Detecting Abandoned Objects in a Multi-Camera Video Surveillance System

Pyke Tin, Thi Thi Zin, Takashi Toriu
Graduate School of Engineering
Osaka City University
Osaka, Japan
pyketin@ip-info.eng.osaka-cu.ac.jp

Hiromitsu Hama
Research Center for Industry Innovation
Osaka City University
Osaka, Japan

***Abstract*—The problems of detecting abandoned packages are receiving increased worldwide attention in many contexts including airports, subways, building lobbies, sporting events, and other public venues. Some recent terrorist attacks with abandoned explosive items illustrate the importance of these problems. In this aspect, we present a multi-camera video surveillance system which can automatically detect abandoned objects in complex environments. By using a multiple-state Markov Model based on the simple foreground mask and the estimation of the size distribution on image plane, the method can handle arbitrary number of cameras and objects. Due to its statistical Markov nature, the proposed method efficiently handles such challenging situations as changes in viewpoint, occlusions due to view variations, and background clutter. We evaluate the performance of the proposed method using our own video sequences taken in real-life visual behavior understanding scenarios and we compare and discuss the obtained results.**

Keywords- abandoned object; multi-camera; video surveillance; multiple-state Markov model

I. INTRODUCTION

Explosive attacks with abandoned packages involving suspicious people are repeatedly concentrated on public places. Hence, more and more surveillance cameras are installed for public space monitoring and used to prevent terrorist attacks or many other possible public dangerous events [1–3]. As the growing number of surveillance systems and cameras, watching surveillance videos and identify specious events have become challenging issues for security companies and personnel. Surveillance applications developed nowadays are part of third generation surveillance systems that cover a wide area using a multi-camera network [4, 5]. Typical watched areas are sensitive public places and infrastructures that are susceptible of being crowded. Since tracking people and objects in a crowded environment is a big challenge in image space, we must deal with merging, splitting, entering, leaving and correspondence. The problem is more complicated when the environment is observed by multiple cameras. One of the key challenges nowadays is to robustly and efficiently extract and present meaningful intelligence from multi-camera networks. Many have been introduced in the literature for event detection in surveillance videos from single camera views; multi-camera event detection still remains an emerging problem in real-life application such as detection and management of large scale of complex events across multiple cameras with overlapping or disjoint fields-of-views (FOV).

This paper will develop an automatic system for recognition of abandoned packages in crowded under unconstrained contexts presenting many challenges. These include occlusions, appearance variations as people move relative to the camera, lighting changes, and other complex interactions between objects in the scene. We address these problems through multiple cameras with overlapping FOV, occlusion detection/compensation, and higher level scene understanding.

II. RELATED WORKS

A. Background

Various approaches have been presented for the abandoned object detection tasks in the past [4, 6–8]. The identifications of suspicious objects through change detection relative to a reference image have been examined in [6, 7]. Moreover, discriminating between suspicious objects, stationary people, persistent lighting changes and structural changes through a combination of position constraints are also outlined. In our previous works [8], a statistical approach to detecting changes from an adaptively constructed background image of a scene has been employed. In this aspect, the three periodic reference background models are established: (i) Short Length periodic updated background model (SL), (ii) Long Length periodic updated Background model (LL), and (iii) Stochastically Varied likelihood image model (SV).

The first frame of the input video image is initialized as SL and LL respectively in our application, and an improved adaptive background updating method is applied by constructing two maps of pixel history: Stable history Map (SM) and Difference history Map (DM). They are defined as follows:

$$\begin{aligned} \mathrm{SM} &= \mathrm{SM}+1 \quad \text{if } |I_n(x,y)-I_{n-1}(x,y)|<Th_s, \\ &= 0 \qquad\quad \text{otherwise.} \end{aligned} \tag{1}$$

$$\mathrm{DM} = n - \mathrm{SM}, \tag{2}$$

where n stands for the total number frames in the sequence. Based on the information from both maps the backgrounds adaptively updated frame-by-frame by:

