Detection of Abandoned Objects in Crowded Environments

by

Medha Bhargava, B.E.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin
May 2007

Detection of Abandoned Objects in Crowded Environments

Approved by
Supervising Committee:

J. K. Aggarwal, Supervisor

Kristen Grauman

For

Dadima and Dadaji,

Mummy and Papa.

Acknowledgements

Time and time again we are told, as researchers and as students of science, that our observations must be clear, objective and devoid of emotion. However, as I go over my dissertation for the umpteenth time, making sure that I have dotted every “i” and crossed every “t”, I cannot help thinking, with more than a little emotion, about how this project has become such a big a part of my life. Memories of the 2.3 reams of paper, 4 erasers, 8 pens, 11 pencils (mechanical, .7mm 2B), 29 weekends, 47 broken dates, 7 missed movies, 1 Spring Break, 2 cancelled camping trips, 3 bottles of Tylenol, 9 tubs of BlueBell Ice Cream, 14 jars of Nutella®, 231 cups of coffee, 2 bottles of Pepto-Bismol®, 7,962 apologies and the countless all-nighters that went into the making of this thesis will affect me profoundly for years to come.

My only hope is that someday, when my daughter decides to write her graduate dissertation, she will be blessed with the support, advice, friendship and encouragement that I was given by my family, my advisors and my peers. Dr. Aggarwal, thank you for your support and your confidence in me. I will always be touched by the kindness you showed to the nervous international student who came by your office, asking if you had a TA position available for the semester. Dr. Grauman, although our association was brief, the valuable insights you provided helped shape the scope and direction of my research.

More importantly, please know that you will always be held as an ideal by this female engineer getting ready to strike out on her own in a male-dominated industry.

George, Elden, Jong Taek, Michael, Goo and Matt – you will always have a very special place in my heart as my friends, my colleagues and as sources of immeasurable support and encouragement. Angelica, Catalina and Limin – you may never fully understand how the crazy things we did actually brought normalcy into my life. Thank you. Rohan – You're the kind of friend everyone longs for, but few ever find. Thank you for always being there and cheering me on at every step. Finally, Papa, Mummy, Aamoo, Dadaji and Dadima – thank you for your love, your faith in me and all those constant reminders that home would never be too far away. This one's for you.

May 4, 2007

Abstract

Detection of Abandoned Objects in Crowded Environments

Medha Bhargava, M.S.E.

The University of Texas at Austin, 2007

Supervisor: J. K. Aggarwal

With concerns about terrorism and global security on the rise, it has become vital to have in place efficient threat detection systems that will identify potentially dangerous situations, and alert the authorities to take appropriate action. Of particular relevance is the case of abandoned objects in highly crowded areas. This thesis describes a general framework that recognizes the event of someone leaving an object unattended in forbidden areas. Our approach involves the recognition of four sub-events that characterize the activity of interest. When an unaccompanied object is found, the system analyzes its history to determine its most likely owner(s). Through subsequent frames, the system keeps a lookout for the owner, whose presence in or disappearance from the scene defines the status of the object and determines the appropriate course of action. The system was successfully implemented and tested on several standardized datasets.

Table of Contents

List of Tables

List of Figures

Chapter 1: An Introduction

The human visual system is one of the greatest marvels of nature. Humans can, in a single quick glance, accurately take in an entire scene, spot the living beings and inanimate objects in it, and summarize ongoing activities. The human eye can automatically adjust itself to focus on objects of interest and blur out the background, while being generally aware of the remaining environment. However, while we seldom doubt our eyes, there are some serious shortcomings of human vision that are largely unnoticed. Therein lies the tremendous advantage of teaching computers to 'see'.

Recent studies have shown that the average human can attentionally track the movements of up to four dynamic independent targets (individual moving objects such as people or cars) simultaneously, and can detect changes to the attended targets but not the neighboring distractors [1]. It appears that there are spatial and temporal limits to the tracking capability of humans. When targets and distractors are too close, it becomes difficult to individuate the targets and maintain tracking. This difficulty in selection of a single item from a dense array, despite clear visibility, has been attributed to the acuity of attention, or, alternatively, to obligatory feature averaging. Also, if targets move too fast, the average person is unable to track them accurately [2]. Further, according to the classical spotlight theory of visual attention, people can attend to only one region of space (i.e. area in view) at a time [3], or at most, two [4]. Simply put, human visual processing capability and attentive capacity required for effective monitoring of crowded scenes or multiple screens within a surveillance system are limited.

And yet, most visual surveillance systems today consist of hundreds of cameras monitored by a small team of human operators. Typically, each operator watches a set of (four) screens that cycle through views of different locations every few seconds. Failure to instantly spot the slightest suspicious action can cost precious lives and millions in repair, but given our visual inadequacies, it is only inevitable. Multiple object tracking is an inherently active task as opposed to the sustained but fundamentally passive vigilance that current 'manned' surveillance systems are reduced to. Thus, more often than not, camera footage at such locations finds greater use in post-event criminal investigation than in crime prevention and security enforcement.

Intelligent video analysis offers a promising solution to the problem of active surveillance. Automatic threat detection systems can greatly assist security personnel by providing better situational awareness and attentional cues, thereby enabling them to respond to critical situations more efficiently.

In this thesis, we present a methodology for detecting objects left behind in public areas such as mass transit centers, sporting events and entertainment venues as well as other high-security sites. Our algorithm is general, and may be readily adapted for several related applications such as the detection of fallen rocks, debris and other obstructions on roads, railway tracks and airport runways, and the monitoring of cargo. It could also be used to facilitate the detection and management of dangerous situations involving vehicular traffic, such as vehicle accidents, run-over pedestrians and cars stalled at unattended level crossings.

Here, we focus on the detection of abandoned baggage at train stations, where an object is defined as '*abandoned*' in a spatio-temporal context i.e. when its owner has left a predefined area for longer than a certain period of time. This particular problem is of prime importance and relevance to modern society. Two recent conferences – the 2006 IEEE Workshop on Performance Evaluation of Tracking and Surveillance (PETS) and the 2007 IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) – featured this same problem, sponsored by the EU Project

Integrated Surveillance of Crowded Areas for Public Safety (ISCAPS) and UK Government's Home Office Scientific Development Branch (HOSDB) respectively. They serve as real-world indicators of the increasing recognition of the potential and viability of intelligent systems, and moreover, the need for such automated assistance in trying to combat omnipresent threats of terrorism.

Our system essentially simulates the working of a human operator. At the first sight of an object that seems suspiciously unattended to, the operator is likely to rewind the tape to determine how it came to be left there and ascertain whether it has been abandoned altogether or if its owner has simply stepped away momentarily. If the person who brought it onto the train platform, presumably the owner, is still present in the scene, there is no reason to worry, but if he or she is nowhere to be seen, there is certainly cause for alarm. Similarly, if a lone object is discovered in the scene, our system traces it back in time to look for its owner. Here, the '*owner*' of the bag is defined as the person who brings the bag into the scene and sets it down where found. Once a candidate owner has been associated with the potentially-abandoned object by inspecting its past history, a search for the owner is initiated in the current and subsequent frames. If the owner is found to be missing from the detection zone for an extended period of time (predefined), the object is deemed as abandoned and an alarm is raised. The system keeps an alert eye out for the owner. If eventually the person returns to the suspicious object, the alarm is defused.

A patch-based approach is adopted for modeling the appearance of candidate owners, and classical normalized cross-correlation is employed as the distance metric for comparison of patches to match people in the scene with the discovered model of potential owners.

There are several complications that make our task daunting. Perhaps the greatest challenge lies in handling the large crowds inevitable at as dynamic a setting such as that considered here. Limited image resolution and large perspective distortion add to the difficulty of the job. Robustness to the inescapable problems of varying illumination, affine distortion, clutter, noise and occlusion is key. Additionally, the tremendous variety in bags and traveling gear as well as the appearance of people add to the complexity of the problem.

The algorithm was successfully tested on two standardized benchmark datasets – the i-LIDS dataset made available for AVSS 2007, and the PETS2006 dataset. Each of these datasets contains very realistic sequences involving multiple bags and multiple 'actors'. Ground truth is provided for each sequence. In all but one case, our results are accurate to within a second of the ground truth.

There are many different ways of looking at this problem, each with its pros and cons. Some of the more popular approaches are discussed next, in Chapter 2, before a detailed description of our methodology is presented in Chapter 3. Strengths and shortcomings of our framework are demonstrated experimentally and some concerns are addressed in Chapter 4. Chapter 5 wraps up the thesis with a summary of our work, its applicability and several interesting directions worth exploring in future.

Chapter 2: Previous Approaches

Over the past few years, owing to the increasingly ubiquitous presence of cameras, the design of automatic surveillance systems for event recognition in crowded public areas has received much attention. The goal is to equip intelligent systems to reliably pick out the slightest possibility of danger, without raising too many false alarms as to render them futile and undependable. Such systems must prove their effectiveness and worth in complex situations involving considerable crowds, clutter and occlusion. They must be economically feasible and practically realizable in real-time, so as to be able to alert the authorities in a timely fashion to avert great potential harm. Additionally, any image processing framework must be able to successfully overcome the problems of lighting, viewpoint changes, noise and other distortions. Several different approaches have been proposed and explored to solve the problem, some of which are summarized here.

Haritaoglu *et al* introduced *Backpack* [4], a system that exploits periodic motion and static symmetry of a person's silhouette to determine if the person is carrying an object. Their proposed shape analysis algorithm segments and tracks the object detected to recognize any subsequent transfer, exchange, setting down or removal of the object. *Backpack* is designed to work under the control of W^4 [5], and is thus a real-time surveillance system. However, while its underlying principle is instinctively convincing, the heavy reliance on silhouette traits and uncluttered background may hinder its practical generality and application in a dynamic environment.

The ultimate solution to the problem of visual surveillance lies possibly in the accurate tracking of every person and object in the scene. Drawing inferences from a precise record of each person's actions and whereabouts would then be much easier and more credible. However, while a tremendous amount of research has been done on tracking, the ideal solution remains elusive. The most sophisticated algorithms falter in the presence of large numbers of entities of interest, and severe occlusion. Nonetheless, many research groups have pursued this approach for the left-luggage detection problem. A few different methods that conform to this approach are documented here.

Spengler and Schiele [6] propose an approach for detecting abandoned objects and tracking people using the CONDENSATION algorithm in a monocular sequence. Lv *et al* [7] combine a Kalman filter-based blob tracker with a shape-based human tracker to detect people and objects in motion. Like particle-based filters, Kalman filters are effective while dealing with short-term occlusion, but require a certain time lag to allow for consistent tracking in the event of occlusion. That is, if Object 2 appears shortly after Object 1 disappears, the Kalman tracker is easily confused. In their system [7], event recognition is set up in a Bayesian inference framework, which compensates for imperfect tracking to some degree. Stauffer and Grimson [9] present an event detection module that classifies objects, including abandoned objects, using a neural network, but is limited to detecting only one abandoned object at a time. The probabilistic tracking model proposed by Smith *et al* [15] is built of a mixed-state dynamic Bayesian network and a trans-dimensional Markov Chain Monte Carlo chain. The results of tracking are fed into a baggage detection process that identifies still bags and their owners, and then checks to see if the alarm criteria are met. Their method is, in essence, somewhat similar to ours, but for a few significant contrasts that will become apparent shortly.

