An Object Detection Based Approach to Abandoned Luggage Identification

Fazail Amin
1821CS01

Under The Supersivion of
Dr. Arijit Mondal
Dr. Jimson Mathew

Progress Report
For Registration Seminar

Department of Computer Science & Engineering
IIT Patna

Contents

List of Figures

List of Tables

An Object Detection Based Approach To Abandoned Luggage Identification

Abstract

Video surveillance is a crucial part of public safety and security systems. We propose a detection based approach for abandoned object localization and owner identification in video surveillance systems. For correctly identifying the abandoned objects and corresponding owner or the owner group an improved method for association has been proposed. State of the art techniques either use back tracing for owner identification or there is no provision of reporting ownership of the objects abandoned. For a technique to be applicable to a wide array of situations it must be tested on such kind of environments, to this end we provide an elaborate surveillance dataset covering many complex cases which are not available in previous datasets. Our method provides a robust way to flag the abandoned objects along with proper localization and identification of the owner and the time of the drop. These information are very crucial for further processing and investigation and can help in alleviating unwanted incidents in public places.

1 Introduction

Abandoned object detection in public places is a critical application which requires robust systems that can run with minimum human intervention and high accuracy. With improved image processing, machine learning, deep learning methods coupled with improvements in computational power, it is possible to deploy such systems, still many open challenges remain to be addressed. Even though it is possible to get a blanket cover of the area under surveillance using more and more number of cameras but analysing the video stream continuously for suspicious event or activity is a tedious job and needs human intervention along with several other challenges. For example, obtaining a generic model which can work for different scenarios is very complex because of the reasons like difficulty in understanding social groups, crowded scenes, variety in objects at different locations, various shapes and sizes of objects. Achieving real time processing of image data is another roadblock for obtaining efficient and practical systems. The randomness in the public spaces and crowd movement poses a very challenging task to model any rule based system which may capture the randomness in its true form. There are many factors associated with these systems like sudden illumination changes, shadows, partially occluded and fully occluded objects make the problem very challenging.

2 Literature Survey

With improved background modelling techniques [10][11][12], static foreground can be detected very precisely by using them in background subtraction methods. It is possible to get stationary objects in the videos very accurately in terms of localisation with very few false detections. Tracking methods [13][14][15] have also helped in improving the solutions for this problem by providing reliable tracking of objects, but these methods still need a lot of improvements in terms of owner identification, tracking in crowded areas, occlusion handling, illumination or brightness changes etc. Also, Convolution Neural Network (CNN) based detection models have shown impressive performance in classification and localization of objects and can be used to replace or improve tracking based frameworks. The biggest hurdle in using deep CNN models is their speed and computational requirements. There are various techniques [16][17] for model compression which provide efficient ways to reduce the size of models such that they can be used in devices with very low memory and processing power.

From the experimental point of view abandoned event needs to be defined with temporal and spatial constraints. As shown in Figure 1, an object is considered abandoned if both the temporal and spatial thresholds are crossed. If the object is unattended for an interval less than temporal threshold for abandonment before the owner re-attends it, thus alarm should not be raised in such cases. Also if the person is within a distance less than spatial threshold δ of the object for more than temporal threshold τ, alarm should not be raised.

Following are the contributions of this work:

- Created an elaborate dataset which captures many complex and challenging abandoned object scenarios which are not covered in previous datasets. This has been made publicly available for testing such systems.
- Designed an integrated framework that achieves the goal of abandoned object detection and owner identification with special emphasis on small object detection.

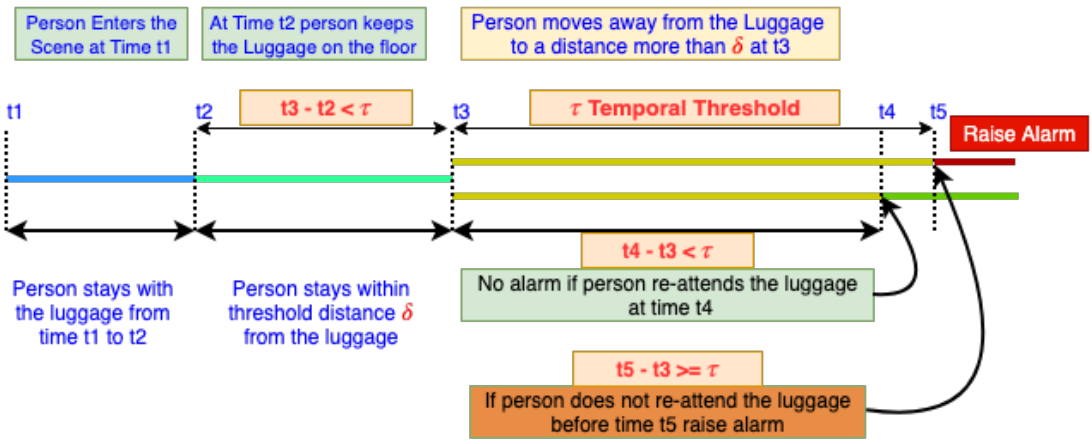

Figure 1: Timeline of abandonment and non-abandonment events.

