Person Detection: Unmanned System and Small Sensor Applications

by

Paul Edward Rosendall

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of
Master of Science in Aeronautics and Astronautics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2008

Author . . . .
Department of Aeronautics and Astronautics
May 23, 2008

Certified by . . . .
Tomaso A. Poggio
Eugene McDermott Professor, Dept. of Brain and Cognitive Sciences
Thesis Advisor

Certified by . . . .
Jeffrey W. Miller
Draper Laboratory Technical Supervisor
Thesis Supervisor

Certified by . . . .
Brent D. Appleby
Draper Laboratory Tactical ISR Division Leader
Aero/Astro Dept. Thesis Reader

Accepted by . . . .
Prof. David L. Darmofal
Associate Department Head
Chair, Committee on Graduate Students

Person Detection: Unmanned System and Small Sensor Applications

by

Paul Edward Rosendall

Submitted to the Department of Aeronautics and Astronautics
on May 23, 2008, in partial fulfillment of the
requirements for the degree of
Master of Science in Aeronautics and Astronautics

Abstract

The ability to quickly and reliably detect people in images and video is highly desired. Several object recognition algorithms have demonstrated successful detection of multi-class objects with varied scale, position and orientation. This study examines the effectiveness of these methods when applied to detecting humans in two distinct domains: A) Leave-behind sensing and B) Aerial surveillance. Using novel image sets that are significantly more realistic and difficult than standard datasets, a variety of tests are conducted to compare the algorithms in terms of classification success rate. Dalal and Triggs' Histogram of Oriented Gradients algorithm, when trained with image samples taken from inside MIT's Stata Center, detects with no false positives all but one person in six minutes of video taken from inside a separate building. An enhanced version of Riesenhuber and Poggio's cortex-like recognition model, trained to detect people, correctly classifies 95% of images taken from a small UAV when trained with an independent set of images. These results illustrate the potential to accurately and reliably determine the presence of people in video from unmanned aircraft and indoor sensors.

Thesis Advisor: Tomaso A. Poggio
Title: Eugene McDermott Professor, Dept. of Brain and Cognitive Sciences

Thesis Supervisor: Jeffrey W. Miller
Title: Draper Laboratory Technical Supervisor

Aero/Astro Dept. Thesis Reader: Brent D. Appleby
Title: Draper Laboratory Tactical ISR Division Leader

Acknowledgments

The countless contributions of people both inside and outside the Draper/MIT community have greatly influenced the design and development of this thesis.

I would first like to thank my supervisor, Jeffrey Miller, for his thorough guidance throughout this project. Our weekly meetings were illuminating, inspiring, and enjoyable. This thesis was made possible by his keen interest in the field of object recognition and his idea to put some amazing algorithms in action.

These algorithms were the very source of this project's success, and I thank every author for their essential contributions. In particular, I would like to thank Jim Mutch, Navneet Dalal, Bill Triggs, and Rob Fergus for providing source code along with assistance.

I am humbly grateful to have learned from and interacted with Tomaso Poggio, Thomas Serre, and Stanley Bileschi, who willingly and graciously allowed me to incorporate their extraordinary body of work into my thesis.

My supervisors and co-workers at Draper have been tremendously helpful and reliable. I would like to thank Brent Appleby and Paul DeBitetto for their ideas and assistance from start to finish; without them, I would not have had this wonderful opportunity. I also greatly appreciate the many Draper Fellows who were more than willing to help, especially when I asked them to pose for so many pictures. To Ben and Josh: I will definitely miss having all those delightfully distracting discussions.

Finally, my family has always played an enormous role in my life. I want to thank my parents, my brother and sister, and my aunts, uncles, cousins, and grandparents for always being there for me. And despite my distance from home, the last two years have been the best of my life because I have shared them with my amazingly wonderful girlfriend Kelly, whose love and support (and great ideas) have carried me through.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

1.1 Motivation

Obtaining accurate information from the vast amount of military sensor data is a tactical necessity. This information is often the foundation upon which entire military operations are planned and performed. The varied nature of these operations demands the logical delivery of information. Troops should receive accurate and pertinent information to maximize the chance of success in dangerous environments.

Military surveillance often entails using sensors to detect the presence of people. Having humans perform this detection process is an effective, yet ultimately inefficient approach considering the remarkable advances in computer vision algorithms. A reliable person detection system would lead to significant savings in personnel hours and greater operational flexibility.

1.2 Problem Statement

Ideal aerial surveillance is the perfect monitoring of events on the ground. One important area of surveillance is tracking people and their actions, where it is critical to first determine the presence of people on the ground. Video surveillance offers the viewer the opportunity to quickly determine this presence and respond appropriately. While effective, this technique undoubtedly wastes the time of the viewer uninterested in the majority of footage. An effective solution would be a system that promptly alerts the user when a person has entered the monitored ground area, giving the user the ability to respond when necessary and perform other important tasks in the meantime.

Another important military operation is checking buildings for the presence of enemy combatants. The US Marine Corps stresses of the importance of units establishing “clearly recognizable and understandable signals for marking cleared rooms and buildings” [25]. These markings are often made directly on the building exterior. One main problem with this current approach is that enemies may likely enter cleared rooms, posing a serious threat to troops who are misled by their own markings. Another problem is the potential for troops to mark rooms cleared when enemies are actually present. Dan Zanini, deputy program manager of the Army’s Future Combat Systems program, says that Israeli Defense Forces recently suffered the most casualties against Hezbollah “from forces that they bypassed and forces that came in from their rear” [12].

These problems would be alleviated by placing a small sensor inside each cleared room, where troops would receive notification when someone is detected. An automated system that quickly and accurately determines human presence would bypass the need for constant human monitoring.

The definition of successful person detection can depend on the application. Momentarily neglecting to notice human presence during one mission may have fewer consequences than overlooking enemy presence in another. An effective person detection system should strive for an extremely low miss rate without generating many false positives. It is important to keep an appropriate balance between these two factors. Finally, each detection system must respond quickly enough to maintain usefulness and reliability.

1.3 Thesis Overview

This thesis displays the full progression of this project. Chapter 2 describes the most significant testing results of this thesis and how they relate to real detection scenarios. Chapter 3 details the review of background literature, beginning with the general development of object recognition and ultimately describing the specific detection methods that were examined. Chapter 4 describes the testing of benchmark algorithms with novel datasets. The problem characteristics, dataset development, and testing progression and results are detailed for both the indoor and aerial domains. Chapter 5 applies object tracking to yield an alternative approach to testing with video, and the performance improvements achieved are evaluated. Finally, Chapter 6 summarizes the overall conclusions and ideas for future research.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Key Results

This section briefly summarizes the two most important test results of this thesis. These tests are simple yet significant, as they illustrate real potential for person detection in multiple domains.

2.1 Leave-Behind Sensing with HOG

In the Indoor Sequence Test (described fully in Section 5.3), Dalal and Triggs' Histogram of Oriented Gradients (HOG) algorithm successfully detected all but one person in six minutes of indoor video, without producing any false positives. This was done by tracking each object in order to incorporate multiple classifications. Figure 2-1 is a plot of detection rate vs. false positive rate for all object sequences in the four test videos. This impressive result is significant for three main reasons:

1. Both the training and test image sets are realistic: they were obtained from authentic videos taken from inside MIT buildings and contain instances of many different people in an abundance of poses and positions.

2. The training and test sets are independent. Because the sets contain videos taken in different locations with different people, these results can be extended to real-world scenarios.

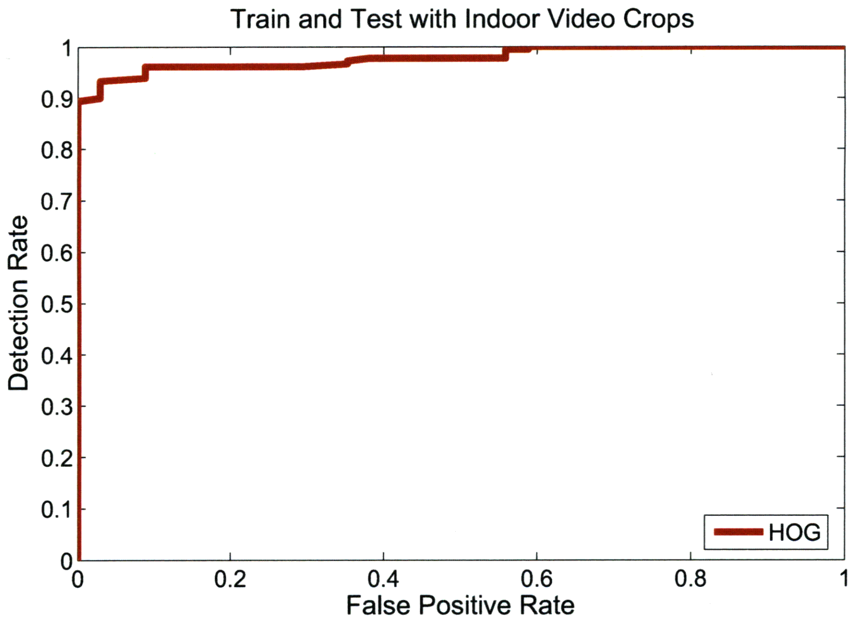

Figure 2-1: ROC curve for testing HOG with indoor video crop sequences.

3. The detection system employed in this test ran faster than the video itself, which means that successful real-time person detection is possible using HOG with image subtraction.

2.2 Aerial Surveillance with HMAX

An enhanced version of Riesenhuber and Poggio's "Standard Model" of cortex-like object recognition, referred to as HMAX, was able to correctly determine whether people were present in 95% of images from a small UAV (as shown in Figure 2-2). This result is significant for three main reasons:

1. For this test (described fully in Section 4.3.3), HMAX was trained with cropped samples of images taken from atop a parking garage. Both the training set and the UAV test set contain authentic images with people and other objects. Furthermore, the complete independence between training and test sets illustrates both the dynamic nature of the algorithm and the realistic nature of this result.

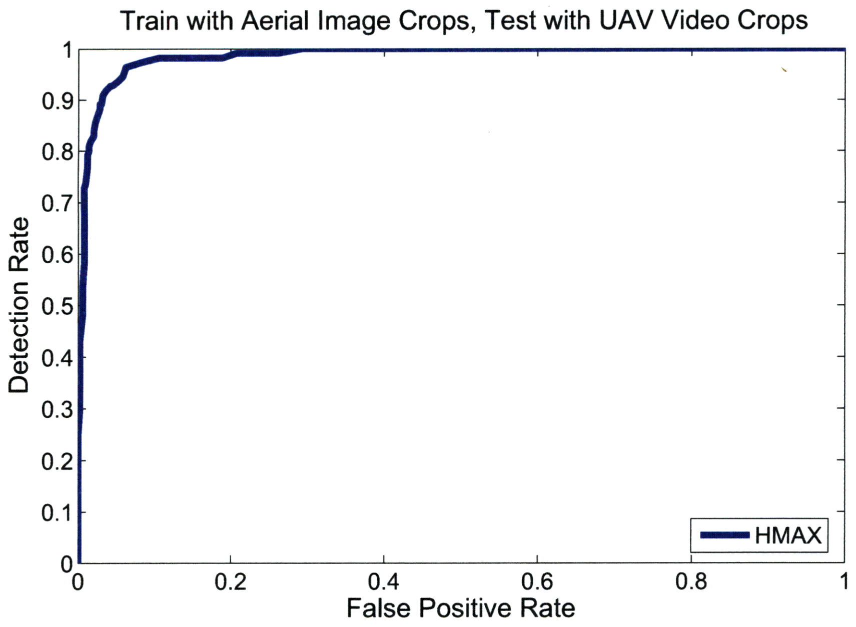

Figure 2-2: ROC curve for testing HMAX with UAV video crops.

2. More than half of the positive UAV samples contain a person who is not fully visible. Such occlusion is realistic, yet not well-handled by many algorithms; this result shows that HMAX is an exception.

3. The positive samples were manually cropped to contain people who were often not centered in the image and sometimes not fully in view. This was done to mimic the nature of image registration techniques designed to automatically isolate foreground objects. This test demonstrates the potential power of combining HMAX with image registration techniques.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Current Techniques

3.1 Overview

Computer object recognition has been an active area of research for nearly five decades. Recognizing objects is an important step towards the ability for computers to "perceive" what surrounds them, the implications of which are exciting and significant. Applications in automation, robotics, consumer engineering, and many other fields are seemingly endless, and as a result, there is an ever-increasing reliance upon effective and efficient object recognition.

There are many approaches to object recognition. Some use explicit or probabilistic shape models of the object, whereas others consider the context in which an object is found or the function an object often serves [18]. Approaches range from recognizing simple 2D objects to complex 3D shapes, and can use intensity images, range information, or a combination as a means for recognition.

The first object recognition systems employed autocorrelation and template matching; 2D pattern classification was a common goal [14]. Curved 2D shapes, 3D structures, and occluded objects with background clutter eventually followed, and researchers continue to examine object geometries and functions as a means to improved detection.

Detecting the presence of people in images and video is an important class of general object recognition. Many intelligence and surveillance applications would benefit greatly from accurate and reliable person detection, bypassing the need for constant human observation.

While face detection systems have demonstrated success with large variations in size, shape and orientation, there is a need for similarly dynamic detection of people when their faces are not visible. Recently, template matching and learning-based techniques have been applied to recognizing people [16]. Gait has also been used to detect people in image sequences. There is an ongoing effort to choose image features that ensure reliable invariance to pose and context variations.

3.2 Detection Methods

There are many ways in which object recognition methods can differ from one another. The multitude of characteristics that define these methods can be placed into one of two groups. First, many characteristics relate to the method's intention, namely what problem the method is attempting to solve. Any remaining characteristics will define the algorithm's implementation, or the ways in which the problem is solved. The algorithms reviewed throughout this study had wide-ranging properties of both intention and implementation.

The main question regarding algorithm intention is which objects are classified. Several methods have shown success in distinguishing among many object categories, whereas others specifically distinguish people from the background or identify people among other foreground objects. Those algorithms that detect people often recognize full bodies, upper bodies or faces. Certain methods are claimed to detect partially missing or occluded objects, whereas for others either no such claims are made, or it is acknowledged that occluded objects will be missed. Also varying is the range of distances at which objects can be detected. Several algorithms are specifically geared for either far-field or close-range detection, whereas for many others neither claims nor limitations are stated regarding detection range.

For the methods reviewed during the study, there are two overarching categories regarding implementation. The first is whether the object recognition is "parts-based," whereby objects are modeled as a probabilistic configuration of "parts," or feature groupings. Among the parts-based algorithms, methods vary according to whether the parts are predetermined. For instance, certain methods model humans as a configuration of body parts (such as face, upper torso, and legs), each of which is modeled from feature groupings. Others model an object as a group of unlabelled parts. The other overarching implementation category is whether motion information is used to help detect an object. While most of the reviewed algorithms operate on static images, some use consecutive image frames as the basic input for classification.

There were three key elements for an algorithm to be chosen for testing. First, it was essential to have the ability to retrieve and properly implement the software. This required both obtaining the same software used to produce benchmark results and acquiring the image sets necessary to verify those results, thus verifying proper implementation. The second priority was demonstrated success with difficult datasets. The leave-behind sensor and aerial surveillance problems are inherently tough; people, often in the presence of extreme clutter and background variation, must be detected consistently and reliably. The last major element considered when determining the testing algorithms was runtime, of which there are two components: offline training and online testing. Minimizing training time becomes important upon retraining, but minimal testing time is essential for any real-time person detection system.

Serre and Poggio's recent extension of Riesenhuber and Poggio's "Standard Model" of object recognition in cortex (HMAX) has recently shown remarkable success in detecting objects from the Caltech101 and MIT-CBCL image sets [22], the latter of which contains objects under extreme illumination conditions [24]. The only algorithm that has been known to consistently outperform HMAX on pedestrian detection is Dalal and Triggs' Histogram of Oriented Gradients (HOG) approach [23]. Viola and Jones' Boosted Cascade detector (V-J) is able to quickly and accurately detect faces among clutter and can be trained on other classes besides faces [26]. Training their system for person detection was another viable option. HMAX, HOG and V-J were the only algorithms that displayed all three key attributes, so each was trained and tested with the abundance of collected images.

3.2.1 HMAX

Humans and other primates still outperform state-of-the-art machine vision systems in almost every measure, and thus a system that emulates object recognition in cortex is very desirable. In the past, there has not been much attention paid to biologically plausible recognition features with higher complexity than using Derivative of Gaussian and Gabor filters [23]. Serre and Poggio have extended Riesenhuber and Poggio's biologically based architecture for object recognition; this HMAX algorithm is founded upon a quantitative theory of the ventral stream of primates' visual cortex [20].

The key element in the HMAX method is a novel set of position and scale-invariant feature detectors that agree with the tuning properties of ventral stream cells and are adaptive to training. HMAX has shown impressive performance on the recognition of objects in cluttered surroundings for multiclass classification.

Standard Model of Visual Cortex

The HMAX algorithm models the feedforward path of object recognition in cortex, which accounts for the first 100-200 milliseconds of ventral stream processing [23]. HMAX is consistent with several generally accepted facts about the ventral stream in visual cortex: (1) visual processing is hierarchical, first generating invariance to position and scale and then to other transformations, (2) both the size of neural receptive fields and the complexity of their optimal stimuli increase along the hierarchy, (3) information processing for immediate recognition is feedforward, and (4) learning and plasticity occur at all stages of the hierarchy.

The model contains four layers of computational units. Simple (S) units combine inputs to increase selectivity, or ability to discriminate between different objects and object classes, whereas complex (C) units combine inputs to increase invariance, or tolerance to transformations such as scaling, translation, and viewpoint changes. The S units combine inputs with a bell-shaped function, whereas C units perform a maximization procedure.

Implementation Details

All images are first converted to grayscale and scaled such that the short edge is 140 pixels in length and the aspect ratio is maintained [15]. Each image is stored at ten different scales, each $2^{\frac{1}{4}}$ times smaller than the last.

The first simple layer (S1) corresponds to V1 simple cells, and is computed by centering 2D Gabor filters with a full range of orientations at every possible position and scale. Thus the S1 layer is a 4D structure, where every position/scale in the 3D structure has multiple oriented units. The Gabor filters are 11×11 pixels in size and are described by:

$$G(x,y) = \exp\left(\frac{x_o^2 + \gamma^2 y_o^2}{2\sigma^2}\right) * \cos\left(\frac{2\pi}{\lambda} x_o\right), \text{ where}$$

$$x_o = x\cos\theta + y\sin\theta \quad \text{and} \quad y_o = -x\sin\theta + y\cos\theta$$

x and y vary between -5 and 5, and θ varies between 0 and π. Appropriate values for γ (aspect ratio), σ (effective width), and λ (wavelength) are taken from Serre and Poggio's 2005 article [24]. Each filter is normalized so that its components have a mean of 0 and the sum of their squares is 1. The response of a group of image pixels X to a Gabor filter G is given by:

$$R(X,G) = \left|\frac{\sum X_i G_i}{\sqrt{\sum X_i^2}}\right|$$

The first complex layer (C1) is modeled after V1 complex cells [15]. The S1 pyramid is convolved with a 3D (10×10 units across, 2 units deep in scale) maximization filter at every orientation. The C1 unit's value is the highest value of any S1 unit that falls within the maximization filter, and the filter is shifted across the S1 pyramid in steps of 5 units across and 1 unit deep in scale.

