.

3.1.1 Adaptive Combination of Classifiers-Based Systems

Once the component classifiers were trained, the next step in evaluating the Adaptive Combination of Classifiers (ACC)-based systems was to train the combination classifier. Positive and negative examples for the combination classifier were collected from the same databases that were used to train the component classifiers. A positive example was obtained by processing each image of a person at a single appropriate scale. The four component detectors were applied to the geometrically permissible areas of the image at the allowable scales. These geometrically permissible areas were determined by analyzing a sample of the training set images, as described in Section 2.2.1. There is no overlap between these images and the testing set used in this experiment. The greatest positive classifier output for each component was recorded. When all four component scores were greater than zero, they were assembled as a vector to form an example. If all of the component scores were not positive, then no vector was formed and the window examined did not yield an example. The negative examples were computed in a similar manner, except that this process was repeated over the entire image and at various scales. The images for the negative examples did not contain people.

We used 889 positive examples and 3,106 negative examples for training the classifiers. First, second, third, and fourth degree polynomial SVM classifiers were trained using the same training set and, subsequently, tested over identical out-of-sample test data.

The trained system was run over a database containing 123 images of people to determine the positive detection rate. There is no overlap between these images and the ones that were used to train the system. The out-of-sample false alarm rate was obtained by running the system over a database of 50 images that do not contain any people. By running the system over these 50 images, 796,904 windows were examined and classified. The system was run over the databases of test images at several different thresholds. The results were recorded and plotted as Receiver Operating Characteristic (ROC) curves.

3.1.2 Voting Combination of Classifiers-Based System

The other method of combining the results of the component detectors that was tested is what we call a Voting Combination of Classifiers (VCC). VCC systems combine classifiers by implementing a voting structure amongst them. One way of viewing this arrangement is that the component classifiers are weak experts in the matter of detecting people. VCC systems poll the weak experts and then based on the results, decide if the pattern is a person. For example, in a possible implementation of VCC, if a majority of the weak experts classify a pattern as a person, then the system declares the pattern to be a person.

In the incarnation of VCC that is implemented and tested in this experiment, a positive detection of the person class results only when all four component classes are detected in the proper configuration. The geometric constraints placed on the components are the same in the ACC- and VCC-based

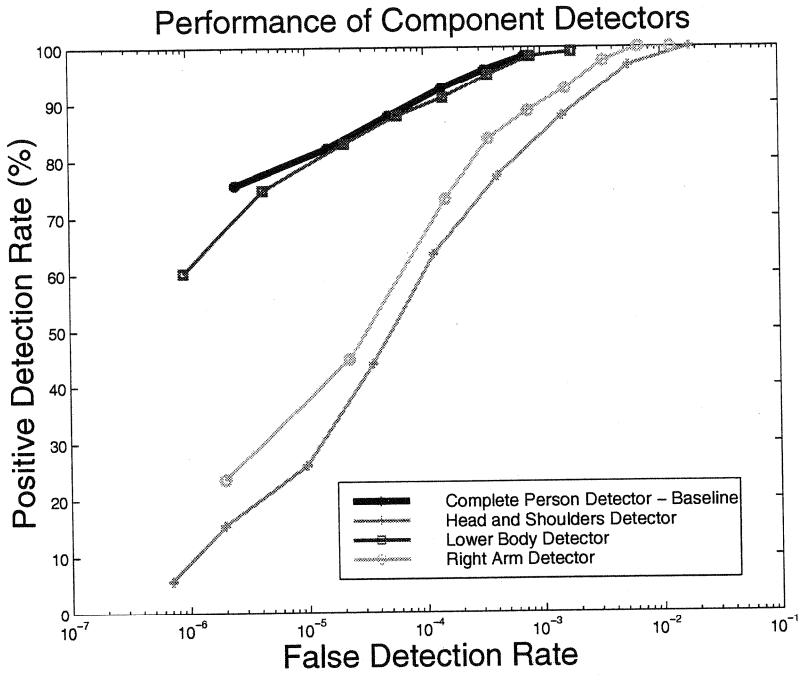


Fig. 6. ROC curves illustrating the ability of the component detectors to correctly indentify a person in an image. The positive detection rate is plotted as a percentage against the false alarm rate which is measured on a logarithmic scale. The false alarm rate is the number of false positive detections per window inspected.

systems. For each pattern that the system classifies, the system must evaluate the logic presented below:

$$Person = Head \ \& \ Legs \ \& \ Left\ arm \ \& \ Right\ arm, \quad (3)$$

where a state of *true* indicates that a pattern belonging to the class in question has been detected.