Adaptive background subtraction (ABS) has been a rather popular choice to detect unknown, changed or removed articles in the foreground. These methods, such as those described in [10] and [11], build and maintain a statistical model of the background, usually implemented in conjunction with an object tracker. The algorithm described in

[10] learns active appearance and spatial models of two different types of contextual background regions (CBRs), and controls the pixel-level background maintenance based on contextual interpretation. The global features of distinctive CBRs provide a strong cue for background perception once some background part is exposed. One of three different learning rates is applied for updating each pixel value based on this contextual reasoning. The method detailed in [11] operates at the block- or patch-level. In [11], Grabner *et al* apply AdaBoost to develop a robust background, and then train an object classifier to distinguish between suspicious events and other 'natural' occurrences. However, in general, ABS-based systems run the risk of integrating stationary foreground objects into the background before they are actually deserted. Their performance also suffers considerably from foreground clutter.

Multi-view surveillance systems for the automatic detection of deserted objects offer the significant advantages of inferring the 3D spatial position of all objects, their depth, size and motion. They also notably alleviate the problem of occlusion. A few recent systems are discussed here.

Martinez-del-Rincon *et al* [12] utilize double background differencing (long-term and short-term backgrounds) for static object detection. They use a multiple-camera modification of the Unscented Kalman Filter (UKF) to combine several measurements from the different views. When an isolated object is detected, their system proceeds to track the person nearest to it as the presumed owner. This hypothesis of ownership may be reasonable in most regular cases, but is susceptible to fail in crowded scenes and severe occlusion which may result in a delay in initial detection of the object. Krahnstoever *et al* [13] exploit the availability of multiple camera views to constrain target tracking results from individual cameras by centralizing tracking in a calibrated metric world coordinate system. Auvinet *et al* [14] follow a similar centralized strategy, using homographic transformation to merge information in the orthographic projection of the objects in the ground plane (also known as the *orthoimage*). Their event inference heuristic relies on the tracking of blobs as spatio-temporal entities, the detection of spatio-temporal forks corresponding to the dropping off of an object and finally, the detection of immobile objects. In [8], Guler and Farrow blend a background subtraction based video tracker with an object detector that focuses on blob splits that could possibly qualify as '*drop-off*' events. Final abandoned object results are obtained by fusing the information from these detectors over multiple camera views. Object tracking and stationary object detection are conducted in parallel, and the potential object drop-off events are correlated with candidate abandoned objects.

It is worthwhile to compare our method to the vastly successful and very promising multiple-camera approach. The methodology presented in this thesis can be easily adapted to the multi-camera case, and its performance would be enhanced by the ability to draw inference based on observation of different views. However, the deployment of multiple cameras per location is usually not practical in vastly spread systems such as the railways. Our goal is to be able to use existing camera networks, such as those at metro train stations in the United Kingdom, for monitoring public areas, demanding little or no changes or additional expense. Thus, we limit our work to monocular image sequences.

Chapter 3: Algorithm for Event Recognition

The proposed algorithm mimics the natural flow of thought of a human operator who decides whether someone has actually left an object at the scene or if the person has only stepped away momentarily. Our current approach to the problem is not based on individually tracking all the people and objects at the scene, which is not the desired goal of this research. Under conditions such as in our case, tracking would not only be unnecessary and computationally wasteful, it would serve to add confusion to the system, and is more likely to hurt system performance rather than help it. Instead of keeping a comprehensive track record of all objects that are present, the system only watches for unattended objects; if found, it proceeds to look for the most likely owner of the bag and determine the owner’s whereabouts. An alarm is triggered if the owner is not found in the area for longer than a set period of time.

Our method is designed to capture and exploit the temporal flow of events related to the abandonment of an object. Figure 1 shows the formal representation of our task, adopted from Allen and Ferguson’s classic temporal interval representation of events [17]. Allen and Ferguson proposed a generic representational framework to describe actions or events in terms of time intervals. Their framework applies temporal interval logic to define the relationships between actions and events, and their effects. An event is defined as having occurred if and only if a given sequence of observations matches the formal event representation and meets the pre-specified temporal constraints. Allen’s representation has been used extensively in a variety of applications, with some very recent work in computer vision by Ryoo and Aggarwal [18] and Nevatia *et al* [19] for activity recognition.

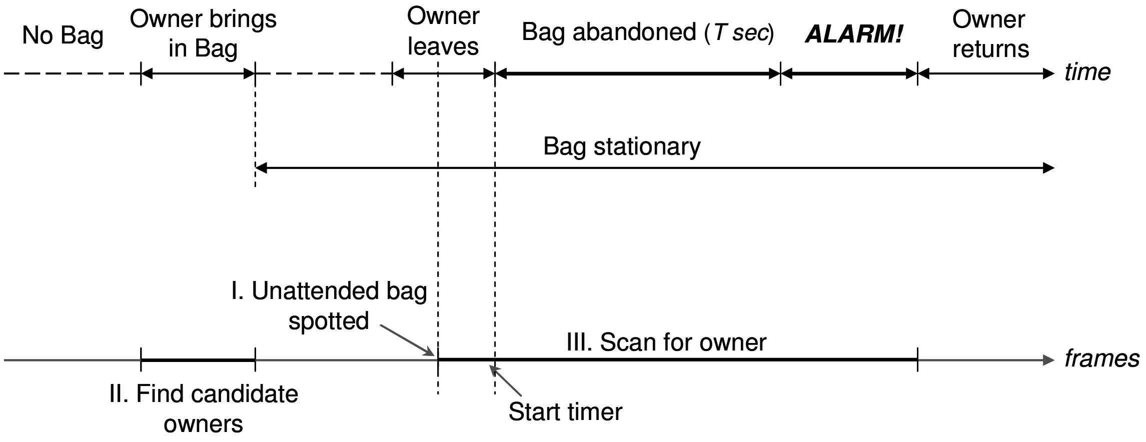


Figure 1: Sequence of events (top two axes) in time and progression of the system algorithm (lowest axis). Module I senses an unattended bag and initiates Module II, which navigates through past frames to the point in time when the owner brought the bag in. Candidate owners are recorded and control is then returned to the present frame and Module III scans the scene for the owner. A timer is set if the owner is not found, and an alarm is triggered if the owner fails to return within *T* seconds.

Here, we define the activity of abandonment of a bag in terms of four sub-events that lead to it – the entry of the owner with the object, departure of the owner without the object, abandonment of the object and subsequent timed alarm, and the possible return of the owner to the (vicinity of the) object. Event inference follows from the detection of each of these sub-events or intervals, as depicted in Figure 1. The sub-events in our case are not independent; the detection of one sub-event triggers the search for the next.

Our algorithm is composed of three computational modules: the detection of unattended object(s), reverse traversal through previous frames to discover the likely owner(s), and the continued observation of the scene. The process is preceded by a basic preprocessing stage that may vary depending on the dataset.

To ensure clarity, most of the algorithm will be described in terms of one abandoned object and one rightful owner. It must be noted that the framework is capable of concurrently handling several abandoned bags with multiple owners, although this has not been experimentally demonstrated as yet. Our current experimental datasets involve only one bag with one owner, or at most, two people traveling together in close proximity of each other and the bag.

LOW LEVEL PROCESSING

Good low level processing, comprising of image preprocessing and segmentation, is critical for any computer vision system. Foreground extraction forms the basis for development of successive modules. The success of subsequent object behavior analysis strongly depends on the reliability of object detection and classification in order to have a firm basis upon which to act. In our case, we find that a few false positives are tolerable, but persistent false positives such as often produced by bright glares or when the object has perceptible shadows, are cause for worry. Extraction of too many false positives is likely to lead the subsequent classifier module astray, and deteriorate the performance of the entire system. Thus, some trade-off is inevitable in search of a satisfactory balance. Here, we choose to err on the side of over-segmentation, and rely on the ensuing models to glean the most useful information from the image segments.

For efficiency and ease of computation, background subtraction is performed on each frame. In practice, it would not be difficult to obtain a ‘true’ background image at an indoor, controlled environment as is the case at metro stations, auditoriums and so on. However, since background information was not available, a background initialization algorithm adapted from [20] was used to build a background model [21]. In [20], at each pixel, stable intervals of time are identified and local optical flow is computed to determine which interval is most likely to capture the background. This method has been shown to yield impressive results when optimal parameters are chosen. In our system, this critical parameter estimation process was automated by analyzing the input sequence [21]. The static background thus extracted is usually impressively close to the true setting. Depending on the dataset, a simple Mixture-of-Gaussians [22] background model might also suffice. Some parts of the background may be masked out if prudent to further facilitate processing.

Background subtraction is performed in the HSV color space since it offers some degree of robustness to changes in illumination, such as the occurrence of shadows. A series of morphological operations is carried out to ‘clean’ up the image, retaining only the most meaningful segments. Next, the mean-shift algorithm [23] is applied for color quantization and image segmentation. Mean-shift is a powerful technique that offers a practical approach to non-parametric cluster analysis of multimodal feature spaces. It seeks to find the modes of the (unknown) probability density function underlying the feature space, which correspond to dense regions or local maxima in the feature space. Once the location of a mode is determined, the cluster associated with it is delineated based on local structural information. Thus, mean-shift is basically an adaptive gradient ascent method that shifts the local mean towards the region of maximum density. The procedure is guaranteed to converge at a point where the normalized density estimate has zero gradient. Mean-shift is a very versatile tool that finds application in a host of key computer vision problems, such as discontinuity-preserving smoothing and segmentation, as used here. For a comparative evaluation, experimentation with k-means clustering in several color spaces was also carried out, but mean-shift emerged as the clear winner. One other obvious advantage that mean-shift offers over k-means, k-medoids and other related methods [22] is that it does not require prior specification of a set number (k) of modes. The only user set parameter required by mean-shift is the feature (range) bandwidth of the analysis.

At the end of this processing stage, we have segmented out the main objects of interest – people, bags and other non-background objects**.** Subsequent processing deals exclusively with these segmented foreground images. By narrowing down the focused area of attention, the computational load on the system is reduced significantly.

DETECTION OF UNATTENDED OBJECTS

The goal of the first module of the algorithm is the detection of any object that seems to have been left by itself. Until such an event occurs, it is unnecessary to track and monitor all ongoing activities in the scene. As mentioned earlier, this not only cuts computational costs but also avoids ambiguities born of inaccuracies in tracking in the presence of much movement and occlusion.

For this thesis, the focus is on the detection of abandoned baggage at a metro station. It is assumed that an abandoned bag may be any bag that can be seen distinctly, separate from all nearby blobs, for at least a short period of time (approximately 8-10 consecutive frames). Baggage could include suitcases, sports bags, rucksacks, backpacks, boxes and so on. The algorithm may be suitably tailored to identify other kinds of objects as well. The manner of constraining the possible assortment of object classes to be expected is site- and application- specific.

We represent bags based on some typical characteristics gleaned from a set of positive and negative examples provided to the system. The k-nearest neighbor classifier is then used to classify foreground blobs in novel frames as belonging to the *Bag* (or *Non-Bag*) class. In general, bags are solid, contiguous entities that usually do not exceed half the height of an average grown human. Thus, classification is based on the size and shape of segmented binary blobs. Useful features that are extracted include compactness, orientation, solidity ratio, eccentricity and size, each of which we define and discuss in turn below.