The S2 layer, which corresponds to the cortical area V4 or posterior IT, is formed upon template matching between patches of C1 units centered at a given location and each of d prototype patches. This matching is performed across all positions and scales of the C1 layer. The d prototype patches are randomly selected from the C1 layers of training images in an initial feature learning stage. The feature learning consists of $n \times n$ ($\times 1$ in scale) patches centered at random positions and scales within the C1 layers of a training image, where n can be 4, 8, 12, or 16. (Multiple feature sizes have been demonstrated useful for effective texture and shape characterization.) After feature learning, every prototype is convolved with a test image's C1 layer, and the S2 pyramid holds d prototype readings for every position and scale within the C1 layer. The response of an image's C1 patch X to a prototype P (size $n \times n$) is given by the Gaussian function:

$$R(X, P) = \exp\left(-\frac{||X - P||^2}{2\sigma^2\alpha}\right),$$

where $||X - P||$ is the Euclidean distance between X and P. The standard deviation σ is often set to 1, and α is a normalizing factor dependent on the patch size n.

The final layer (C2) is a d-dimensional vector, elements of which are the maximum response from the S2 layer over all positions and scales. As a result, full position and scale invariance is attained. The C2 features are then classified using a linear support vector machine (SVM).

Recent Improvements

Mutch and Lowe made four improvements to the original HMAX model for improved functionality and efficiency [15]. First, S2 inputs were reduced to one per C1 position/scale, only using the dominant orientation instead of storing the value for every orientation. This change allowed for the number of Gabor orientation filters to increase from four to twelve without increasing computational burden. The second improvement was also designed to ignore non-dominant orientations: instead of sparsifying S2 inputs, the S1/C1 unit outputs are suppressed to generate a clearer representation of the dominant orientations at each position.

The third improvement limits the position and scale invariance of the model for two reasons: complete invariance is inconsistent with the V4 and IT neurons' lack of complete invariance and leaves the system vulnerable to co-occurrence of features from other objects. The improved model restricts the area of the visual field in which an S2 feature can be found relative to its location in the sample image. Lastly, the improved algorithm drops S2 features with low SVM weight because many features are not necessarily related to the object in question. The desired number of features are selected from an initial set of 12,000 by training the SVM in layers and eliminating up to half of the features each time. This change improves effectiveness and efficiency.

Results

Serre and Poggio trained and tested the "Standard Model" with the Caltech101 database [22], which contains images from 101 categories of objects in a large variety of sizes, positions and orientations [5]. The model outperformed several benchmark algorithms when trained and tested with airplanes, cars, faces, leaves, and motorcycles. The model also outperformed benchmark algorithms when tested with the MIT-CBCL Cars and Faces image sets.

In later work, Serre and Poggio evaluated HMAX performance for object recognition in clutter, object recognition without clutter, and recognition of texture-based objects [23]. For object recognition in clutter, HMAX outperformed several leading algorithms on various Caltech101 and MIT-CBCL datasets. The StreetScenes database was used to test object recognition without clutter; detectors were trained to recognize bicycles, cars, and people. Compared to four leading algorithms, HMAX achieved the best results for bicycles and cars. HMAX was outperformed only by HOG in person detection. HMAX consistently outperformed benchmark algorithms for recognition of four texture-based objects: buildings, roads, skies, and trees.

Mutch and Lowe applied their enhanced HMAX model to the Caltech101 database and the UIUC car dataset, achieving state-of-the-art performance for both image sets [15]. HMAX received the highest average of per-category classification rates from eight independent runs on Caltech101 with fifteen training images per category. For the UIUC car detection/localization task, a sliding window was added to the framework. HMAX achieved the highest precision rates from eight independent runs with single-scale (99.4%) and multi-scale (90.6%) test sets.

3.2.2 HOG

Dalal and Triggs' Histogram of Oriented Gradients (HOG) algorithm aims to discriminate the human form against varied cluttered backgrounds [4]. The authors show that normally localized HOG descriptors form a robust feature set that outperforms other leading feature sets. These descriptors are similar to edge orientation histograms, Scale Invariant Feature Transformation (SIFT) descriptors, and shape contexts, but are computed on a dense grid of uniformly spaced cells with overlapping contrast normalizations for superior performance. Dalal and Triggs used a linear SVM for classification throughout their study.

Overview

This method is based on the premise that object shape and appearance can be sufficiently characterized by a distribution of local edge orientations or intensity gradients, without knowledge of the edge or gradient positions [4]. The procedure evaluates a dense grid of well-normalized local histograms of image gradient orientations.

A given image is divided into small regions, each with a 1-D histogram of edge orientations or gradient directions among that region's pixels. Contrast-normalizing local responses yield HOG descriptors that are invariant to lighting effects. Humans are detected by tiling the detection window with an overlapping grid of HOG descriptors and classifying the combined feature vector with a linear SVM.

The use of orientation histograms reached maturity when Lowe's SIFT approach combined them with local spatial histogramming and normalization, thus providing image patch descriptors for matching scale-invariant keypoints [4]. Dalal and Triggs' study suggests that the leading keypoint-based methods have higher false positive rates than HOG by at least one order of magnitude. No keypoint detector seems able to reliably recognize human body forms. HOG detects humans best with a combination of coarse spatial sampling, fine orientation sampling, and strong local photometric normalization.

Performance Analysis

Dalal and Triggs trained and tested their algorithm with two image sets: the standard MIT pedestrian database (only front or back views of people) and INRIA, which contains 1805 64×128-pixel images of fully visible humans cropped from a set of images [4]. (It is important to note that both of these datasets contain few, if any, images with people occluded by objects or frame boundaries.) For INRIA testing, the initial training set consisted of 2478 positive samples and 12,180 patches sampled randomly from 1218 negative images. After preliminary training, the model was retrained using the initial training set along with any false positives found in the 1218 negative images; this technique, commonly referred to as "bootstrapping," significantly improves performance. HOG generally outperformed the other leading detection methods for both the MIT and INRIA databases.

To find the algorithm's optimal performance for person detection, Dalal and Triggs experimented with input pixel representation, gradient computation, spatial and orientation coarseness, size and normalization of descriptor blocks, detector window proportions, and choice of classifier:

- Performance was reduced when input pixels were reduced to grayscale. As a result, color information is used when available.

- For computing gradients, 1-D discrete derivative masks without smoothing performed best. The larger derivative masks, such as 2×2 diagonal masks and 3×3 Sobel masks, significantly worsened performance.

- Each image pixel generates a weighted vote for an edge orientation histogram channel depending on the orientation of the gradient element centered on that pixel. The algorithm fared best when the vote was simply the gradient magnitude at that pixel, and not some other function of the magnitude. Votes are then accumulated into "orientation bins" over local regions (cells). The votes are linearly interpolated between neighboring bin centers in both orientation and position. Optimal performance was obtained with fine orientation binning and coarse spatial binning.

- Lighting and contrast variations necessitate effective local contrast normalization. Normalization schemes typically group cells into larger spatial blocks and contrast-normalize every block; blocks can be either rectangular or circular log-polar. For the rectangular geometry, 3×3 cell blocks of 6×6-pixel cells performed best. The circular descriptors have been suggested because the transformation field to V1 cortex in primates is logarithmic. Small descriptors with few radial bins gave the best performance. Dalal and Triggs compared four different normalization methods to find the preferred technique.

- Decreasing the background margin (typically 16 pixels around person on each side, for 64×128-pixel images) or increasing the person's size tended to lower detection effectiveness.

- A soft linear SVM trained with SVMLight was used by default. Using a Gaussian kernel SVM improved performance at the expense of a significantly higher runtime.

3.2.3 Viola and Jones' Boosted Cascade

Viola and Jones' Boosted Cascade algorithm (V-J) can efficiently detect the presence of objects in an image after generating a statistical feature-based classifier [26]. Statistical models have greatly aided the world of video surveillance and object recognition in particular, as simple heuristic-based approaches are often insufficient for complex detection. V-J has been proven effective for frontal face recognition and has the potential to aid the more daunting task of detecting human presence in images.

Method Overview

This algorithm primarily relies on an image set used to train the multi-layer classifier [3]. This set must consist of many images that contain the desired object and an abundance of images that do not contain the object. In training, specific features are extracted from the samples with the intention of isolating the features that distinguish a given object from everything else. This information is condensed into a statistical model, each successive layer of which is improved by accounting for known false positives and undetected positives of previous layers.

Features

Haar-like features are fundamental to V-J [3]. Every feature is described by its shape, size, and location relative to the search window. The original Boosted Cascade algorithm used five features, and Lienhart and Maydt extended this set to fourteen, allowing for rotated features [11].

Every feature has black and white rectangles [3]. A feature's value is calculated as a weighted sum of two components: (1) the pixel sum over the black rectangle and (2) the pixel sum over the whole feature. These two components are weighted with opposite sign, and their absolute values are inversely proportional to their respective areas. (e.g., the pixel sum from the black rectangle in a given feature would be multiplied by a factor of 9 if it comprised $\frac{1}{9}$ of the feature's area.)

Integral Image

Because each classifier can contain hundreds of features, computation of pixel sums over multiple rectangles could drastically affect runtime [26]. In order to reduce computation time, V-J creates an "integral image" ii for every image i, where:

$$ii(x,y) = \sum_{x' \leq x, y' \leq y} i(x',y')$$

Thus the pixel sum over any upright rectangle with origin (x,y), width w, and height h is a simple calculation: $ii(x+w,y+h) - ii(x,y+h) - ii(x+w,y) + ii(x,y)$.

Classifier

Each computed feature value $x_\imath$ is given to a simple decision tree classifier that is typically of the following form:

$$f_\imath(x) = \begin{cases} +1 & \text{if } x_\imath \geq t_\imath \\ -1 & \text{if } x_\imath < t_\imath \end{cases} \quad \text{or} \quad f_\imath(x) = \begin{cases} +1 & \text{if } t_{\imath,0} \leq x_\imath < t_{\imath,1} \\ -1 & \text{otherwise} \end{cases},$$

where $f_\imath = +1$ corresponds to the desired object and $f_\imath = -1$ corresponds to everything else [3]. These "weak classifiers" cannot alone determine an object's presence in an image. but rather react to a simple feature that hopefully relates to the object.

Boosting is used to build a complex and robust classifier from the weak classifiers, and a variant of AdaBoost is used to select the appropriate feature sets and train the complex classifier. It has been shown that such a robust classifier can attain an arbitrarily high detection rate and arbitrarily low false positive rate given sufficiently large sets of training images and weak classifiers. V-J chains classifiers together with increasing complexity, and a search window must pass each successive classifier to be classified as a positive sample.

Experiments with face detection have shown that generally 70-80% of search windows are rejected by the first two classifiers, effectively speeding up and making productive use of the detection time.

Achievements

The Boosted Cascade algorithm has been proven effective in the domain of frontal face recognition. Viola and Jones trained a 38-layer cascaded classifier to detect frontal upright faces [26]. The positive training set consisted of 9832 hand labeled faces (4916 faces with their reflections), each resized to a 24×24-pixel resolution. The negative training set was comprised of approximately 350 million 24×24-pixel subwindows selected randomly from 9544 images without faces.

Each classifier in the cascade was trained with all 9832 face samples and 10,000 negative subwindows using the AdaBoost procedure. The initial classifier used 10,000 negative subwindows randomly selected from the vast training set. Each successive classifier used subwindows that the previous classifiers determined to be false positives.

The 38-layer classifier was tested with the MIT+CMU frontal face test set, consisting of 130 images with 507 frontal faces. As it is critical to detect objects with unknown scale, the detector is scaled throughout the image (each scale is a factor of 1.25 larger than the previous scale). Impressive results were achieved by scanning the detector across test images with a step size of 1 pixel [26]. All image subwindows were variance-normalized to minimize lighting effects, and overlapping detections were combined into a single detection.

The Boosted Cascade algorithm features an impressively low detection time. Evaluated on a 700 Mhz Pentium III processor, the face detector processed a 384×288-pixel image in 0.067 seconds on average, which is several orders of magnitude faster than other leading detectors.

Extensions

Viola and Jones found that results were improved at many false positive rates with the addition of two other cascade detectors along with the 38-layer classifier [26]. By picking the majority vote among these classifiers, detection rates were increased by about 1% for most false positive rates. It has been suggested that the improvements would have been greater if the three classifiers were more independent.

Lienhart and Maydt extended the feature set used by Viola and Jones from five features to fourteen [11]. This set that included rotated features was compared with the basic feature set using the MIT+CMU frontal face test set. At comparable detection rates, the extended set demonstrated false positive rates 10% lower on average than the basic set.

3.2.4 Other Methods

Fergus, Perona, and Zisserman's Constellation object recognition framework models objects as flexible assemblies of parts, where the probabilistic assemblies account for shape variation and the possibility of occlusion [6]. Every object model consists of around 3 to 7 parts and each image contains up to 30 "features." Each part is composed of appearance, relative scale, and a binary occlusion factor, all of which are modeled by probability density functions. An object's shape is represented by the relative position of the parts and the Expectation-Maximization algorithm is used to learn each object's parts.

A certain number of interesting features (N) are found in each image. At every point within the image, the intensities within a circular region around the point are determined, and the N regions with the highest saliency (function of the region's entropy) provide the features for learning and recognition. Every test image then contains N features with locations, scales and appearances. The image is classified based on the probability that the object is present given the N features, with every possible feature-parts assignment factored into the probability.

Fergus, Perona, and Zisserman's Heterogeneous Star Model (HSM) is a translation and scale-invariant representation of an object as an organization of parts [7]. HSM is an extension of the Constellation model, which has several shortcomings. Mainly, its exponential computational cost limits the number of parts per model and regions per image that can be dealt with, thus forcing the model to learn from a sparse image representation. HSM's lower complexity for learning and recognition allows it to handle more parts per model and more features per image.

The main reason for HSM's lower complexity is the way in which parts depend upon one another. In the Constellation model, the location of all parts are dependent upon each other; an HSM object has one landmark part upon which all other part locations are dependent. Thus given the landmark part, all other parts are independent from each other. The downside is that the landmark part must be present for an object to be recognized, whereas the Constellation model can detect an object despite any part's absence. HSM also provides a number of different feature types for broader object recognition, where the optimal combination of feature types is determined with an object's initial validation set.

Mikolajczyk, Schmid, and Zisserman's person detection model is able to detect full bodies and close-ups in the midst of clutter and occlusion [13]. As is the case with the Constellation model and HSM, humans are modeled as flexible assemblies of parts, with parts represented as co-occurrences of local features. The human body is modeled as a probabilistic configuration of body parts, with seven parts used: frontal head, frontal face, profile head, profile face, frontal upper body, profile upper body, and legs. The geometric relationship among the body parts is represented by a training-generated Gaussian function. Each body part is represented with orientation-based features and groupings of those features.

The body parts are detected with a cascade-like approach: a succession of strong classifiers where the fastest and most accurate classifiers are applied first. Each strong classifier is an AdaBoost-generated linear combination of weak classifiers. For a given object, a weak classifier is a log-likelihood ratio relating the probability of feature occurrence on the object to the probability of feature occurrence on a non-object, with the probabilities based on feature occurrences in the training data. The final three-step process is performed: (1) individual features are detected at multiple scales within an image, (2) individual parts are detected based on those features, and (3) bodies are detected from configurations of the parts.

Schneiderman and Kanade's system is another parts-based approach. In this system, a trainable object detector determines the presence of faces and cars at any size, pose or location; multiple classifiers are used to cope with different object orientations [21]. Each classifier is based on the statistics of parts, where each part is a transform from wavelet coefficients to a distinct set of values. The statistics of these part values are obtained from positive and negative training examples, and an AdaBoost-trained classifier is used to minimize classification error. The classifier computes the part values within a test window and makes a categorization decision based on a likelihood ratio test derived from the probabilities determined during training. Similar to Viola and Jones' Boosted Cascade system, efficiency is increased by using a series of classification stages.

Gavrila and Munder's PROTECTOR is an integrated detection and tracking system, where detection is performed with a cascade of modules: stereo-based region of interest (ROI) generation, shape-based recognition, texture-based detection, and stereo-based verification [9]. ROI generation acts to reduce the effect of lens distortion away from the image center. In recognition, pedestrians are represented by a series of generated templates that ideally cover multiple people and varying pose and scale. Recognition is also robust to missing features or occlusion. Template matching is based on the chamfer distance transform, where matching a test sample with a template is performed with depth-first search of a template tree.

A supplemental texture-based recognition approach utilizes a richer set of intensity features to distinguish between person and non-person. Interestingly, neural networks were found to outperform SVM with Principle Component Analysis (PCA) for this approach in both detection rate and processing requirements. Finally, pedestrian verification is used to filter out false detections: a pedestrian shape template is applied as a filter for dense cross-correlation. PROTECTOR ultimately uses bounding box position, extent, depth, and their derivatives to track objects.

Bose and Grimson developed a far-field surveillance system for detecting and tracking people and vehicles [2]. The system employs background subtraction and clutter-removing preprocessing. The object-class detector is much simpler than those used in the majority of recognition systems because it only needs to distinguish between foreground objects, and not between object and background.

The low-resolution video often does not provide enough information to accurately determine and detect parts-based features; thus, an object's temporal features are used for classification. SVM was chosen as the classifier, with an emphasis on generating scene-invariant classifiers. This was done by simply training with scene-invariant features as opposed to scene-specific features. Scene-invariant features included orientation, variation in area (second derivative of number of pixels over time) and percentage occupancy (number of silhouette pixels divided by the area of the bounding box). Scene-specific features are helpful for reducing classification error when training and testing in similar scenes, and include location, pixel area, speed, motion direction and aspect ratio. Further adaptation of the classifier is described in their article [2].

Jhuang and Poggio's biologically motivated system recognizes actions of primates in video sequences [10]. The model extends Riesenhuber and Poggio's "Standard Model" for object recognition by modeling motion processing in the visual cortex. The original model categorizes images following the biologically plausible steps that ensure position and scale-invariance. A linear SVM classifier categorizes images based on their C2 features. C2 features are obtained by computing the global max of each of the S2 feature maps. S2 feature maps derive from the template matching performed from the C1 maps, which are computed as the global max over the S1 feature maps.

The key change between Jhuang and Poggio's model and the original is in the S1 feature representation; S1 features are now sensitive to motion direction. The motion-based S1 features can be computed in three different ways: (1) spatial gradients along the horizontal and vertical axes, as well as the temporal gradient across successive image frames, as inputs to S1 features, (2) optimal flow-based S1 features, and (3) 3D space-time filters. Experiments showed that S1 features using space and time gradients and those using 3D space-time filters saw the most consistent success.

