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

A.I. Memo No. 1664 June, 1999
C.B.C.L Paper No. 178

Object Detection in Images by Components

Anuj Mohan

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1500-1999/AIM-1664.ps

Abstract

In this paper we present a component based person detection system that is capable of detecting frontal, rear and near side views of people, and partially occluded persons in cluttered scenes. The framework that is described here for people is easily applied to other objects as well.

The motivation for developing a component based approach is two fold: first, to enhance the performance of person detection systems on frontal and rear views of people and second, to develop a framework that directly addresses the problem of detecting people who are partially occluded or whose body parts blend in with the background.

The data classification is handled by several support vector machine classifiers arranged in two layers. This architecture is known as Adaptive Combination of Classifiers (ACC).

The system performs very well and is capable of detecting people even when all components of a person are not found. The performance of the system is significantly better than a full body person detector designed along similar lines. This suggests that the improved performance is due to the components based approach and the ACC data classification structure.

This report describes research done by the author to fulfill the thesis requirement for the Master of Engineering degree in Electrical Engineering and Computer Science. The work was performed within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and at the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology.

This research is sponsored by grants from the Office of Naval Research under Contract No. N00014-93-13085, the Office of Naval Research-MURI (Defense Advanced Research Projects Agency) under Contract No. N00014-95-1-0600, and the National Science Foundation under Contract Nos. IIS-9800032, No. SBR-960-1828 and No. DMS-9872936 Additional support is provided by Eastman Kodak Company, DaimlerChrysler, Siemens Corporate Research, Inc., AT&T, Digital Equipment Corporation, Honda R&D Co., Ltd., ATR, and Central Research Institute of Electric Power Industry.

1 Introduction

Object detection algorithms are very important because they form the backbone of a wide variety of image understanding applications. A great deal of research has been performed to advance the field and to enhance the capability and robustness of object detection systems. Our goal in this study is to further this work by developing and investigating the performance of a component based object detection system.

1.1 Problem Statement

The problem that we address in this paper is object detection in static digital images. In particular, we pay special attention to the more focussed problem of detecting people in still images.

This is an important issue to address because of the many applications of people detection systems. They can be used to search and label image databases. If this idea is extended to the internet domain, then one could create a visual search engine for the web around such a system. Person detection systems are being used in state of the art surveillance systems and their use in driver assistance systems in cars is being actively explored. These are but some of the uses of a person detection system, and it is clear that if a system that performed better than existing solutions was created then its effect would be felt in several fields.

Moreover, where we focus on a person detection system in this paper, the approach employed is easily applied to detect any other object comprised of distinct identifiable parts that are arranged in a well defined configuration such as cars and faces.

Detecting people in static digital images is a very difficult problem to address because of some characteristic properties of the object of interest, i.e. people. First of all, people are articulate bodies and thus, it is very difficult to define a single model that describes all people. Secondly, people dress in different colors and garment types (skirts, slacks, etc.), which increases the already high intra-class variation amongst people due to their non-rigid structure. Developing a tight model for the person class that captures its salient features and distinguishes it from all other objects becomes a very difficult task because the model has to account for the high intra-class variation. Also, where the high intra-class variability makes it difficult enough to separate people from all other classes, the problem is compounded by the fact that one cannot use color detectors to isolate regions where one is likely to find a person nor can one use a color based scheme to represent the image effectively. An edge detection scheme cannot be used to represent people either because of the variation in garment types. Such a scheme would collect too much spurious data. Lastly, images of people are rarely ever perfect uninterrupted frontal views. Rather, often times the person is partially occluded or some part of the person's body has very little contrast with the background making it hard to distinguish. Also, people can be walking or running in an image. To complicate the issue further, the image may capture a side-view or a slightly rotated (in depth) view of a person, which changes the outline of a person's body.

The pictures of people in Figure 1 illustrate some of the issues outlined above.

Figure 1: *These images demonstrate some of the challenges involved with detecting people in still images. To begin, people are non-rigid objects. They 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). To compound the problem, the background is usually cluttered.*

1.2 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.

1.2.1 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 (Yuille, 1991[25]). 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 (Sinha, 1994[20]). The final set of object detection systems are characterized by their example based learning algorithms (Papageorgiou and Poggio, in preparation [13]; Oren, Papageorgiou, Sinha, Osuna and Poggio, 1997 [9]; and Rowley, Baluja and Kanade, 1998[16]). The learning devices used by the systems listed above differ, but the central idea behind their architecture is similar. They all learn the salient features of a class from positive and negative examples without user intervention.

People Detection in Images

People detection in static images, as explained in the preceding section, is a challenging incarnation of the object detection problem. Papageorgiou *et al.* have developed a person detection system that accounts for some of the difficulties associated with people detection outlined earlier (Papageorgiou and Poggio, in preparation [13]; Oren, Papageorgiou, Sinha, Osuna and Poggio, 1997 [9]).

Papageorgiou's system detects people in cluttered scenes without assuming any *a priori* scene structure. The system uses Haar wavelets to represent the images. Wavelets are a multi-resolution function approximation that allow for the hierarchical decomposition of a signal (Mallat, 1989[8]). Since the Haar wavelets (which are applied to the image at two different scales) encode the local intensity variations in the image, they result in a multi-scale representation of the images, recording the relevant boundary and color information in a computationally efficient manner. This method of image representation maintains a high inter-class and low intra-class variability. Thus, it captures the defining details of the person class while distinguishing it from all other object classes.