The compactness measure of a shape, also known as the shape factor, is defined as the ratio of area to squared perimeter of a shape. It is therefore, independent of scale and orientation, and largely undisturbed by a few outliers or artifacts of imperfect segmentation. Once normalized appropriately (by multiplying by 4π), the compactness measure agrees with intuitive notions of what makes a shape compact. Mathematically,

$$compactness = 4\pi \frac{Area}{\left(Perimeter\right)^2}$$

Compactness has been a classical favorite for (geometric) shape recognition, and with good reason – it substantiates itself to be the single most distinctive attribute for our purpose of bag identification.

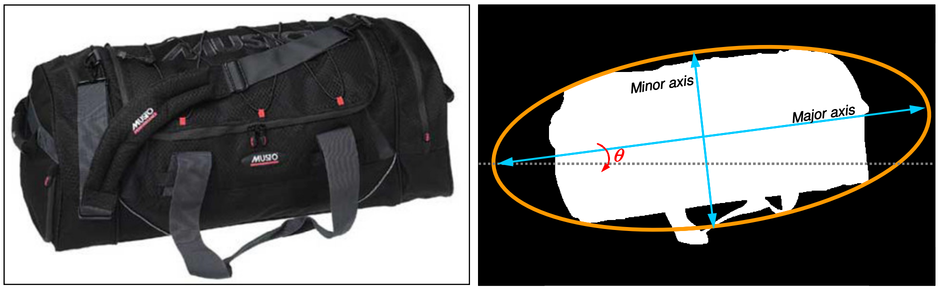


Figure 2: An example training image. Original image (left) is binarized and features are extracted (right). Major and minor axes, as well as the angle of orientation are marked (right).

Solidity is defined as the proportion of the pixels of a blob that lie within its convex hull. The eccentricity and orientation measures are computed using the ellipse that encloses the blob, as shown in Figure 2. Eccentricity is computed as the ratio of the major elliptical axis to the minor, while orientation denotes the angle (θ) that the major axis makes with the horizontal x-axis. Each of these measures collaborate with each other nicely to map a four-dimensional feature space for effective representation and classification of bags, humanoids, other articles, noise and various artifacts of segmentation. All features are normalized to range between 0 and 1.

The size of each binary blob is normalized to account for the effects of perspective projection. Each object is weighted to estimate its size if placed right before the camera i.e. with its base lined up with the lower edge of the image. Normalizing weights (for height and width) are determined empirically by observing the change in floor patterns or other steady objects across the scene. Normalized height/width of a blob is then given as the weighted product of the visible height/weight and the ratio of image height to the blob's lower y-coordinate. Normalization is coarse, but is satisfactory for our purpose. Extremely small blobs are filtered out as noise; blobs larger than a set threshold are excluded from consideration as possible bags. Thresholds for this size filter are determined empirically by analyzing the dataset.

For our purposes, a simple 3-nearest neighbor classifier was found to be sufficient for picking out the solitary bag-like objects in the scene. Training sets are dataset-specific and may vary significantly between different settings. The training sets used here for the experiments were rather small, owing to the limited data available for testing and the constraints imposed on the variety of anticipated baggage. The classifier was trained off-line using 9 positive bag instances and 12 negative examples. The positive training examples were collected through Google Image Search, and subsequently binarized. The negative examples include humanoid blobs and irregularly shaped segments (extracted shadow blobs, etc) selected from the data sequences.

The performance of our simple baggage detection setup is very good. Owing to the simplicity of the binary classifier and the features used, execution time is minimal. However, some false positives may be obtained. Typically, smooth surfaces reflecting changes in illumination (as the train pulls in, for example) and segmented heads, torsos or lower bodies of people are sometimes mistaken as possible bags. To verify the decision of the classifier, each suspect blob is tracked over a set number of consecutive frames (usually 8-10 frames) to check for consistency of detection, before moving on to look for the potential owner of the detected unattended bag. This filters out a considerable number of false positives that appear from time to time. Moving objects incorrectly detected are filtered out by virtue of their motion. Any persistent false positives are eliminated in subsequent stages of processing. Once an object is locked onto as being unattended, its patch-based appearance model is built and stored along with its positional information. The patch-based modeling scheme is detailed in the next section.

A binary classifier is used here to sort all blobs as belonging to either the *Bag* or the *Non-Bag* class. A more sophisticated classifier may be employed for a more refined categorization of individual blobs, for example, suitcases, sports bags, humans, rucksacks, etc.

REVERSE TRAVERSAL FOR FINDING CANDIDATE OWNERS

In crowded scenarios where a bag appears to have been abandoned, a human operator is likely to rewind the video to around the ‘drop-off’ point, when the bag was first brought to and placed at the detected location, and carefully observe the movement and behavior of the owner from that point on, to gauge the threat level of the situation. This module of our system acts in much the same way. Once the system latches onto an unattended bag, it traces it through previous frames to detect the event of the owner setting down the bag.

Most of the backtracking stage is implemented in a straightforward manner to facilitate speedy traversal to the frames of interest, i.e. when the bag was first visibly introduced in the immediate neighborhood of its detected location. Initial tracking is based solely on the location and size of the blob, regardless of appearance. The presence of any blob of approximately the same size or larger, occupying the same area as the suspicious baggage is assumed to indicate the presence of the bag. While this supposition may result in erroneous overshooting of the desired frames, it accounts for instances of severe and even complete occlusion of the bag, thereby reducing the chances of mistaking the wrong person(s) as the possible owner(s).

When no valid blob is found at the anticipated location, it is inferred that the bag was in motion and ought to be present elsewhere in the neighborhood. Note that while tracking in reverse time, the movement of the bag corresponds to the past event of the owner arriving at the location with the bag. The algorithm then performs template-matching using normalized cross-correlation [24] to search for the bag in the nearby region (using the previously stored appearance model of the bag).

Image matching is performed at the patch level. There are several different ways of extracting patches from an image. Most popular among them is the growing of patches around special points retrieved by one or more interest operators. However, our experiments showed that Harris corners [26] and SIFT feature points [27] were inconsistent across images and generally missed the more perceptually appealing and reliable features. This is not a shortcoming of these interest operators *per se*, but a consequence of the CCTV data itself. Most interest operators are designed to work on dense, texturally rich images. Owing to inadequate resolution and small segment size of the image patches in consideration, the points extracted by these operators are sparse and not suitably descriptive. The uniform sampling technique was found to be more effective for our purposes. Rectangular image patches are extracted from the desired neighborhood using uniform grids of different sizes, devised to coarsely account for perspective distortion if necessary. A large, comprehensive pool of patches is extracted at first, but only a small subset (approximately 50) of the most meaningful patches is retained. Patches that contain none or a very small fraction of the segments (less than 50% of patch area) are discarded. A record of the centroid of the parent blob from which each patch was derived is also maintained, and serves as a simple way of incorporating some basic positional information into the model of each candidate owner, as will be clear shortly.

Normalized cross-correlation [24] is one of the simplest but most effective template-matching distance metrics, and been used extensively for patch- and fragment-based recognition. Although several more sophisticated similarity measures have been put forth, normalized cross-correlation (NCC) still remains a very popular choice, due to its invariance to linear brightness and contrast variation. It is a self-sufficient measure and requires no empirical tuning of parameters. The biggest pitfall of NCC is its sensitivity to changes in scale and rotation. However, given the circumstantial viewing conditions in our case, neither the unattended bag nor the people in its neighborhood (the specified region of interest) are likely to exhibit any sizeable change in size or orientation within the neighborhood – thus, scale and rotation invariance is not as critical. There are ways to instill these virtues into the basic NCC measure [25], should the application call for them. In our method, the HSV (Hue, Saturation, Value) components of each patch are found and NCC coefficients are computed for the three planes of the HSV domain. The NCC measure between patches at multiple scales is handled by moving the smaller patch within the bigger one, correlating at each shifted position. A weighted sum of the three coefficients is then used to define the degree of matching. The weights are set experimentally, and tend to be dataset-specific. In most cases, greater weight is attributed to the intensity(V) and hue(H) values than the saturation (S). Traversal from this point on, where we are following a particular unattended bag, follows the normalized cross-correlation matching method instead of the initial rudimentary blob-based process.

Two situations can arise from the outcome of correlation in a frame *n*: either the bag is found nearby or it is not. The methods used for handling the two possibilities are discussed below, augmented illustratively by Figure 3.

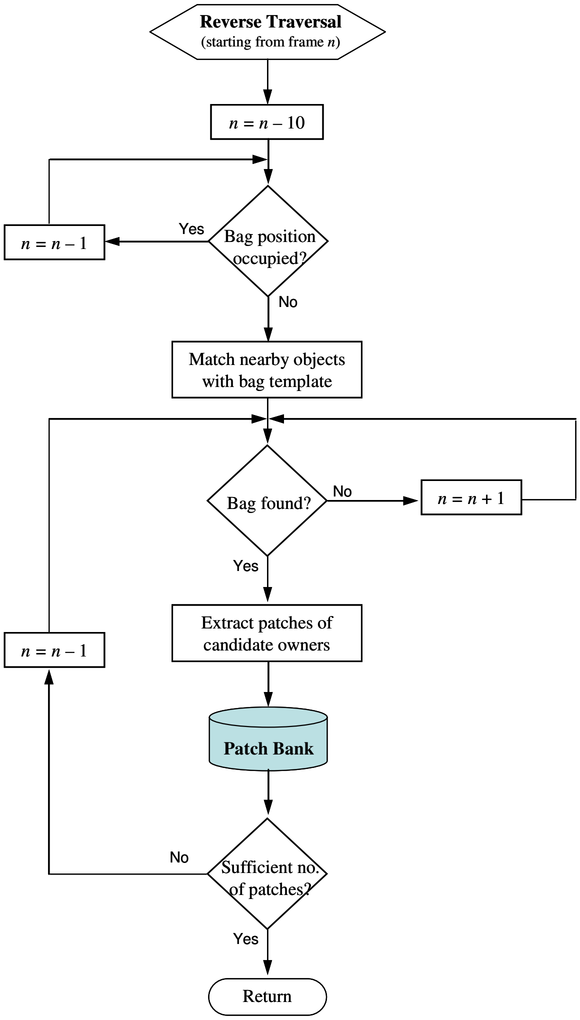


Figure 3: Flowchart description of module.

Situation 1: If the bag is found, it may be inferred that the bag was being moved or carried at the time, presumably by its owner. The extracted foreground blobs in its neighborhood can then be considered representative of the candidate owners. Patch-based appearance models are selectively extracted using the previously described uniform sampling technique to represent each of the valid candidates in the frame and saved. Until a sufficient number of patches (approximately 50-60) is collected or the beginning of the video stream is reached, the system continues backtracking, i.e. traversing from frame *n* to frame *n*-1, and the patch gathering process is repeated (within the updated neighborhood of the bag).