This work is partially supported by SCOPE: Strategic Information and Communications R&D Promotion Program (102307010).

$$SL_n(x,y)=\begin{cases} I_n(x,y) & \text{if } SM(x,y)>Th_f \text{ and } DM(x,y)>Th_f, \\ SL_{n-1}(x,y) & \text{if } SM(x,y)>Th_f \text{ and } DM(x,y)=0, \\ (1-\alpha)SL_{n-1}(x,y)+\alpha I_n(x,y) & \text{if } SM(x,y)=0. \end{cases} \quad (3)$$

$SL_n(x, y)$ and $SL_{n-1}(x, y)$ represent the short periodic updated backgrounds pixel value at position (x, y) in current and previous frames. In the same way, $LL_n(x, y)$ and $LL_{n-1}(x, y)$ represent the long length periodic updated background at position (x, y) and the corresponding updating rules are

$$LL_n(x,y)=\begin{cases} SL_n(x,y) & \text{if } SM(x,y)>Th_f \text{ and } DM(x,y)>Th_f, \\ LL_{n-1}(x,y) & \text{if } SM(x,y)>Th_f \text{ and } DM(x,y)=0, \\ (1-\beta)LL_{n-1}(x,y)+\beta SL_n(x,y) & \text{if } SM(x,y)=0, \end{cases} \quad (4)$$

where $I_n(x, y)$ is the pixel value at position (x, y) in current image frame, Th_f is the threshold value and α, β are the learning rate of two backgrounds. At every frame, we estimate the short length periodic foreground (SF) and long length periodic foreground (LF) by comparing the current frame I by the background models SL and LL. In addition, we construct a stochastically varied likelihood image based on the two previous updating models as follows.

$$SV(x,y)=\begin{cases} SV(x,y)+1 & \text{if } \mathrm{P}(SF(x,y)=1 \cap LF(x,y)=0)=1, \\ SF(x,y)-k & \text{if } \mathrm{P}(SF(x,y)\neq 1 \cup LF(x,y)\neq 0)=1, \\ 0 & \text{if } \mathrm{P}(SV(x,y)<0)=1, \\ \max SV(x,y) & \text{otherwise,} \end{cases} \quad (5)$$

where max and k are positive numbers and $P(.)$ is the probability measure of an event. For each pixel, the likelihood image collects the evidence of being an unattended item. Whenever this evidence elevates up to a present level, we mark the pixel as an unattended item pixel for further analysis.

B. *Multiple Cameras*

In recent years, there have been many interesting work developed for detecting events or activities across multi-cameras [9–15]. A multi-camera environment is not just a collection of cameras that perform their respective tasks independently. Indeed, there is the opportunity to coordinate and integrate information on activities across all cameras in order to improve the performance of the overall system. After object extraction, data originating from individual cameras can be brought into one coherent common frame of observation so that the automated surveillance system can infer a global analysis to understand the scene and accordingly allow appropriate decisions. Single camera approaches do not always work well due to lack of information utilizable. Multiple camera preparation is promising because it provides additional clues to refine segmentation and has robustness against noises. Importantly, it has the potential of providing a complete record of an object's activity in a complex scene, allowing a global interpretation of the object's underlying behaviour. In most multiple-camera surveillance systems, disjoint cameras with non-overlapping FOV are more prevalent, due to the desire to maximize spatial coverage in a wide-area scene whilst minimizing the deployment cost. An illustration of different degrees of overlap between FOV is described in Fig. 1.