- Proposed an algorithm for providing stable identities to overcome occlusion and person re-entry problems in the scene.
- For small object detection deep CNN model has been fine tuned and retrained specially for detecting luggage items including backpack, handbag, suitcase, rucksacks and similar items.

The work done in the area of abandoned object detection can be categorised as shown in Table 1.

2.1 Background Subtraction Based Techniques

Classical methods for abandoned luggage detection systems are primarily based on background subtraction techniques [18] [19] [20] [21]. Here the scene under observation consists of background plus active foreground (parts of the scene which are not stationary) and the goal is to find the static foreground, i.e, parts of the scene which are stationary for a certain interval of time. Static foreground contains the regions which were earlier part of active foreground, i.e, non stationary objects, have stopped moving. Once the static foreground is obtained, contours can be drawn around it and passed to the classifier to filter persons and the rest can be assumed abandoned objects if they meet the temporal and spatial threshold criteria set for the abandoned items.

In [22] proposed a method to find stationary foreground by combining resultant of two different background models with Bayesian update. Another method [29] used sequence of frames and select the ones where intensity values are almost constant as background. Using optical flow, it determines which interval of frames represents the actual background. Foreground mask sampling proposed in [30] uses background subtracted frames and their intersection to find the abandoned objects. Mixture of Gaussian model in [14] [25], uses a combination of three Gaussians to model the distribution of pixels in each frame. The method of [23] uses two background models, utilises initial frame as reference background model for first background model and uses consecutive frames to model the second background. The

Table 1: Techniques and Associated Problems.

S.No.	Category	Approach	Problems
1.	Background modeling and subtraction	Static background models [18]	Can not consider changed background. Affected by clutter and lighing change. Occlusions cannot be handled.
		Dynamic background models [19][20][21][14]	Rosbust to dynamic backgrounds but occlusions cannot be handled
		Dynamic dual models [22][23]	Can handle changing backgounds robustly but sensitive to update rate of thresholds
2.	Tracking based models	Kalman filter based approaches [10] [13]	Difficult to create hypthosis for traking when backgound clutter is very high
		Bayessian inference approaches [24]	Occlusion handling and owner associaiton is difficult to handle
3.	Background subtraction and tracking	Background modelling coupled with tracking [25]	Helps alleviate problem of clutter but owner association and occlusion handling are difficult
4.	Detection based systems	Object detector based approaches [26][27] [28]	Depends highly on the robustness of detector and may miss very small or difficult to detect objects
5.	Detection and tracking based systems	Object detection combined with tracking [15][29]	Heavily depends upon accuracy of detector and to some extent. Can handle occlusion but may miss difficult ot detect objects.

intensity difference is taken between corresponding pixels of the two models, a low value represents that the pixel is part of background. The technique proposed by [31] combines the short-term and long-term background models to extract the foreground objects.

2.2 Object Tracking Based Techniques

This approach involves application of one of the background subtraction methods to find out static regions of the image and then detect the objects from the static foreground region. Detected objects are then tracked for abandoned event by analysing the tracks continuously. As in [13] tracks are analysed for splits and if one of the split objects remain at a location with zero velocity it is assumed to be stationary and it is tracked for abandonment. Instead of detecting the objects in the foreground, simply a blob tracker can be used to track the objects based on the location, size and aspect ratio as in [24]. If a moving blob stays stationary for an interval above the threshold, it is flagged as abandoned.

2.3 Detection Based Techniques

This category of methods rely upon object detection and feature extraction techniques to find objects of interest as in [26][32]. The approach of [32] is based on features extracted using SIFT (Scale Invariant Feature Transform). PCA is then applied to extract significant features. For classifying the objects as luggage and non luggage various classifiers have been used, like [26] applied HOG (Histogram of Oriented Gradients) learned through luggage images. This approach looses its accuracy due to missed detections in some cases.

In [27] static objects are obtained using background subtraction and motion estimation, then static objects are classified by using a cascade to two deep CNNs, as attended or unattended. First CNN classifies objects as luggage and non luggage whereas the second CNN classifies whether it is attended or not. Another method by [28] uses background subtraction and then luggage detection using lightweight CNN. This method achieves real time performance. CNN based systems look promising since they can provide a way to identify objects

in more natural way instead of classifying objects as person and non person only, as done in previous approaches.