Situation 2: In the case where the bag cannot be found, it may be inferred that the desired set of frames has been overshot. Such an eventuality may arise when the actual arrival of the bag on the scene occurs in the presence of occlusion, or if someone (or something) else was standing beside the bag when the real owner left. Conversely, in terms of reverse traversal, this is the situation where the movement of the bag under inspection from its detected location goes unnoticed by the simple blob tracker due to severe obstruction by another sufficiently large object blob, or its merging into another blob visually near it. In such a case, the system flips the direction of traversal (so as to be moving towards along the positive time axis) and attempts to locate the bag in frame (*n*+1). Another possible way of handling this situation would be to enlarge the defined neighborhood area for searching until the bag is found, but with such an approach, scale and rotation invariant normalized cross-correlation would be necessary for reliable recognition of the bag. The variation in scale and rotation is expected due to the numerous possible ways of handling a bag - the bag may be hung from the owner's shoulder or being rolled along at an angle or lifted by hand, and moving in the direction the owner approaches from. Not only is the change in orientation and scale bound to complicate matching, variation in viewpoint could seriously deteriorate performance as well. Therefore, the former method is preferred here. Usually, in the event of missing the ‘drop-off’ point, the overshoot is not more than a few frames. Therefore, the additional tracking does not affect performance terribly.

A pseudo-code representative of the basic working of this module is shown below, in Figure 4.

```
main()
{ ...
 lookForUnattendedBaggage()
 if (suspicious bag detected in frame[n])
 {
   backtrack to frame[n-1] and check
   if (bag found)               ⇒ bag stationary
      continue backtracking till not found

   if (bag not found)
      traverse(n)
 }...
}

traverse(n)
{
 attempt to match bag template in neighborhood

 if (matched)
   ⇒bag in motion, presumably with owner
   findCandidateOwners()
   while (more patches needed)
      continue iterative reverse traversal
      traverse(n-1)

 if (not matched)
   ⇒bag not in scene
   ⇒point of entry overshot in traversal
   continue iterative traversal in forward direction
   traverse(n+1)

   when (matched)
      findCandidateOwners()
      while (more patches needed)
         continue iterative forward traversal
         traverse(n+1)
}
```

Figure 4: Pseudo-code for detecting candidate owners.

The difficulty of pinpointing exactly which observed blob corresponds to the true owner of the bag in the presence of several people must be appreciated. The odds of making a mistake in assigning specific ownership in a crowded scenario are rather high, so any attempts to zero in on a single individual right away are best avoided. It could possible that the true owner came in alongside another person (or more), and given the view, it may not be possible (even humanly so) to discern the real owner reliably. In fact, in such a case, it would probably be desirable to attribute possession to all possibly involved persons for later inspection and investigation, should any foul play be suspected.

The patch bank collected over a short image sequence during traversal is expected to contain several redundant patches. This redundancy is intentional and useful to bias subsequent comparisons to match the actual owner, since it is only reasonable to expect that the true owner would stay by the bag for at least a short period of time, while others might pass by. Ambiguity may arise if another person enters the scene at around the same time, and remains stationary nearby, perhaps while waiting for a train. In this case, both persons are treated as possible candidates. The redundancy could also potentially serve as a means of handling (slight) viewpoint variation in the subsequent stage of processing.

As a final step of creating owner hypotheses, the patch bank is consolidated by taking advantage of the positional relationships between patches. Since a few successive images in the sequence are considered for building appearance models of candidate owners, large translation of blobs across the frames is highly improbable. Patches collected across different frames but originating from blobs that lie in the same locality are clustered together as potentially belonging to the same person(s). The redundancy of the patch bank together with this spatial clustering results in an inherent probabilistic distribution of likelihood of ownership per blob within the prescribed window of interest around the bag. The patch bank may be finally be pruned by eliminating clusters that are too small to be meaningful for comparison.

CONTINUED SCENE MONITORING

Once a representative patch bank is constructed, we return to the point when the bag was first identified as unattended, i.e. the present frame. Looking forward in time from then on, our intention is to keep track of the actions of all possible owners of the bag. The system maintains a watchful eye to detect the event of their departure from the neighborhood of the bag, observes the area for the possible eventuality of their return and sounds the alarm if they are missing longer than a predefined *T* seconds.

In order to look for candidate owners, every color blob in the (predefined) vicinity of the bag is cross-correlated with the complete patch bank. Only a fraction of the most similar patches is retained. The spatial coherence of the top hits for each color blob is then analyzed to see if the blob closely matches any single appearance model in the patch bank. If it does, then ownership of the bag is assigned to the corresponding blob. This step is taken to safeguard against the possibility that parts of the blob match random patches in the patch bank that were originally extracted from different blobs. This kind of match is meaningless and confusing. Thus, adding the spatial configuration parameter, we add to the uniqueness of each patch and the robustness of the system. Despite measures taken to ensure that the patches are discriminative, their distinction cannot be guaranteed, especially in the case when several people may share some commonality in attire (dark trousers or overcoats, for example). Thus, the presence or absence of the owner is established based on both appearance and spatial constraints.

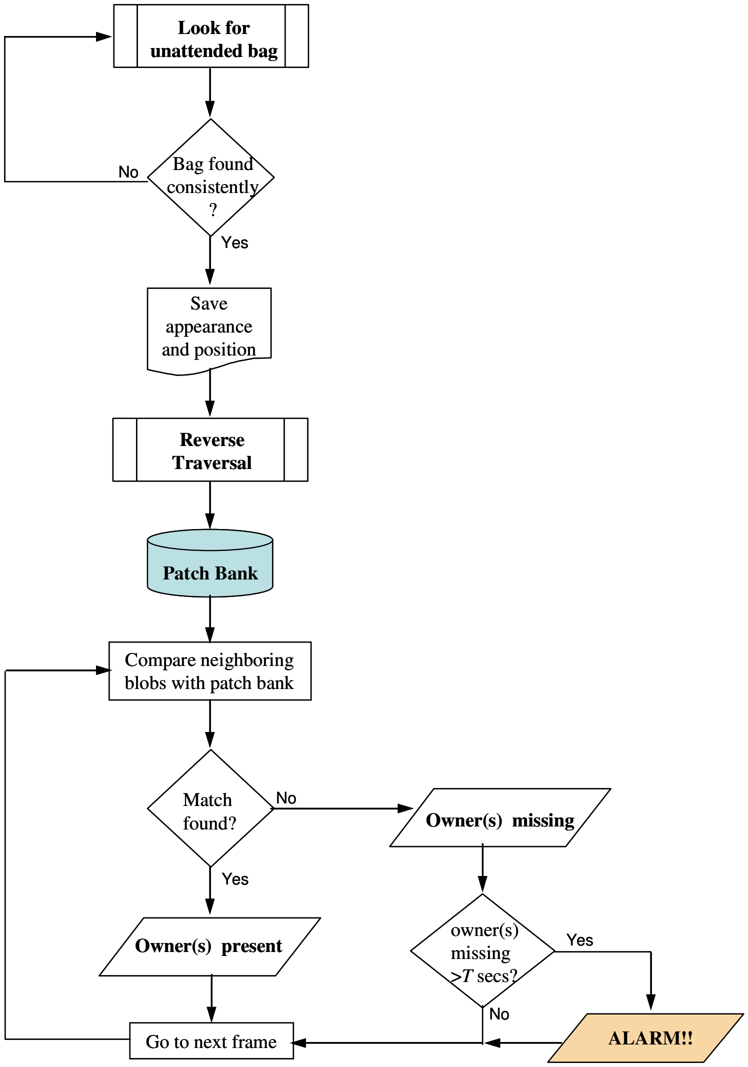


Figure 5: Flowchart outlining the overall algorithm.

If a likely owner who meets both matching requirements is found in the detection area, no action is taken. For as long as the bag remains ‘unattended,’ the system continues scanning for the owner in the area. However, if the owner steps outside the predefined detection zone or is not visible at all, a timer is set. To insure against inaccurate feature matching, the conclusion of possible abandonment is reached over a sequence of successive frames (usually, 3 frames). The system carries on its scrutiny of the scene for the possible event of the owner returning to the bag, in which case the timer is deactivated. Once again, this decision is made over several frames to add to its confidence. In the event that the timer ticks on for *T* seconds, an alarm is triggered, and lasts for as long as the owner is away. The alarm persists until the owner returns to the bag or the alarm is manually reset.

Chapter 4: Experimental Results

We tested our system on two standardized datasets [28, 29] representative of real operating conditions in train stations, filmed in London, UK. All footage is recorded in an uncontrolled public environment. Both datasets offer a set of video clips that showcase different plausible ways in which people may abandon baggage without being very noticeable. Bags used for the staging of the event range from mid-sized backpacks and sports bags to large suitcases and even ski carriers. Complications arise from the inevitable occlusion of entities, differing depths of view and perspective distortion. Ground truth data, comprising of the ideal alarm start and stop times (or frames), is provided in all cases for an accurate evaluation of methodology.

As with most computer vision applications, there are several parameters and features of the system that are scene-specific and require supervised tweaking for good performance. While this is unfortunate, it is practically unavoidable. However, parameter initialization is a one-time cost, and not very labor-intensive since there are only a few values to set. The low level processing stage uses several parameters that have to be tailored per scene, viz parameters for morphological operations and connected component analysis, and bandwidth selection for mean-shift procedure. Next, the object detection module is trained using data designed according to the relevant dataset. Size normalization weights vary significantly as well, based on the location of the camera and depth of view. Finally, possibly the most crucial parameter that must be carefully set is the threshold for matching patches. Once initialized, these parameters are not changed per sequence of each dataset, as long as the background view remains the same. If the position of the camera is changed between sequences, the thresholds for background subtraction and morphological operations have been accordingly varied.

No background is made available, and is thus modeled by the background initialization algorithm [21] discussed previously. Certain irrelevant sections of the images have been masked out to facilitate computation.

i-LIDS DATASET

The Imagery Library for Intelligent Detection Systems (i-LIDS) dataset [28] was made available for academic research by UK Government's Home Office Scientific Development Branch, in association with the 2007 IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS 2007). i-LIDS is the UK Government's benchmark for evaluating smart video-based detection systems (VBDS) using CCTV imagery, and comprises of nearly 24 hours worth of images. It has been developed in partnership with UK's Center for the Protection of National Infrastructure.

The dataset provided for the Abandoned Baggage Detection Challenge comprises of three video clips from the original i-LIDS dataset, featuring scenarios of temporarily abandoned baggage shot at Westminster underground train station, labeled by their projected level of difficulty. Videos are captured in MJPEG format CIF-4 resolution of 576 x 704 (4:3 aspect ratio) with 25 fps (interlaced) and 8 bit color quantization. Zero padding framing the image makes the final resolution 576 x 720 pixels. Image sequences are decimated by 4 to facilitate processing.

The three videos involve varying degrees of scene density, baggage size and type. The railway platform is designated as the desired detection area, and divided into three zones for reference and indexing purposes, as shown in Figure 6.

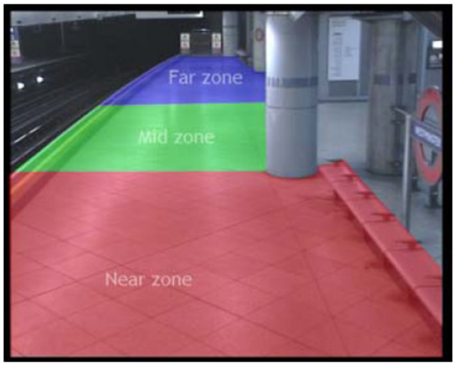


Figure 6: Demarcated detection area for the i-LIDS Abandoned Baggage detection challenge (Special session, AVSS 2007).