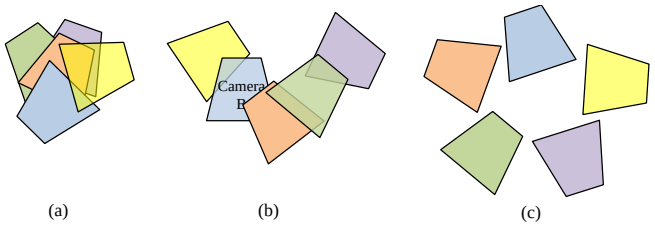

Figure 1. An illustration of different degrees of overlap between the field of view (FOV) of five cameras: (a) all cameras partially overlap with each other, (b) cameras partially overlap with adjacent cameras, (c) non-overlapping camera network.

C. *Our Approach*

Our approach differs in at least two major ways from previously reported work. First, we analyse the relationships between objects. The owner of each abandoned object is determined and tracked using distance and time constraints through a multi-state Markov model. This higher level context and understanding is necessary for avoiding false alarms in realistic scenarios. Second, we have exploited multiple cameras with overlapping fields of view to cope with occlusions of various types as illustrated in Fig. 2. While the general approach is similar, our system has some additional features. In particular, our system does not assume that the cameras are synchronized, making camera deployment simpler. Also, our system is capable of operating with only a single camera, though missed detections and false alarms will decrease as additional cameras are deployed.

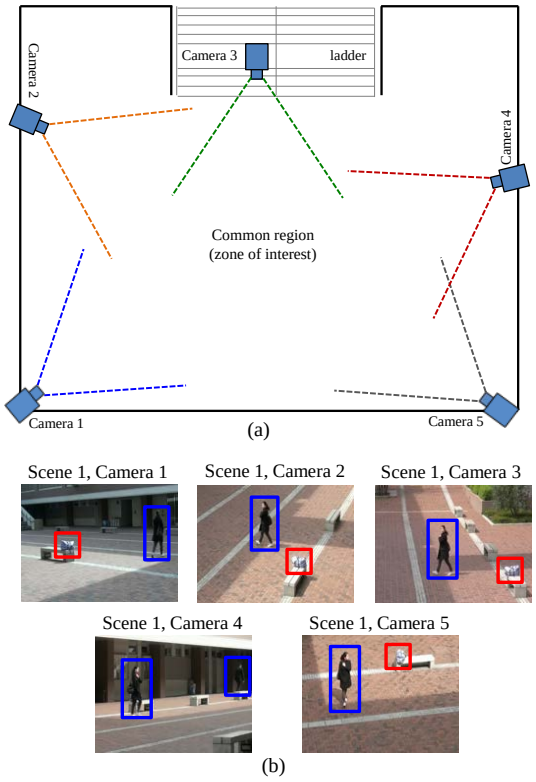

Figure 2. Multi-camera system: (a) multi-camer setting and (b) partial observations of abandoned object detecion from different camera views.

The paper is structured as followed: An overview of our approach is described in section III. The multi-camera multi-state Markov model is established in section IV. Finally, results will be presented in section V followed by a short discussion in section VI.

III. Propose System Architecture Overview

In this section, we shall describe an overview of the proposed system architecture for abandoned object detection in multi-camera environment as illustrated in Fig. 3. In this system, the video from each camera is separately processed as in a single camera system before a combined processing phase. In this case, the video processed data from a single camera unit for each video frame may include the following information: the camera time stamp, list of targets with their ids and image properties such as bounding box, centre of gravity, footprint, and classification label. Then, fusion module combines the video processed-data from all cameras into a common coordinate system, but still maintaining the video processed-data format, so that the fused-data can be fed to the event detection module in decision making process. This event detection module is identical to the one used in the single camera system.

For a single camera the coordinates are relative to a single frame, while for the fusion camera they are relative to the global map. This means that the combined data is the same whether it is generated by a view or a map. This design has many benefits. The time consuming video processing is distributed among the single camera sensors, communication bandwidth requirements are low due to transmitting only the view processing outputs. The architecture is simple: the system running on the individual camera is almost identical to the single camera system. It is still possible to have single camera event detection running on the individual camera, if required. The only difference is that the video processed-data is streamed out to the fusion process to enable multi-camera event detection. The fusion process is more different, but still has a lot in common to single camera process. The main difference is the front end: it ingests multiple video processed-data streams instead of video, and uses the data fusion module to convert it into fused-data. Then, by using these fused-data, objects are classified as human, package, or unknown using evidence from multiple video channels. The object tracker state is monitored by the observer module and alerts are generated should an abandoned package be detected.