The detection threshold of the VCC-based system is determined by selecting appropriate thresholds for the component detectors. The thresholds for the component detectors are chosen such that they all correspond to approximately the same positive detection rate, estimated from the ROC curves of each of the component detectors shown in Fig. 6. These ROC curves were calculated in a manner similar to the procedure described earlier in Section 3.1.1. A point of interest is that these ROC curves indicate how discriminating the individual components of a person are in detecting the full body. The legs perform the best, followed by the arms and the head. The superior performance of the legs may be due to the fact that the background of the lower body in images is usually either the street, pavement, or grass and, hence, is relatively clutter free compared to the background of the head and arms.

3.1.3 Baseline System

The system that is used as the "baseline" for this comparison is a full-body person detector. Details of this system, which was created by Papageorgiou et al. are presented in [16], [14], and [15]. It has the same architecture as the individual component detectors used in our system, described in Section 2.2.1, but is trained to detect full-body patterns and not separate components. The quadratic SVM classifier was trained on 869 positive and 9,225 negative examples.

3.2 Experimental Results

We compare the ACC-based system, the VCC-based system, and the full-body detection system. The ROC curves of the person detection systems are shown in Fig. 7 and explicitly capture the tradeoff between accuracy and false detections that is inherent to every detector. An analysis of the ROC curves suggest that a component-based person detection system performs very well and significantly better than the baseline system at all thresholds. It should be emphasized that the baseline system uses the same image representation scheme (Haar wavelets) and classifier (SVM) that the component detectors used in the component-based systems. Thus, the improvement in performance is due to the component-based approach and the algorithm used for combining the component classifiers.

For the component-based systems, the ACC approach produces better results than VCC. In particular, the ACC-based system that uses a linear SVM to combine the component classifier is the most accurate. This is related to the fact that higher degree polynomial classifiers require more training examples in proportion with the higher dimensionality of the feature space to perform at the same level as the linear SVM. During the course of the experiment, the linear SVM-based system displayed a superior ability to detect people even when one of the components was not detected, in comparison to the higher degree polynomial SVM-based systems. A possible explanation for this observation may be that the higher degree polynomial classifiers place a stronger emphasis on the presence of combinations of components, due to the structure of their kernels. The second, third, and fourth degree polynomial kernels include terms that are products of up to two, three, and four elements (which are component scores).