The data vectors that are obtained by applying the Haar wavelets to different areas of the image are classified as either "persons" or "non-persons" by a Support Vector Machine (SVM). An SVM is an example based learning mechanism that was proposed by V.Vapnik (V.Vapnik, 1995[23]). Since the SVM classifier is example based, it learns the features of a class from examples which eliminates the need to explicitly model the person class.

Papageorgiou's system has reported successful results detecting frontal, rear and side views of people which indicate that the wavelet based image representation scheme and the SVM classifier perform well for 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

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. In this manner, the problem of face detection is reduced to the problem of finding facial features and their configuration. This component based approach to object detection has been used in face detection systems (Shams and Spoelstra, 1996[19]; Leung, Burl, and Perona, 1995[7]; and Yow and Cipolla, 1997[24]), but its application to the problem of finding people in images has not been reported.

It is worth mentioning that a component based 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 intra-class variation concerning the configuration of body parts, but also that it is more difficult to detect the body parts because their appearance changes as a person moves. The example of a walking man illustrates this point well. When a person walks, the configuration of his legs, and arms vary continuously. This translates directly to a large number of possible body part configurations, which makes it increasing difficult to define a model that captures all of these possibilities. As facial features do not normally have the same degree of freedom as legs and arms there are relatively fewer permissible configurations of the mouth, eyes and nose. Also, since the limbs are moving their appearance changes, which makes them increasingly difficult to detect. This problem is encountered to a significantly less degree in face detection, where the mouth is the only feature that changes shape appreciably.

Presented hereafter are brief outlines of several component based face detection systems. The system of Shams and Spoelstra uses a neural network to generate confidences for possible left and right eye regions which are paired together to form all possible combinations (Shams and Spoelstra, 1996[19]). The confidences of these pairings are weighted by their topographic suitability which are then thresholded to classify the pattern. These weights are defined by a 2-D Gaussian function.

The system of Leung, Burl, and Perona uses a probabilistic model to score potential matches (Leung, Burl, and Perona, 1995[7]). The feature detectors are model based. Based on the location of features with high confidence ratings, they select geometrically suitable facial features. Candidate constellations are formed using only these chosen features. The final stage of detection is posed as a binary hypothesis testing problem. The first hypothesis is that the vector of feature distances is a face and the second is that it is not. To detect faces, the conditional probability for both hypotheses are calculated and their ratio is compared to a suitable threshold. The system has the capability to explicitly deal with occlusions.

Yow and Cipolla have also developed a component based approach to detecting faces (Yow and Cipolla, 1997[24]). In their system they categorize potential features into candidate groups based on topographic evidence and assign probabilities (that they are faces) to these groups. The probabilities are updated using a Bayesian network. If the final probability measure of a group is above a certain threshold, then it is declared as a "detection." The features are initially identified using an image invariance scheme.

Where the above systems take different approaches to detecting faces in images by components, they have two similar features:

- They all have *component detectors* that identify candidate components in an image.
- They all have a means to integrate these components and determine if together they define a face.

1.2.2 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 datasets (Breiman, 1996[2]; Freund and Schapire, 1996[4]; and Quinlan 1996[14]). Despite the well documented practical success of these algorithms, the reasons why bagging and boosting work so well is still open to debate. One theory proposed by Schapire likens boosting to support vector machines in that both maximize the minimum margin over the training set (Schapire et.al., 1998[18]). However, his definition of "margin" differs from Vapnik's (Vapnik, 1995[23]). Bauer and Kohavi present a study of such structures including bagging and boosting, oriented towards determining the circumstances under which these algorithms are successful (Bauer and Kohavi, 1998[1]).

1.3 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 attempts to detect 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 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 results of previous work in the field.

The fundamental design of the system is similar to the component based face detection systems described in Section 1.2.1, in that it has detectors at one level for finding components of a person and a means at the next level to combine the component detector results.

The system introduces a new hierarchical classification architecture to visual data classification. Specifically, it is composed of *distinct* example based *component classifiers* trained to detect different objects at one level and a similar example based *combination classifier* at the next. This type of architecture, where example based learning is conducted at two levels, is called Adaptive Combination of Classifiers (ACC). The component classifiers detect separately, components of the "person" object, i.e. heads, legs, and arms. The combination classifier takes the output of the component classifiers as its input and classifies the entire pattern under examination as a "person" or a "non-person." The notation concerning *component* and *combination classifiers* is used throughout this paper.

Despite its relative complexity in comparison to a full body detection algorithm, a component based approach to detecting people is appealing. This is because it allows for the use of the geometric information concerning the human body to supplement the visual information present in the 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 explicitly take advantage of the known geometric properties of the human body.

Also, 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.

Another reason to adopt a component based approach to people detection is that the framework directly addresses the issue of detecting people that are partially occluded or whose body parts have little contrast with the background. This can be 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 component detectors are patterned after the full body person detector developed by Papageorgiou, described in detail in [13] and [9], and briefly in Section 1.2.1, that has yielded excellent results. This allows us to use the full body person detector as a basis for judging the benefits of component based detection and of ACC to combine the components.

Haar wavelets are used to represent the images in the component detectors. Wavelets are a computationally efficient manner to encode intensity and color differences in local regions within an image (Mallat, 1989[8]). The representation scheme results in a multi-scale edge representation of the components that maintains high inter-class and low intra-class variation. This allows for the development of tight class models that still capture all of the defining characteristics of the components. Most importantly, they are free of some of the problems associated with pixel and edge based representation which are outlined earlier in Section 1.1.