IV. Combined Processing for Abandoned Object Detection

Object detection across cameras is used to interpret the combined sets of time-stamped foreground marks segmented from each video stream. Given a set of known objects in the scene and the camera field of view, a foreground marks could be: a single object, multiple objects that overlap, camera sensor noise not captured by the background model, background motion, mirrored reflections, bright reflections, etc. The object detection must deal with unreliable input, and maintain a consistent state that can be used to understand the objects and their interactions within the scene. We use several mechanisms to build and maintain such a consistent state.

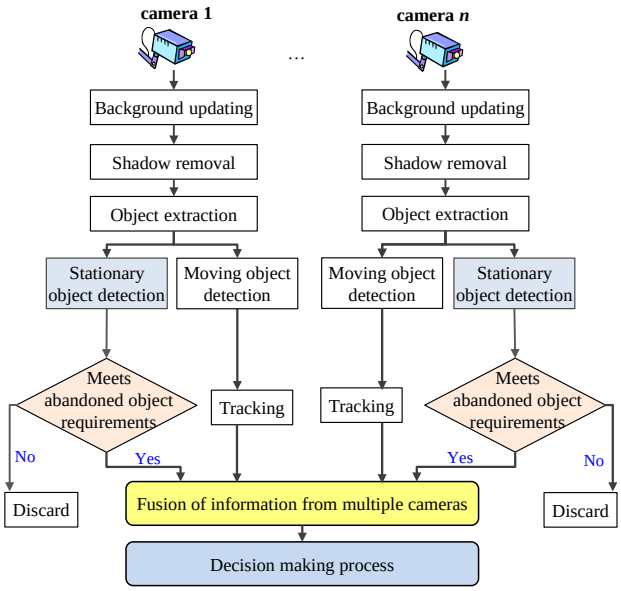

Figure 3. The overview of the prposed muli-camera system.

A. *Classification*

Detected objects are classified as human, package, or unknown. The unknown class represents foreground marks resulting from lighting changes, reflections, shadowing, etc. We are using two features to classify objects: area and compactness. The area feature is the number of pixels belonging to the object. The compactness feature denoted by C represents how "stretched out" a shape is, and is defined as $C = \text{area}/P^2$. The perimeter P is the number of pixels that are part of an object and have at least one 4-connected neighbour that is not in the object. A circle has the minimum perimeter for a given area, hence exhibits the highest compactness. Decision regions in a two dimensional feature space are defined by a set of quadratic discriminant functions. Classification is based on selecting the maximum.

B. *Data Association*

Association between measurements and detected objects is a fundamental step in object classification. Measurements are the segmented blobs from the current video frame, and detected objects are the set of existing objects in the scene. We use a linear assignment problem algorithm for performing an optimal association simultaneously for all measurements and detected objects [15]. This efficient algorithm uses a cost matrix of measurement-detect object pairs, and minimizes the total cost of the selected associations. An individual measurement-detect object association cost is the weighted sum of position, size, and colour information. There will always be situations when data association will make mistakes, so systems must be robust and be able to recover gracefully. Moreover, the recovery from mistakes should happen quickly, so that historical information will be consistent when used for higher level processing such as finding abandoned packages. Some specific example scenarios causing mistakes are:

1) Consider two cameras: Camera 1 and camera 2 with overlapping field of views. There is one person in the scene, which is detected in camera 1, and just entering camera 2's FOV. Suppose there are many differences in the segmented view of the object in camera 2 relative to camera 1. This can often times result in a new object being created for the camera 2 view, because the color histogram comparison indicates they are different. Now a single person is represented by two distinct track objects.