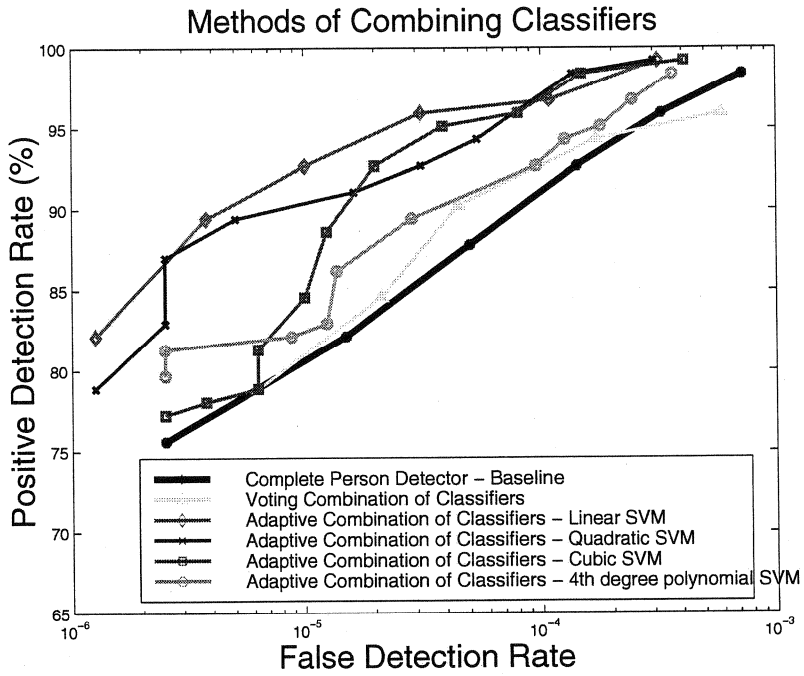


Fig. 7. ROC curves comparing the performance of various component-based people detection systems using different methods of combining the classifiers that detect the individual components of a person's body. The positive detection rate is plotted as a percentage against the false alarm rate which is measured on a logarithmic scale. The false alarm rate is the number of false positives detections per window inspected. The curves indicate that the system in which a linear SVM combines the results of the component classifiers performs best. The baseline system is a full-body person detector similar to the component detectors used in the component-based system.

It is also worth mentioning that the database of test images that were used to generate the ROC curves did not just include frontal views of people, but also contained a variety of challenging images. Included are pictures of people walking and running, occluded people, people where portions of their body has little contrast with the background, and slight rotations in depth. Fig. 8 is a selection of these images.

Fig. 9 shows the results obtained when the system was applied to images of people who are partially occluded or whose body parts blend in with the background. In these examples, the system detects the person while running at a threshold that, according to the ROC curve shown in Fig. 7, corresponds to a false detection rate of less than one false alarm for every 796,904 patterns inspected. Fig. 10 shows

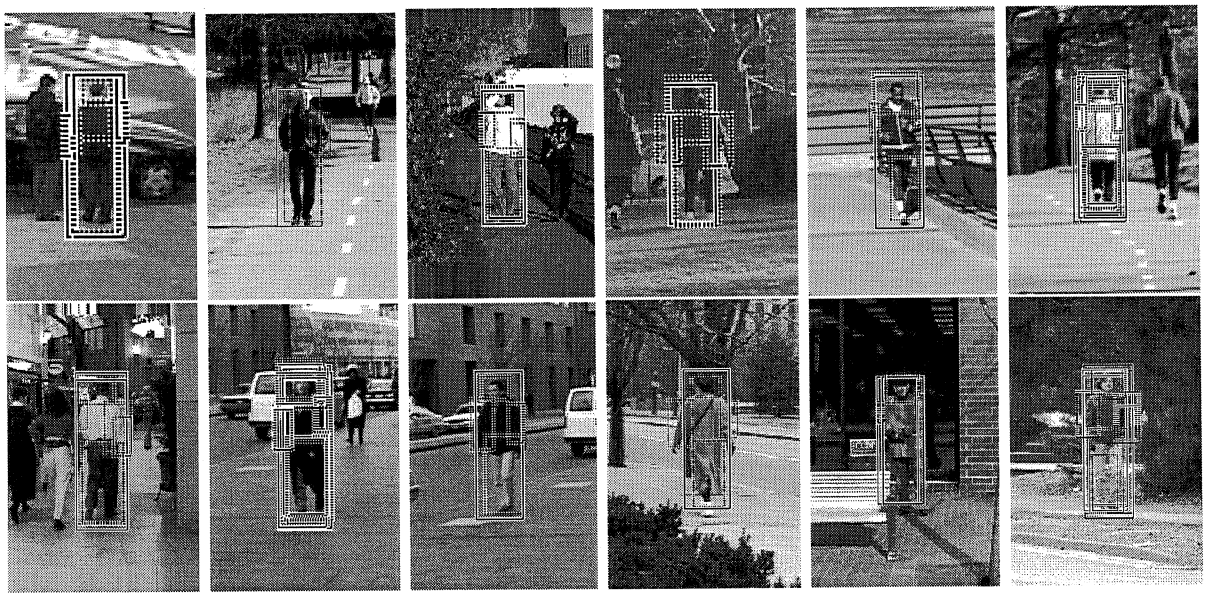


Fig. 8. Samples from the test image database. These images demonstrate the capability of the system. It can detect running people, people who are slightly rotated, people whose body parts blend into the background (bottom row, second from right—the person is detected even though the legs are not), and people under varying lighting conditions (top row, second from left—one side of the face is light and the other dark).