As evident from the image, the train tracks and the extreme top right corner area need not be scrutinized for suspicious bags. Thus, some part of it is masked out with the help of a manually designed binary mask, as shown in Figure 8. However, this simplification comes at the cost of losing some part of people standing along the tracks or by the bench. In our case, this loss did not cause affect system performance adversely. If necessary, one could attempt to overcome this limitation to some extent by extending blobs cut off using available region-growing schemes along with knowledge of the estimated background.

The key challenges of this dataset stem from severe and frequent occlusion, acute perspective distortion and difficulties in segmentation. Our system is able to successfully surmount these difficulties to yield impressive results, as shown in Table 1. All alarm times match within one second of the provide ground truth.

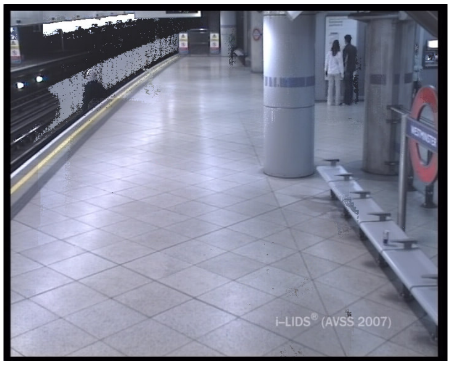


Figure 7: Initialized Background Model.

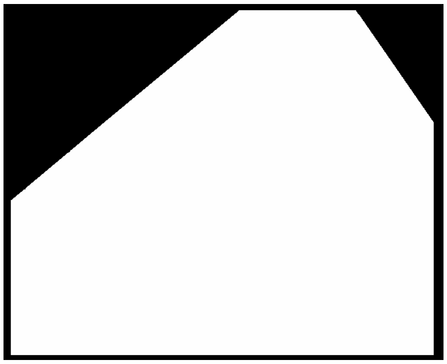


Figure 8: Binary mask used to highlight active region in consideration.

Sequence	Start time		Alarm Description	Alarm Duration		Object Zone
	Ground Truth	Our Result		Ground Truth	Our Result	
easy_AB	00:03:00	00:02:59	Abandoned Bag	00:00:12	00:00:12	Near
medium_AB	00:02:42	00:02:42		00:00:18	00:00:18	Mid
hard_AB	00:02:42	00:02:42		00:00:24	00:00:25	Far

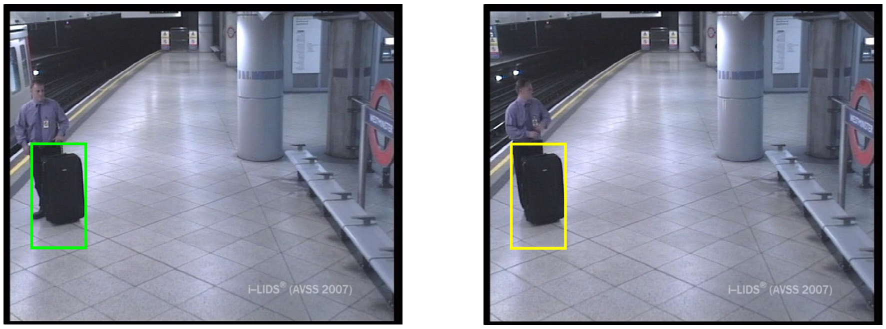
Table 1: Performance on the i-LIDS dataset.

A brief description of each individual video clip and details of the processing involved as well as the results obtained follows. The color of the bounding boxes of the bags indicates their status – Green marks a bag that *seems* to be left unattended; yellow denotes that the bag has been deemed as *unattended* and the timer has been set; red highlights a bag that has been judged as *abandoned*, and signals the alarm event. For this dataset, the warning time interval before the alarm, *T*, is defined as one minute.

Sequence easy_AB

This video clip features a large black suitcase in the possession of a gentleman who temporarily 'abandons' it in the near zone after shuffling about beside it for a while. The platform is very sparsely crowded, and the bag is never fully occluded. Even though much of his upper body is masked out, it is still relatively easy to identify the bag's owner, owing to the unique color of his shirt and the clear view of his hand.

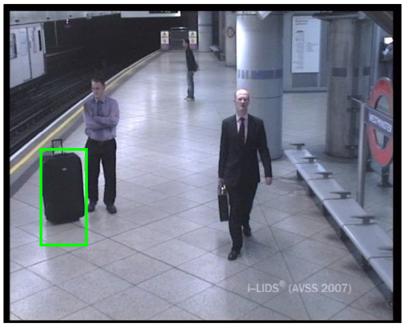
The difficulty of processing this clip arises from the nature of the mask, prominent shadow of the bag and the shuffling behavior of the owner. As the owner moves farther out towards the edge of the platform, more and more of his upper body is masked out, and at times the system sees his black pants as part of the bag. Thus, the bag is falsely suspected to be unattended at a number of times. At one point, the bag is deemed as unattended although the owner is beside it, and reverse traversal is initiated. Continued observation finds the owner within the next couple seconds and deactivates the warning. The situation can be seen in Figure 9. Another instance of where the system misconstrues the bag as unattended is when the owner steps away from it slightly while shifting about, as shown in Figure 10. When the owner does leave, the process is initiated all over again, this time accurately finding that the bag has indeed been abandoned.



Frame 2012

Frame 2040

Figure 9: Examples of errors made by algorithm. Bag misjudged as unattended due to recognition of bag merged with person’s trousers as a single bag entity. Warning is disabled in Frame 2268

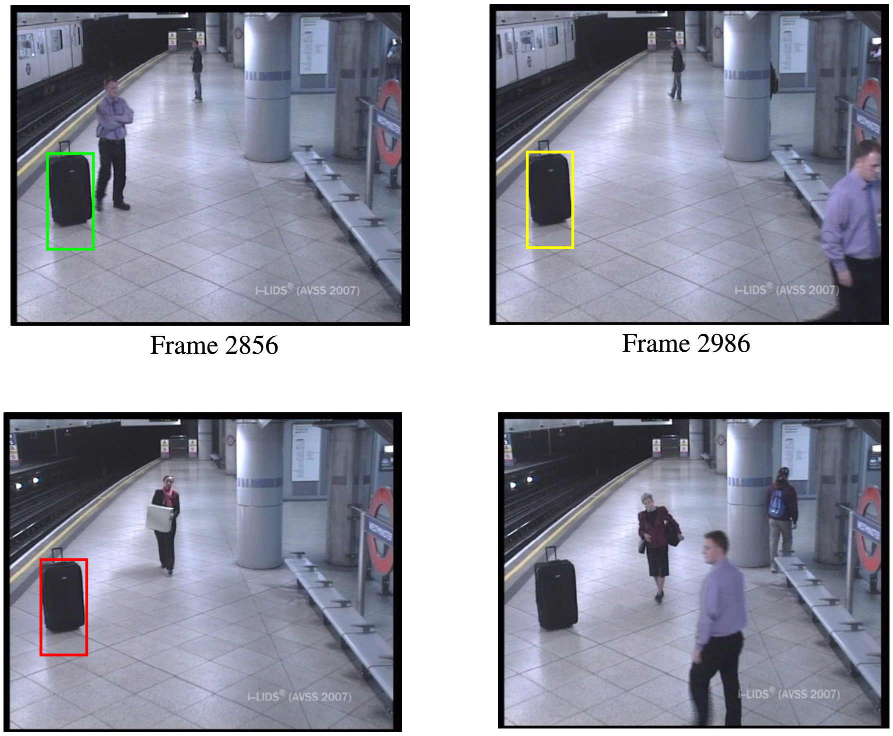


Frame 2768

Figure 10: System considers the bag as possibly unattended but rectifies its decision shortly.

At this point, there is no provision for storing previously computed bag-owner association. It is certainly most desirable and could probably be realized using motion information and other cues.

Results of event detection for this sequence are illustrated in Figure 11.



Frame 2856

Frame 2986

Frame 4484

Frame 4732

Figure 11: Results of processing sequence *easy_AB*. Frame 2856: Possibly unattended bag detected; Frame 2986: Bag deemed as unattended; Frame 4484: Alarm triggered; Frame 4732: Alarm disabled.

Sequence medium_AB

This video clip was, for this system, the hardest. A black sports bag is brought into the mid zone by a woman wearing white trousers that blended into the background completely. Hence the woman's complete body is never accurately segmented. This effectively reduces her corresponding blob size by half. Due to this segmentation problem, on several occasions, her upper body is misclassified as a possibly unattended object. At other times, even as it lays by her feet, the sports bag is detected as left alone. This An example of this can be seen in Figure 12. In this particular sequence, a lot of false positives are detected sporadically, but do not persist long enough to be categorized as unattended objects.

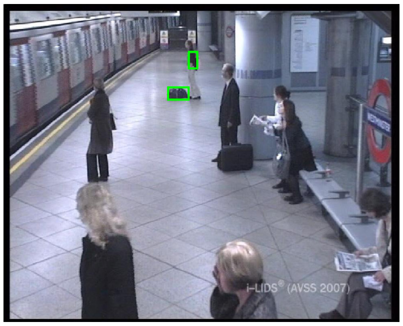
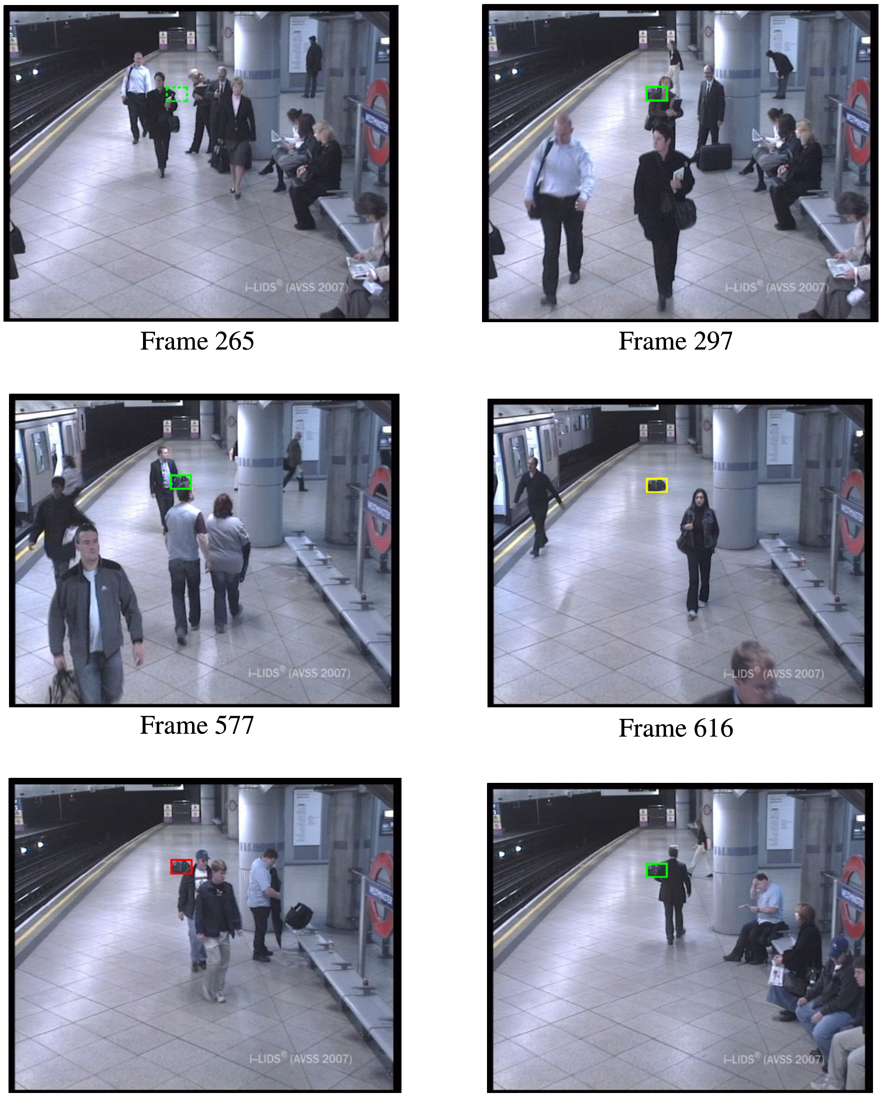


Figure 12: False positives obtained owing to the blending in of the off-white color of the owner's trousers with the initialized model of the platform floor.