2) One object occludes another object, resulting in a single measurement for two objects. The linear assignment algorithm will have at most one object associating with a given measurement; hence one of the two objects will have no association or incorrectly associating with another nearby object (and other cascading association errors).

C. Abandoned Package Detection

Determination of an abandoned package event requires a precise definition of what it means for a package to be abandoned. The most basic characteristic is lack of motion. This describes a situation where a package was initially being carried, and is later left in the scene. However, consider an airport scenario where travellers standing in line repeatedly drop and pick up their packages. This would cause false alarms, because there are times the package is not moving, but it is clearly not abandoned. Now add a constraint that the package must be a minimum distance away from the nearest person. While this helps for the standing in line scenario, it will cause a missed detection if a person drops a package near another person and then leaves. To address this, the minimum distance should be between the package and its apparent owner. This final definition of an abandoned package is more realistic, and will reduce false alarms and missed detections in similar scenarios. The definition of abandoned we use is a stationary package that has no discernible owner, or the owner is sufficiently far away for some amount of time. Given the package velocity v, elapsed time Δt, distance between owner and package d, and corresponding thresholds V_0, T_a and D, this can be expressed as:

$$\text{Abandoned} = (v < V_0) \ \wedge (\text{un-owned} \vee d > D) \wedge (\Delta t > T_a).$$

Fig. 4 shows the multiple-state Markov model representing this definition of an abandoned package, with additional transitions for when the conditions only partially hold.

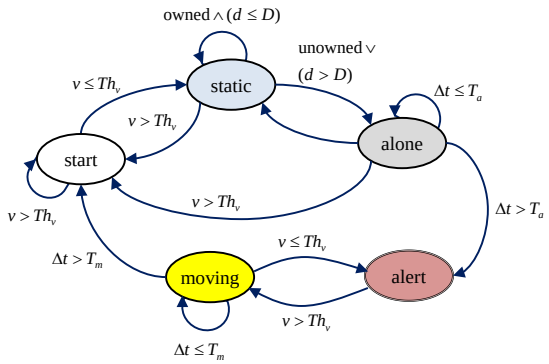

Figure 4. Multiple-state Markov model for detecting abandoned packages.

When an object appears that is classified as a package, it begins in the Start state. The static state is entered when the velocity of the package becomes low enough. If the package does not have an owner, or if the distance between the owner and the package exceeds a threshold (e.g., 2.0 meters), we enter the alone state. When a threshold amount of time has passed and the package object has remained stationary and isolated from its owner, we enter the Alert state, and an operator is notified. When the owner returns, or if the package starts moving, we leave the Alert state and turn off the notification. Our determination of ownership avoids a missed detection for the specific scenario where a package is dropped next to an unrelated person.

D. Package Ownership

The ownership criterion in the definition of an abandoned package is crucial to avoid false alarms and missed detections. Specifically, for a non-moving package, which human is its owner? The object tracker retains historical snapshots of the current state to facilitate history examination. To reduce the size of the history, a maximum of K past frames is stored, which are spread over an s second window. When a newly idle package is first detected, history of all cameras is examined to determine ownership. Frames are checked to find human objects that contain pixels matching the current view of the idle package. This continues back through the entire history window, and the object with the most matches is selected as the owner of the idle package. This allows for ownership determination even for cases when the actual package drop is temporarily occluded. The owner determination performs well for the test cases we have used, but we expect enhancements will be required. First, if package hand-offs happen frequently, it is unclear who the correct owner should be. Perhaps in this case, we should indicate all strong matches as co-owners. Second, for complex scenes with many people carrying similar packages, we could also look forward from the point the package appears, and make sure that only humans that no longer match the histogram of the package are considered candidates for ownership. This makes sense because if the true owner drops the package, it should look different than it did before the drop.