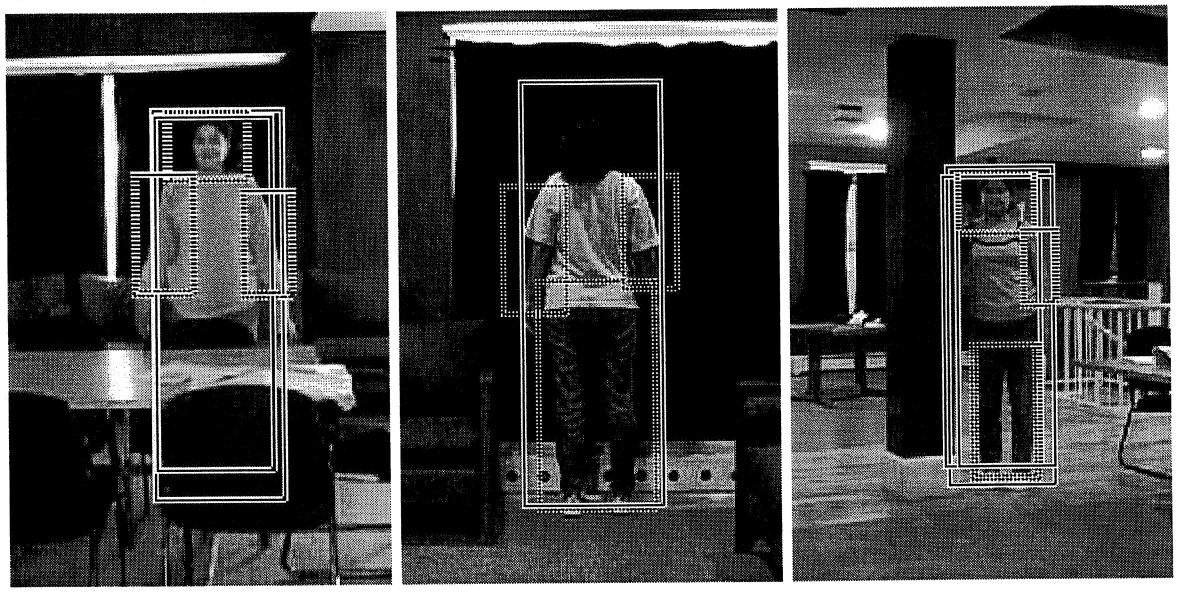


Fig. 9. Results of the system's application to images of partially occluded people and people whose body parts have little contrast with the background. In the first image, the person's legs are not visible, in the second image, her hair blends in with the curtain in the background, and in the last image, her right arm is hidden behind the column.

the result of applying the system to sample images with clutter in the background.

3.3 Extension of the System

In the component-based object detection system presented in this paper, the constraints that are placed on the size and relative location of the components of an object are determined manually. As explained in Section 2.2.1, the constraints were calculated from the training examples. While this method produced excellent results, it is possible that it may suffer from a bias introduced by the designer. Therefore, it is desirable for the system to learn the geometric constraints to be placed on the components of an object from examples. This would make it easier to apply this system to other objects of interest. Also, such an object detection system would be an initial step toward a more sophisticated component-based object detection system in which the components of an object are not predefined.

We created a component-based object detection system that learns the relative location and size of an object's components from examples in order to explore the viability and performance of such a system. In the new system, the geometrically permissible areas are learned by SVM classifiers from training examples. Thus, instead of checking the candidate coordinates of a window against the constraints listed in Table 1, the coordinates are fed into an SVM classifier. The output of the each *geometric classifier* determines whether the window is permissible for the particular component. The coordinates that are fed into the geometric classifiers are the location of the top left corner and bottom right corner of the window, relative to the top left corner of the 128×64 pixel window, i.e., four dimensional feature vectors.

The kernel function K in (2) that is used in the geometric classifiers is a fourth degree polynomial and has the form $K(\mathbf{x}, \mathbf{x}_i^*) = (\mathbf{x} \cdot \mathbf{x}_i^* + 1)^4$. We trained the geometric classifiers for each component on 855 positive and 9,000 negative examples, from the same databases of images used to train the component classifiers.