The owner leaves the platform at a time of great activity, soon after the train pulls in. Her disappearance is not noticed until several frames afterwards, since the unattended bag is not detected as such until the scene has considerably cleared up. The patch bank needed for inspecting the scene for the owner is found correctly, and the timer is set instantly. However, since backtracking was initiated after a slight delay, the timer is also off by the same difference.

One way to remedy such a situation (where the owner is found missing in the very frame that initiated reverse traversal) would be to traverse in forward-time direction from the deduced drop-off point, until the exact point when the owner leaves. Control would then jump back to the present frame, and the timer would be set accordingly. Extending the same train of thought - it might even be possible to adjust system performance to within a predefined tolerance. For instance, if it were acceptable to have the alarm triggered up to say, 10 seconds late, the system could be designed to abandon its search for the exact time-point of the owner's departure once it reached within 10 seconds worth of frames past.



Frame 265

Frame 297

Frame 577

Frame 616

Frame 949

Frame 1059

Figure 13: Results of processing sequence *medium_AB*. Frame 616 shows the identified unattended bag, which initiates reverse traversal up to Frame 265, where the bag is not found. Tracking in forward direction to Frame 297 rediscovers the bag and candidate owners are recorded. The timer is set at Frame 577. 1 minute later, the alarm is sent off in Frame 949 and eventually disabled in Figure 1059, when the owner returns.

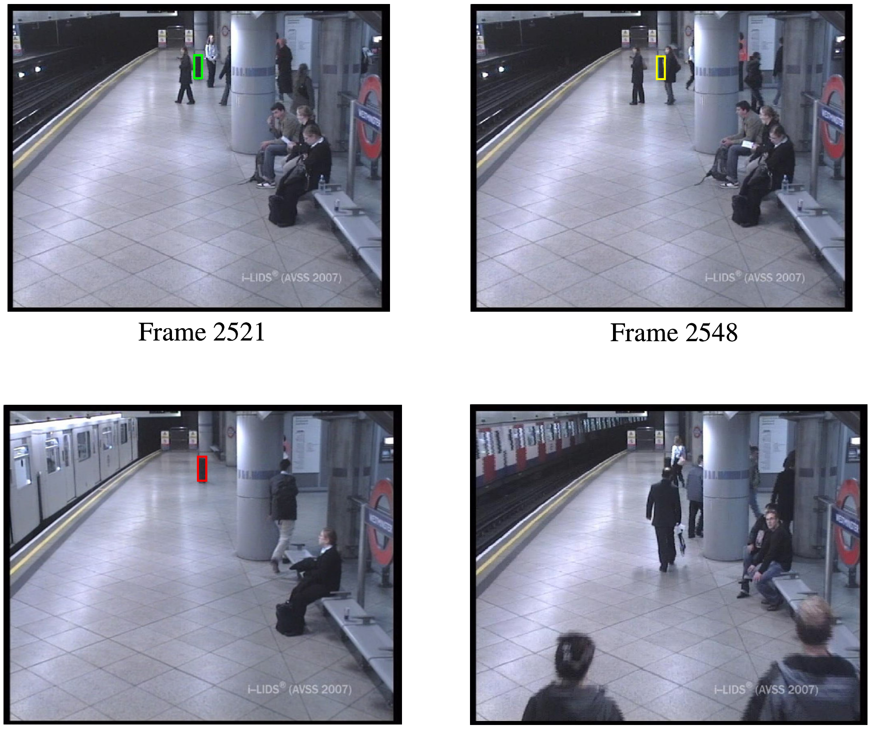
This solution, however, is rather idealistic in that it expects to trace the owner accurately for a substantial length of time in a crowded environment. The fact that the event detection was delayed is in itself an indicator of the difficulty of the situation. So while the method may improve the result by a few seconds, it may be more feasible to accept the slight delay in notification than to risk further error. Also, the additional complexity and computation are undesirable.

The results of processing this sequence are shown in Figure 13.

Sequence hard_AB

This sequence was found to be surprisingly problem-free. A black suitcase is temporarily abandoned by its female owner in the far zone of the platform. Objects in this area are very hard to discern even with the human eye. The owner is barely noticeable as she leaves the bag, even to the human observer, owing to her distance from the camera and the crowd. Similarly, her return to the bag is not obvious. In both cases, there is a lot of activity on the platform, and there is a lot of occlusion of both the owner and the bag.

The main difficulty encountered in this case is the fact that the bag aligns perfectly with the woman's trousers, in terms of both position and color. Thus the owner's blob, once again, bears fewer patches and risks matching accuracy. However, the results found were satisfactory, as displayed in Figure 14.



Frame 2521

Frame 2548

Frame 4048

Frame 4636

Figure 14: Results of processing sequence *hard_AB*. The lone bag is detected in Frame 2521 and put under observation in Frame 2548. After 60 seconds, in Frame 4048, it is deemed as abandoned until Frame 4636 when the owner returns.

PETS 2006 DATASET

The PETS 2006 benchmark data [29] consists of seven datasets recorded at Victoria Station in London, UK. The datasets comprise multi-sensor sequences containing left-luggage scenarios with increasing scene complexity. The workshop was sponsored by the EU project ISCAPS (Integrated Surveillance of Crowded Areas for Public Security), in collaboration with the British Transport Police and Network Rail. This IEEE International Workshop is yet another indicator of the recent acknowledgement for the need of technology in the battle against crime and for the efficient management of transport networks and public facilities.

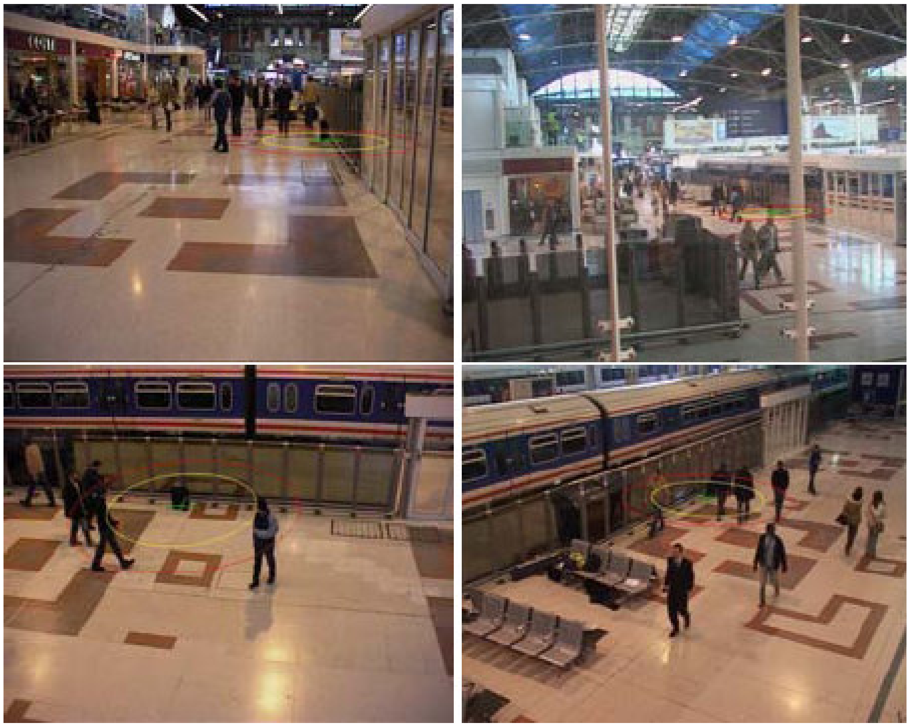


Figure 15: A sample of the different camera views, overlaid with zones of interest. Top row: Camera1 (left), Camera2; Bottom row: Camera3 (left), Camera4.

The event recognition challenge at PETS 2006 involves the identification of unattended and abandoned baggage using a network of four cameras at a train station that provide four distinctly different viewpoints. For our experiments, only the sequences filmed using Camera3 are used. The corresponding sight is the top-front view of the activity scene. The problem defines luggage to include all kinds of baggage that can be carried by hand, for example, trunks, bags, rucksacks, backpacks, parcels and suitcases. The dataset involves five types of bags, specifically, a briefcase, a suitcase, a 25 liter rucksack, a 70 liter backpack and ski gear carrier. Elliptical zones of different radii define the regions of detection.

All datasets were filmed at a resolution of 768 x 576 with 25 frames per second (PAL standard resolution) and compressed as JPEG image sequences.

The problem posed at PETS 2006 is very similar to that at AVSS 2007. The problem is defined more precisely, using three rules that are used to determine the status of a bag. A contextual rule states that a bag is owned and attended to by a person who enters the scene with it, for as long as it is in physical contact with the person. This corresponds directly to the definition of the owner that was adopted previously. A spatial rule designates the area within a radial distance *a* meters of the bag as the region within which the owner must be present for the bag to be considered as attended to. Finally, a spatio-temporal rule deems the bag as abandoned if the owner steps out of an area within radial distance *b* (such that $b > a$) for longer than *t* seconds. Figure 16 illustrates these zonal limits set in images taken by Camera3, with $a = 2$, $b = 3$ and $t = 30$. More details of the challenge and the workshop can be found in [29, 30].