Viola, Jones, and Snow developed a person detection system that integrates intensity information with motion information [27]. This implementation is able to detect people at very small scales (down to 20-pixel height) and has demonstrated success on low-resolution images under challenging environmental conditions. However, this system is unable to detect occluded or partial figures.

The detector is based on the simple rectangle filters used by Viola and Jones' Boosted Cascade system. Motion information is extracted from successive image information, with five "shifted images" generated: one temporally shifted image and four directionally and temporally shifted images (motion images). There are three types of filters that operate on these shifted images: (1) comparing sums of absolute differences between images, (2) comparing sums within the same motion image, and (3) measuring the magnitude of motion within the motion image.

The AdaBoost learning algorithm chooses from the range of motion/appearance filters to construct the optimal classifier for separating positive and negative samples. Scale-invariant detection is performed, and a boosted cascade approach is used: each successive stage acts to decrease both the detection rate and the false positive rate, where ideally the latter decreases faster than the former.

3.3 Algorithm Attributes Summary

Many object recognition algorithms were considered over the course of this project. While only three methods were used to test the novel indoor and aerial datasets, it is important to dissect every examined algorithm with regard to key desirable attributes. Table 3.1 gives a comprehensive summary in which each row corresponds with a benchmark algorithm and the attributes are listed by column. For the algorithms that were not used for testing, properties were established based on the results and claims made in the respective articles. The attributes of HMAX, HOG, and V-J were verified through testing. Ultimately, this section is intended to assist future researchers and developers, whether looking for a quick overview or attempting to locate the algorithm with the right mix of attributes.

	Attributes						
Algorithms	Fast Training	Fast Testing	Eye-Level Success	Aerial Success	Broad Range	Occluded Success	Human Focus
Poggio "HMAX" [20]	N	N	Y	Y	Y	Y	N
Dalal "HOG" [4]	Y	Y	Y		Y	Y	Y
Viola (V-J) [26]	N	Y	Y		Y	Y	N
Fergus "Constellation" [6]	N	N	Y			Y	N
Fergus "HSM" [7]			Y			Y	N
Mikolajczyk [13]			Y		Y	Y	Y
Schneiderman [21]		N	Y		Y		N
Gavrila "PROTECTOR" [9]		Y	Y		N	Y	Y
Bose & Grimson [2]				Y	N		Y
Jhuang & Poggio [10]		Y	Y				N
Viola, Jones & Snow [27]		Y		Y	Y	N	Y

Table 3.1: Overview of algorithm attributes, where Y signifies that the attribute has been demonstrated/claimed, N signifies that the contrary has been demonstrated/claimed, and no marking signifies that there has been no definitive determination.

The attributes given in Table 3.1 are defined as follows:

- An algorithm possesses "fast training/testing" if its current implementation is ready for real-time use. **Fast Training** means that retraining with a sufficiently large image set (over 100 images) can occur within a reasonable time frame (under 10 minutes). **Fast Testing** is defined as the ability to process a typical 640×480-pixel image in under one second.

- **Eye-Level Success**: Demonstrated correct-classification rates of at least ~90% for objects at eye-level (e.g., indoor testing).

- **Aerial Success**: Demonstrated correct-classification rates of at least ~90% for objects with altitude at least 50 feet below camera altitude.

- **Broad Range**: Adaptable to both close-range and far-field object detection.

- **Occluded Success**: Demonstrated successful detection of objects less than 75% visible. HMAX, HOG, and V-J were given this attribute due to their ~90% correct-classification rates for the Indoor Video Set samples that included occluded people. However, this attribute was not explicitly verified because the impact of occlusion was not statistically examined.

- **Human Focus**: Algorithm is intended and/or specifically tuned for person detection.

Chapter 4

Image and Video Testing

4.1 Methodology

4.1.1 Algorithm and Dataset Concerns

Algorithm testing was performed to establish an indication of how well the algorithms will perform in real-world scenarios and to provide a fair comparison of the methods. A fair comparison necessitates consistency among both the datasets used and how they are used to train and test each algorithm. Several steps were taken to ensure this consistency.

For every test, there were several ways in which the selected algorithms needed to be consistently applied. First, the same training and test sets were used for each algorithm. Additionally, emphasis was given to ensuring that all algorithms were run with the techniques instrumental to any single algorithm's success. For instance, HOG is trained with bootstrapping, whereby secondary training is conducted with negative samples that are incorrectly classified by the initial training. Though specific bootstrapping implementation varied among the algorithms, it was important that all algorithms were trained with bootstrapping. In addition, some algorithms thoroughly scan each test image to classify subwindows, whereas others classify the entire image. These inconsistencies were removed upon testing.

For every algorithm, measures were taken to ensure the legitimacy of test results.

Algorithm restrictions regarding image size and proportion were considered when generating the image sets. For instance, HOG and V-J must be trained with uniformly sized positive samples; the aspect ratio of all generated positive samples was thus held constant. Also, because the HMAX implementation classifies the entire image rather than classifying subwindows, its performance is predicated on the need for consistent size among all positive and negative training/test images.

It was also essential that training and test image sets were as independent as possible. Real-world testing will likely occur in several disparate locations with conditions different than those during training. Measures taken to ensure this independence between the training and test sets generated for the indoor and aerial applications are discussed in Sections 4.2.2 and 4.3.2.

Another priority was to train algorithms with datasets large enough to ensure valid results. Dalal and Triggs trained their HOG model with over 1000 positive person samples. Mutch and Lowe trained the HMAX model with 30 samples from each of 101 object categories. Viola and Jones used about 5000 positive and 10,000 negative subwindows for training. Generating appropriate datasets was essential for this study; because only so many images could reasonably be generated, training was sometimes performed with fewer samples than potentially necessary for optimal results. With larger training sets, algorithm results will likely improve. Test set size was also an important consideration, as sufficient sample size is necessary for valid conclusions. (One way to achieve statistical validity with smaller datasets is to average results over multiple random or stratified training/test splits. This technique would have certainly substantiated the test results in this thesis; unfortunately, time constraints precluded its use.)

The tests conducted in this thesis are "system-level" comparisons of the algorithms. This means that each algorithm was implemented with non-feature characteristics (such as the bootstrapping technique or the classifier) that were specifically chosen by the algorithm's developer. (The exception is HMAX bootstrapping, for which pseudocode is given in Appendix C.) As a result, no conclusions can be made about which algorithm has superior features, or which particular characteristics were responsible for any given results. Instead, these tests are intended to show readers which systems work well given a particular image domain.

4.1.2 Algorithm Implementations

This section briefly describes how each algorithm was implemented for both indoor and aerial testing. Detailed descriptions of each algorithm are given in Section 3.2.

Mutch and Lowe's enhanced version of the "Standard Model" of cortex-like object recognition (HMAX) was used for testing. For each test, C2 features were classified using a linear SVM from Franc and Hlavac's Statistical Pattern Recognition Toolbox for Matlab [8]. For every training set, 1500 features were selected from among 12,000 random C2 features using successive layers of feature elimination.

The HOG algorithm was implemented with the following parameter values:

- Detector Window Size: most often 64×128 pixels
- Cell Size: 8×8 pixels
- Block Size: 2×2 cells
- Number of Orientation Bins: 9
- Descriptor Stride in Window: 8×8 pixels

These parameters are fully described in Dalal and Triggs' article [4]. A soft linear SVM trained with SVMLight was used for classification.

Viola and Jones' statistical algorithm for face detection (V-J) was retrained for person detection using OpenCV.

4.1.3 Accuracy Conventions

There are several standard methods for characterizing and comparing the accuracies of given algorithms. Generally, each algorithm is trained and tested with the same image sets, and an algorithm's accuracy is defined by two values: detection rate (DR) and false positive rate (FPR). These rates are defined as follows:

$$DR = \frac{TP}{\#P} \quad \text{and} \quad FPR = \frac{FP}{\#N}\ , \text{ where}$$

$$\begin{aligned}
TP &= \text{number of positive test images that are correctly classified} \\
\#P &= \text{total number of positive test images} \\
FP &= \text{number of negative test images that are incorrectly classified} \\
\#N &= \text{total number of negative test images}
\end{aligned}$$

One of two main approaches is typically used to quantify classification accuracy. The first is to plot a Receiver Operating Character (ROC) curve for each algorithm. The horizontal axis corresponds with false positive rate and the vertical axis corresponds with detection rate. A curve is then generated by varying the classifier threshold that acts to distinguish between positive and negative classification.

Many of the leading object recognition methods, including HMAX and HOG, use a Support Vector Machine (SVM) to classify query images. SVMs are linear classifiers that separate the p-dimensional training data points with a $(p-1)$-dimensional hyperplane. A test image's SVM output, which is related to the distance between the image's data point and the separating hyperplane, is usually used as a proxy for classification confidence. In a two-class SVM, all test images with SVM output above a certain threshold are placed into one classification category, whereas all images with output below the threshold are placed into the other group.

Sometimes various aspects of the ROC curve are used for characterizing accuracy, such as the area under an ROC curve or the "equilibrium point," where false positive rate equals miss rate. Miss rate is the opposite of detection rate, and is the number of incorrectly classified positive samples divided by the total number of positive samples. The terms "correct-classification rate" and "success rate" are used in this thesis to denote the detection rate at the equilibrium point. To compare two algorithms, points from the ROC curves will often be compared; the preferred method has the higher detection rate for the same false positive rate.

The other main approach taken to determine accuracy is to plot a Recall-Precision curve for each algorithm. The horizontal axis corresponds with Precision, or the percentage of the images classified as positive that are actually positive. The vertical axis corresponds with Recall, the equivalent of detection rate. Like the ROC curve, points from the curve are often used to measure accuracy. There are other variations to the ROC and Recall-Precision curves, many of which deal with absolute numbers instead of rates.

This study is concerned with determining the presence of one or more people in an image. As a result, the goal is to classify the entire image, not to classify subwindows of the image. Thus if any subwindow is classified as positive, then the entire image is labeled positive. This is a noteworthy distinction because of the following implication: false positive subwindows in positive images will not lower the algorithm's defined accuracy. This is justified because false positive subwindows in negative images will decrease the accuracy. In other words, if there are a significant number of false detections, it stands to reason that they will be distributed between both the positive and negative images, thus lowering accuracy.

The standard ROC curve is an appropriate selection for this study. The ultimate goal is to obtain a perfect classification curve, where a range of classifier thresholds exist with which every positive sample is correctly classified and there are no false positives. Because this is highly unlikely, the realistic goal becomes consistently correct classification. Acceptable rates of correct classification are subjective, but the goal of this study is to obtain high correct-classification rates (preferably above 90%) for a sufficiently large sample set of query images. In order for the results to be generalized to real-world scenarios, there should be high statistical confidence that new images will be correctly classified at a rate that is hopefully around 90%.

Statistical convention dictates that if a percentage P of the test images (assumed to be a random sample of all appropriate images) are classified correctly, then a different, randomly selected image will be correctly classified at least $(P - \frac{Z}{2\eta})$ percent of the time at confidence C. Z depends on C (e.g., $Z = 1.9599$ for $C = 95\%$), and η is the number of images in the test set.

4.2 Leave-Behind Sensing

4.2.1 Problem Characteristics

There are a few characteristics of the leave-behind sensing problem that help to shape both the expectations of successful person detection as well as the process of achieving this. First, indoor surveillance necessitates that people are detected at a range of distances from the camera. In this sense, indoor detection is more complex than detection from the air, where all people have much more consistent size and pose with respect to the camera. Because of the multitude of detection ranges and the fact that close-range surveillance provides for a large range of human pose, achieving consistently high accuracy from a given training set may be more difficult for indoor sensing.

One major benefit to indoor object detection is the ability to use a stationary camera. This allows the possibility of using established motion detection algorithms as a first pass for specific object recognition techniques. Because a person must move into the frame to be detected, effective motion detection will determine nearly every instance where a person is present. Complex detection techniques are still necessary to protect against false positives, but using these techniques alone will not be as cost-effective as coupling them with motion detection.

A prominent yet simple technique for detecting motion is background subtraction. First, the difference in pixel values between a current video frame and a background frame (in which no people are present) is taken. Then, any region in the current frame where that difference is large enough is regarded as evidence of motion. Of course, this "motion" may have nothing to do with a person: lighting changes, shadow movements, and other object movement are some examples.

The algorithm written to detect motion has several components (pseudocode is provided in Appendix B). For each current frame, background subtraction is applied to produce a "difference frame" with pixel values equal to the difference between the pixel values of the current and background frames. Pixel values range from 0 (black) to 255 (white). Pixels with values above a selected threshold are then turned to white, whereas those with values below that threshold are turned to black. Every motion region, or region with white pixels, is cropped for testing, eliminating the need to scan every frame for test crops. These crops are then tracked from frame to frame, removing the need to test the same object more than once.

4.2.2 Datasets

Every image set used to conduct leave-behind sensing tests is described in this section. Table 4.1 gives the general characteristics of each dataset.

Indoor Image Set

The primary goal upon generating the Indoor Image Set was to collect four groups of images: (1) training images containing a person, (2) training images not containing a person, (3) test images containing a person, and (4) test images not containing a person. Restrictions were then placed on these datasets for the sake of consistency. First, it is important that training images are as independent from test images as possible, because real systems will most likely be trained with conditions independent from those in which testing will occur. Second, positive and negative images were to be different only by the presence of a person, and not by the presence of any other object. For this reason, every positive image has a corresponding negative image, where the only change is the presence of a person. In other words, every non-person object is present in equally as many positive images as negative images.

Dataset	**Type**	**# Pos.**	**# Neg.**	**Locations**
Indoor Image Set	Training Images	100	100	One Kendall Draper Bldg.
	Test Images	100	100	Guggenheim Lab, Koch Bldg.
Indoor Image Set	Training Crops	112	500	One Kendall Draper Bldg.
	Test Crops	114	500	Guggenheim Lab, Koch Bldg.
Indoor Video Set	Training Crops	284	162	Stata Center
	Test Crops	481	154	77 Mass. Ave. Dome

Table 4.1: Characteristics of each indoor image set: number of positive images, number of negative images, and locations.

The Indoor Image Set is comprised of 400 480×640-pixel images, containing 200 training images and 200 test images. The training and test sets each contain 100 positive and 100 negative images. Sample positive and negative images from both the training and test sets are shown in Figures 4-1 and 4-2.

In order to make the training and test groups independent, several steps were taken. First, all training images were taken in different locations than the test images. The training images were taken in the One Kendall Draper Building in different offices, hallways and conference rooms. The test images were taken in classrooms in two MIT buildings: the Guggenheim Laboratory and the Koch Biology Building. There were independent groups of people between training and testing, and the two sets were taken on different days. There are seemingly endless ways to make the groups independent, and thus improve the validity of the sets, like

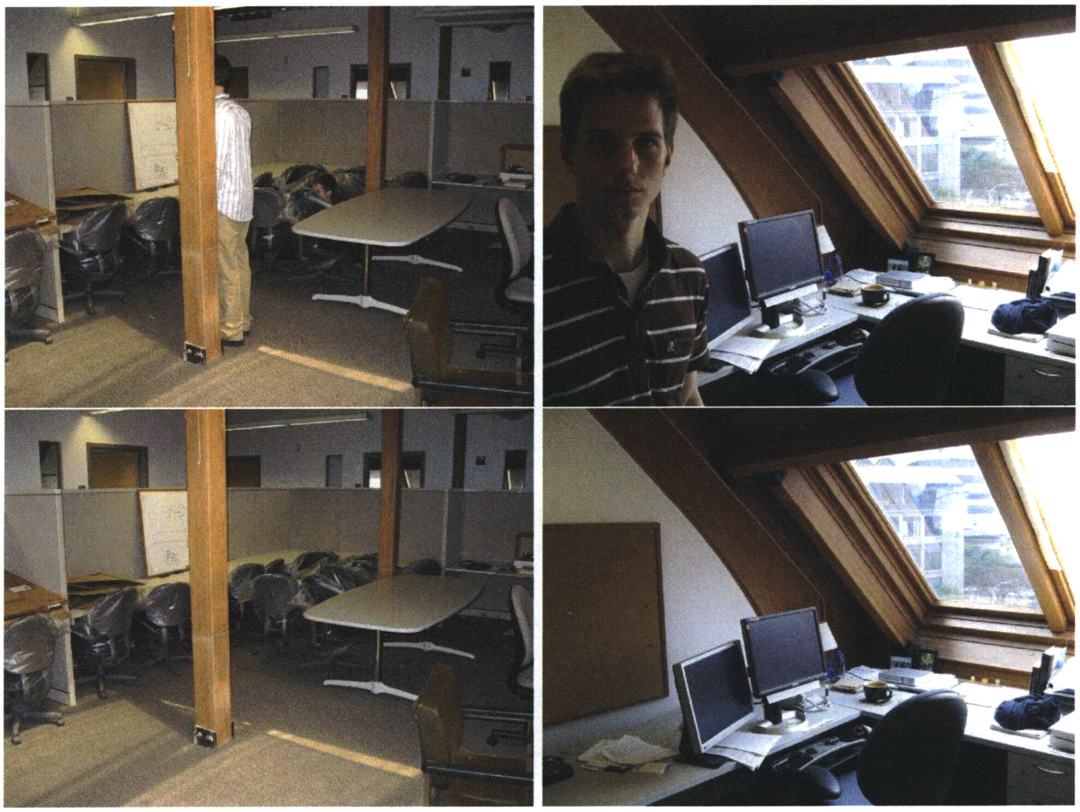

Figure 4-1: Sample positive and negative indoor images from the training set.

changing the camera/camera settings between training and testing, or even changing the campus/area between the groups. However, the measures taken were adequate for generating an appropriate dataset for this study.