3 Motivation and Objectives

Most of the techniques rely on the robustness of background subtraction method applied. Adaptive background subtraction methods are quite effective in detecting unknown, removed, or changed objects in a video. It is used in conjugation with object tracking as in [10][11]. Approach of [22] is reasonably good in this regard, it uses dual foregrounds for finding static regions which are probable candidate for abandoned objects. The problem with this approach is that, it is heavily dependent on the rate of update of the two foregrounds which needs to be carefully tuned and there is no precise way, except heuristic methods, to come up with reliable values. Adaptive background subtraction based methods pose problems where stationary foreground objects may get updated into the background before they are actually processed for an abandoned event. Performance of these methods are also affected by foreground clutter. Lv et al. [24] used Kalman filter based blob tracker to track objects in the image and corresponding event is detected using Bayesian inference model, tracking all objects of the scene is very complex for crowded scenarios. Stauffer and Grimson in [15] present a model which classifies the object using neural network but it can detect only one object at a time. Bhargava et al. [29] uses a model where constituent sub-events adds up to object abandonment. Though their algorithm works well under various levels of occlusions, perspective distortion and it can detect multiple abandoned objects, but it fails in the cases where people are wearing dark or texture-less clothes.

Most of the systems do not not consider association of objects with persons, they assume those in a blob who appeared first with the luggage to be the owners. There is no way of tracking them further in future and the information of owner is not used efficiently to infer the abandonment. Tracking based methods have also been used but maintaining tracks for each and every object in a scene is not a practical approach where there is high occlusion and scene variations due to randomness in the movement of crowd.

4 Proposed Integrated Framework for Abandonment Detection

We have proposed an integrated framework for abandoned object detection and owner identification which can handle complex scenarios like person re-entry, short term occlusions, ownership establishment in groups.

4.1 Objectives of the Proposed System:

- To find and localize any abandoned object in a given scenario.
- Identify the owner or the group of owners of the abandoned object.
- Reduce false alarms in cases where a person leaves an object and re-attends the same after some time.

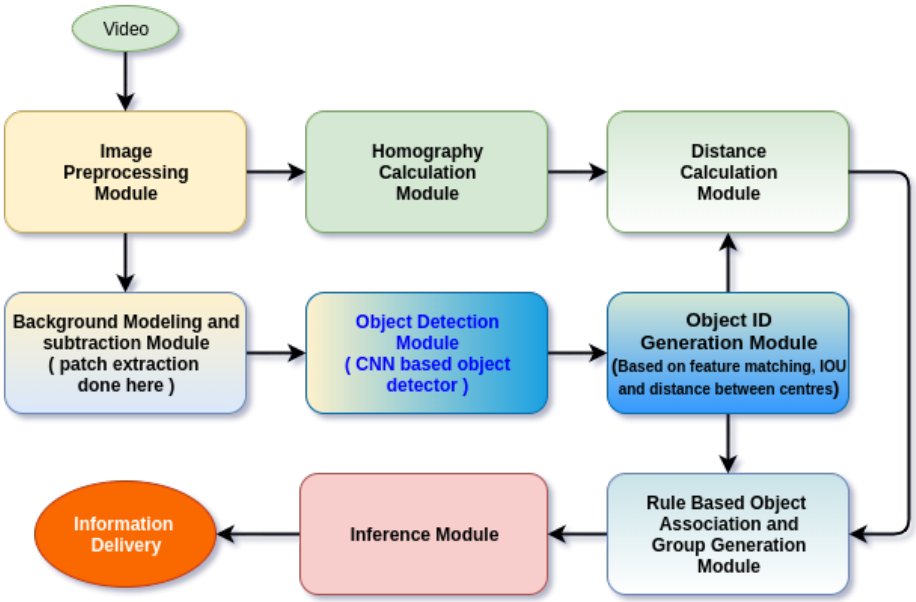

Figure 2: The Whole Abandoned Luggage Localization and Owner Identification Pipeline.

- Designing reliable associations between object and persons in the scene to establish ownership correctly.

4.2 General Outline of the Proposed Method

To achieve our target of detecting and localizing the abandoned objects and corresponding owners in a scene, we need to find static foreground objects. Identify and keep track of owners of the objects in the scene. If an object is not attended by its owner or the group of owners for the temporal and spatial threshold, then flag that object as abandoned along with the owner's detail.

The proposed framework is shown in Figure 2, where, after basic preprocessing background subtraction method is used to find the portions of frames which are not the part of the background. Once this foreground is obtained, detection module can be run on the patches of the images obtained from the foreground. These detected objects are then passed to another CNN which extracts the features for cosine similarity matching, so that stable identities can be provided. Persistent objects are possible candidates for abandoned object. Using distance between object and persons, an association table is generated which serves as the way to establish ownership of the detected object. Since people may come in groups in the scene, we need to take care of the cases where a luggage is owned by a group instead of a single person so that false alarms can be avoided. Using distance between person and objects we can find the occurrences of object in the scene where there is no owner near it within a threshold distance.