SVM classifiers are used as the classification devices within the ACC architecture of the system. One of the motivating reasons for settling on the ACC architecture for the system is that within such an architecture all of the classifiers are example based machines. The example based modeling of the component classes and the person class is desirable because example based classifiers learn the salient features of a class from examples and hence, are free from any bias associated with a hand-crafted model. Biases are introduced into hand-crafted models when designers include parameters that they believe are significant, but which in reality are not required to describe the class. Use of example based devices are also advantageous because it allows the system to be applied to different objects of interest relatively easily. SVM's are chosen as the example based classification device not only for their demonstrated superior performance and sound mathematical foundation but also because they produce a raw output along with the binary class when they classify a data vector. The raw output produced when a data vector is classified by an SVM classifier is a rough measure of how "well" the vector matches its designated class. This is important, since the raw output can be employed as a confidence rating without any further processing.

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

2 System Details

This section describes the structure and operation of our person detection system.

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. Figure 2 is a graphical representation of this procedure.

The system starts detecting people in images by selecting a 128×64 [1] window from the top left corner of the image as an input. This input is then classified as either a "person" or a "non-person", 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. There are four distinct component detectors in this system which operate independent of each other and are trained to find separately the four components of the human body - the head, the legs, and the left and right arms.

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. The training of the component and combination classifiers is described in detail in 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. 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.

[1]All dimensions are in pixels.

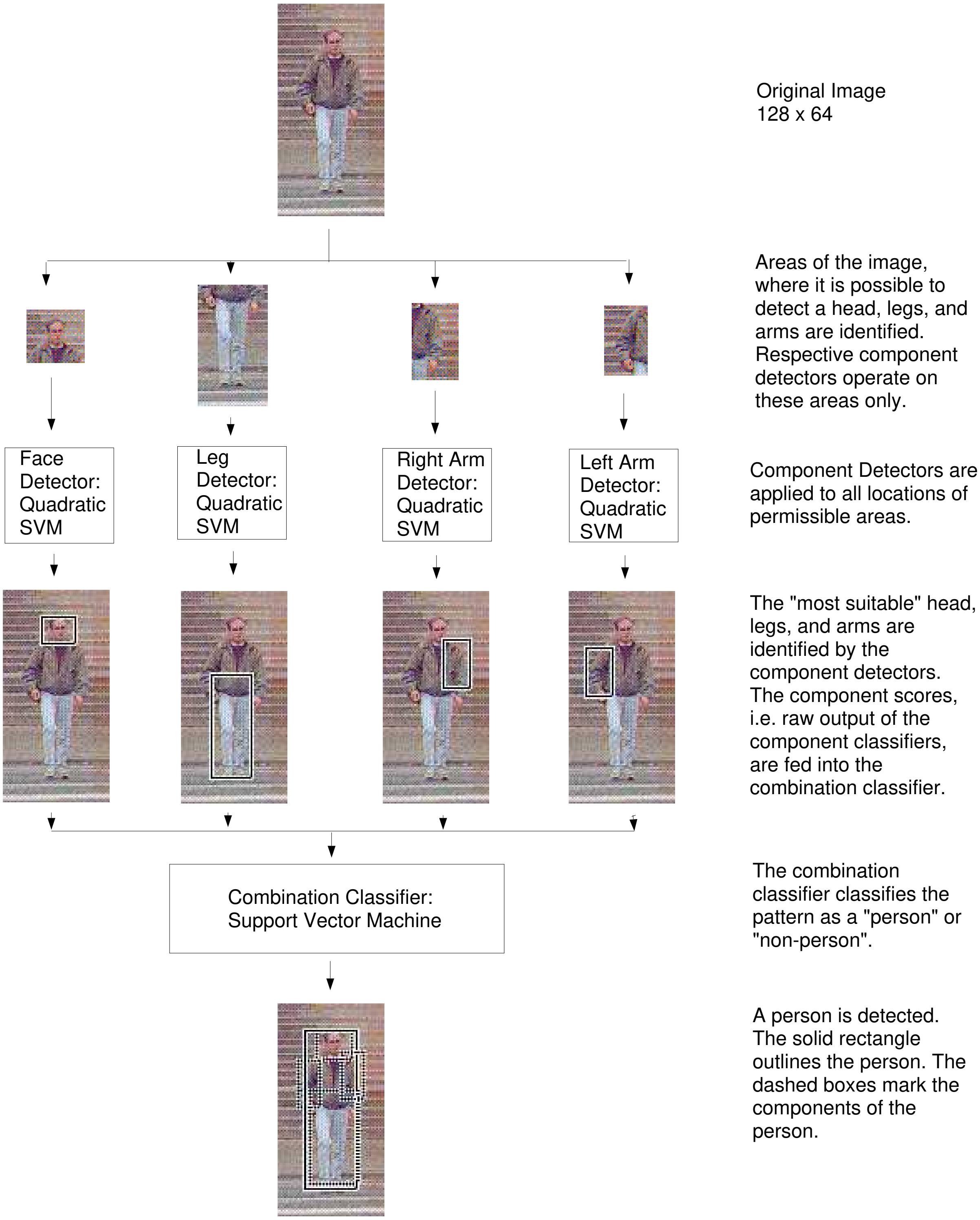

Figure 2: *Diagrammatic description of the operation of the system.*

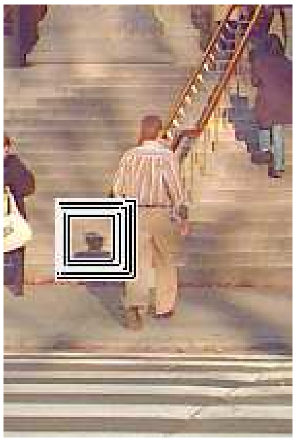

Figure 3: *It is very important to place geometric constraints on the location and scale of component detections. Even though a detection may be the strongest in a particular window examined, it might not be located properly. In this figure, the shadow of the person's head is detected with a higher score 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.*

The highest component score for each component is fed into the combination classifier which is a linear SVM. 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 combination classifier processes the set of scores received from the component classifier 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 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

This section outlines the details of the component detectors and the combination classifier.