As this image set was designed to test for a real-life indoor surveillance system, it was essential to vary the positions and orientations of the people visible in the images, as well as their distances from the camera. In order to reduce bias, there was also an emphasis on varying the gender, race, clothing, and apparel of the people in the images. Because a reliable system would detect the presence of people in any condition, this image set contained training and test images with people partially occluded by other objects and only in part of the image frame. Also, the room lighting was often varied, as was the position of the camera within the room. This range of variations makes the Indoor Image Set considerably more difficult than commonly

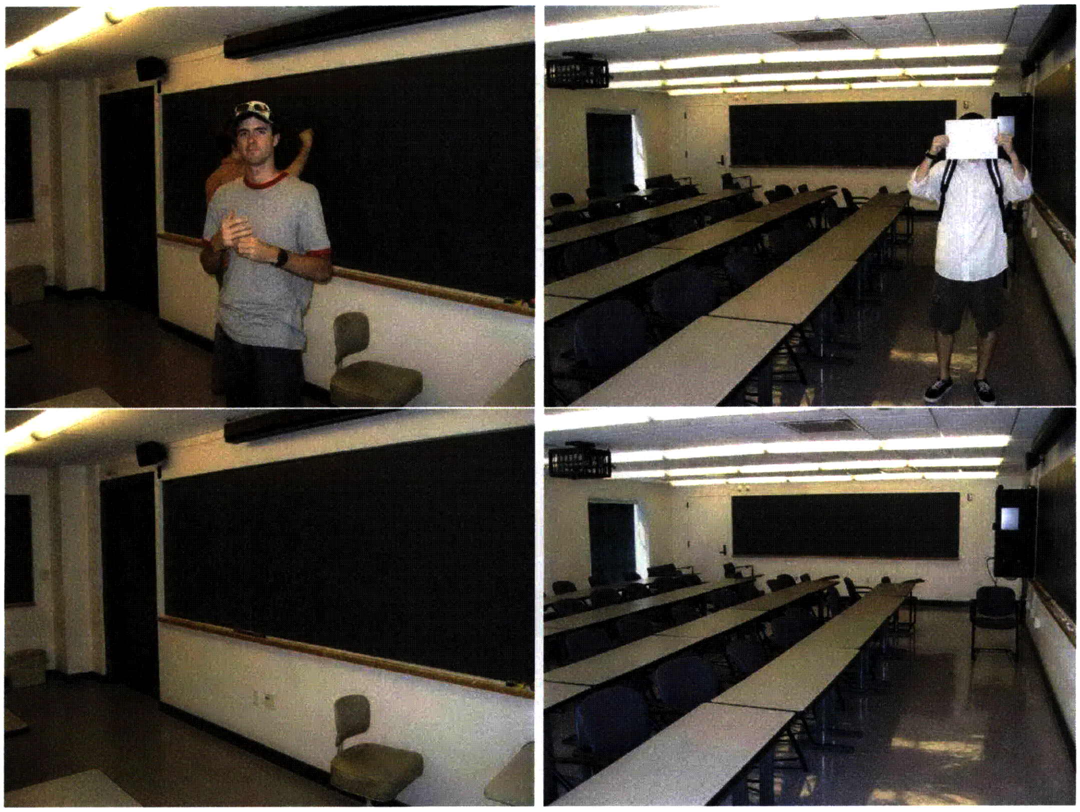

Figure 4-2: Sample positive and negative indoor images from the test set.

used sets for testing, such as MIT Pedestrian Database and the INRIA Dataset.

From the 400 images in this set, another image set was generated that contained only cropped subsamples (crops) of the original images. There were 112 crops of people taken from the 100 positive training images and 114 crops of people taken from the 100 positive test images. Each crop has a fairly consistent background border surrounding the person, and every training crop has a 1:1.5 aspect ratio. (HOG and V-J require a consistent aspect ratio among all training positives.)

There were 500 negative crops taken randomly from both the negative training and test image sets, with random width between 60 and 200 pixels and height between 100 and 300 pixels. Figure 4-3 shows sample crops from the training set and Figure 4-4 shows sample crops from the test set.

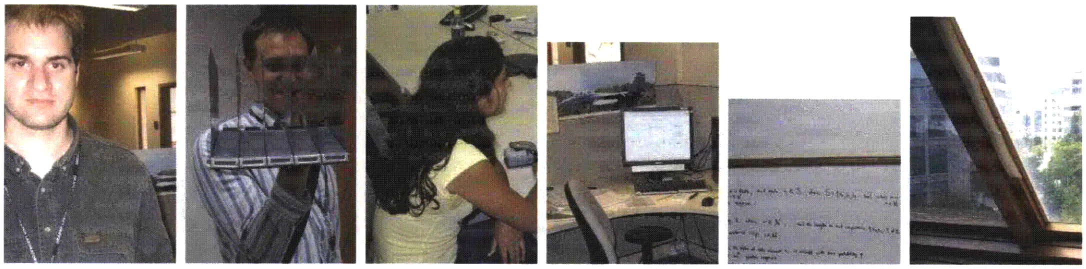

Figure 4-3: Sample positive and negative indoor image crops from the training set.

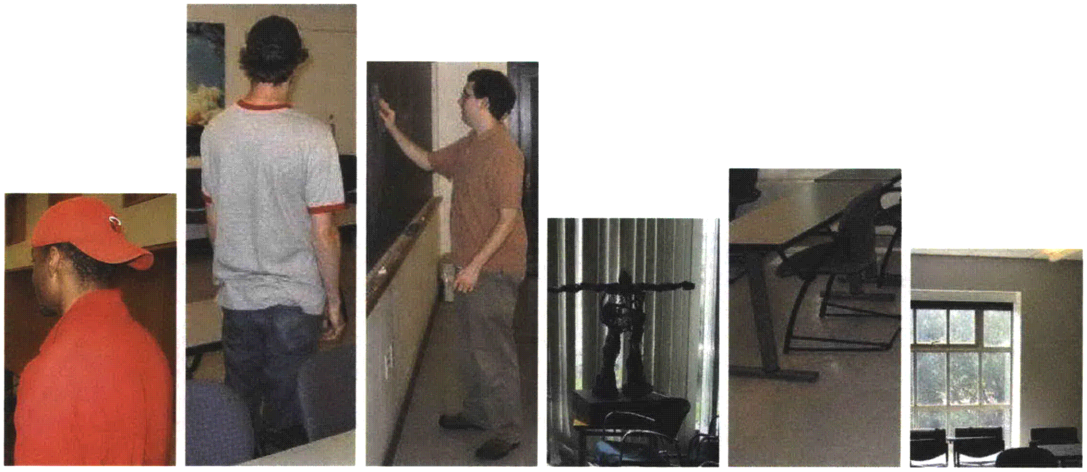

Figure 4-4: Sample positive and negative indoor image crops from the test set.

Indoor Video Set

An Indoor Video Set was obtained by taking 640×480-pixel video at 30 frames per second in multiple locations on the MIT campus. For training data, three videos (totaling 3 minutes and 34 seconds) were taken with different views from within the Stata Center. The test set was comprised of five videos (totaling 5 minutes and 46 seconds) taken with various views from inside the 77 Massachusetts Avenue dome. While many indoor settings naturally yield few false positives (the vast majority of moving objects are people), the Indoor Video Set locations were chosen to include non-person motion, such as shadows, moving doors, or outside vehicles.

The motion detection algorithm generated a 1:2-aspect ratio crop around any object that was significantly different than a background frame. Thus any crop wider than 240 pixels was rejected because its height would exceed the available 480 pixels. (This is not an ideal approach for a true detection system; in later sequence testing, each wide crop was subsampled with 1:2-aspect ratio crops.) All crops were uniformly resized before training or testing. By taking crops from every tenth frame, 483 training crops were generated and 803 test crops were generated. These crops were then manually placed into one of four categories: (1) the crop contains a person and it is apparent without context clues, (2) the crop does not contain a person and it is apparent without context clues, (3) the crop contains a person and it is apparent only with context clues, and (4) it is not apparent whether the crop contains a person. It is not fair to train or test with crops from categories (3) and (4) because the algorithms cannot interpret the context clues that humans instinctively use. There were 284 positive crops (first category) and 162 negative crops (second category) for training and 481 positive and 154 negative crops for testing. Sample crops from the training and test sets are shown in Figures 4-5 and 4-6, respectively.

The crops generated from the indoor images and video frames are in many ways more difficult than many of the standard image sets. Compared with these standard sets, the indoor image and video crops contain people with a much larger variation in pose. People in the indoor sets can be found standing, sitting, walking, leaning, crouching, or even laying down. By contrast, the MIT-CBCL Pedestrian database only contains views of upright persons centered in each image [17]. Another difficult aspect is the presence of heavily occluded people; in a sizable portion of the crops from both images and video, people are partially to mostly occluded or are only partially in the image frame. Many standard image sets only contain positive images in which people are mostly visible. For example, the StreetScenes database contains only people who are at least "75% visible" [1].

The indoor video crops are a particularly challenging set. Both the training and test sets contain images without a consistent background border (margin around the person). The INRIA dataset used by Dalal and Triggs is one of many sets with such consistency; each person has around 16 pixels of margin on all four sides [4]. The set of human faces used by Viola and Jones also has very consistent object-to-background

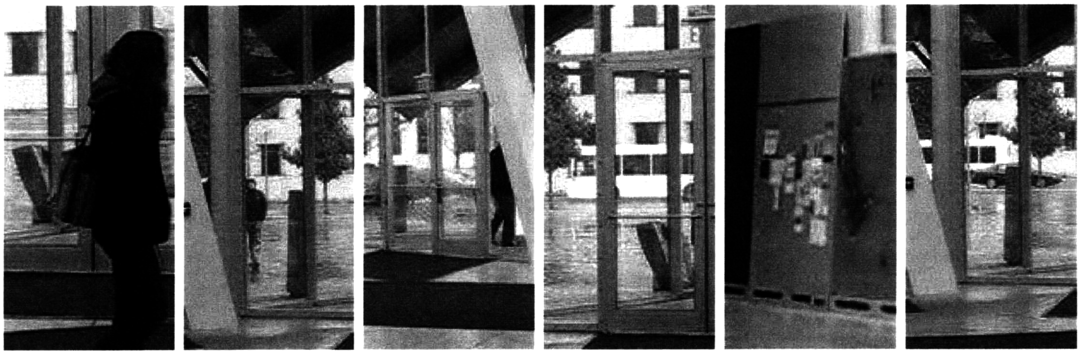

Figure 4-5: Sample positive and negative indoor video crops from the training set.

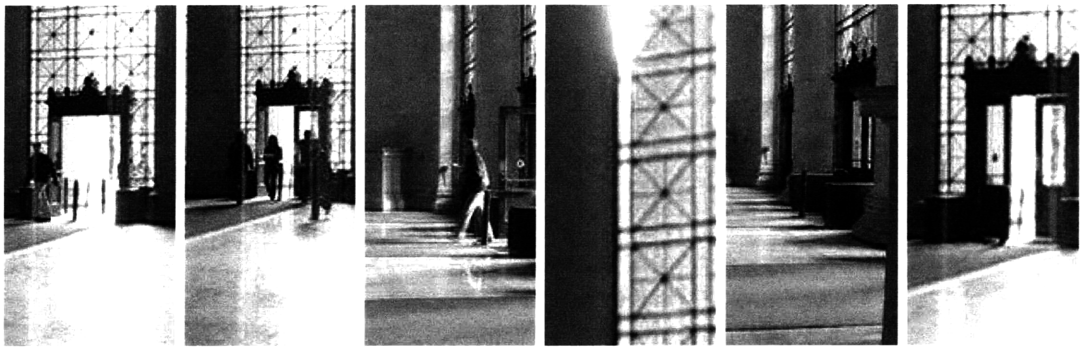

Figure 4-6: Sample positive and negative indoor video crops from the test set.

proportions. In addition, the indoor video crops contain close-ups of people inside as well as far-field views of people outside. There are very few, if any, standard image sets with such a situational variety among the training and test images. The goal in generating these image sets was to encompass the variety of situations in which people can be found, and the results of testing with these sets must be viewed not only in comparison to other studies, but also within the context of these complexities.

4.2.3 Test Progression

It was essential to apply the datasets fairly and consistently when training and testing the selected algorithms. Designing the appropriate types and amount of testing is an inexact science, so emphasis was placed on conducting a wide range of image categorization tests. Some test procedures are ultimately feasible, whereas others serve more as proof of concept. Certain procedures are more convenient to perform than others. There are many tradeoffs involved, as the most convenient and feasible procedures are often met with lower expectations and ultimately lower performance. These tradeoffs provide the context for the following discussion of the implementation and results of this study's most significant tests. Table 4.2 gives an overview of

	Test Datasets			
Training Datasets	Full Images	Image Crops	Video Crops	Video Crop Sequences
Full Images	HOG: 70% HMAX: 65% (No Boot.)			
Image Crops		HOG: 90% HMAX: 88% (No Boot.)		
Video Crops			HOG: 89% V-J: 85% HMAX: 83%	HOG: 95% (Sect. 5.3)

Table 4.2: Overview of indoor tests organized by training and test datasets. Each algorithm's approximate correct-classification rate is listed, with the best performance highlighted. All tests are described in this section unless otherwise noted. The HMAX bootstrapping technique was implemented for all tests unless otherwise noted.

the indoor tests discussed in this thesis. (For a complete list of all tests conducted throughout this study, see Appendix A.)

Training and Testing with Full Indoor Images

Many object recognition methods employ a similar strategy: an object is recognized by testing an exhaustive list of subwindows within an image. While this approach has been successful with standard image sets, there is appeal in avoiding subwindowing and instead testing the entire image at once. Testing hundreds of windows just to classify one image is both cumbersome and costly, where the majority of tests are ultimately unnecessary.

Training algorithms with full positive and negative images and testing with entire images is an extremely convenient and feasible procedure. The first major test was conducted with 100 full indoor images in each of the four categories: positive training, negative training, positive testing, and negative testing (sample images are shown in Figures 4-1 and 4-2). Only HMAX and HOG were used for this test because V-J is not meant for training upon and categorizing 480×640-pixel windows (a typical size

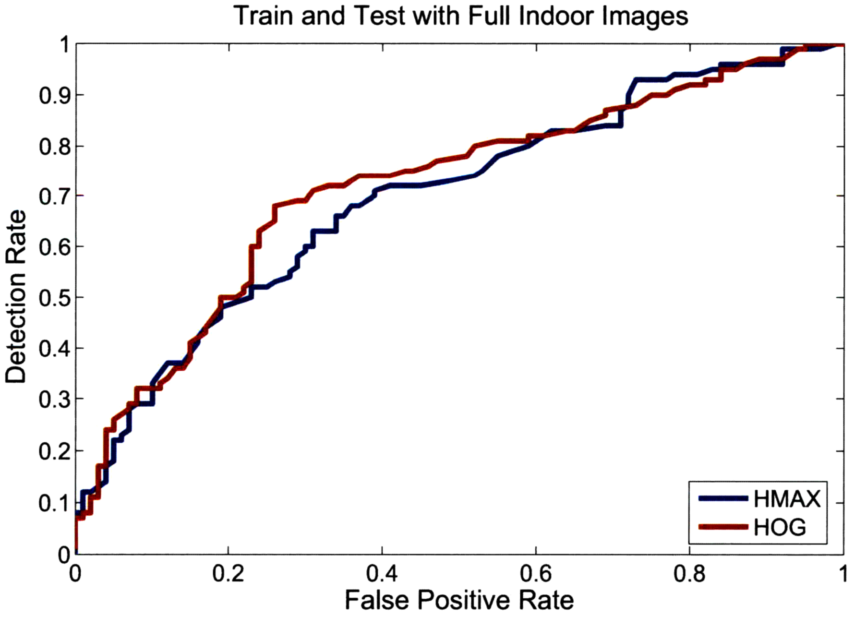

Figure 4-7: ROC curves for training and testing with full indoor images.

is 20×20 pixels).

Though HOG outperformed HMAX, correct-classification rates of around 70% (as shown in Figure 4-7) are unreliable by any realistic standards. Not surprisingly, there is good reason for using the subwindowing approach instead of categorizing entire images. Small windows mostly occupied by the object intuitively provide a more effective positive training set than large images mostly occupied by the background. The next tests were conducted with crops from the Indoor Image Set.

Training and Testing with Indoor Image Crops

The 112 positive training crops and 114 positive test crops were taken manually from the indoor images. As for the negatives, there were 500 training and 500 test crops taken randomly from the respective image sets. (Sample crops from the training and test sets are shown in Figures 4-3 and 4-4, respectively.) The algorithms were trained and tested with these image sets, and results are impressive despite not using uniformly sized crops or enhancing HMAX with the bootstrapping technique.

Figure 4-8 shows that both HMAX and HOG achieved very promising correct-

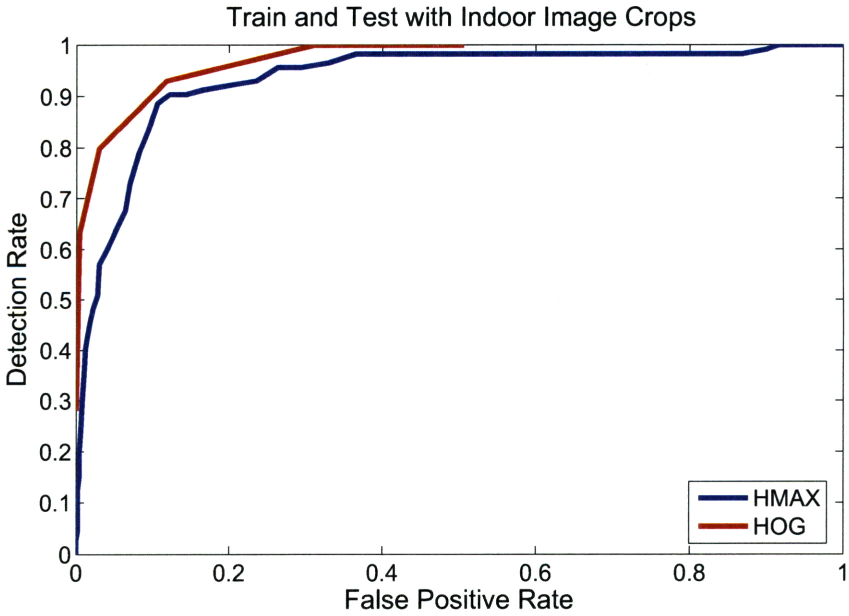

Figure 4-8: ROC curves for training and testing with indoor image crops.

classification rates of around 90%. However, these results are deceiving for two main reasons. First, while manually cropping people for training is acceptable, doing the same for testing is infeasible. Fortunately, simple image subtraction-based techniques help to generate crops of moving objects in video frames. These crops will not be as uniform as the manual crops, but feasibility will have been attained. The second misleading part of this procedure was testing with random negatives; it is possible that the algorithms simply distinguished between background-bounded objects (i.e., those well-centered in the image with a relatively uniform background frame) and random patterns, not between people and non-people. (In fact, the HMAX correct-classification rate decreased by around 20% when cropped non-person objects were used as training and test negatives.) An appropriate alternative is for the subtraction-based technique used to generate the test positives to also generate the test negatives.

Training and Testing with Indoor Video Crops

Using the crops generated with simple image subtraction for training and testing is convenient and feasible. In fact, the following results are essentially a lower bound on the potential of this process, as more sophisticated motion detection techniques would inevitably provide cleaner crops. The indoor training video generated 284 positive and 162 negative crops (samples are shown in Figure 4-5), and each algorithm was trained in a similar manner.

The bootstrapping technique was used with all algorithms, whereby initial training is followed by training on samples that were initially false positives. The 284 positive samples, along with their horizontally flipped counterparts, formed the 568 positive samples for HMAX, HOG, and V-J. For HMAX, the positives were resized to 64×128 pixels, and the 162 negative crops were resized to 128×256 pixels. The initial training set for HMAX consisted of 568 positives and 810 negatives (five random 64×128 windows from each of the 162 negative crops). Then, 2430 negative crops (15 per negative crop) were tested by the initial classifier, and the 400 most positively classified samples (the false positive candidates) were added to the negative training set for secondary training on 568 positives and 1210 negatives, all 64×128 pixels.