4.3 Technical Details of the Proposed Method

4.3.1 Video Preprocessing Module

Videos from various surveillance cameras can have different resolutions, frame rates, distortions etc. These parameters must be obtained from the input videos, to make necessary provisioning for feeding them to the system. CNN models expect the input to be of a particular dimension so the video frames need to be resampled to correct shape. Also radial distortion in the images from surveillance videos is very common because of the usage of fish eye cameras to get more coverage, this distortion should be corrected before the video is processed by the CNN module. Any kind of distortion will cause inaccuracies during distance calculation.

4.3.2 Background Subtraction Module

Background subtraction is essential component to find active areas in a surveillance video. In this module, background model is maintained using dynamic approach. As shown in Figure 3, when a new frame comes in, it is compared against the background model and the difference image is calculated. After thresholding the difference image, binary foreground mask is obtained. Using foreground mask contours are drawn around probable objects. From these contours bounding boxes are drawn and top left and bottom right of each bounding box is returned. Patches which are very small in size or completely contained within other bigger patch should be dropped and only the larger patch will be kept as shown in Figure 3.

For dynamic background Gaussian mixture model as implemented in [12] is used. Since running the detection directly on the image itself is very likely to miss the objects because models are not capable of detecting objects which are very small when run as whole image. Since the image is downsampled before passing to the detector so the small objects loose a lot of information and are not detected consistently. Hence feeding the patches helps in improving the detection quality.

4.3.3 Object Detection Module

This module consists of a deep CNN RetinaNet [9] which is state of the art in object detection and it has been retrained further to detect luggage items. Since detecting small sized objects is very challenging as there are very low number of discriminating features which can be obtained from it. Customizations have been done to the CNN model to make it suitable for this framework. Last dense layer is retrained on downsampled images and training is done with reduced anchor sizes to detect small objects.

Since running detection directly on images will result in poor performance, we use the patches obtained from background subtraction module to feed the neural network for detecting the objects. Feeding small image patches cause another problem, since the patches are rescaled to full size image before feeding to the detector causes poor detection which is alleviated significantly by introducing some context area around the patches. Heuristically we found that 3 times the size of patch works best in our experiments. One of the major requirements from the detection module is that it should provide tightest possible bounding boxes because the base of these boxes are used to measure the distance between objects in the image.

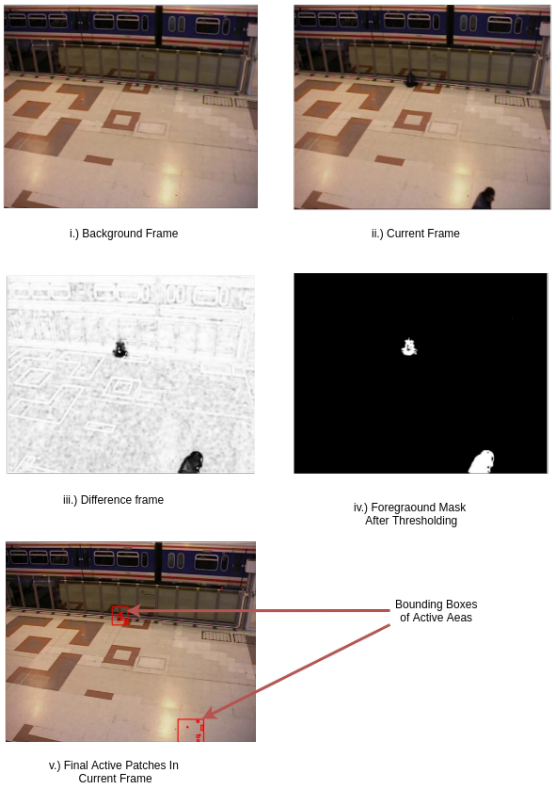

Figure 3: Patch Extraction Using Background Subtraction.

4.3.4 Feature Extraction Module for Similarity Matching

This is a CNN based feature extractor [33] and it is a pre-trained model. This model is capable of learning metric representation and the features obtained from this are resistant to small changes in background, lighting and articulations. It returns feature vector of length 128. This CNN takes images of size 128×64, images are rescaled before feeding into the network. While passing through the network, size of the feature map reduces to 16×8 from which a vector of length 128 is obtained by final dense layer. Each convolution layer has 3×3 filters only, including max pooling layer which is also implemented as 3×3 filter with stride length of 2. Feature extractor for similarity matching is a pre-trained model [33] which is used without retraining.