V. Experimental Results

The proposed system for abandoned object detection is tested using (i) PETS 2006 dataset and (ii) our owned dataset taken in an international airport and Osaka City University Campus. In our experiments, the test bed for developing and evaluating our approach consists of three cameras setting and five cameras setting. The image resolution is 320x240 (QVGA) and the frame rate is 10fps. We have described the results in details for three cameras setting since the outcomes are similar.

Fig. 5 shows that video frames from each camera are separately processed before a combined processing phase. The per camera view processing outputs foreground regions that are obtained by using our previous works for single camera setting developed in [6-8]. To be specific, some qualitative results are shown in Fig. 5 in which the results obtained from cameras 1, 2, and 3 are displayed respectively in rows 1, 2, and 3. In the figure, the first three columns show the current image frame, our periodic background models and foreground.

As can be seen, foreground masks yielded by our proposed method are noticeably better and the targeted static object detection is successfully made as displayed in the final column. Moreover, the additional results are shown in Fig. 6 in which camera 1 (top row) has a clear view of the people and the package, whereas camera 2 (middle row) has some occluded views of the scene. Multiple cameras help to resolve the occlusion in camera 3. Sufficient camera coverage, combined with an ownership and unattended luggage criteria, allows us to detect abandoned packages in these realistic scenarios. The common frame for the output results are shown in Fig. 7.

As shown in Fig. 6, the system is able to detect people and abandoned objects, and associate the abandoned objects with their owners. In the case of Fig. 6(b-ii), the abandoned object is occluded for a while because the new person stands in front of the view. As the person moves away, the remainder of the object is detected and flagged. Fig. 6(b-iv) illustrates the system ability to maintain the location of the abandoned object during occlusions. The abandoned object is occluded several times, yet remains correctly detected. The use of feedback into the motion detection allows the abandoned object to be held out of the background and remain detected. Fig. 7 along with Fig. 6 show abandoned object detection results. The red shaded area indicates an abandoned object, with the shaded yellow person indicating the owner of the object. According to the result of Fig. 6(a-iv), only the camera 1 cannot detect the event of carrying another bag when the owner leaving out. The proposed system can make the correct decision due to the fusion of information from multiple cameras.

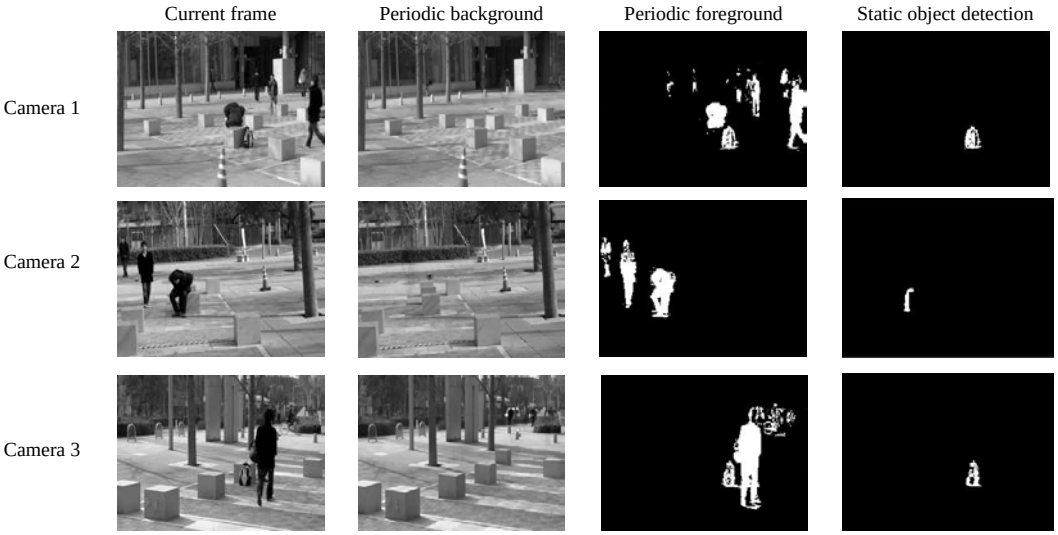

Figure 5. Periodic background substraction results and static object detection for three cameras.