Pseudocode for this technique is provided in Appendix C.

HOG was initially trained with the same 568 positive and 810 randomly chosen negative samples, and subsequently trained with false positives from an exhaustive set of negative windows. V-J was given an initial training set of 568 positive and 810 randomly chosen negative samples, and performed nineteen layers of retraining, each with the same 568 positives and a maximum of 810 false positives from exhaustive scanning of the 162 original negative crops. (For computational reasons, all crops used for training V-J are 32×64 pixels instead of 64×128 pixels.)

Each algorithm was tested with the 481 positive and 154 negative crops generated from the video test set, all consistently sized with the training crops (64×128 pixels for HMAX and HOG, 32×64 pixels for V-J). Figure 4-6 shows sample positive and negative crops from the test set.

Figure 4-9 shows that all three algorithms achieved successful categorization rates for the video-generated crops. HOG, which demonstrated a correct-classification rate of 88%, outperformed HMAX and V-J. This success rate was statistically extended (using the methodology in Section 4.1.3) to account for the broad range of potential

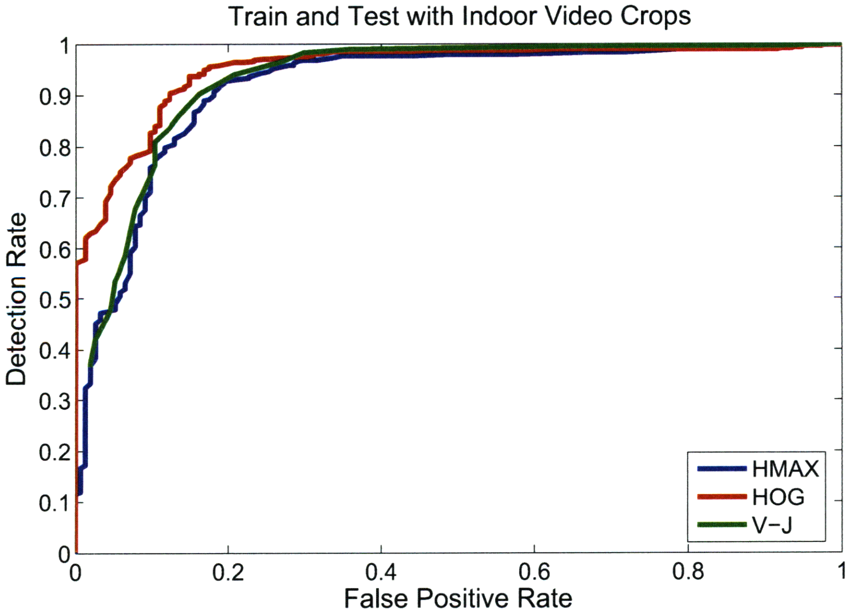

Figure 4-9: ROC curves for training and testing with indoor video crops.

test samples. Assuming that the 635 classified test samples comprise a random selection of similar indoor video samples, it can be said that new test images will be correctly categorized more than 84% of the time at 95% confidence.

4.3 Aerial Surveillance

4.3.1 Problem Characteristics

Successfully determining the presence of people on the ground from aerial imagery is difficult for several reasons. Many leading object recognition algorithms show impressive results for objects of a certain (pixel) size, but are untested for the inevitably smaller sizes of people in aerial images. While accuracy may suffer as the object size decreases, there is an increase in the conformity of samples, since greater distance between camera and person decreases the variability of the person's pose with respect to the background. These aspects of far-field aerial detection are worth considering when training and testing the numerous benchmark methods.

Another difficult aspect of detecting people from aircraft imagery is the camera's instability. Unlike the leave-behind sensing problem, aerial detection will be done from a moving platform. This means that any tool used to improve the runtime of a detection system, such as a background subtraction or filtering-based approach, must be enhanced before being applied to the aerial domain. Image registration is a standard technique for placing successive images from a moving platform onto one common reference frame [19]. Applying this type of technique to the shaky succession of images would enable the use of background subtraction or an alternative form of motion detection. While more effort is required, the benefits of adding motion detection to the recognition process can still be achieved.

While using image registration is a viable option, the method was not applied in this work. Because image registration is not the focus of the thesis, selecting and applying such an algorithm to the small sample of available aerial video was deemed unnecessary in the presence of a viable alternative. Instead, simulated aerial video was taken with a stationary camera atop the Draper parking garage and then divided into training and test sets. This allowed for training the detection models on an image set that is independent from the UAV video test set, and for testing with the aid of background subtraction.

4.3.2 Datasets

For the aerial surveillance problem, the initial goal was to use a small UAV to collect enough video capturing people from a suitable altitude in order to create the same groups as in the Indoor Image Set (positive/negative training and testing). The available footage from the UAV consisted of about twenty minutes of video at 30 frames per second, generating about 30,000 still images. While this number of images should have been sufficient, this dataset was ultimately inadequate for both training and testing. Only about 30 seconds of the video contain definitive human presence and there are only a few different people in the video. Because the training and test sets should be independent and because of the sparsity of samples, the UAV video alone did not suffice for both training and testing.

One way to generate more images for aerial surveillance training and testing was to simulate aerial images with those taken from rooftops. The major concern was making sure that images taken from rooftops are appropriately consistent with those taken in authentic aerial surveillance missions. It was essential that the pixel-size of persons in the simulated images correspond with the pixel-size of persons in authentic imagery. The characteristics of a typical small camera were used to get an estimate of authentic pixel-size. Using a 20° vertical field of view and 30° inclination angle from an altitude of 100 feet, a 6-foot person would be 37 pixels high on a 640×480-pixel display. While that altitude is very low compared to that of a typical surveillance mission (about 500 feet in altitude), cameras chosen specifically for far-field surveillance would likely yield more information than the typical small camera. Using rooftop imagery with people ranging from 40 to 100 pixels for training and testing was deemed appropriate.

This section describes every image set used to conduct aerial surveillance tests. Table 4.3 gives the size and location of each image set.

Dataset	Type	# Pos.	# Neg.	Locations
Aerial Image Set	Training Images	101	100	Draper Parking Garage
	Test Images	100	100	
Aerial Image Set	Training Crops	100	500	Draper Parking Garage
	Test Crops	100	500	
Aerial Video Set	Training Crops	195	708	Draper Parking Garage
	Test Crops	300	474	
UAV Video Set	Test Crops	110	500	Fort Devens, Mass.

Table 4.3: Characteristics of each aerial image set: number of positive images, number of negative images, and locations.

Aerial Image Set

Similarly to the Indoor Image Set, the Aerial Image Set is comprised of training positives, training negatives, test positives and test negatives. There are 401 480×640-

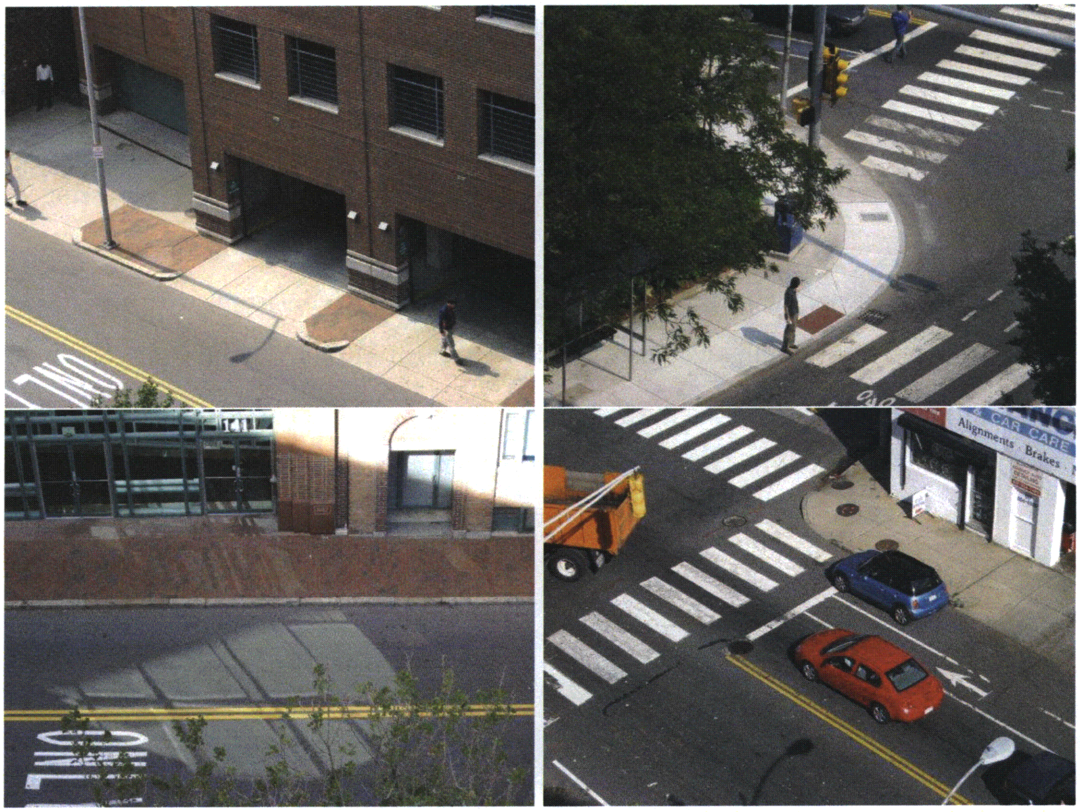

Figure 4-10: Sample positive and negative aerial images from the training set.

pixel images in the Aerial Image Set, containing 201 training images and 200 test images. Of the training images, 101 are positive and 100 are negative. Of the test images, 100 are positive and 100 are negative. The training and test groups were made independent in several ways. First, training and test images were taken on different days and at different times of day. This created a noticeable difference in overall lighting between the two groups. Second, the same person does not appear in both the training and test groups. Third, the camera zoom and inclination is not consistent between the two groups. Sample positive and negative images from both the training and test sets are shown in Figures 4-10 and 4-11.

For the Aerial Image Set, it was essential to have as much variation as possible to simulate authentic testing. Images were taken from the Draper parking garage at varying elevations and locations. As in the Indoor Image Set, the people captured in

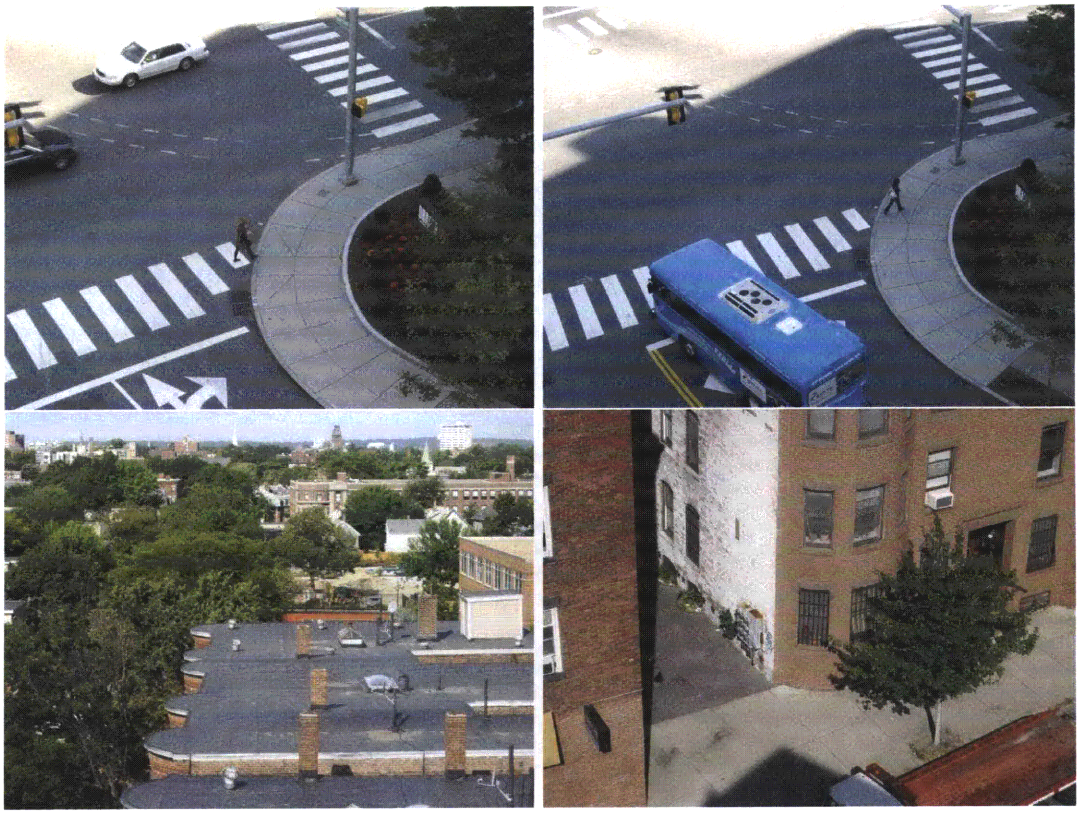

Figure 4-11: Sample positive and negative aerial images from the test set.

the images have varying positions, orientations and distance from the camera. The gender, age, race, and apparel of the people also vary, and there are instances of people walking dogs and riding bicycles and motorcycles. The types and range of variation, specifically the varying camera inclination and elevation, make the Aerial Image Set considerably more challenging than the INRIA Dataset and the MIT-CBCL StreetScenes Database.

For further training and testing, crops were taken from the 401 images in this set.

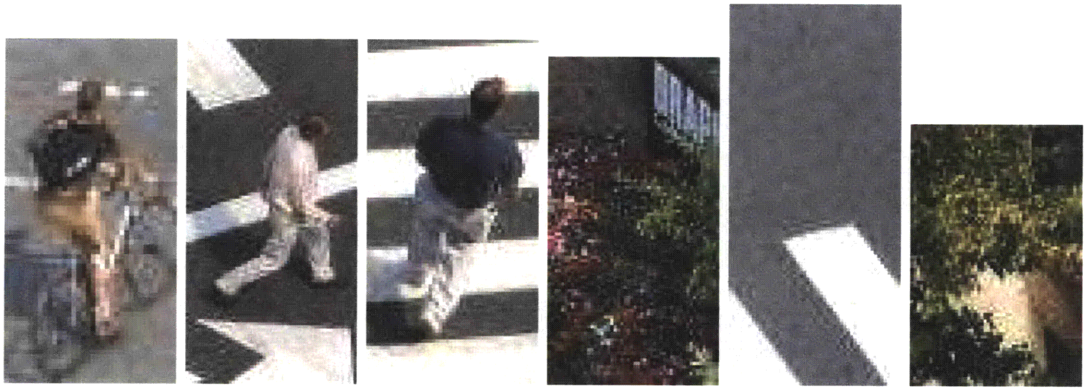

Figure 4-12: Sample positive and negative aerial image crops from the training set.

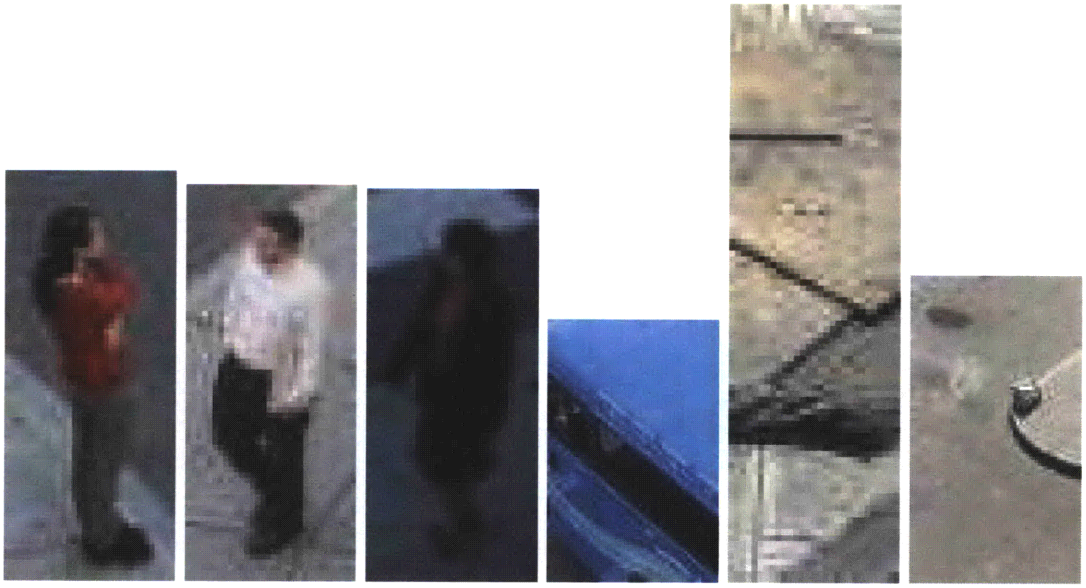

Figure 4-13: Sample positive and negative aerial image crops from the test set.

There were 100 samples of people cropped from the first 47 positive training images and 100 samples of people cropped from the first 71 positive test images. Each crop has a fairly consistent background border surrounding the person and every training crop has a 1:2 aspect ratio. The negative samples were randomly cropped from the positive images (500 training and 500 testing). The negative training crops were given a random width between 20 and 200 pixels and height between 40 and 300 pixels. The negative test crops were given random dimensions that were more consistent with the positive crops: width is a random value between 20 and 60 pixels and height is a random value between 40 and 120 pixels. Figure 4-12 shows sample crops from the training set and Figure 4-13 shows sample crops from the test set.

Aerial Video Set

While the relatively simple background subtraction technique does not suffice for video from a moving platform, more advanced background subtraction with image registration could serve to detect motion in more complex aerial platforms. Nonetheless, stationary video from atop the Draper parking garage was adequate for proof of concept testing; actual UAV video substantiated the results.

The Aerial Video Set consisted of crops from four training videos (totaling 3 minutes) and four test videos (totaling 5 minutes and 16 seconds). While all videos were taken from the Draper parking garage, training and test sets were obtained from different sides of the garage, with different weather, time of day, and camera inclination and zoom. The methodology and categorization used for generating Indoor Video Set crops was repeated for the Aerial Video Set. Originally, there were 971 crops generated from the training videos and 855 crops generated from the test videos. There were 195 positive and 708 negative crops for training, and 300 positive and 474 negative crops for testing. Sample crops from the training and test sets are shown in Figures 4-14 and 4-15, respectively.