Apart from using the extracted features, shift between centres and IOU of two objects being matched are also used. Using these three parameters it is possible to obtain stable IDs for the detected objects.

4.3.5 Single Camera Distance Calculation and Homography Module

One of the most important parts of our surveillance system is to measure the distances between objects. Distance measurement becomes more challenging if only a single camera view is available. Usage of single camera reduces the computations to a great extent as less

number of images need to be processed. Since the surveillance systems need to work on live video feeds, a multi camera feed can have different delays and a careful preprocessing will be required before it can be further analysed. For this purpose ground plane homography [18] is used. As shown in Figure 4 first row images, it is not possible to calculate the distance between objects on the floor correctly but once the images are transformed using ground plane homography as shown in second row images of Figure 4, distance calculation on the ground plane becomes feasible. These distances between pixels can be mapped back to actual distances since the real dimensions of the floor are known. First we calculate the homography matrix, H, using set of eight points (four points in pixel coordinates from image and four corresponding points known from the actual geometric information of the floor) as follows:

$$w = Hx$$

where $x \in Z^3$ is the pixel coordinate in the image and $w \in R^3$ is the corresponding point in the transformed coordinate system. w is further rounded off to integer coordinates. The homography matrix H arising from the equations obtained from 8 selected points is calculated using the method of [34].

4.3.6 Object Association and ID Generation Module

It is crucial to create association between luggage items and persons. In real life scenario, the cases are very common where single luggage may be owned by multiple persons or multiple luggage may be carried by a single person. To handle these kind of situations, association table is maintained where luggage and corresponding owners are kept in one group with unique ID.

In most of the traditional approaches, a person is associated with an object in the scene if that person and object appear together in the scene. This approach is not very reliable because there may be other persons also who will enter at the same time as the owner of the object. This is improved by using simultaneous appearance within a threshold number of frames from the first appearance of the object. Since we are using detection based method, chairs and benches can be detected and if an object is detected whose base is completely inside the chair then the base of the chair is considered to be the base of the object. In our framework owner will be the person who first appears with the object, stays within close distance from it and moves in the same direction as the object. Using these assumptions ownership will be established where a person needs to be within the spatial threshold distance δ for an association temporal threshold τ_{asso} from the first appearance of the object and both move in the same direction. The person will be marked as owner and a unique group ID will be given to this pair. In our experiments δ is chosen as 2 meters and τ_{asso} is selected as 10 seconds. Using the same method groups are created. All the persons who satisfy the criteria will be added to the same group.

4.3.7 Inference Module

When a frame is processed, distance between objects and persons are obtained. If no member from the owner group is present near the object, it will be moved to a tentative list of abandoned objects. This list will be monitored and if the occurrence of items exceeds a predefined temporal threshold, then this item will be added to the list of abandoned objects.

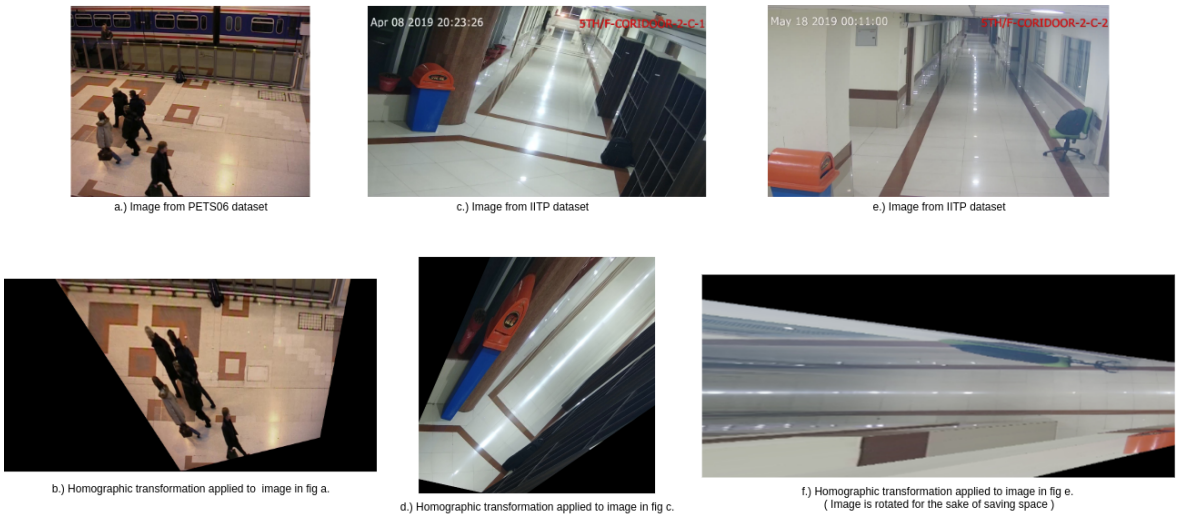

Figure 4: Homographic transformation on sample images from PETS06 dataset and IITP dataset. First row shows the original images and second row shows the the images obtained by applying homographic transformations.