2.2.1 First Stage - Identifying Components of People in an Image

When a 128×64 window is evaluated by the system, the component detectors are applied only to specific areas of the window and only at particular scales. This is because the arms, legs, and head of a person have a defined relative configuration, i.e. the head is found above the legs, with left and right arms to either side. The components must also be proportioned correctly. By placing these geometric constraints on the location and scale of the components, we ensure that they are arranged in the form of a human body, and thus improve the performance of the object detection system. This is necessary, because even though a component detection is the strongest in a particular window under examination (i.e. it has the highest component score), it does not imply that it is in the correct position, as illustrated in Figure 3.

Since the component detectors operate on rectangular areas of the image, the constraints placed on the location and scale of component detections are expressed in terms of the properties

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

Table 1: *Geometric constraints placed on each component. All coordinates are in pixels and relative to the upper left hand corner of a* 128×64 *rectangle. Dimensions are also expressed in pixels.*

of the rectangular region examined. For example, the centroid and boundary of the rectangular area determines the location of a component detection and the width of the rectangle is a measure of a component's scale. All coordinates are relative to the upper left hand corner of the 128×64 window.

We calculated the geometric constraints for each component from a sample of the training images. The constraints themselves are tabulated in Table 1 and shown in Figure 4. The values of quantities such as the location of the centroid and top and bottom boundary edges of a component were determined by taking the mean of the quantities 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 multi-resolution function approximation that allow for the hierarchical decomposition of a signal (Mallat, 1989[8]). 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 invariance scheme (Sinha, 1994[20]). Haar wavelets of two different scales (16×16 and 8×8) are used to generate a multi-scale representation of the images. The wavelets are applied to the image such that they overlap 75% with the neighboring wavelets in the vertical and horizontal directions. 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 a 32×32 window for the head and shoulders and 954 features for 48×32 windows representing the lower body and the left and right arms. This method results in a

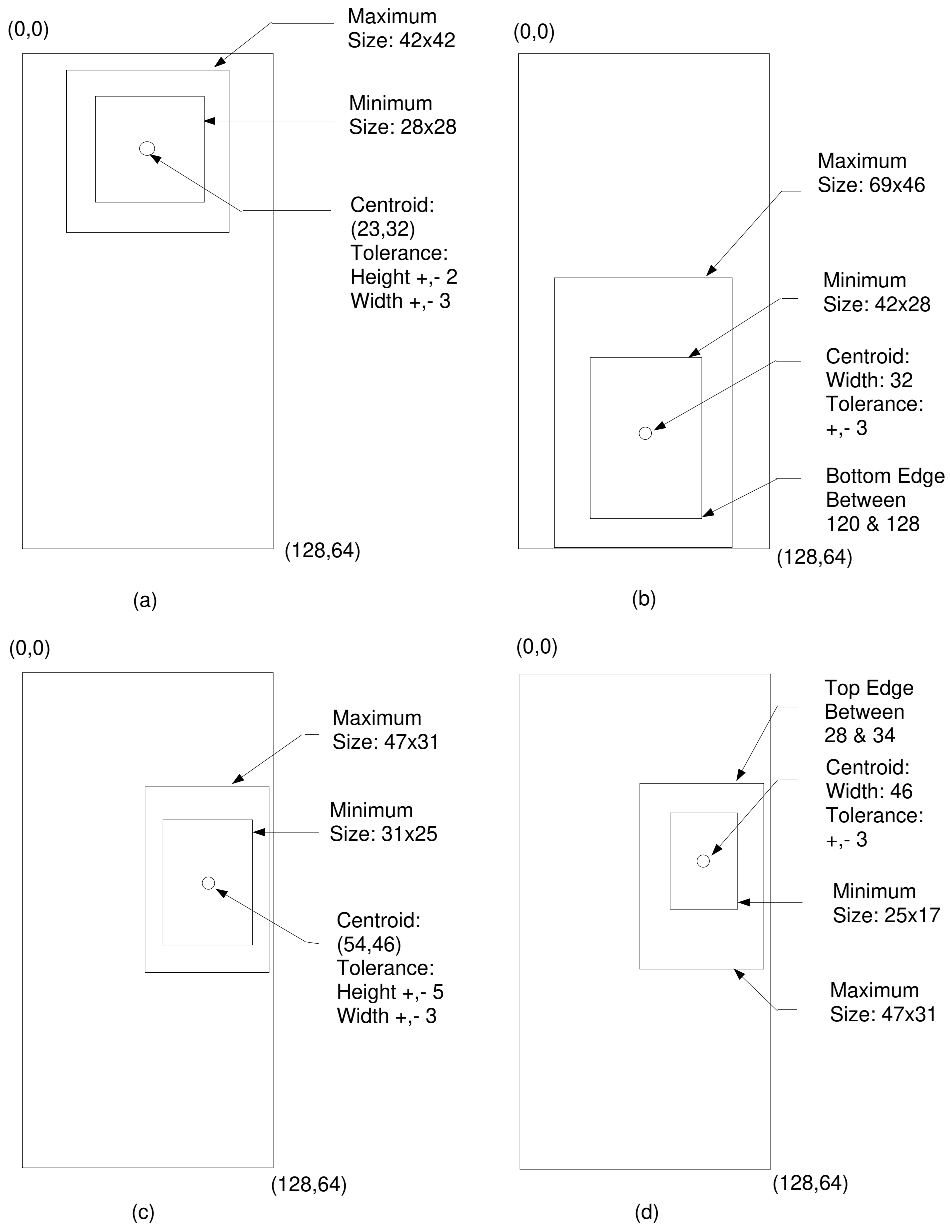

Figure 4: *Geometric constraints that are placed on different components. All coordinates are in pixels and relative to the upper left hand corner of a* 128×64 *rectangle. Dimensions are also expressed in pixels. (a) illustrates the geometric constraints on the head, (b) the lower body, (c) an extended right arm, and (d) a bent right arm.*