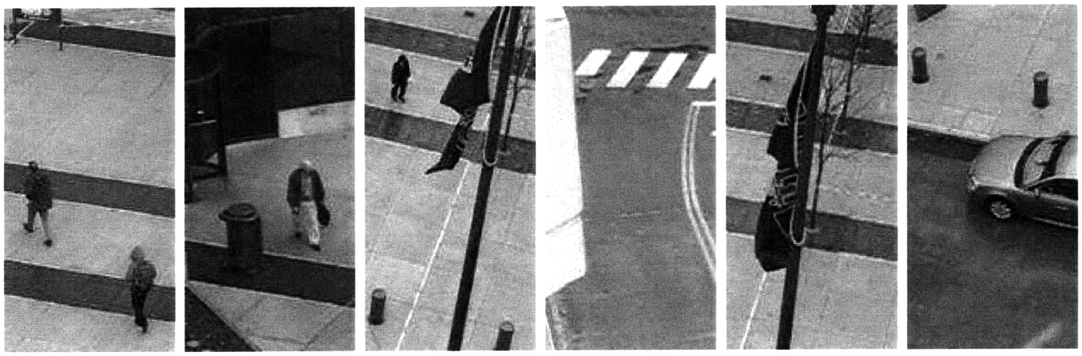

Figure 4-14: Sample positive and negative aerial video crops from the training set.

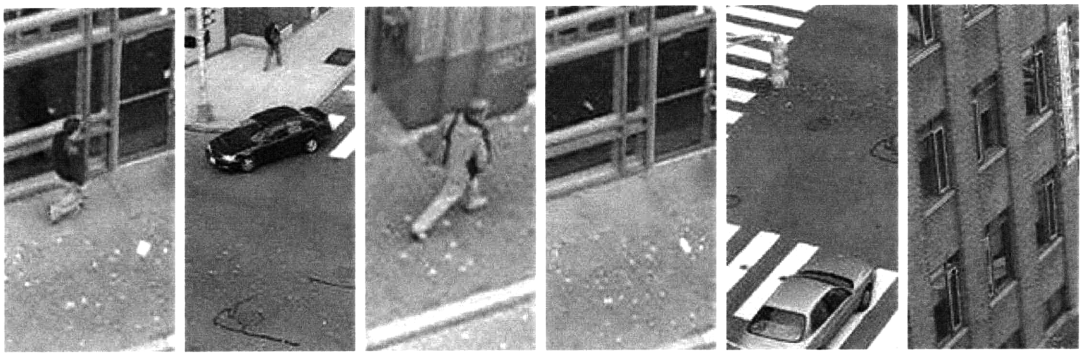

Figure 4-15: Sample positive and negative aerial video crops from the test set.

UAV Video Set

Draper received 16 minutes and 32 seconds of 30-frames per second video from a small UAV with varying altitude above 100 feet. The video contains views of fields, trees, roads, sky, vehicles, and people. From the total footage, there are only approximately 30 seconds of footage in which one or more people were definitively present. Sample positive and negative video frames are shown in Figure 4-16. (Note that the image quality was sometimes reduced due to dropouts of wireless communication.) In the absence of an advanced motion detection algorithm incorporating image registration, positive 1:2-aspect ratio crops were manually taken from the 30 seconds of video. The crops were intentionally taken roughly (people not centered and often not fully in the crop) in order to mimic the difficulty of motion detection algorithms in such a complex domain.

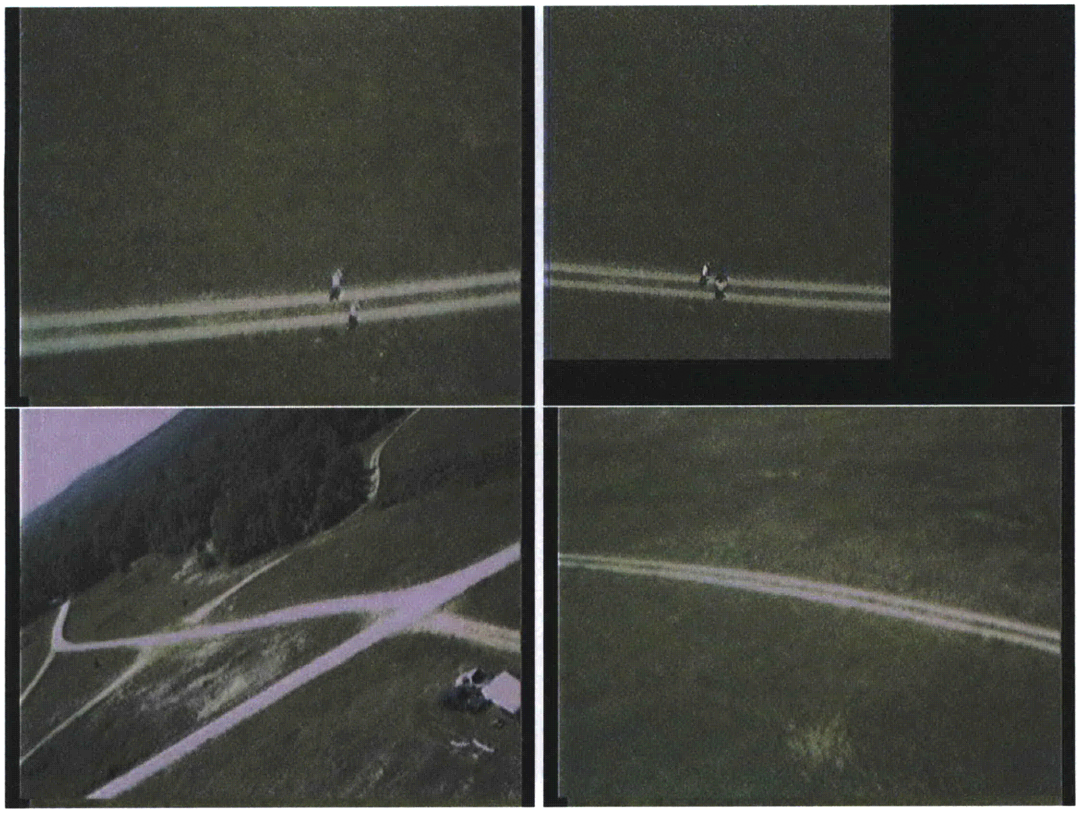

Figure 4-16: Sample positive and negative images from the UAV video set.

The set of crops taken from the video provide an authentic test set and the opportunity to train and test in completely separate environments. The positive test set is comprised of 110 manually cropped samples and the negative test set contains 500 randomly cropped samples. Sample positive and negative crops are shown in Figure 4-17.

Like the indoor datasets, the aerial image and video sets are inherently difficult to train and test with. The image and video crops do not have the same margin and pose consistency as many of the standard object image sets, including the Caltech101 set and the UIUC Image Dataset. The aerial testing results become more impressive upon recognizing the uniquely challenging nature of the training and test images.

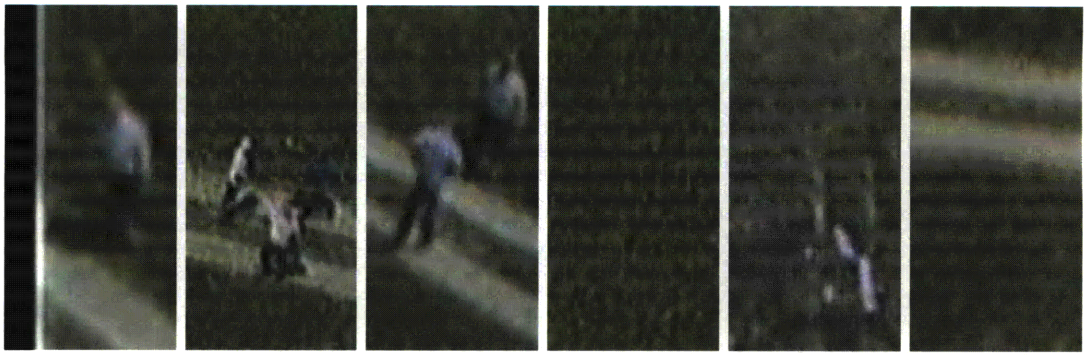

Figure 4-17: Sample positive and negative UAV video crops.

4.3.3 Test Progression

A logical progression of tests was developed for the aerial surveillance application. Like the indoor testing, algorithms were trained and tested with full images, manually selected crops, and subtraction-based crops from video frames. Additional tests were conducted to improve results and apply them toward the UAV video. Table 4.4 gives an overview of the aerial tests discussed in this thesis.

	Test Datasets				
Training Datasets	Full Images	Image Crops	Video Crops	Video Crop Sequences	UAV Crops
Full Images	HOG: 67% HMAX: 61% (No Boot.)				
Image Crops		HMAX: 90% (No Boot.) HOG: 88%	HMAX: 65% HOG: 65% V-J: 60%		HMAX: 95% HOG: 71% V-J: 71%
Video Crops			HMAX: 68% HOG: 65% V-J: 65%	HOG: 70% (Sect. 5.4)	

Table 4.4: Overview of aerial tests organized by training and test datasets. Each algorithm's approximate correct-classification rate is listed, with the best performance highlighted. All tests are described in this section unless otherwise noted. The HMAX bootstrapping technique was implemented for all tests unless otherwise noted.

Training and Testing with Full Aerial Images

HMAX and HOG were trained and tested with the full images taken from atop the Draper parking garage (samples are shown in Figures 4-10 and 4-11). While this approach is extremely convenient, the unacceptably low 60-70% correct-classification rates (as shown in Figure 4-18) are not surprising. Only a very small portion (smaller than indoor) of each positive image is occupied by a person, and it was evident that training and testing with background-bordered crops of people would result in more accurate detection.

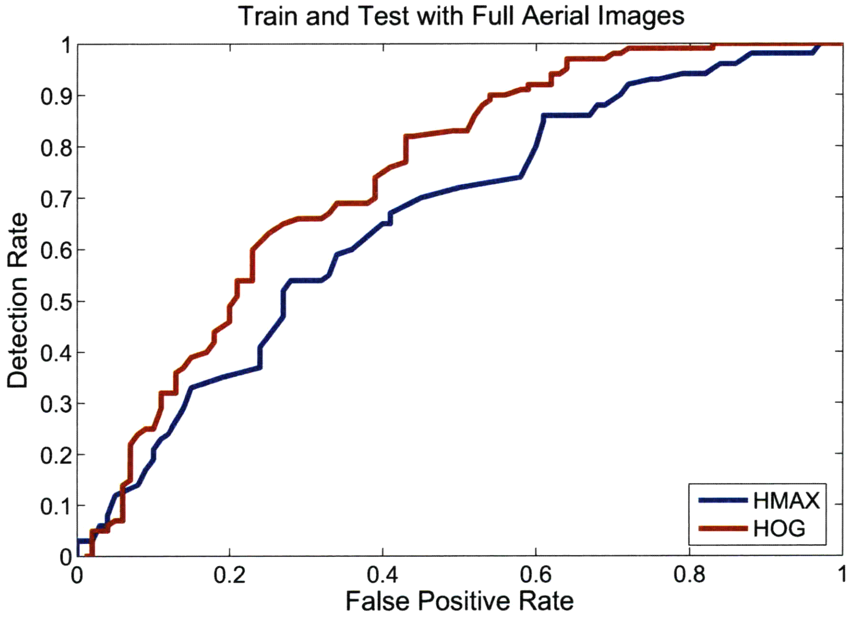

Figure 4-18: ROC curves for training and testing with full aerial images.

Training and Testing with Aerial Image Crops

The 112 positive training and 114 positive test samples were manually cropped from the aerial images. There were also 500 training and 500 test crops taken randomly from the image sets. Sample images from the training and test sets are shown in Figures 4-12 and 4-13, respectively. HMAX and HOG were trained and tested with these sets. The results are impressive despite the fact that HMAX bootstrapping was not yet added and the fact that the crops are not uniformly sized.

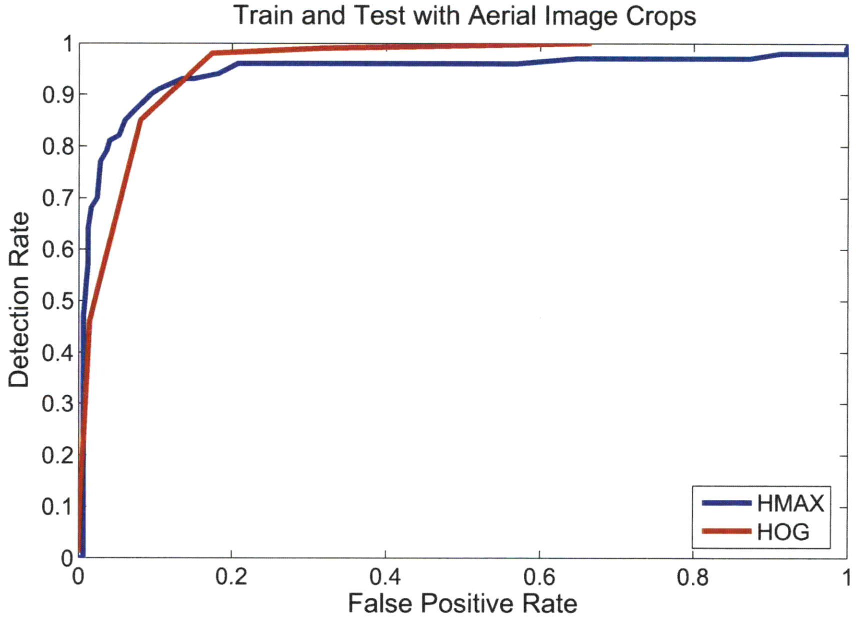

Figure 4-19: ROC curves for training and testing with aerial image crops.

Figure 4-19 shows that both HMAX and HOG demonstrated correct-classification rates of around 90%. While this is impressive, there are some feasibility concerns. Since positive test crops were manually selected, the high classification rates may be unrealistic. Also, because the negative training and test crops were randomly generated, the algorithm could succeed by merely recognizing the presence of any background-bordered object, not necessarily a person. Using the background subtraction-generated crops for training and testing alleviates these concerns.

Training and Testing with Aerial Video Crops

This approach is certainly more realistic than training and testing with manually selected image crops, but is still not altogether feasible. The results from this test must be viewed with a caveat: crops generated by applying image subtraction to stable video may result in unrealistically high classification rates compared with shaky video.

The process of testing with aerial video crops generally mirrored that of testing with indoor video crops; only the size of the image sets changed. HMAX was initially

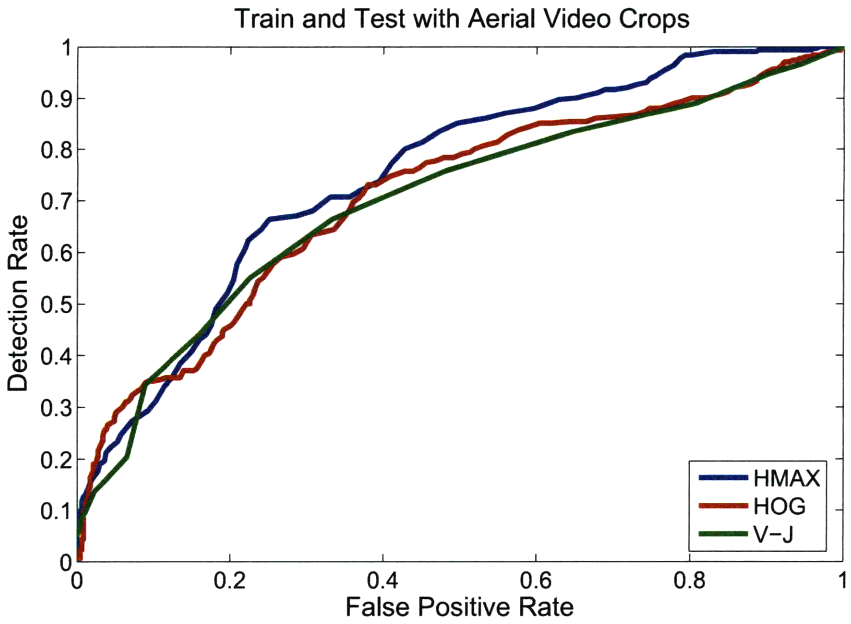

Figure 4-20: ROC curves for training and testing with aerial video crops.

trained with 390 positive crops (195 crops with their flipped counterparts) and 810 negative crops (5 random windows from each of 162 subtraction-generated crops), all 64×128 pixels. (The set of 162 negative crops is the original set of 708 negatives with most redundancies eliminated.) The algorithm was then retrained with the same images plus the 400 "most false positive" crops from among the 2430 exhaustively generated negative set. HOG was trained with 390 positive crops, 810 randomly chosen negative crops, and additional false positives. V-J generated 20 successive classifiers, each trained with negative samples that were incorrectly categorized by the previous classifier.

Each algorithm was tested with the 300 positive and 474 negative crops generated from the test videos, all consistently sized with the training crops (64×128 pixels for HMAX and HOG, 32×64 pixels for V-J). Sample images from the training and test sets are shown in Figures 4-14 and 4-15, respectively.

The algorithms achieved fairly successful categorization rates for the video-generated crops. As shown in Figure 4-20, HMAX generally outperformed HOG and V-J, achieving higher detection rates by 5-10% for most false positive rates.

HMAX correctly classified about 68% of the test images. Assuming that the 774 test samples comprise a random selection of similar aerial video samples, new test images will be correctly categorized more than 64% of the time at 95% confidence. (The applied statistical convention is described in Section 4.1.3.) This is 20% lower than the corresponding indoor success rate and is unacceptable for any real-world detection system. While the success rate is low, the size and difficulty of the training and test sets must be reemphasized. The test images are difficult but realistic; the training set, however, is limited in size when compared to traditional sets used to train these algorithms.

Training with Aerial Image Crops and Testing with Aerial Video Crops

In order to improve results from the last procedure, the positive training set was replaced by the manually cropped positive samples. The idea was that the manual crops are more consistent than the subtraction-based crops, and thus would provide for better algorithm training. The 200 positive image crops were used to train the

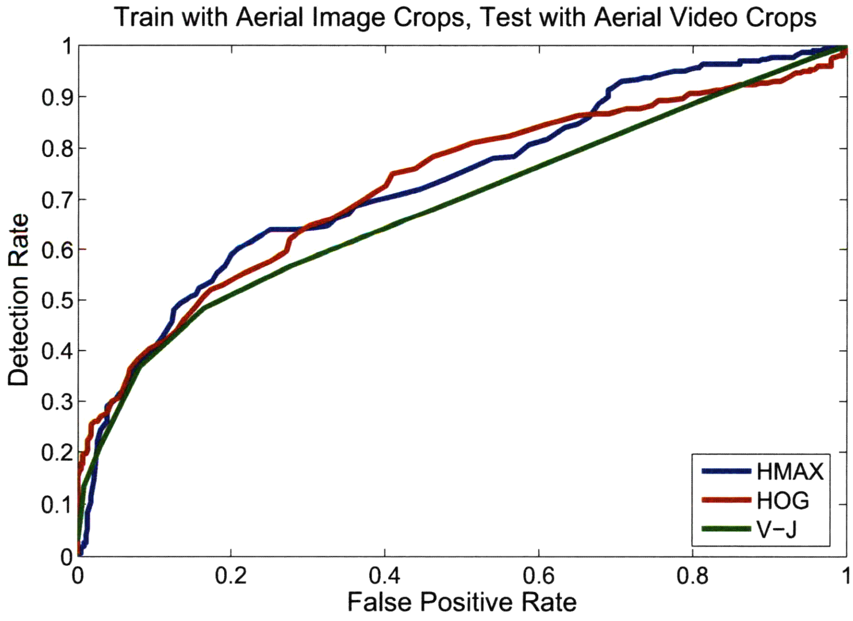

Figure 4-21: ROC curves for training with aerial image crops and testing with aerial video crops.

algorithms and the negative training crops were taken randomly from the Aerial Image Set (sample training crops are shown in Figure 4-12). The test dataset was comprised of the background subtraction-generated crops from the Aerial Video Set.