This new system was tested on the same database as the system presented earlier. Fig. 11 compares the ROC curves for the two systems. Where the performance of the two system is very similar, the system that learns the geometry of an object performs better at higher thresholds. An added advantage of the system that learns the relative location and size of the components of an object is that one can change the size of the geometrically permissible area by varying the penalty parameters, *Cpos* and *Cneg*, for the misclassification of positive and negative examples during training [25], [3]. This results in different geometric classifiers and, hence, different geometrically permissible areas. ROC curves corresponding to different penalty terms are shown in Fig. 11.

4 Conclusions and Future Work

In this paper, we have presented a component-based person detection system for static images that is able to detect frontal, rear, slightly rotated (in depth) and partially occluded people in cluttered scenes without assuming any a priori knowledge concerning the image. The framework described here is applicable to other domains besides people, including faces and cars.

A component-based approach handles variations in lighting and noise in an image better than a full-body person detector and is able to detect partially occluded people and people who are rotated in depth, without any additional modifications to the system. A component-based detector looks for the constituent components of a person and if one of these components is not detected, due to an occlusion or because the person is rotated into the plane of the image, the system can still detect the person if the component detections are combined using an appropriate *hierarchical* classifier.

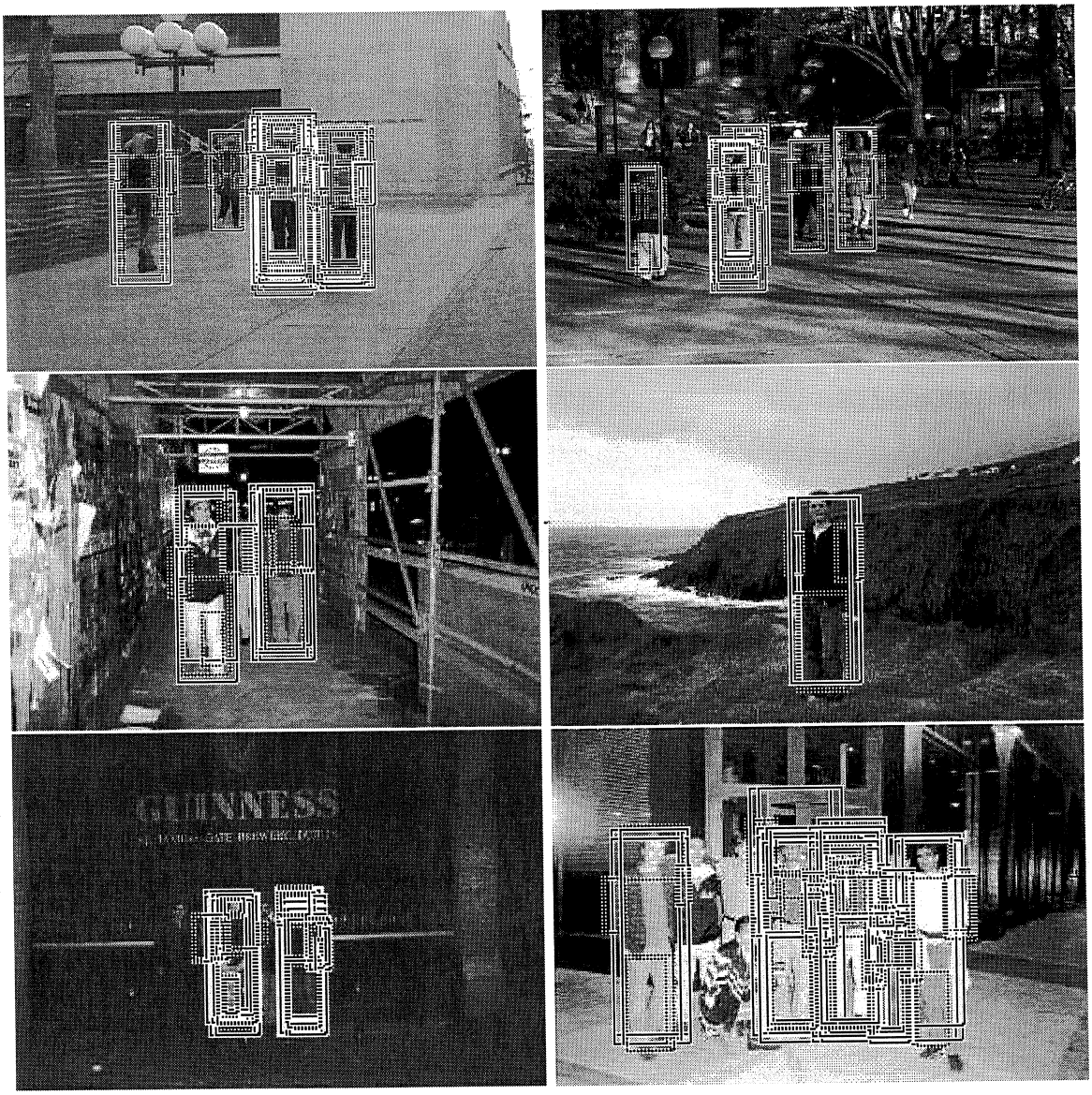


Fig. 10. Results from the component-based person detection system. The solid boxes outline the complete person and the dashed rectangles identify the individual components. People may be missed by the system because they are either too large or too small to be processed by the system (top right—person on the right) because several parts of their body may have very little contrast with the background (bottom left—person on the left) or because several parts of their body may be occluded (bottom right—person second from the left).