thorough and compact representation of the components, with high inter-class and low intra-class variation.

We use SVM's to classify the data vectors resulting from the Haar wavelet representation of the components. SVM's were proposed by Vapnik and have yielded excellent results in various data classification tasks, including people detection (Papageorgiou and Poggio, in preparation [13]; Oren, Papageorgiou, Sinha, Osuna and Poggio, 1997 [9]) and text classification (Joachims, 1998[5]). Traditional training techniques for classifiers like multilayer perceptrons use empirical risk minimization and lack a solid mathematical justification. The support vector machine 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:

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

where,

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

In Equation 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 fall closest to the separating hyperplane. Finally, the coefficients α_i and b are determined by solving a large-scale quadratic programming problem. The kernel function K that is used in the component classifiers is a quadratic polynomial and has the form shown below:

$$K(\mathbf{x}, \mathbf{x_i^*}) = (\mathbf{x} \cdot \mathbf{x_i^*} + 1)^2 \tag{3}$$

$f(\mathbf{x}) \in \{-1, 1\}$ in Equation 1 is referred to as the *binary class* of the data point $\mathbf{x}$ which is being classified by the SVM. Values of 1 and -1 refer to the classes of the positive and negative training examples respectively. As Equation 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.

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 Figure 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 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

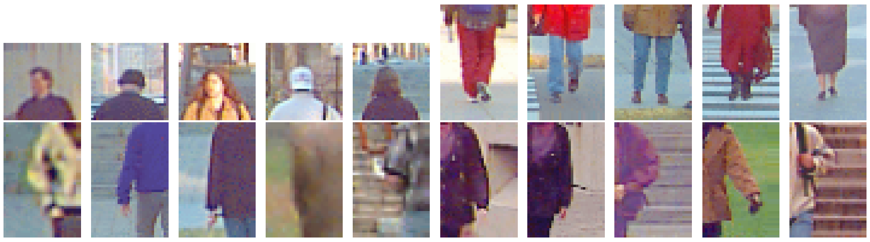

Figure 5: *The top row shows examples of "heads and shoulders" and "lower bodies" of people that were used to train the respective component detectors. Similarly, the bottom row shows examples of "left arms" and "right arms" that were used for training purposes.*

hyperplane. This distance is 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 window, then zero is placed in the data vector. A component score of zero refers to a test point that is classified as neither a "component" nor a "non-component" because it lies on the hyperplane.

The combination classifier is a linear SVM classifier. The kernel K that is used in the SVM classifier and shown in Equation 2 has the following form:

$$K(\mathbf{x}, \mathbf{x_i^*}) = (\mathbf{x} \cdot \mathbf{x_i^*} + 1) \tag{4}$$

This type of hierarchical classification architecture where learning occurs at multiple stages is termed Adaptive Combination of Classifiers (ACC).

Positive examples were generated by processing 128×64 images of people at one scale, and taking the highest component score (from detections that are geometrically allowed) for each component type. Table 2 shows examples of data vectors that were used to train the combination classifier.

3 Results

In this section we present the results of an experiment that was conducted to determine the performance of our person detection system. The performance of this system is compared to that of other component based person detection systems that combine the component classifiers in a different way and the full body person detection system that is described in [13] and [9] and reviewed in Section 1.2.1. This framework allows us to determine the strengths of the component based approach to detecting objects in images and the performance of various methods of combining the component classifiers.

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

Head and Shoulder Scores	*Lower Body Scores*	*Right Arm Scores*	*Left Arm Scores*
Positive Examples			
2.415	3.152	3.233	3.145
1.861	1.855	2.339	2.280
4.184	2.332	3.258	3.994
2.871	1.691	2.311	1.221
Negative Examples			
0.677	0.694	0.817	1.020
4.530	0.231	0.252	0.824
0.105	0.021	0.002	0.560
1.869	0.010	0.718	1.746

Table 2: *Examples of positive and negative data points used to train the combination classifier. The entries are component scores. The component scores of the positive examples are generally higher.*

Component Classifier	*Number of Positive Examples*	*Number of Negative Examples*
Head and Shoulders	856	9315
Lower Body	866	9260
Left Arm	835	9260
Right Arm	838	9260

Table 3: *Number of positive and negative examples used to train the different component classifiers.*

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 number of positive and negative examples that were used to train the different component classifiers are presented in Table 3.

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 and at the allowable scales. The greatest positive classifier output for each component, i.e. the component score, 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 tested.

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 which do not contain any people. These images are pictures of natural scenery and buildings. 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 known as 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."