For the most part, the changes in classification rates from video crop training to image crop training were small. However, Figure 4-21 shows that the expected increase in detection rate occurred in all three algorithms for low false positive rates. Thus in order to test with the UAV crops, the image crops were used for training.

Training with Aerial Image Crops and Testing with UAV Video Crops

This procedure almost reaches full feasibility. While the training set remained the same as it was for the previous procedure, the test set was comprised of 1:2-aspect ratio crops from actual UAV video frames. The positive samples were manually cropped from video frames, and the negative samples were randomly selected. Like previous tests, the manual crop selection compromises the feasibility of the procedure; however, these positives were roughly cropped: the background borders are not consistent among the test positives, and many samples do not contain the full person. This was an attempt to mimic the inexact nature of the more complex image registration techniques necessary for unstable video. This procedure gives key insight into the potential for these leading object recognition algorithms to classify authentic UAV video samples.

HMAX, HOG, and V-J were trained using bootstrapping with 200 positive and 800 negative crops that were generated from the Aerial Image Set. The algorithms were tested with 110 positive and 500 negative crops from the UAV video. Figure 4-17 shows sample positive and negative UAV video crops. All training and test crops were resized to 64×128 pixels (32×64 for V-J).

The results in Figure 4-22 show a stunning contrast between HMAX and the other two algorithms. HMAX obtained a successful classification rate of about 94%, whereas HOG and V-J only correctly classified about 70% of the 610 test images. This large disparity is most likely due to the fact that neither HOG nor V-J have traditionally succeeded in detecting occluded objects. In this dataset, 68 of the 110

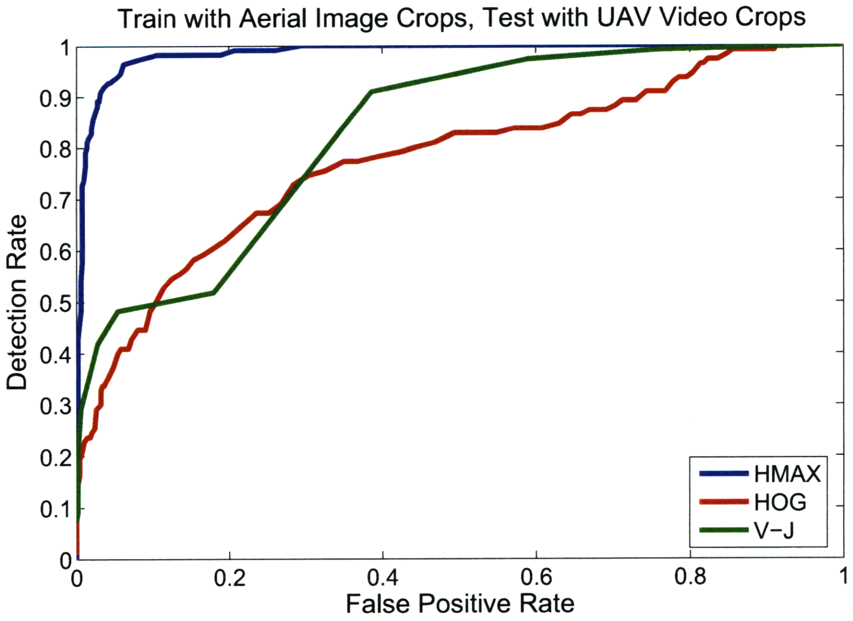

Figure 4-22: ROC curves for training with aerial image crops and testing with UAV video crops.

positive samples contain one or more occluded people. The fact that the positive samples were manually extracted may call into question the applicability of these results, but the high classification rates suggest that coupling HMAX with suitable image registration and background subtraction methods could result in a very reliable person detector.

4.4 Timing Analysis

This section describes the trends and contrasts associated with the algorithm runtimes. Training and testing for HMAX, HOG, and V-J were conducted on a 3.6 GHz Pentium 4 processor. HMAX was implemented using Matlab R2006b to run numerous Matlab/C++ components [15]. HOG training and testing were performed using a series of C executables [4]. V-J was implemented using OpenCV Workspace MSVC6 [3].

Tables 4.5 and 4.6 give the algorithm training and testing runtimes for every dataset. For each of the seven training image sets (three indoor, four aerial), HOG

Dataset	**HMAX**	**HOG**	**V-J**
Full Indoor Images	2 hrs, 51 min	5 min	N/A
Indoor Image Crops	15 hrs, 57 min	5 min	N/A
Indoor Video Crops	6 days, 10 hrs	2 min	4 days, 21 hrs
Full Aerial Images	3 hrs, 29 min	5 min	N/A
Aerial Image Crops (Original Set)	12 hrs, 34 min	35 sec	N/A
Aerial Video Crops	9 days, 3 hrs	1 min, 30 sec	1 day, 18 hrs
Aerial Image Crops (Uniform Set)	1 day, 23 hrs	7 min	12 hrs, 38 min

Table 4.5: Training times for indoor and aerial datasets.

Dataset	**HMAX**	**HOG**	**V-J**
Full Indoor Images	15 min	4 min	N/A
Indoor Image Crops	1 hr	8 min	N/A
Indoor Video Crops	1 hr, 5 min	18 sec	26 sec
Full Aerial Images	16 min	4 min	N/A
Aerial Image Crops	1 hr, 5 min	43 sec	N/A
Aerial Video Crops (Video Crop Training)	1 hr, 22 min	23 sec	20 sec
Aerial Video Crops (Image Crop Training)	1 hr, 19 min	22 sec	9 sec
UAV Video Crops	1 hr, 3 min	19 sec	12 sec

Table 4.6: Testing times for indoor and aerial datasets.

was the fastest to train. HOG was faster than HMAX by between one and four orders of magnitude for every dataset. For all datasets used to train V-J, HOG was faster than V-J by between two and three orders of magnitude. HOG was the fastest algorithm to test for all three indoor datasets and two of the five aerial datasets (V-J was fastest for the other three aerial datasets). The runtime disparity between HOG and the other algorithms was not as large for testing as it was for training; HOG and V-J demonstrated comparable testing speed. That said, HOG was faster than HMAX by up to two orders of magnitude for some test sets.

These results demonstrate that HOG has the potential for real-time person detection and retraining. In addition, V-J has shown the potential for real-time testing. While HMAX was not the fastest to train or test for any dataset and was often significantly slower than both HOG and V-J, this comes as no surprise. Riesenhuber and Poggio's primary goal in developing HMAX was to establish an appropriate model of the visual cortex; the algorithm's processing speed is limited by initial template matching and maximum pooling operations [23].

Chapter 5

Sequence Testing

5.1 Problem Characteristics

The image background subtraction module tracks each object for the duration of time in which the object is visible. Thus there is a series of crops containing each moving object. This approach is beneficial for two main reasons. First, it is not computationally efficient to repeatedly retest an object that has already been classified. However, in order to avoid this retesting, there should be high confidence that the object's classification is indeed correct. This leads to the second reason for tracking objects: testing with a sequence of object images allows for more confident classification than simply testing with a single image.

Every tracked object has a certain lifespan during which the object remains within the image frame. For every image sample in an object's lifespan, the detection algorithm generates the (classifier) confidence that the sample contains a person. The series of confidences for all test samples throughout an object's lifespan is then converted into an overall confidence that the object is a person. This is done by taking the average of the individual confidences. (Other methods designed to convert from confidences to probabilities were attempted, but simply averaging worked best.) Another approach would be to penalize high deviation among the confidences. There are other, more statistically based methods for determining an overall confidence.

With this overall confidence in mind, a new success curve can be generated with axes similar to the ROC curve. The major difference is that rates now depend on the total number of positive and negative sequences instead of the number of positive and negative object crops. The horizontal axis represents the false positive rate, or the number of sequences without a person that are classified as positive. The vertical axis represents the detection rate, or the correct-classification rate of sequences containing at least one instance of a person. Finally, instead of generating the curve by varying the classifier confidence threshold, the "overall confidence threshold" is varied. This threshold is the average confidence above which a sequence is classified as positive. The advantage of using multiple crops per object should result in higher correct-classification rates than before.

5.2 Algorithm Tradeoffs

The image and video testing results were used to determine the best algorithm to employ for the final sequence testing. The most important traits to consider were detection speed and success rate, with simplicity of implementation also considered. The HOG approach stood out in all three categories.

Relatively speaking, HOG was the most consistently successful algorithm. In all three indoor tests, HOG outperformed (had a higher detection rate for most false positive rates) the other algorithms. However, HOG only outperformed the other algorithms in two of the five aerial tests (HMAX won three). This is where algorithm runtime was considered. The time necessary to train and test HOG was significantly less than the runtimes of HMAX and V-J (sometimes by three orders of magnitude). As the sequence tests were designed to determine the feasibility and effectiveness of real-time testing, HOG's runtime dominance could not be ignored.

Simplicity of implementation was the final consideration when deciding which algorithm to use. Each method was relatively simple to run, with HOG having the edge because of its ready-made module for classifying individual samples. With all three factors considered, HOG was chosen for sequence testing and analysis.

5.3 Leave-Behind Sensing

5.3.1 Methodology

Similar to the previous video crop testing, object crops (portions of the video frame with substantial difference from the background frame) were obtained from every tenth frame and tracked. An object history is the series of SVM confidence scores from crops containing the tracked object, where the overall confidence for a sequence is the average of the confidence scores in the history. For this test, each object history had a maximum of five samples. If an object was visible for longer than the duration needed for five samples, another history was created for the same object. This helped to mitigate the problem of an isolated series of outlier scores impacting the object's overall classification. The history length was also limited because the merging and splitting of object regions made it difficult to accurately generate the full history for each object.

For cases in which two independent object histories were combined (e.g., two people move together), the combined object history was the individual history with the greater overall confidence, or higher likelihood of containing a person. If one of the objects of interest is indeed a person, it is important that the combined score reflect the classification already achieved. For any case where an object history was split (e.g., two people move apart), a new history began for each object. This was done to avoid making assumptions about the individual objects based on their combined confidence scores.

Each object sequence was first grouped into one of four categories: (1) one or more sequence samples definitively contains a person, (2) no samples contain a person, (3) one or more samples contains a person as determined only from context clues and the other samples do not contain a person, and (4) one or more samples may contain a person and the other samples do not contain a person. An ROC curve was generated by varying the overall confidence above which a sequence was classified as positive, with positive and negative sequences taken from groups (1) and (2), respectively. It is important to note that no object histories with fewer than three samples were classified, which helped to avoid the problem of outlier SVM scores contributing to sequence classification.

Among the five indoor test videos, there were 251 positive sequences and 94 negative sequences. After eliminating all sequences with fewer than three samples, there were 179 positive sequences and 34 negative sequences.

5.3.2 Results

Direct comparison between the original test and the Indoor Sequence Test shows a dramatic improvement in classification. Figure 5-1 shows that there exists an overall confidence threshold such that 160 of the 179 positive sequences (89%) were correctly classified and 100% of the negatives were correctly classified. Even more impressive is the fact that of the 19 incorrectly identified positive sequences, 18 contain one or more persons that were present in correctly classified sequences. Also, all 72 positive sequences that were eliminated because of their small sample size contain one or more persons that were present in correctly classified sequences. Thus HOG, trained with

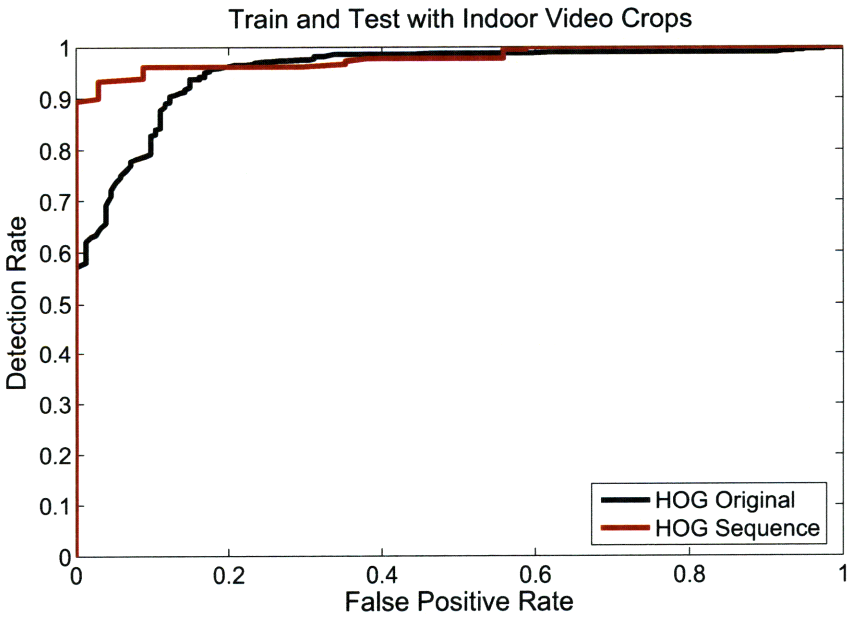

Figure 5-1: ROC curves comparing original indoor video testing to testing with indoor video crop sequences.

background subtraction crops from three different videos from one building, was able to detect, without any false positives, all but one person in five separate videos from another building.

5.4 Aerial Surveillance

5.4.1 Methodology

The setup for the Aerial Sequence Test was very similar to that of the Indoor Sequence Test. Object crops were obtained from every tenth frame of each of the four aerial test videos (taken from atop the parking garage). Each object was tested over the duration of its in-frame visibility, with each sample's classifier confidence recorded in the object's history (maximum length of five samples). The ROC curve was then generated by varying the overall confidence threshold above which each sequence is classified as positive. The overall confidence is taken as the average of the classifier confidences for the individual samples.

Methodology adjustments from indoor to aerial testing were minor. The merging and splitting of objects were handled exactly as they were for the indoor test. The only change made was decreasing the gap width within an object region necessary for the region to be split (e.g., two people moving away from one another). The necessary gap width was made smaller because objects were further away from the camera.

Object sequences were labeled positive, unsure, or negative depending on whether a person was definitively, not definitively, or never seen in the sequence, respectively. Among the four aerial test videos, there were 264 positive sequences and 564 negative sequences. Like the indoor testing, all sequences with fewer than three samples were eliminated, leaving 195 positive and 193 negative sequences for HOG to classify.

5.4.2 Results

Classifying object sequences instead of individual samples improved the correct-classification rate. As shown in Figure 5-2, detection rates improved for most false positive rates, including notable improvements of around 20% for false positive rates between 10% and 20%. For the Indoor Sequence Test, some potential doubts about the methodology and results were settled, as every positive sequence eliminated due to small sample size and all but one incorrectly identified positive sequence contained people that were correctly classified in a prior or subsequent sequence. The same idea likely applies to this test, strengthening the existing detection results.

The detection improvement seen by classifying aerial sequences was not as noteworthy as the improvements seen with the Indoor Sequence Test. There are several components of the methodology that, if changed, would have most likely yielded better results:

- Crops with high person-to-background ratio were more consistently detected as positive. The relatively crude image subtraction technique, combined with the simple way in which object regions were merged and split, often generated crops with inordinately large background space. Simply put, better processing techniques would have almost certainly yielded better results.

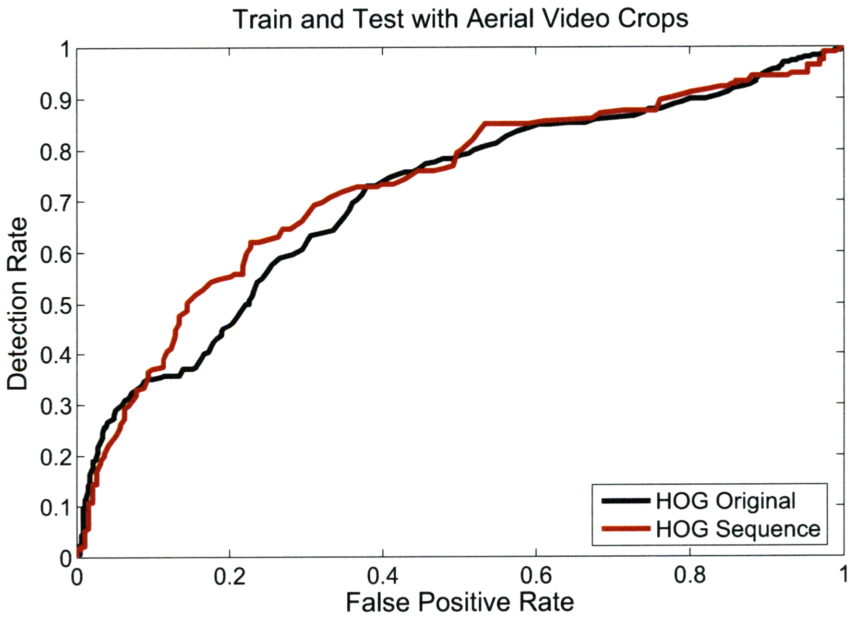

Figure 5-2: ROC curves comparing original aerial video testing to testing with aerial video crop sequences.

- The current overall confidence was taken as the average of individual confidences. If a better approach had been taken, such as penalizing high deviation among confidences in a sequence, more consistent detection would have likely resulted.

- With only one positive sample needed for an entire sequence to be labeled positive, any correctly labeled negative samples in such a sequence were not rewarded. Even worse, they were penalized, decreasing the overall confidence for the positive sequence. Limiting the sequence history length helped to combat this problem, but there is room for improvement.

- Likewise, many positive sequences contained ambiguous samples (those in which a person is present, but only recognizable through context clues). Though highly tedious, it would have been fair to eliminate the inclusion of these individual confidences in overall confidences.

5.5 Timing Considerations

In addition to detection accuracy, system runtime is also a major consideration. HOG has not only provided better detection results than HMAX and V-J, but the method has greatly outperformed the other methods with respect to runtime for both training and testing. The main concern for the sequence testing was whether HOG's impressive testing runtime results could translate to real-time detection without much delay.

The quickest way to determine the feasibility of online sequence testing is to compare the testing runtime with the length of the video itself. The Indoor Sequence Test used five videos totaling 5 minutes and 46 seconds. Running the background subtraction and tracking module at 3 frames per second and categorizing each object sample took a total of 4 minutes and 20 seconds. The four aerial videos totaled 5 minutes and 16 seconds, with the Aerial Sequence Test taking 7 minutes and 7 seconds to complete.