This temporal threshold is obtained using the time that is assumed for flagging the abandoned event and the frame skipping (frame skipping is used to increase the processing speed based on the assumption that in videos very small changes occur in consecutive frames) used. If threshold for abandoned event is assumed to be τ seconds and frame skipping of s is used, the value of temporal threshold will be τ/s. Items in the tentative list will be updated continuously as as the incoming frames are processed and if the owner re-attends the object then it will be removed from the tentative list. Timing for abandonment will be obtained from frame rate and the number of frames processed. A window mechanism has been used to process the videos such that cumulative history of the detections can be maintained and this helps in keeping short term track of the objects.

4.3.8 Information Delivery Module

Raising an alarm on occurrence of abandonment event is also crucial. The event should be reported with precise location of object and the identification of the owner or the owner group. Location of the abandoned object will be marked by bounding box in the currently processed frame. Feature vector corresponding to the owner who has left the luggage will be returned, which can be used further to identify the owner if that person reappears somewhere in the scene.

5 Experimental Setup and Results

5.1 Datasets

We have implemented the proposed system and experimented on both public domain dataset as well as a newly created dataset which is also available freely.

5.1.1 PETS06 Dataset

PETS06 dataset provides 7 videos of abandoned luggage scenarios. All the videos consist of luggage abandoning event except the third video sequence which shows a person who keeps his luggage on the floor for a short duration and picks it up but does not abandon it.

5.1.2 IITP Dataset

PETS06 dataset has been one of the highly used datasets for testing and evaluating abandoned object detection systems and contains scenarios which are very basic in nature. A system should not only perform robustly in these kind of controlled scenarios but in actual cases also, these are complex cases yet occur very commonly. For an implementation to be practically applicable it must cover more complex cases of abandoned object scenario. Our dataset contain 58 videos for abandoned case and 18 videos for non abandonment, recorded from campus surveillance cameras. Each video is recorded in full HD resolution at 25 frames per second. All the sequences are 3 to 5 minutes long. These videos are categorised in easy, medium and hard classes depending upon the complexity of the abandonment event.

5.2 Experimental Setup and Assumptions

5.2.1 Object detector

A CNN based model, RetinaNet [9] is used as object detector, which is a variant of single shot detector. The model is used with ResNet50 backbone. This network is fine tuned for detecting objects of class luggage which includes items like suitcase, backpack, and handbag. It is fine tuned on 5000 images of luggage from OID (Open Images Dataset). Since number of images is very less, various image transformations such as rotation, flipping and translation have been applied to generate more images. Only the last layer is retrained keeping all other layers non-trainable and using a learning rate of 0.00001 with Adam optimizer.

5.2.2 Assumptions for experiments

An object is considered to be unattended if the object is stationary and there is no person within a radius of δ for a time interval more than τ. Ownership of the object is established by assuming that if a person is within a radius δ of the object for more than τ_{asso} time and the entry of object and person is in the same frame or in nearby frames here it is assumed to be within n_{asso} frames. In our tests τ_{asso} is set to 10 seconds and n_{asso} is set at 10 frames. Also re-entry of the owner is considered and the re-identification is done using cosine similarity matching, intersection over union (IOU) and centre shift between the object of current frame and the previous frame. Performance of the proposed model is evaluated on IITP and PETS06 datasets. Precision and recall are used used to measure the performance of the system. Also mean error in alarm timing of the abandoned event is also shown.

5.3 Results on PETS06 Dataset

Table 2 shows the evaluation results on PETS06 dataset. Third video sequence does not contain abandoned object event hence alarm should not be raised. Our method successfully flagged the abandoned event in all the cases along with the feature vector of the owners. Since

third video does not contain abandonment, our method does not report any abandonment in this case which shows the effectiveness of our approach. It can also be seen from Table 2 that in all the cases owner has been associated correctly, the corresponding features for the owner has been returned. Though there is an average delay of around 9 seconds in triggering the alarm but the system successfully detects the abandonment. Table 3 shows the comparison of precision, recall and F-measure on PETS06 dataset with various available methods. It can be observed that our method performs at par with state of the art works [31] [35], but additionally it gives owner information without back tracing as opposed to other available methods.

Table 2: Abandoned event detection on PETS06 dataset.

Vid. Seq.	**Ground Truth**	**Alarm Time**	**Delay in Alarm**	**Owner**
seq 1	113.7	121	7.3	yes
seq 2	91.8	105	13.2	yes
seq 3	na	na	na	na
seq 4	104.0	115	11	yes
seq 5	110.6	118	7.4	yes
seq 6	96.9	105	8.1	yes
seq 7	93.9	102	8.1	yes

Table 3: Comparison with available methods on PETS06 dataset.