We tried VCC as an approach to combining the component classifiers since it is one of the simplest classes of classifier combination algorithms and hence afforded the best opportunity to judge the strengths of a component based object detection system that is not augmented with a powerful classifier combination algorithm. Experimenting with the VCC based system was also an opportunity to compare it with an ACC based system and determine the benefits of more sophisticated classifier combination methods. Since the computational complexity of these methods are known, and the experiment described in this section determines their performance, this framework characterizes the tradeoff involved between enhanced performance and greater computational complexity for these systems. The person detection systems which are evaluated here, in decreasing order of computational intensity, are: the ACC based systems, the VCC based system, and the full body system (the baseline) described in [13], [9], and Section 1.2.1. As the results show, this is also the order of the systems when sorted by performance, with the best performing system listed first.

In the incarnation of VCC that is implemented and tested in this experiment, a positive

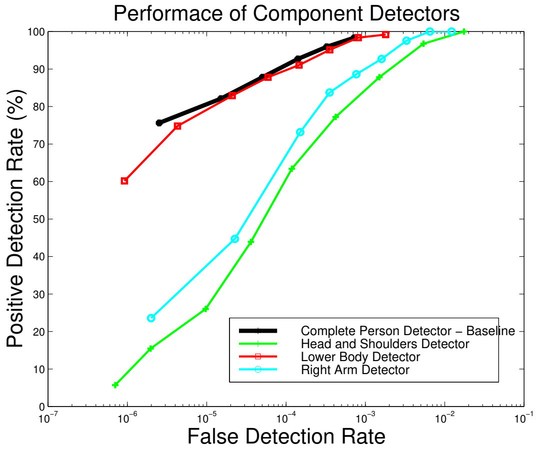

Figure 6: *ROC curves illustrating the ability of the component detectors to correctly flag a person in an image. The positive detection rate is plotted as a percentage against the false alarm rate which is measured on a logarithmic scale. The false alarm rate is the number of false positive detections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

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 systems and are described in Section 2.2.1. For each pattern that the system classifies, the system must evaluate the logic presented below:

$$Pattern\ class \;=\; Head\ class\ \&\ Legs\ class\ \&\ Left\ arm\ class\ \&\ Right\ arm\ class \qquad (5)$$

where a logic 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. This information was estimated from the ROC curves of each of the component detectors that are shown in Figure 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 the process of 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 [13], [9], [10], [12], and [11]. It has the same architecture as the individual component detectors used in our system and which are described in Section 2.2.1. The only difference between the baseline system and the component detectors is that the baseline system is trained to detect the pattern of an upright person in an image instead of an arm, legs, or a head. The baseline system uses Haar

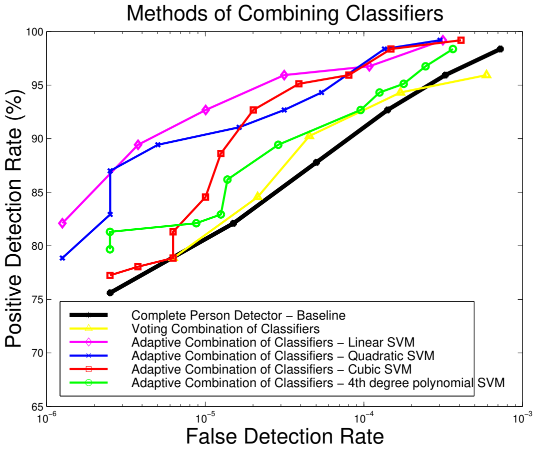

Figure 7: *ROC curves comparing the performance of various component based people detection systems. The systems differ in the method used to combine the classifiers that detect the various components of a person's body. The positive detection rate is plotted as a percentage against the false alarm rate 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 a 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.*

wavelets to represent the images and a quadratic SVM classifier to classify the patterns. The SVM classifier was trained on 869 positive and 9,225 negative examples.

3.2 Experimental Results

The ROC curves of the person detection systems are shown in Figure 7. ROC curves are the most suitable framework for analyzing the different object detection systems because they explicitly capture the tradeoff between accuracy and false detections that is inherent to every detector. This is an important property of detectors because one may wish to sacrifice a certain degree of accuracy for the possibility of less false alarms or vice versa for a particular application. Thus, a complete characterization of performance over a range of detection thresholds is essential.

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. This is noteworthy because the baseline system has produced very accurate results. 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 baseline system is a very true control case for this experiment and the improvement in performance must be due to the component based approach and the algorithm used for combining the component classifiers. It should be noted that all of the component based systems were comparable to or better than the baseline system. This fact suggests that the additional information concerning the geometric properties of the human body, incorporated in the component based systems and absent in the full body person detector, improves the performance of a person detection algorithm.