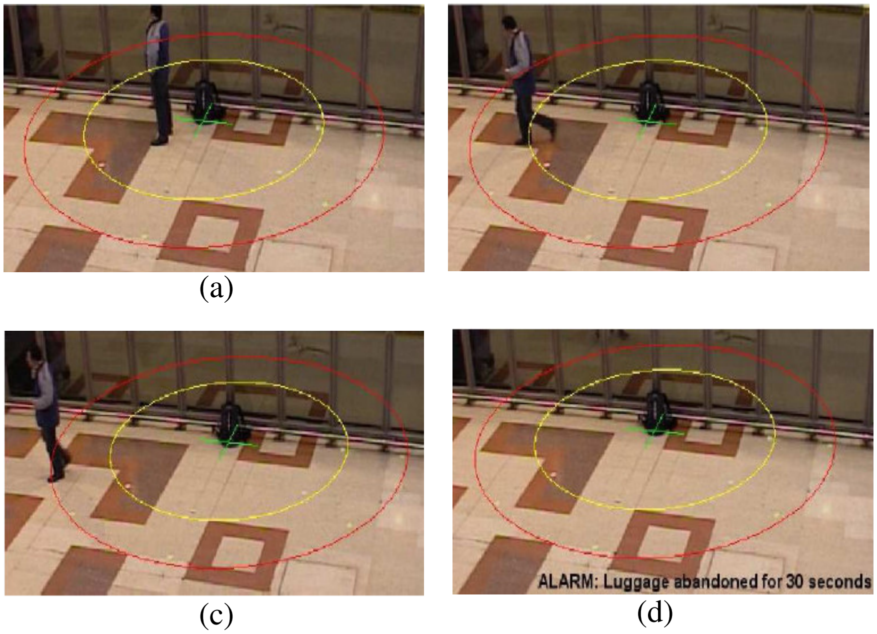


Figure 16: Zonal limits of detection. (a) Owner is within $a = 2$m radius; bag is attended to; (b) Owner steps into the intermediate warning zone outside radius *a* but within *b*; (c) Owner crosses the $b = 3$m radius around the backpack; (d) Bag left unattended for $t = 30$ seconds, at which point the alarm is triggered.

For our implementation, a coarsely estimated mask was used to model the critical b = 3m radial zone, centered at the centroid of the detected bag. The size of the mask was adjusted according to the position of the bag to account for perspective distortion.

While the problems posed at the two conferences are essentially equivalent, the datasets designed exhibit some very obvious dissimilarities. The most obvious distinction between the two lies in the number of people present in the scenes – the PETS sequences seem relatively bare as compared to the i-LIDS setup. Augmented by the fact that the selected view (from Camera3) exhibits significantly less perspective distortion, this makes the PETS tasks seem a lot simpler. Unfortunately, while some major issues are alleviated, the PETS dataset poses its own problems. The biggest challenge while working with this dataset arises from the dark silhouette-like appearance of most people on the scene. This is partly due to the fact that most people are wearing dark overcoats and pants, and partly due to the lighting and viewing angle. Thus, color and texture information is limited and not reliably discriminative. During initial tests, it was observed that the patch-based normalized cross-correlation module tends to assign ownership incorrectly to the largest (generally dark) blob that enters the demarcated zone, thereby inappropriately deactivating the timer for the alarm. Another instance of system failure is highlighted when the timer is reset upon detection of a fraction of the real owner (usually upper body) passes through the zone of detection. This event occurs, for instance, in Dataset S7, when the person loops halfway around the bag at a distance as he is leaving, i.e. the person walks outwards away from the bag and then turns around and exits the scene through the opposite side of the scene. This results in some portion of his corresponding blob being detected within the zone, and the bag is mistaken as attended to.

To avoid these glitches, we exploit the available spatial and contextual information to introduce an additional condition for recognition. Since foreground blob extraction is fairly reliable, we check to see whether the 'feet' of the blob fall within the interest zone. This simple clause enhances system accuracy significantly for sequences such as S7.

The seven sequences in the PETS 2006 dataset showcases a variety of different ways in which a bag may be abandoned. For this thesis, four representative sequences are analyzed. Each of these sequences is unique and demonstrates the capability of the designed algorithm in diverse settings. Results obtained are documented in Table 2, and illustrated per sequence as well. The legend for the bounding boxes is as follows: green indicates a possibly unattended bag detected, cyan encases baggage under observation, yellow marks the bag as unattended, red denotes alarm condition for abandoned baggage.

Sequence	Alarm Start Time		Baggage Description
	Ground Truth	Our Result	
S3	No alarm	No alarm	No bag abandoned
S5	110.56	No alarm	Ski equipment
S6	96.88	96.32	Rucksack
S7	93.96	93.76	Suitcase

Table 2: Performance on the PETS 2006 dataset.

Dataset S6

This scenario contains two people who enter the scene together. One person places a rucksack on the ground, before both people leave together (without the bag). The difficulty that arises in this dataset is in the precise identification of the owner of the deserted bag. Both people are treated as equally likely candidates for bag ownership. However, the system is biased towards the person in the blue sweater, owing to the significantly more texture and color information available for him as compared to the other, who is seen as almost uniformly black. In terms of event detection and alarm time, the system still does very well. This sequence is a demonstration of the less-than-perfect owner retrieval, and how it is not critical for overall performance.

The background model used for this sequence was obtained using [21] is as shown in Figure 17. Some parts of the scene can be omitted for processing; the corresponding mask used is displayed alongside the initialized background in Figure 16. Results obtained are shown in Figure 18.

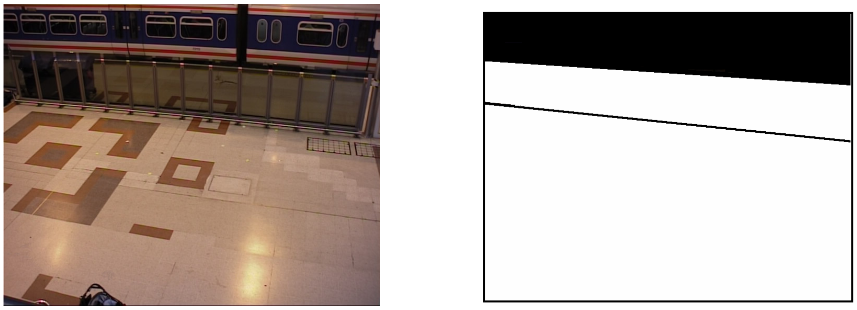
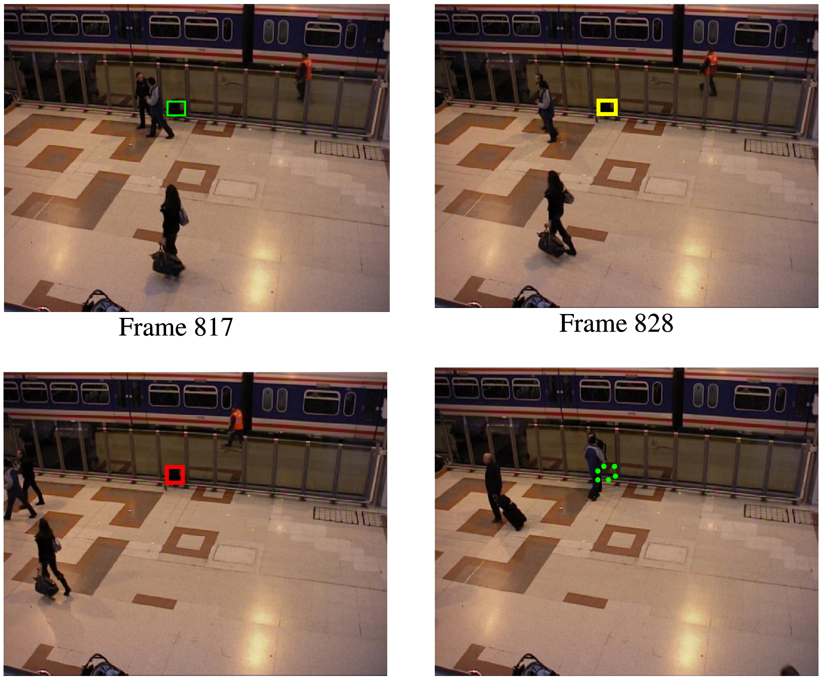


Figure 17: Initialized background model (left) and the corresponding mask used (right).



Frame 817

Frame 828

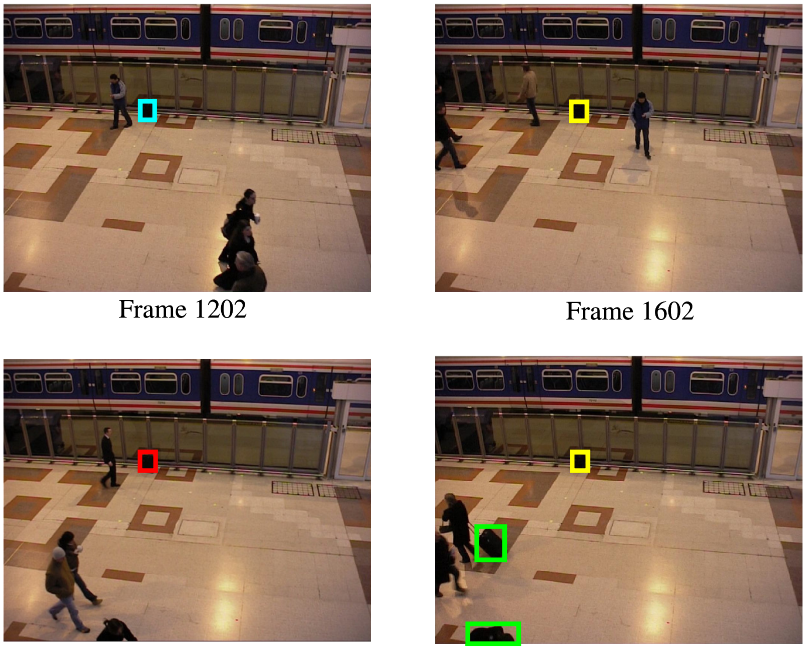
Frame 1203

Frame 666

Figure 18: Results of processing S6.

Dataset S7

This sequence is rated at the highest level of difficulty (5/5), and features the most activity in the complete corpus. The scenario contains a single person with a black suitcase who loiters in the scene, outside the defined neighborhood of the suitcase, before leaving the item of luggage unattended. As discussed previously, in this sequence, the owner traces a partially (roughly) circular clockwise route away from the bag before leaving the scene at the left edge of the image. During this event, five other people move in close proximity to the item of baggage. At two isolated instances, a passer-by is mistaken as the owner, but since classification decisions are made by analyzing results over 3 consecutive frames, these false positives are eliminated.



Frame 1202

Frame 1602

Frame 2328

Frame 2196

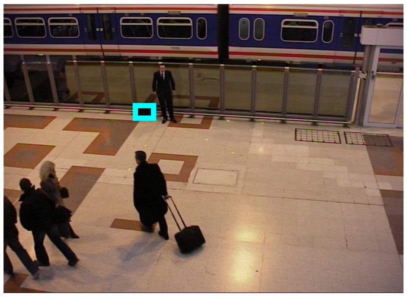
Figure 19: Results of processing S7: The bag is put under observation at Frame 1202, and is left unattended in Frame 1602. 30s later, an alarm is set off at Frame 3628. Frame 2198 is a good example of false positive results extracted by the system as potentially unattended

The true background was exposed for a short period during this sequence, hence background initialization was unnecessary. A median image over the relevant sequence served as an accurate background. Results obtained are shown in Figure 19.

Dataset S3

This sequence contains a person who enters the scene with a black briefcase which he then sets down beside himself temporarily. The briefcase is visible as a distinct entity, not conjoined with any other blob. The owner stands by for a while before picking it back up and leaving. It is important for any intelligent system to distinguish between such a routine occurrence and the suspicious event of bag abandonment, hence the significance of this test. Our results is shown in Figure 20.