The hierarchical classifier that is implemented in this system uses four distinct component detectors at the first level, that are trained to find, independently, components of the "person" object, i.e., heads, legs, and left and right arms. These detectors use Haar wavelets to represent the images and Support Vector Machines (SVM) to classify the patterns. The four component detectors are combined at the next level by another SVM. We call this type of hierarchical classification architecture, in which learning occurs at more than two levels, an Adaptive Combination of Classifiers (ACC). It is worth mentioning that one may use classification devices other than SVM's in this system; a comparative study in this area to determine the performance of such implementations would be of interest.

The system is very accurate and performs significantly better than a full-body person detector designed along similar lines. This suggests that the improvement in performance is due to the component-based approach and the ACC classification architecture we employed. (Further work in this area to quantitatively determine how much of the improvement can be attributed to the component-based approach and how much is due to the ACC classification architecture would be useful.) The superior performance of the component-based approach can be attributed to the fact that it operates with more information about the object class than the full-body person detection method. Specifically, where both systems are trained on positive examples of the human body (or human body parts in the case of the component-based system), the component-based algorithm incorporates explicit knowledge about the geometric properties of the human body and explicitly allows for variations in the human form.

This paper presents a valuable first step but there are several directions in which this work could be extended. It

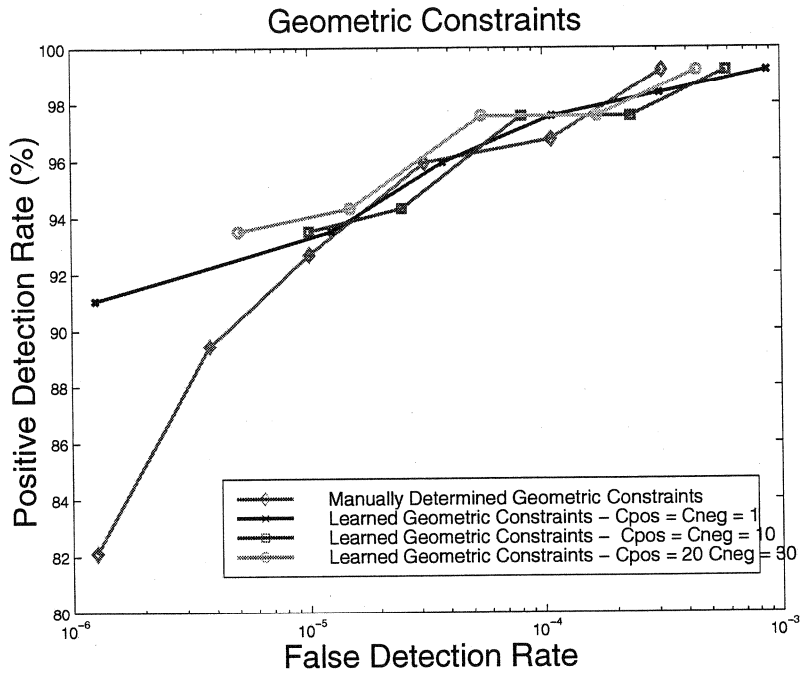


Fig. 11. ROC curves comparing the performance of various methods of defining and placing geometric constraints on components of objects. The core component-based object detection algorithm is the same for all the systems tested here—a linear SVM-based ACC system. The positive detection rate is plotted as a percentage against the false alarm rate which is measured on a logarithmic scale. The false alarm rate is the number of false positives detections per window inspected. The curves indicate that the systems that learn the geometric constraints perform slightly better than the one that uses manually determined values. The graphs also shows that changing the penalty parameters for misclassifications of the geometry (Cpos and Cneg) alters the overall system's performance.