Method	**Precision**	**Recall**	**F-measure**
Porikili et. al. [22]	0.03	1.0	0.05
Auvinet et. al. [18]	0.58	1.0	0.71
Liao et. al. [36]	0.75	1.0	0.86
Li et. al. [11]	1.0	0.71	0.83
Tian et. al. [25]	0.85	1.0	0.92
Fan et. al. [35]	0.95	0.80	0.87
Lin et. al. [31]	1.0	1.0	1.0
Proposed	**1.0**	**1.0**	**1.0**

5.4 Results on IITP Dataset

Abandoned event set is categorised into three parts namely hard, medium and easy based on the difficulty and complexity of the scene. Evaluation results on IITP dataset as shown in Table 4 establishes the effectiveness of our method. It misses some of the events particularly 5 cases of hard category which have multiple owners and crowded environment and misses 3 case in medium difficulty category.

5.5 Ablation Study

The whole pipeline is tested with varying configurations of the constituent modules. Observations from the experiments are elaborated below:

Table 4: Abandoned event detection accuracy on IITP dataset: With retrained detector.

Category	**Total Cases**	**Detected Cases**	**Detection Accuracy**	**Avg. Delay in Detection**	**Average IOU**
Hard	19	14	0.737	15	0.622
Medium	20	17	0.850	15	0.681
Easy	19	19	1.000	12	0.794

Table 5: Abandoned event detection accuracy on IITP dataset: Without retraining detector but using patch feed with context area.

Category	**Total Cases**	**Detected Cases**	**Detection Accuracy**	**Avg. Delay in Detection**	**Average IOU**
Hard	19	11	0.579	22	0.458
Medium	20	15	0.750	17	0.517
Easy	19	17	0.895	16	0.662

- In Table 5 results show that, when pre-trained object detector is used without retraining and fine tuning, the performance of the whole system degrades sharply as compared to the results shown in Table 4. This can be attributed to the failure of detector to identify the objects in the scene.
- It was observed that accuracy of the system showed no improvement when only patch feeding was used, instead it degrades when patch feeding was used without context area as shown in part (a) of Table 6. Context area is provided by including the surrounding region so that the size of patch becomes $2x, 3x, 4x$ and $5x$, where x is the size of patch. It was observed that $3x$ patches worked best with highest detection rate and covers the various scales of practical objects. Results shown in Table 6 are obtained with $3x$ patches wherever patch feed with context is used.
- Also since the whole model is based on the accuracy of the detector, similar trend was observed in case of average delay in raising alarm and IOU of the detected objects with the truth value as shown in part $(b) and (c)$ of Table 6.

6 Conclusion

Automated video surveillance poses many complex challenges which needs multiple objectives to be handled simultaneously. It is very difficult to model the rules for associating persons with luggage since there are people moving around the objects in actual scenario. Our approach handles these situations reasonably well by establishing ownership of objects in the scene and filtering out non-owners which consists of the people randomly moving around. Another challenge is distance calculation using single camera setup because there is no depth information. This problem is tackled by using homographic transformations on the ground plane. In image processing, occlusion has always been a big challenge and the randomness of crowd movement makes it even more challenging. To handle occlusion, context information

Table 6: Accuracy, Avg. Delay and IOU on IITP dataset with different configurations (**M1:**Retrained detector, **M2:**Patch feed method without retrained detector, **M3:**Patch feed with context without retrained detector).

(a) **Accuracy** with different configurations.

Category	**Without M1, M2, M3**	**With only M1**	**With only M2**	**With only M3**	**With M1, M2, M3**
Hard	0.421	0.579	0.368	0.526	0.737
Medium	0.450	0.750	0.450	0.650	0.580
Easy	0.632	0.895	0.579	0.736	1.000

(b) **Average Delay** with different configurations.

Category	**Without M1, M2, M3**	**with only M1**	**With only M2**	**With only M3**	**With M1, M2, M3**
Hard	25	22	23	22	15
Medium	19	18	19	17	15
Easy	18	18	18	16	12

(c) **IOU** with different configurations.

Category	**Without M1, M2, M3**	**With only M1**	**With only M2**	**With only M3**	**With M1, M2, M3**
Hard	0.382	0.458	0.394	0.417	0.622
Medium	0.441	0.488	0.405	0.419	0.681
Easy	0.496	0.643	0.521	0.612	0.794

has been used, for example a person can reappear within a small region of the image since the speed of person will most likely be very small. The system is based on assumptions like temporal and spatial threshold for abandonment and ownership establishment which may not hold. There are still many open challenges in the areas including occlusion, crowd behaviour, accurate distances calculations, re-entry, re-identification, owner identification etc., which needs a lot of improvement. These enhancements will be incorporated in future works.