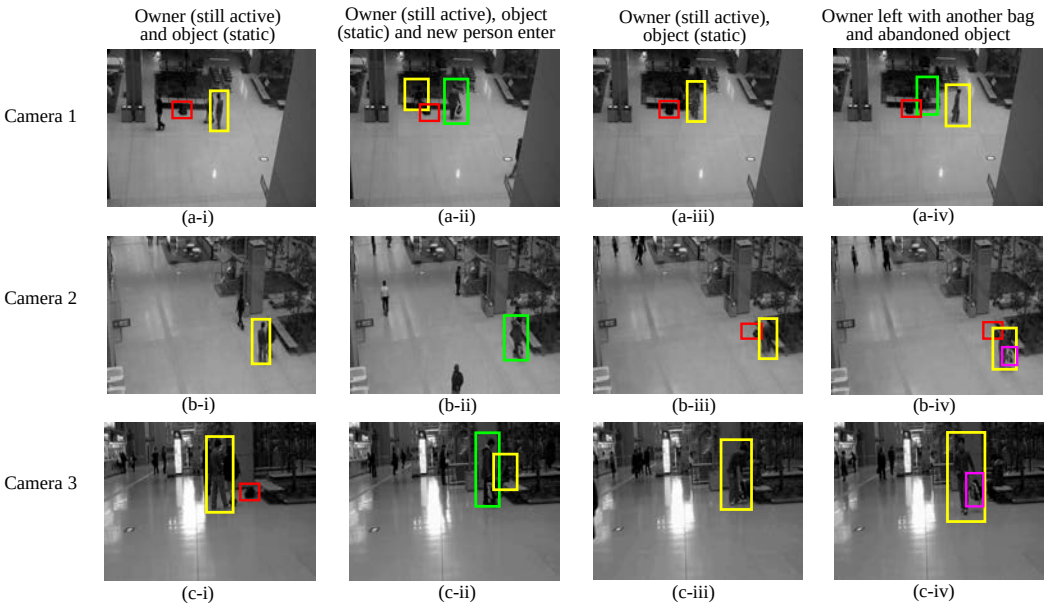

Figure 6. Abandoned object detection process from a multi-camera system: (a) camera 1, (b) camera 2 and (c) camera 3.

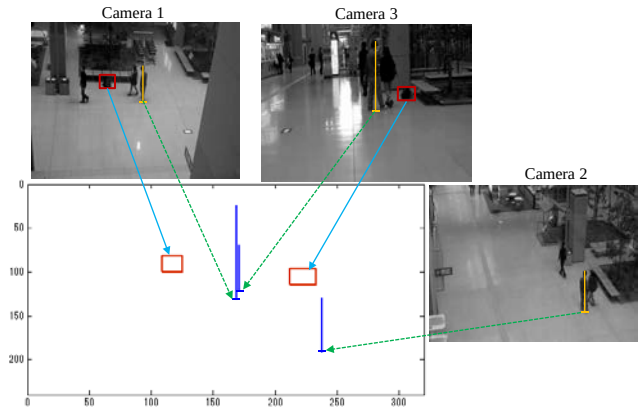

Figure 7. Example of the estimated common occupnacy with some object's corresponding locations.

VI. CONCLUSIONS

We have presented a multi-camera video surveillance system that detects abandoned packages automatically. A multiple-state Markov model of an abandoned package enables detection of realistic abandoned package events using ownership criteria. Our future work involves completing a thorough quantitative evaluation using a larger corpus of video datasets. There are many scenario complications for future work. Has the owner transferred the package to a new owner? How similar to benign objects can the package appear? Can concealment of a package be detected? Can the package owner be tracked backward and forward in time across a crowed environment? Can we detect theft of a package?