Once again the true background information is available. Requisite processing is relatively simple and straightforward. The system promptly identifies the briefcase as a potentially unattended object, but since the owner is in close proximity the whole time, it is never even given the status of unattended, and no alarm is triggered.



Frame 951

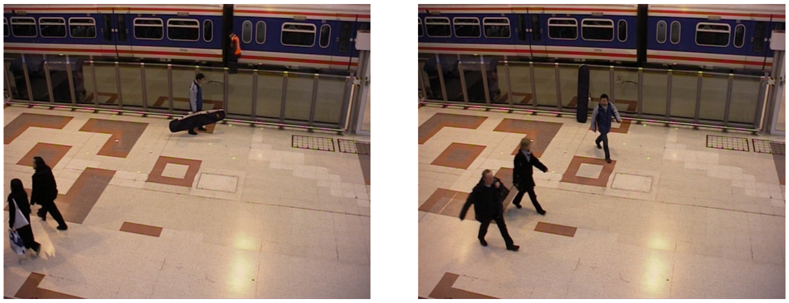
Figure 20: Result of processing S3: Briefcase and owner monitored.

Dataset S5

This sequence has a person carrying skiing equipment who loiters around for a little while before abandoning the item of luggage. By itself, the sequence is relatively uncomplicated since there are few people and little occlusion within the interest zone. However, the unattended object detection module fails to identify the ski carrier that is left behind. This is because the training data that was made available to the nearest neighbor classifier did not include any instances of such baggage. On the contrary, the size (height) of the bag and its slender shape match closely with that of several humanoid training samples provided.

This failure is a great illustration of the need for a carefully selected data-specific training set. It also demonstrates the need for a richer bag representation that can handle more unusual types of baggage as well.

An image showing the owner walk away from the skiing equipment is shown in Figure 21.



Frame 1103

Frame 2002

Figure 21: Sample sequences extracted from Dataset S5. System fails to detect the skiing equipment due to insufficiently representative training data.

Chapter 5: System Evaluation and Discussion

This thesis describes a unique framework to detect objects abandoned in a busy scene. The algorithm is, to the best of our knowledge, novel and unique. The proposed algorithm is appealing in its simplicity and intuitiveness, and is demonstrated experimentally to be conceptually sound. It is well-equipped to handle the concurrent detection of multiple abandoned objects swiftly, in the presence of occlusion, noise and affine distortion. The algorithm lends itself naturally to the recognition of a vast variety of related activities, ranging from surveillance and corridor observation to traffic management and cargo monitoring. Its modular structure allows the flexibility for integrating more functionality and sophisticating various sub-modules without disturbing the remaining framework.

The current algorithm derives its strength from its simplicity of concept and implementation, and performs well even in the presence of substantial occlusion and activity. Experimental results obtained by testing on two benchmark data corpuses, each comprising of multiple diverse scenarios, attest to its effectiveness and potential. Nearly every experiment yielded impressive results which are within a second of the ground truth provided. Instances of imprecision or failure are reasonable, and can probably be overcome using more advanced techniques and due experimentation.

The methodology presented in this thesis can be easily adapted to the multi-camera case, and would certainly benefit from the collaborated observation from different views, provided that the cameras are strategically positioned. 3D object recognition would significantly enhance the capability of the system. Some of the current uncertainty associated with the identification of the rightful owner of the abandoned object would possibly be ironed out. However, the deployment of multiple cameras per location is usually not practical in vastly spread systems such as the railways. Our goal is to be able to use existing camera networks for monitoring public areas, demanding little or no changes or additional expense. Thus, this work currently focuses on event recognition from monocular image sequences.

The performance and success of our methodology is promising, but much remains to be done. The current MATLAB implementation is computationally sub-optimal. The system can certainly be easily parallelized. A binary search may be performed to look for the 'drop-off' point of the baggage. Several more advanced methods can be used to patch selection and object recognition. For more reliable segmentation, it would be worth exploring ways of periodically updating background, or even using different backgrounds in different contexts (for example, a background with the train at the station).

A thorough evaluation of the algorithm would not be complete without addressing some of its limitations and pitfalls. Unfortunately, there are several cases where the system is likely to fail. One prime example is the situation where the view or circumstance is such that everyone appears uniformly dark or textureless, so that patches extracted from the candidate owners would be terribly ineffective at modeling the appearance of the person. No known patch-based recognition method is capable of handling the lack of information reliably. The only possible way of overcoming this problem would be the strategic repositioning of the camera, so that details are sufficiently visible. That may still not solve the issue of distinguishing between people dressed identically, especially in uniform color. A second, less practical solution would be to use high-resolution images that can capture smaller details that would otherwise go unnoticed.

Another problem with the current system was discussed previously in the case of the i-LIDS *medium_AB* sequence. At times when there is no clear view of the object, owing either to occlusion or merging with another object, the system's observation of the lone bag is likely to be delayed. Such an eventuality could prove rather dangerous and would certainly undermine the capability of automated visual surveillance. The forward traversal solution proposed earlier involves further computational expense and can still not guarantee success. Individual tracking was avoided in this project due to the immense obstacles offered by frequent and severe occlusion and perspective distortion. Tracking and object recognition are prone to failure in unconstrained environments such as the train stations considered here. It may however, be necessary to incorporate some form of tracking and motion information to back the system up in times of uncertainty.

There are certainly many possible directions for future work on this project. Other than adding measures to counter some of the shortcomings discussed above, one obvious extension would be the refinement of the object recognition module. Presently, the system is designed to detect abandoned baggage only. It would fail to detect other objects and more unusual types of luggage such as ski equipment carriers, tripod holders, trolleys, etc. In order to handle more general kinds of objects that may be left around, the system would need to be equipped with more advanced object detection and recognition techniques. Also, a device to separate merged or partially covered objects is needed.

Automated video summarization is another interesting arena that would be very useful to explore. Because the operation of our system was meant to mimic the human monitor, it would be a natural extension of its functionality to extract such information automatically, thereby saving precious reaction time. The algorithm is designed to hunt for certain key events in time that are significant in decision making, namely, the appearance of a suspicious-looking object, identification of the person(s) responsible for leaving it, and the behavior of the person over the period. To ascertain the threat level of the situation, security personnel are most likely to base their judgment on these same events. It would benefit them tremendously to be able to automatically zoom into these relevant frames of critical activity. Concatenating snippets of these sub-event intervals from the video would not be a difficult task.

There is tremendous scope for experimentation and improvement of the current system. It is nonetheless, a step towards effective, efficient monitoring of objects in challenging public environments.

References

[1] C. Sears and Z. Pylyshyn, Multiple Object Tracking and Attentional Processing, *Canadian Journal of Experimental Psychology*, Vol. 54, pp. 1-14, 2000.

[2] P. Cavanaugh and G. Alvarez, Tracking Multiple Targets with Multifocal Attention, *Trends in Cognitive Sciences*, Vol. 9(7), pp. 349-354, 2005.

[3] C. Eriksen and St. James, Visual Attention Within and Around the Field of Focal Attention: A Zoom Lens Model, *Perceptual Psychophysiology*, Vol. 40, pp. 225-240, 1986.

[4] F. Tong, Splitting the spotlight of visual attention, *Neuron*, Vol. 42, pp. 524-526, 2004.

[5] I. Haritaoglu, R. Cutler, D. Harwood and L. Davis, Backpack: Detection of People Carrying Objects using Silhouettes, *Proc. IEEE International Conference on Computer Vision*, Vol. 2, pp. 102-107, 1999.

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

[7] M. Spengler and B. Schiele, Automatic Detection and Tracking of Abandoned Objects, *Proc. Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2003.

[8] Lv, X. Song, B. Wu, V. Singh and R. Nevatia, Left-Luggage Detection using Bayesian Inference, *Proc. IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 83-90, 2006.

[9] S. Guler and M. Farrow, Abandoned Object Detection in Crowded Places, *Proc. IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 99-106, 2006.

[10] C. Stauffer and W. E. L. Grimson, Learning Patterns of Activity Using Real-time Tracking, *Proc. IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 22(8), pp. 747-757, 2000.

[11] L. Li, R. Luo, W. Huang and H. Eng, Context-Controlled Adaptive Background Subtraction, *Proc. IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 31-38, 2006.

[12] H. Grabner, P. Roth and M. Grabner, Autonomous Learning of a Robust Background Model for Change Detection, *Proc. IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 39-54, 2006.

[13] J. Martinez-del-Rincon, J. Herrero-Jaraba, J. Gomez and C. Orrite – Urunuela, Automatic Left Luggage Detection and Tracking Using Multi-Camera UKF, *Proc. IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 59-65, 2006.

[14] N. Krahnstoever, P. Tu, T. Sebastian, A. Perera and R. Collins, Multi-View Detection and Tracking of Travelers and Luggage in Mass Transit Environments, *Proc. IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 75-82, 2006.

[15] E. Auvinet, E. Grossmann, C. Rougier, M. Dahmane and J. Meunier, Left-Luggage Detection using Homographies and Simple Heuristics, *Proc. IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 51-58, 2006.

[16] K. Smith, P. Quelhas and D. Gatica-Perez, Detecting Abandoned Luggage Items in a Public Space, *Proc. IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 75-82, 2006.

[17] J. Allen and G. Ferguson, Actions and Events in Interval Temporal Logic, *Journal of Logic and Computation*, Vol. 4(5), pp. 531-579, 1994.

[18] M. S. Ryoo and J. K. Aggarwal, Recognition of Composite Human Activities through Context-Free Grammar Based Representation, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1709-1718, 2006.

[19] R. Nevatia, T.Zhao and S. Hongeng, Hierarchical Language-based Representation of Events in Video Streams, *Proc. IEEE Workshop on Event Mining*, 2003.

[20] D. Gutchess, M. Trajkovic, E. Cohen-Solal, D. Lyons and A. K. Jain, A Background Model Initialization Algorithm for Video Surveillance, *Proc. IEEE International Conference on Computer Vision*, pp. 733-740, 2001.

[21] Chia-Chih Chen and J. K. Aggarwal, An Adaptive Background Model Initialization Algorithm for Video Surveillance, (*submitted for publication*), 2007.

[22] R. Duda, P. Hart and D. Stork, *Pattern Classification*, Wiley Interscience, 2nd Edition, ISBN: 0-471-05669-3, 2000.

[23] D Comaniciu and P. Meer, Mean Shift Analysis and Applications, *Proc. IEEE International Conference on Computer Vision*, pp. 1197-1203, 1999.

[24] J. Lewis, Fast Normalized Cross-Correlation, *Vision Interface*, 1995.

[25] F. Zhao, Q. Huang and W. Gao, Image Matching by Normalized Cross-Correlation, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 729-732, 2006.

[26] C. Harris and M. Stephens, A Combined Corner and Edge Detector, *Fourth Alvey Vision Conference*, pp. 147-151, 1988.

[27] D. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Vol. 60(2), pp. 91-110, 2004.

[28] i-LIDS dataset for AVSS 2007, http://www.avss2007.org.

[29] Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, http://www.pets2006.net.

[30] D. Thirde, L. Li and J. Ferryman, Overview of the PETS2006 Challenge, *Proc. IEEE International Workshop on Performance Evaluation and Tracking for Surveillance*, pp. 47-50, 2006.