References

[1] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[3] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," July 2017.

[4] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[6] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2874–2883, 2016.

[7] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[10] H. Grabner, P. M. Roth, M. Grabner, and H. Bischof, "Autonomous learning of a robust background model for change detection," in *Proc. PETS*, pp. 39–46, 2006.

[11] L. Li, R. Luo, W. Huang, and H.-L. Eng, "Context-controlled adaptive background subtraction," in *Proceeding Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), New York, USA*, pp. 31–38, 2006.

[12] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 2, pp. 28–31, IEEE, 2004.

[13] N. Krahnstoever, P. Tu, T. Sebastian, A. Perera, and R. Collins, "Multi-view detection and tracking of travelers and luggage in mass transit environments," in *In Proc. Ninth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, vol. 258, 2006.

[14] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 2, pp. 246–252, IEEE, 1999.

[15] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[17] C. Bucilu, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, ACM, 2006.

[18] E. Auvinet, E. Grossmann, C. Rougier, M. Dahmane, and J. Meunier, "Left-luggage detection using homographies and simple heuristics," in *Proc. 9th IEEE International Workshop on Performance Evaluation in Tracking and Surveillance (PETS06)*, pp. 51–58, Citeseer, 2006.

[19] N. Bird, S. Atev, N. Caramelli, R. Martin, O. Masoud, and N. Papanikolopoulos, "Real time, online detection of abandoned objects in public areas," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pp. 3775–3780, IEEE, 2006.

[20] M. Spengler and B. Schiele, "Automatic detection and tracking of abandoned objects," in *Proceedings of the Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Citeseer, 2003.

[21] S. Ferrando, G. Gera, and C. Regazzoni, "Classification of unattended and stolen objects in video-surveillance system," in *2006 IEEE International Conference on Video and Signal Based Surveillance*, pp. 21–21, IEEE, 2006.

[22] F. Porikli, Y. Ivanov, and T. Haga, "Robust abandoned object detection using dual foregrounds," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, p. 30, 2008.

[23] A. Filonenko, K.-H. Jo, *et al.*, "Unattended object identification for intelligent surveillance systems using sequence of dual background difference," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 6, pp. 2247–2255, 2016.

[24] F. Lv, X. Song, B. Wu, V. K. Singh, and R. Nevatia, "Left luggage detection using bayesian inference," in *Proc. of IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 83–90, Citeseer, 2006.

[25] Y. Tian, R. S. Feris, H. Liu, A. Hampapur, and M.-T. Sun, "Robust detection of abandoned and removed objects in complex surveillance videos," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 5, pp. 565–576, 2010.

[26] W.-S. Zheng, S. Gong, and T. Xiang, "Quantifying and transferring contextual information in object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 762–777, 2011.

[27] S. Smeureanu and R. T. Ionescu, "Real-time deep learning method for abandoned luggage detection in video," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1775–1779, IEEE, 2018.

[28] S. Sidyakin and B. Vishnyakov, "Real-time detection of abandoned bags using cnn," in *Automated Visual Inspection and Machine Vision II*, vol. 10334, p. 103340J, International Society for Optics and Photonics, 2017.

[29] M. Bhargava, C.-C. Chen, M. S. Ryoo, and J. K. Aggarwal, "Detection of object abandonment using temporal logic," *Machine Vision and Applications*, vol. 20, no. 5, pp. 271–281, 2009.

[30] J.-Y. Chang, H.-H. Liao, and L.-G. Chen, "Localized detection of abandoned luggage," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 1, p. 675784, 2010.

[31] K. Lin, S.-C. Chen, C.-S. Chen, D.-T. Lin, and Y.-P. Hung, "Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1359–1370, 2015.

[32] A. F. Otoom, H. Gunes, and M. Piccardi, "Feature extraction techniques for abandoned object classification in video surveillance," in *2008 15th IEEE International Conference on Image Processing*, pp. 1368–1371, IEEE, 2008.

[33] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 748–756, IEEE, 2018.

[34] S. J. Prince, K. Xu, and A. D. Cheok, "Augmented reality camera tracking with homographies," *IEEE Computer graphics and Applications*, vol. 22, no. 6, pp. 39–45, 2002.

[35] Q. Fan, P. Gabbur, and S. Pankanti, "Relative attributes for large-scale abandoned object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2736–2743, 2013.

[36] H.-H. Liao, J.-Y. Chang, and L.-G. Chen, "A localized approach to abandoned luggage detection with foreground-mask sampling," in *2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*, pp. 132–139, IEEE, 2008.