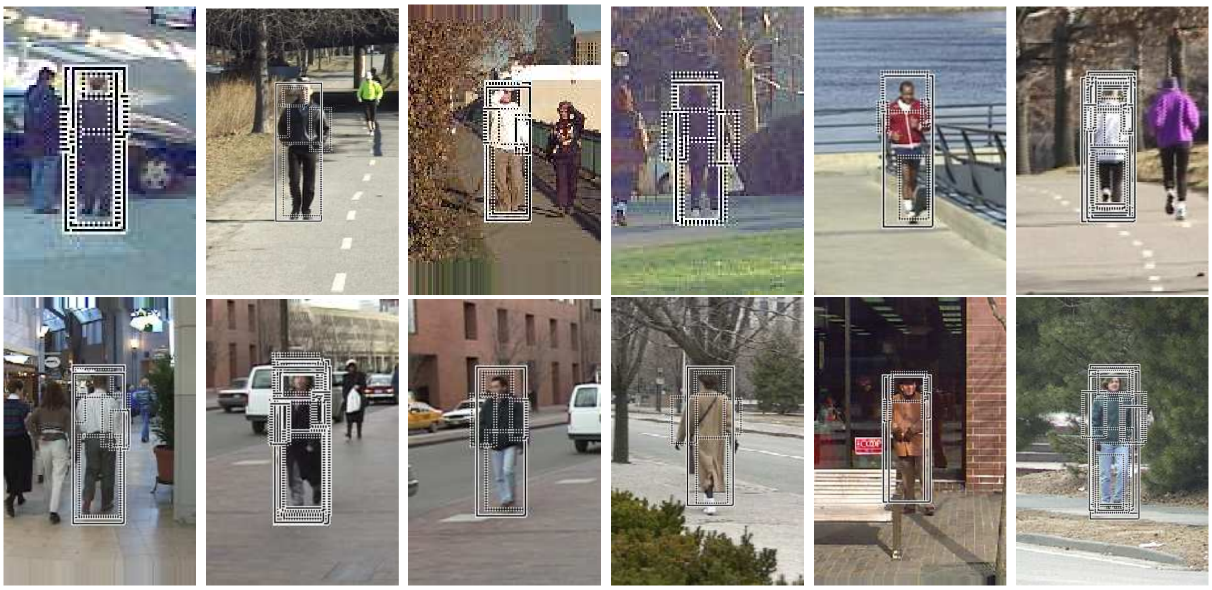

Figure 8: *Samples from the test image database. These images demonstrate the capability of the system. It can detect running people, people who are slightly rotated, people whose body parts blend into the background (bottom row, second from right - person detected even though the legs are not), and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

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. 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). This suggests that all of those elements must be person-like for the pattern to be classified as a person. The emphasis placed on the presence of combinations of components increases with the degree of the polynomial classifier. The results show that the performance of the ACC based systems decreases with an increase in the degree of the polynomial classifier. In fact, the ROC curve for the ACC based system that employs a fourth degree polynomial classifier is very similar to the VCC based system. Interestingly, both of the above systems search for all four components in a pattern. The VCC based system explicitly requires the presence of all four components where as the ACC based system that uses the fourth degree polynomial classifier makes it an implicit requisite due to the design of its kernel. It is also possible that the higher degree polynomial classifiers may require more training examples in proportion with the higher dimensionality of their feature space to perform at the same level as the linear SVM.

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

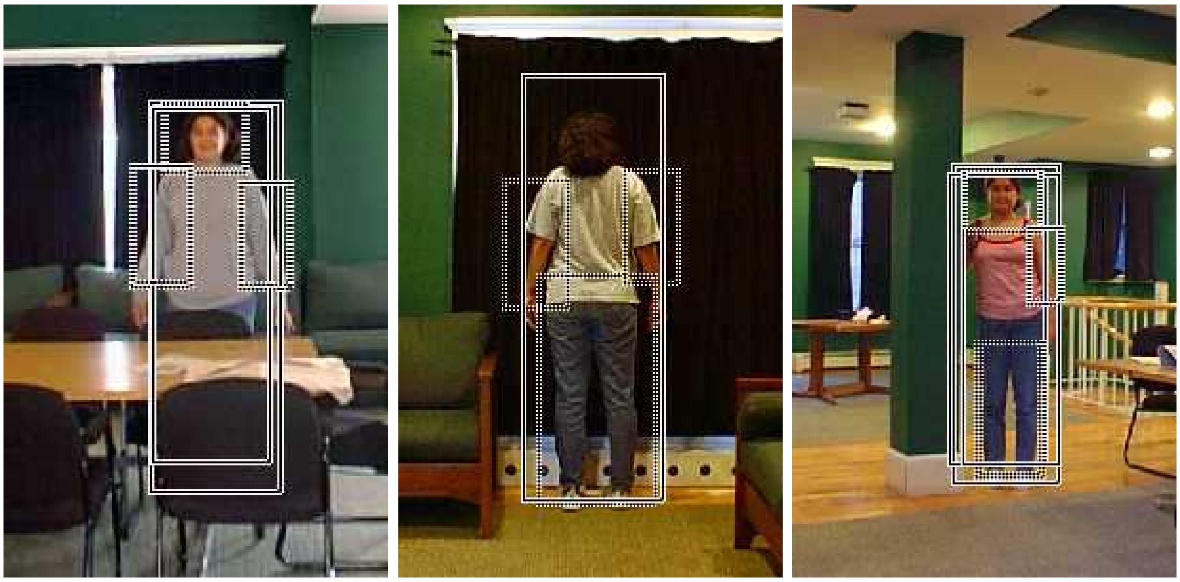

Figure 9: *Results of the system's application to images of partially occluded people and people whose body parts have little contrast with the background. In the first image, the person's legs are not visible; in the second image, her hair blends in with the curtain in the background; and in the last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

challenging images. Included are pictures of people walking and running. In some of the images, the person is partially occluded or a part of their body has little contrast with the background. A few of the images depict people who are slightly rotated in depth. Figure 8 is a selection of these images.

Figure 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 Figure 7, corresponds to a false detection rate of less than 1 false alarm for every $796,904$ patterns inspected.

Figure 10 shows the result of applying the system to sample images with clutter in the background. Even under such circumstances the system performs very well. The lower four images were taken with different cameras than the instruments used for the training set images. The conditions and surroundings for these pictures are different too.