would be useful to test the system described here in other domains, such as cars and faces. Since the component-based systems described in this paper were implemented as prototypes, we could not gauge the speeds of the various algorithms accurately. It would be interesting to learn how the different algorithms compare with each other in terms of speed. It would also be interesting to study how the performance of the system depends on the choice of the SVM kernels and the number of training examples. While this paper establishes that this system can detect people who are slightly rotated in depth, it does not determine, quantitatively, the extent of this capability; further work in this direction would be of interest. Along similar lines, it would be useful to investigate if the approach described in this paper could be extended to detect objects from an arbitrary viewpoint. In order to accomplish this, the system would have to have a richer understanding of the geometric properties of an object, that is to say, it would have to be capable of learning how the various components of an object change in appearance with a change in viewpoint and also how the change in viewpoint affects the geometric configuration of the components.

Acknowledgments

The research described in this paper was conducted within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. The research is sponsored by grants from the US Office of Naval Research under contract no. N00014-93-1-3085 and contract no. N00014-95-1-0600 and the US National Science Foundation under contract No. IIS-9800032 and contract No. DMS-9872936. Additional support is provided by: AT&T, Central Research Institute of Electric Power Industry, Eastman Kodak Company, Daimler-Chrysler, Digital Equipment Corporation, Honda R&D Co., Ltd., NEC Fund, Nippon Telegraph & Telephone, and Siemens Corporate Research, Inc.

References

[1] E. Bauer and R. Kohavi, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants," *Machine Learning*, 1998.

[2] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, pp. 123-140, 1996.

[3] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Proc. Data Mining and Knowledge Discovery*, U. Fayyad, ed., pp. 1-43, 1998

[4] D. Forsyth and M. Fleck, "Body Plans," *Computer Vision and Pattern Recognition*, pp. 678-683, 1997.

[5] D. Forsyth and M. Fleck, "Finding Naked People," *Int'l J. Computer Vision*, 1998. (pending publication.)

[6] Y. Freund and R. Schapire, "Experiments with a New Boosting Algorithm," *Machine Learning: Proc. 13th Nat'l Conf.*, 1996.

[7] I. Haritaoglu, D. Harwood, and L. Davis, "W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People," *Face and Gesture Recognition*, pp. 222-227, 1998.

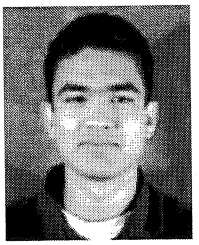
[8] D. Hogg, "Model-Based Vision: A Program to See a Walking Person," *Image and Vision Computing*, vol. 1, no. 1, pp. 5-20, 1983.

[9] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Proc. 10th European Conf. Machine Learning (ECML)*, 1998.

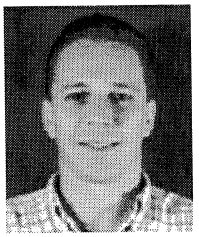
[10] M.K. Leung and Y.-H. Yang, "A Region Based Approach for Human Body Analysis," *Pattern Recognition*, vol. 20, no. 3, pp. 321-39, 1987.

[11] T. Leung, M. Burl, and P. Perona, "Finding Faces in Cluttered Scenes Using Random Labeled Graph Matching," *Proc. Fifth Int'l Conf. Computer Vision*, pp. 637-644, June 1995.