These impressive runtimes become even more promising upon considering two other factors. First, these test videos contained an abundance of objects; real surveillance scenarios will likely contain much fewer objects, effectively decreasing runtime. Furthermore, the detection system does not have to categorize an object sample once its status has been determined (overall confidence has reached a defined threshold). This could significantly improve runtime. In the end, these results seem to strongly indicate that person detection in live video is possible using Dalal and Triggs' HOG approach.

Chapter 6

Conclusion

6.1 Findings

The ability to detect people in a variety of situations is a critical component of many surveillance applications. Automating this process has many benefits, and this thesis examines how state-of-the-art object recognition algorithms fare in this context. These algorithms were tested with novel datasets that were designed to reflect the inherent difficulties of the task. Numerous tests seem to suggest that the examined algorithms (most notably Dalal and Triggs' Histogram of Oriented Gradients approach) have potential for successful, reliable, and easily implementable person detection. There are several significant outcomes from this work, each enumerated in this section:

1. **Even when faced with the significantly challenging Indoor Video Set, HOG delivered highly impressive results when combined with tracking and categorizing object sequences.**

   With independence between training and test sets, HOG was able to detect all but one person in five separate videos without any false positive sequences. This is very impressive considering the limited size and consistency of the training set, as well as the disparate test set. With more appropriate training sets and image subtraction techniques, this result could extend to a number of different leave-behind sensing scenarios.

2. **An effective and efficient person detector can be easily adjusted to situational needs.**

   The Indoor Sequence Test showed the promise of training HOG with datasets similar in nature to the test set. A detector was created that can be adjusted to either ensure that every person is detected or that no false positives are generated, without significantly compromising either priority. This important distinction can be made with the tuning of one parameter (classifier confidence threshold). For some scenarios (e.g., searching for enemy presence), correctly detecting every person is critical, whereas for other situations, an abundance of false positives could be more detrimental than missing a few people. The person detection system applied in this study has the flexibility to encounter this range of scenarios.

3. **HOG and V-J were efficient enough for real-time object detection.**

   With background subtraction reducing the workload, these algorithms took only a fraction of a second to categorize each image window. The original goal was to quickly alert users for the presence of people; the results suggest that this can be done successfully and reliably.

4. **HMAX and V-J achieved impressive results considering their original intentions.**

   These two algorithms fared only slightly worse than HOG for most of the image and video tests, and both outperformed HOG when categorizing UAV video samples. This is significant because whereas HOG was specifically tuned for human detection, HMAX and V-J were not. The HMAX implementation was designed to distinguish among general object instances, not to detect people from the background. Viola and Jones' original model was trained to detect frontal faces and was not intended to detect objects such as humans with large variation in appearance and pose.

5. **Motion detection algorithms (e.g., background subtraction) greatly aid systems for reliable, real-time person detection.**

   The background subtraction and tracking algorithm used in this work generated mostly uniform object crops and eliminated the need to test with full frames. While this simple technique worked well, more robust and reliable methods could raise success rates even higher. Local filtering to reduce noise, keeping more current backgrounds to allow for lighting changes, and using image registration to account for motion effects are just a few possibilities.

6.2 Future Work

Over the course of this project, several research extensions came to mind that were not possible to fully examine under the given time constraints. In the hopes that others following in this line of work take interest, these ideas are listed in this section:

1. **Examine the impact of expanding or improving the training sets for these tests.**

   The test results in this study were impressive considering the relatively small training sets that were used. The sets were on the order of several hundred images, while each detection method had achieved state-of-the-art results with training sets composed of thousands of samples. This thesis was intended to examine the application of benchmark algorithms to realistic image sets, and as such the training sets were limited to what could be collected. It would be revealing to see how the methods fare when faced with larger and perhaps more effective training sets.

2. **Attempt to combine benchmark algorithms to form superior detection systems.**

   This thesis did not examine the effect of combining algorithms, though certain situations (e.g., occlusion and poor lighting) may have warranted this. Combining algorithms could mean using both in a given system or using aspects of each method to construct a hybrid approach. Of course, runtime must always be considered.

3. **Generate an effective multi-object detection system.**

   There are many other objects besides people that must be detected in a number of practical situations. For instance, vehicle detection was briefly examined during this study. It would be very interesting to look at ways in which several different object categories could be detected and distinguished in real-time.

4. **Enhance the "motion detection" capability employed in this work while considering these improvements' costliness.**

   The detection system developed in this project used simple background subtraction, along with a relatively crude form of tracking, to identify the presence of objects before classifying them. There are entire fields of research devoted to special image processing techniques that would be of great use to object detection/recognition. Finding the right combination of these techniques that would both improve upon this system and still maintain runtime feasibility would be an interesting challenge.

5. **Perform testing with more varied examples of live video.**

   With all of the webcam and video streams available online, there is no shortage of relevant test data. As a result of this work, the capability is in place to read these live videos and detect the presence of people.

Appendix A

Completed Tests

1. HMAX - Indoor Image Set - Train/Test with Full Images
2. HMAX - Indoor Image Set - Train with Full Images, Test with Pos. Crops & Neg. Full Images
3. HMAX - Indoor Image Set - Train with Pos. Crops & Neg. Full Images, Test with Full Images
4. HMAX - Indoor Image Set - Train/Test with Pos. Crops & Neg. Full Images - Random 100 Train, 100 Test Crops
5. HMAX - Indoor Image Set - Train/Test with Pos. Crops & Neg. Full Images - Different Random Set
6. HOG - Indoor Image Set - Train with Pos. Crops & Neg. Full Images, Test with Full Images
7. HOG - Indoor Image Set - Train/Test with Pos. Crops & Neg. Full Images - 1010 Neg. Tr. Samples, 60×90 Detector Window (DW)
8. HOG - Indoor Image Set - Train/Test with Pos. Crops & Neg. Full Images - 10,100 Neg. Tr. Samples, 60×90 DW
9. HOG - Indoor Image Set - Train/Test with Pos. Crops & Neg. Full Images - 1010 Neg. Tr. Samples, 40×60 DW
10. HOG - Indoor Image Set - Train/Test with Pos. Crops & Neg. Full Images - 1010 Neg. Tr. Samples, 40×80 DW
11. HOG - Indoor Image Set - Train/Test with Pos. Crops & Neg. Full Images - 101,000 Neg. Tr. Samples, 60×90 DW

12. V-J - Indoor Image Set - Train Upright-Facing Pos. Crops & Neg. Full Images, Test Full Images - 20 stages

13. V-J - Indoor Image Set - Train Upright-Facing Pos. Crops & Neg. Full Images, Test Full Images - 15 stages

14. HOG - Indoor Image Set - Train/Test with Pos. Crops & Neg. Full Images - Also Train with Flipped Pos. Crops

15. HMAX - Indoor Image Set - Train/Test with Pos. Crops & Neg. Full Images - Test Pos. are Neg. Object Crops

16. HMAX - Indoor Image Set - Train/Test with Pos. Crops & Neg. Object Crops

17. HMAX - Indoor Image Set - Train with Pos. Crops & Neg. Full Images, Test with Pos. Crops & Neg. Object Crops

18. HMAX - Indoor Image Set - Train/Test with Pos. Crops & Neg. Random Crops - No Train with Flipped Pos. Crops

19. HMAX - Indoor Image Set - Train/Test with Pos. Crops & Neg. Random Crops - Also Train with Flipped Pos. Crops

20. HOG - Indoor Image Set - Train/Test with Pos. Crops & Neg. Random Crops - 58×87 DW

21. HOG - Indoor Image Set - Train/Test with Pos. Crops & Neg. Random Crops - 40×60 DW

22. HMAX - Aerial Image Set - Train/Test with Pos. Crops & Neg. Random Crops

23. HMAX - Aerial Image Set - Train/Test with Full Images

24. HOG - Aerial Image Set - Train/Test with Pos. Crops & Neg. Random Crops

25. HOG - Aerial Image Set - Train with Pos. Crops & Neg. Random Crops, Test with Full Images

26. HMAX - Indoor Video Set - Train/Test with Crops

27. HOG - Indoor Video Set - Train/Test with Crops - 20×40 IM, 19×38 DW

28. HMAX - Aerial Video Set - Train/Test with Crops

29. HOG - Aerial Video Set - Train/Test with Crops - 64×128 IM, 40×80 DW

30. HOG - Aerial Video Set - Train/Test with Crops - 64×128 IM, 64×128 DW (Invalid: Some Test Images < 64×128)

31. HOG - Aerial Video Set - Train/Test with Crops - 35×70 IM, 35×70 DW

32. HOG - Indoor Video Set - Train/Test with Crops - 35×70 IM, 35×70 DW

33. HOG - Aerial Video Set - Train/Test with Crops - 64×128 IM, 64×128 DW (Valid: All Test Images > 64×128)

34. HMAX - Indoor Image/Video Set - Train with Image Crops, Test with Video Crops

35. HOG - Aerial Image/Video Set - Train with Image Pos. Crops & Neg. Full Images, Test with Video Crops

36. HMAX - Aerial Image/Video Set - Train with Image Pos. Crops & Neg. Full Images, Test with Video Crops

37. HOG - Indoor Image/Video Set - Train with Image Pos. Crops & Neg. Full Images, Test with Video Crops

38. HMAX - Indoor Image/Video Set - Train with Image Pos. Crops & Neg. Full Images, Test with Video Crops

39. HMAX - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with UAV Crops

40. HOG - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with UAV Crops

41. V-J - Aerial Image/Video Set - Train with Image Pos. Crops & Neg. Full Images, Test with Video Crops

42. V-J - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with UAV Crops

Retesting (All Crops 64×128, HMAX with Bootstrapping)

43. HMAX - Indoor Video Set - Train/Test with Crops

44. HOG - Indoor Video Set - Train/Test with Crops - 810 Neg. Tr. Samples

45. HOG - Indoor Video Set - Train/Test with Crops - 16,200 Neg. Tr. Samples

46. HOG - Aerial Video Set - Train/Test with Crops - Full Training Set

47. HMAX - Indoor Video Set - Train/Test with Crops - No Bootstrapping (to see effect)

48. HMAX - Aerial Video Set - Train/Test with Crops

49. V-J - Indoor Video Set - Train/Test with Crops

50. HMAX - Aerial Image Set - Train/Test with Full Images

51. HOG - Aerial Video Set - Train/Test with Crops - Streamlined Training Set

52. HOG - Aerial Image/Video Set - Train with Image Pos. Crops & Neg. Video Crops, Test with Video Crops

53. HOG - Aerial Image/Video Set - Train with Image Pos. Crops & Neg. Full Images, Test with Video Crops

54. HOG - Indoor Image Set - Train/Test with Full Images

55. HOG - Aerial Image Set - Train/Test with Full Images

56. HMAX - Aerial Video Set - Train/Test with Crops - Streamlined Training Set

57. HMAX - Aerial Image/Video Set - Train with Image Pos. Crops & Neg. Full Images, Test with Video Crops

58. HMAX - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with UAV Crops

59. HOG - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with UAV Crops (DEFAULT - Cell Size: 8×8 pix., Cell Block: 2×2 cells)

60. HOG - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with Full UAV Frames

61. HOG - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with UAV Crops (Cell Size: 6×6 pix., Cell Block: 3×3 cells)

62. HOG - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with UAV Crops (Cell Size: 8×8 pix., Cell Block: 3×3 cells)

63. HOG - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with UAV Crops (Cell Size: 10×10 pix., Cell Block: 3×3 cells)

64. HOG - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with UAV Crops (Cell Size: 10×10 pix., Cell Block: 2×2 cells)

65. HOG - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with UAV Crops (Cell Size: 6×6 pix., Cell Block: 2×2 cells)

66. V-J - Aerial Video Set - Train/Test with Crops - Streamlined Training Set

67. V-J - Aerial Image/Video Set - Train with Image Pos. Crops & Neg. Full Images, Test with Video Crops

68. V-J - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with UAV Crops

69. V-J - Aerial Image Set/UAV - Train with Image Pos. Crops & Neg. Full Images, Test with Full UAV Frames

70. HOG - Indoor Video Set - Train with Crops, Test with Crop Sequences

71. HOG - Aerial Video Set - Train with Crops, Test with Crop Sequences

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix B

Background Subtraction & Tracking Pseudocode

```
• Obtain "background image" that does not contain any people
• Convert background image to grayscale
• For every tenth frame:
    – Convert current frame to grayscale
    – Generate "difference image," in which each pixel value, between 0 (black) and 255 (white), equals the absolute value of the pixel difference between the current frame and the background image
    – Generate "working image," which is the difference image with all pixel values above a certain threshold turned to white, and all below that threshold turned to black (helps to reduce noise)
    – For every "motion area" (previously detected window with enough white pixels):
        ∗ Adjust the motion area based on any motion since the last frame
        ∗ Merge two motion areas if they overlap
        ∗ Split motion area given any sufficiently large gaps of black pixels
        ∗ If there are no white pixels in the motion area, remove area from consideration
    – Generate "search image," which is the working image with all pixels in designated motion areas turned to black (this allows an easy search for new white pixels)
    – Find bounding box on any new white pixels
```

- If width or height of bounding box is too large, split into multiple bounding boxes given any sufficiently large gaps of black pixels
- For every bounding box:
    - Add pixel margin to all sides (to account for object's future motion)
    - If bounding box does not overlap any motion areas, make bounding box into new motion area
    - If bounding box overlaps any motion area, merge the two and redefine all motion areas given any overlaps
- Remove any motion area with sufficiently small dimensions
- For every motion area:
    - Generate 1:2-aspect ratio crop sample around motion area
    - Resize the sample to 64×128 pixels (i.e., consistent with size of training samples)
    - Classify sample with recognition algorithm
    - Add classifier confidence to history and update overall confidence

Appendix C

HMAX Bootstrapping Pseudocode

- Begin with positive (64×128) and negative (128×256) training images
- Initial Training Set: All positive images and N random 64×128 windows from negative training images
- Train initial classifier
- Generate Z windows from an exhaustive scan of the negative training images
- Classify all Z windows and obtain the F most positively classified windows ($F < Z$)
- Final Training Set: Same positive images and $N + F$ negative windows
- Train final classifier and categorize test set

  Note:

- HOG performs the same sequence, but the number of false positives used for secondary training can vary
- V-J performs a similar version of bootstrapping:
  - Initial classifier is trained with positive images and an initial set of random negative windows generated from the negative training images
  - Initial classifier is used to categorize subwindows obtained from scanning the negative training images
  - All false positives (up to a given maximum number) are then used, along with the same positive set, to train the next classifier
  - Subsequent classifiers are trained with the false positives generated using the previous classifier (all positive training samples remain the same)

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

[1] Stanley Bileschi. *StreetScenes: Towards Scene Understanding in Still Images.* PhD thesis, Massachusetts Institute of Technology, 2006.

[2] Biswajit Bose and Eric Grimson. Improving Object Classification in Far-Field Video. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:II–II, 2004.

[3] Gary Bradski, Adrian Kaehler, and Vadim Pisarevsky. Learning-Based Computer Vision with Intel's Open Source Computer Vision Library. *Intel Technology Journal*, 9, May 2005.

[4] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:886–893, June 2005.

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[6] Rob Fergus, Pietro Perona, and Andrew Zisserman. Object Class Recognition by Unsupervised Scale-Invariant Learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[7] Rob Fergus, Pietro Perona, and Andrew Zisserman. A Sparse Object Category Model for Efficient Learning and Exhaustive Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[8] Vojtech Franc and Vaclav Hlavac. Statistical Pattern Recognition Toolbox for Matlab. *Prague, Czech: Center for Machine Perception, Czech Technical University*, 2004.

[9] Dariu M. Gavrila and Stefan Munder. Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle. *International Journal of Computer Vision*, 73(1):41–59, 2007.

[10] Hueihan Jhuang, Thomas Serre, Lior Wolf, and Tomaso Poggio. A Biologically Inspired System for Action Recognition. In *ICCV '07: Proceedings of the Eleventh IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[11] Rainer Lienhart and Jochen Maydt. An Extended Set of Haar-like Features for Rapid Object Detection. In *ICIP '02: Proceedings of the International Conference on Image Processing*, volume 1, pages 900–903, September 2002.

[12] Stew Magnuson. Soldiers Test Tools for Urban Surveillance. *National Defense*, May 2007.

[13] Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman. Human Detection Based on a Probabilistic Assembly of Robust Part Detectors. In *ECCV '04: Proceedings of the Eighth European Conference on Computer Vision*, volume 1, pages 69–82, 2004.

[14] Joseph L. Mundy. Object Recognition in the Geometric Era: A Retrospective. In *Toward Category-Level Object Recognition*, pages 3–28, 2006.

[15] Jim Mutch and David G. Lowe. Multiclass Object Recognition with Sparse, Localized Features. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 11–18, Washington, DC, USA, 2006. IEEE Computer Society.

[16] Chikahito Nakajima, Massimiliano Pontil, Bernd Heisele, and Tomaso Poggio. Full-Body Person Recognition System. *Pattern Recognition*, 36:1997–2006, 2003.

[17] Constantine Papageorgiou and Tomaso Poggio. A Trainable System for Object Detection. *International Journal of Computer Vision*, 38(1):15–33, 2000.

[18] Arthur R. Pope. Model-Based Object Recognition - A Survey of Recent Research. Technical Report TR-94-04, The University of British Columbia - Department of Computer Science, January 1994.

[19] Richard J. Radke, Srinivas Andra, Omar Al-Kofahi, and Badrinath Roysam. Image Change Detection Algorithms: A Systematic Survey. *IEEE Transactions on Image Processing*, 14(3):294–307, March 2005.

[20] Maximilian Riesenhuber and Tomaso Poggio. How Visual Cortex Recognizes Objects: The Tale of the Standard Model, June 2002.

[21] Henry Schneiderman and Takeo Kanade. Object Detection Using the Statistics of Parts. *International Journal of Computer Vision*, 56(3):151–177, 2004.

[22] Thomas Serre, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, Gabriel Kreiman, and Tomaso Poggio. A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex. Technical Report CBCL-259, MIT Artificial Intelligence Laboratory, December 2005.

[23] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust Object Recognition with Cortex-Like Mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, March 2007.

[24] Thomas Serre, Lior Wolf, and Tomaso Poggio. Object Recognition with Features Inspired by Visual Cortex. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 994–1000, Washington, DC, USA, 2005. IEEE Computer Society.

[25] United States Marine Corps. *Military Operations on Urbanized Terrain (MOUT): Student Handout*, b0386 edition.

[26] Paul Viola and Michael Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *CVPR '01: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–511 - I–518, 2001.

[27] Paul Viola, Michael J. Jones, and Daniel Snow. Detecting Pedestrians Using Patterns of Motion and Appearance. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 734, Washington, DC, USA, 2003. IEEE Computer Society.