4 Conclusions and Future Work

In this paper we presented a component based person detection system for static digital 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 system for detecting people in images had not been successfully developed prior to this project. We chose to take a component based approach to the problem because such a solution promised to handle variations in lighting and noise in an image better than a

Figure 10: *Results from the component based person detection system. The solid boxes outline the complete pedestrian, where the dashed rectangles are the components.*

full body person detector. We also anticipated that a component based system would be able to detect partially occluded people and people who were rotated in depth, without any additional modifications to the system. A full body person detector is unable to do this because it searches for the *complete* pattern of the human body in an image, which is often distorted by an occlusion or a rotation. A component based detector, on the other hand, looks for components of a person, i.e. a head, legs, and arms, and if one of these components was not detected, due to an occlusion or because a person was rotated into the plane of the image, the system could still detect a person if the component detections were combined using an appropriate algorithm. Another reason we decided on the component based approach was that it lends itself conveniently to the use of a hierarchical classifier to classify the patterns. Previous research suggests that a hierarchical classification system performs better than a simple single layer classifier for a particular data classification task.

The hierarchical classifier that is implemented in this system uses four distinct component detectors at the first level, which 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. This type of architecture, in which learning occurs at more than two levels, is relatively new, and is known as Adaptive Combination of Classifiers (ACC).

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 employed. The superior performance of the component based approach can be attributed to the fact that it operates with more information about the problem 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 knowledge about the geometric properties of the human body. It uses this additional information concerning the relative configuration of body parts to increase accuracy in terms of a lower false alarm rate for a given positive detection percentage.

One drawback of the component based person detection system is that it is currently slower than a system that detects the full body. This is because the system involves multiple detectors that search for components of a person and which are subsequently combined. The relatively complex architecture of the component based system makes it more computationally intensive than a full body object detection system, and thus slower.

4.1 Suggestions for Future Work

In this system, we place manually determined constraints on the relative location and size of the component detections. While this method of ensuring that the detections are in the proper configuration produces excellent results, it may suffer from a bias introduced by the designer. Therefore it is desirable for the system to learn the geometry of an object from examples. This would also make it easier to apply this system to other objects of interest. Such an object detection system would be a step towards a more sophisticated component based detection system in which the components of an object are not predefined by a user.

It would be useful to test the system described here in other domains, such as, cars and faces.

While this report 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.

In summary, the component based object detection system presented here produces encouraging results. This can be attributed partially to the idea of detecting objects by locating their components and partially to the use of ACC architecture to classify the patterns.

5 Acknowledgments

Thanks are due to Professor Tomaso Poggio and Constantine Papageorgiou for their assistance in this study. Constantine Papageorgiou is primarily responsible for developing the object detection system that is at the core of the component based object detection system. Thanks to Massimiliano Pontil and Theodoros Evgeniou for their valuable suggestions.

References

[1] Eric Bauer and Ron Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning*, 1998.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[3] Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. In U. Fayyad, editor, *Proceedings of Data Mining and Knowledge Discovery*, pages 1–43, 1998.

[4] Y. Freund and R. E. Schapire. *Machine Learning: Proceedings of the Thirteenth National Conference*, chapter Experiments with a new boosting algorithm,. Morgan Kaufmann, 1996.

[5] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings 10th European Conference on Machine Learning (ECML).*, 1998.

[6] N. Kehtarnavaz and F. Rajkotwala. Real-time vision based detection of waiting pedestrians. *Real-Time Imaging*, 3(6):433–40, Dec 1997.

[7] T.K. Leung, M.C. Burl, and P. Perona. Finding faces in cluttered scenes using random labeled graph matching. In *Proceedings of the Fifth International Conference on Computer Vision*, pages 637–644, Jun 1995.

[8] S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, Jul 1989.

[9] Micheal Oren, Constantine Papageorgiou, Pawan Sinha, Edgar Osuna, and Tomaso Poggio. Pedestrian detection using wavelet templates. In *Proceedings Computer Vision and Pattern Recognition*, pages 193–199, Puerto Rico, Jun 1997.

[10] Constantine Papageorgiou. Object and pattern detection in video sequences. Master's thesis, Massachusetts Institute of Technology, May 1997.

[11] Constantine Papageorgiou, Theodoros Evgeniou, and Tomaso Poggio. A trainable pedestrian detection system. In *Proceedings of Intelligent Vehicles*, pages 241–246, Oct 1998.

[12] Constantine Papageorgiou, Micheal Oren, and Tomaso Poggio. A general framework for object detection. In *Proceedings of the International Conference on Computer Vision*, Bombay, India, Jan 1998.

[13] Constantine Papageorgiou and Tomaso Poggio. A general trainable system for object detection. In preparation.

[14] J.R. Quinlan. Bagging, boosting and c4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1996.

[15] L.A.D. Reading, G.L. Wan, and N.W. Dickinson. Pedestrian detection at road crossings by computer vision. In *Ninth International Conference on Road Transport Information and Control*, pages 139–43, Apr 1998.

[16] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, Jan 1998.

[17] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Rotation invariant neural network-based face detection. In *Proceedings 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 38–44, Jun 1998.

[18] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. The Annals of Statistics, 1998. To appear.

[19] Ladan Shams and Jacob Spoelstra. Learning gabor-based features for face detection. In *Proceedings of World Congress in Neural Networks. International Neural Network Society.*, pages 15–20, Sep 1996.

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

[21] Kah-Kay Sung and Tomaso Poggio. Example-based learning for view-based human face detection. In *Proceedings from Image Understanding Workshop*, Monterey, CA, Nov 1994.

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

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

[24] Kin Choong Yow and R. Cipolla. Feature-based human face detection. *Image and Vision Computing*, 15(9):713–35, Sep 1997.

[25] Alan L. Yuille. Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70, 1991.