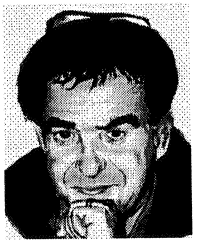
[12] S. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674-693, July 1989.
[13] H. Murase and S. Nayar, "Visual Learning and Recognition of 3D Objects from Appearance," *Int'l J. Computer Vision*, vol. 14, no. 1, pp. 5-24, 1995.
[14] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian Detection Using Wavelet Templates," *Proc. Computer Vision and Pattern Recognition*, pp. 193-199, June 1997.
[15] C. Papageorgiou, M. Oren, and T. Poggio, "A General Framework for Object Detection," *Proc. Int'l Conf. Computer Vision*, Jan. 1998.
[16] C. Papageorgiou and T. Poggio, "A Trainable System for Object Detection," *Int'l J. Computer Vision*, vol. 38, no. 1, pp. 15-33, 2000.
[17] J. Quinlan, "Bagging, Boosting, and C4.5," *Proc. 13th Nat'l Conf. Artificial Intelligence*, 1996.
[18] H. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23-38, Jan. 1998.
[19] H. Rowley, S. Baluja, and T. Kanade, "Rotation Invariant Neural Network-Based Face Detection," *Proc. Computer Vision and Pattern Recognition*, pp. 38-44, June 1998.
[20] L. Shams and J. Spoelstra, "Learning Gabor-Based Features for Face Detection," *Proc. World Congress in Neural Networks, Int'l Neural Network Soc.*, pp. 15-20, Sept. 1996.
[21] P. Sinha, "Object Recognition via Image Invariants: A Case Study," *Investigative Ophthalmology and Visual Science*, vol. 35, pp. 1735-1740, May 1994.
[22] K.-K. Sung and T. Poggio, "Example-Based Learning for View-Based Human Face Detection," *Proc. Image Understanding Workshop*, Nov. 1994.
[23] K.-K. Sung and T. Poggio, "Example-Based Learning for View Based Human Face Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 39-51, Jan. 1998.
[24] R. Vaillant, C. Monrocq, and Y. Le Cun, "Original Approach for the Localisation of Objects in Images," *IEE Proc. Vision Image Signal Processing*, vol. 141, no. 4, pp. 245-50, Aug. 1994.
[25] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer Verlag, 1995.
[26] K. Yow and R. Cipolla, "Feature-Based Human Face Detection," *Image and Vision Computing*, vol. 15, no. 9, pp. 713-35, Sept. 1997.
[27] A. Yuille, "Deformable Templates for Face Recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 59-70, 1991.



Anuj Mohan received the SB degree in electrical engineering and the MEng degree in electrical engineering and computer science from Massachusetts Institute of Technology in 1998 and 1999, respectively. His research interests center around engineering applications of machine learning and data classification algorithms. He is currently a software engineer at Kana Communications where he is working on applications of text classification and natural language processing algorithms.



Constantine Papageorgiou received the BS degree in mathematics/computer science from Carnegie Mellon University in 1992. After receiving his degree, he worked in the Speech and Language Processing Department at BBN until starting graduate school in 1995. He received the doctorate in electrical engineering and computer science from Massachusetts Institute of Technology in December 1999. His research focused on developing trainable systems for object detection. He has also done research in image compression, reconstruction, and superresolution, and financial time series analysis. Currently, he is working as a research scientist at Kana Communications where his focus is on natural language understanding and text classification.



Tomaso Poggio received the doctorate degree in theoretical physics from the University of Genoa in 1970. From 1971 to 1981, he held a tenured research position at the Max Planck Institute, after which he became a professor at Massachusetts Institute of Technology (MIT). Currently, he is the Uncas and Helen Whitaker Professor in the Department of Brain and Cognitive Sciences at MIT and a member of the Artificial Intelligence Laboratory. He is doing research in computational learning and vision at the MIT Center for Biological and Computational Learning, of which he is co-director. He has authored more than 200 papers in areas ranging from psychophysics and biophysics to information processing in man and machine, artificial intelligence, machine vision, and learning. His main research activity at present is learning from the perspective of statistical learning theory, engineering applications, and neuroscience. He has received a number of distinguished international awards in the scientific community, is on the editorial board of a number of interdisciplinary journals, a fellow of the American Association for Artificial Intelligence as well as the American Academy of Arts and Sciences, and an Honorary Associate of the Neuroscience Research Program at Rockefeller University. Dr. Poggio is a member of the IEEE.