ACKNOWLEDGMENT

We thank Scope project members and the students of Physical Electronics and Informatics, for their participation in producing tested videos.

REFERENCES

[1] J. Black, D. Makris and T. Ellis, "Hierarchical database for multi-camera surveillance system," Pattern Analysis Application, pp. 430–446, 2005.

[2] R. T. Collins, A. J. Lipton and T. Kanade, "Introduction to the special section on video surveillance," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 745–746, Aug. 2000.

[3] Thi Thi Zin, Pyke Tin, T. Toriu and H. Hama, "A Markov random walk model for loitering people detection," in the 6th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Darmstadt, Germany, pp. 680–683, Oct. 2010.

[4] E. Auvinet, E. Grossmann, C. Rougier, M. Dahmane and J. Meunier, "Left-luggage detection using homographies and simple heuristics," Proc. of the 9th IEEE International Workshop on Performance Evaluation in Tracking and Surveillance (PETS'06), New York, NY, USA, pp. 51–58, Jun. 2006.

[5] C. Fookes, S. Denman, R. Lakemond, D. Ryan, S. Sridharan and M. Piccardi, "Semi-supervised intelligent surveillance system for secure environments," Proc. of the IEEE International Symposium on Industrial Electronics, Bari, Italy, pp. 2815–2820, 2010.

[6] Thi Thi Zin, Pyke Tin, H. Hama and T. Toriu, "Unattended object intelligent analyzer for consumer video surveillance," IEEE Trans. on Consumer Electronics, vol. 57, no. 2, pp. 549–557, May. 2011.

[7] Thi Thi Zin, H. Hama, Pyke Tin and T. Toriu, "Evidence fusion method for abandoned object detection," ICIC Express Letters (Part B: Applications): An International Journal of Research and Surveys, vol. 2, no. 3, pp.535–540, Jun. 2011.

[8] Thi Thi Zin, Pyke Tin, H. Hama, S. Nakajima and T. Toriu, "Effective multiple stochastic background modeling for stationary objects detection in complex environments," ICIC Express Letters: An International Journal of Research and Surveys, vol. 5, no. 10, pp. 3767–3772, Oct. 2011.

[9] T. Ahmedali and J.J. Clark, "Collaborative multi-camera surveillance with automated person detection," in the 3rd Canadian Conference on Computer and Robot Vision (CRV2006), Quebec, Canada, Jun. 2006.

[10] V. Kettnaker and R. Zabih, "Bayesian Multi-camera surveillance," IEEE Conference on. Computer Vision and Pattern Recognition, vol. 2, 1999.

[11] S. Lim, L.S. Davis and A. Elgammal, "A Scalable Image-Based Multi-Camera Visual Surveillance System," IEEE Conf. on Advanced Video and Signal Based Surveillance (AVSS2003), Miami, Florida, Jul 21-22, 2003.

[12] K. Kim and L. S. Davis, "Multi-Camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," European Conference on Computer Vision (ECCV), LNCS, pp. 98–109, 2006.

[13] P. Remagnino, A.I. Shihab and G.A. Jones, "Distributed intelligence for multi-camera visual surveillance," Pattern Recognition, vol. 37, no. 4, pp. 675–689, Apr. 2004.

[14] G.Wu, Y.Wu, L. Jiao, Y.Wang and E.Y. Chang, "Multi-camera spatio-temporal fusion and biased sequence-data learning for security surveillance," Proc. of ACM International Conference on Multimedia, November, pp. 528–538, 2003.

[15] H. Zhou and D. Kimber, "Unusual event detection via multicamera video mining," Proc. of the 18th International Conference on Pattern Recognition (ICPR), Hong Kong, pp. 1161–1166, Aug. 2006.